

# Package ‘vimp’

July 22, 2025

**Type** Package

**Title** Perform Inference on Algorithm-Agnostic Variable Importance

**Version** 2.3.3

**Description** Calculate point estimates of and valid confidence intervals for nonparametric, algorithm-agnostic variable importance measures in high and low dimensions, using flexible estimators of the underlying regression functions. For more information about the methods, please see Williamson et al. (Biometrics, 2020), Williamson et al. (JASA, 2021), and Williamson and Feng (ICML, 2020).

**Depends** R (>= 3.1.0)

**Imports** SuperLearner, stats, dplyr, magrittr, ROCR, tibble, rlang, MASS, boot, data.table

**Suggests** knitr, rmarkdown, gam, xgboost, glmnet, ranger, polspline, quadprog, covr, testthat, ggplot2, cowplot, cvAUC, tidyselect, WeightedROC, purrr

**License** MIT + file LICENSE

**URL** <https://bdwilliamson.github.io/vimp/>,  
<https://github.com/bdwilliamson/vimp>,  
<http://bdwilliamson.github.io/vimp/>

**BugReports** <https://github.com/bdwilliamson/vimp/issues>

**RoxygenNote** 7.2.3

**VignetteBuilder** knitr

**LazyData** true

**Encoding** UTF-8

**NeedsCompilation** no

**Author** Brian D. Williamson [aut, cre] (ORCID:  
<<https://orcid.org/0000-0002-7024-548X>>),  
Jean Feng [ctb],  
Charlie Wolock [ctb],  
Noah Simon [ths] (ORCID: <<https://orcid.org/0000-0002-8985-2474>>),  
Marco Carone [ths] (ORCID: <<https://orcid.org/0000-0003-2106-0953>>)

**Maintainer** Brian D. Williamson <brian.d.williamson@kp.org>

**Repository** CRAN

**Date/Publication** 2023-08-28 21:10:02 UTC

## Contents

average_vim . . . . .	3
bootstrap_se . . . . .	4
check_fitted_values . . . . .	6
check_inputs . . . . .	7
create_z . . . . .	8
cv_vim . . . . .	8
estimate . . . . .	15
estimate.predictiveness_measure . . . . .	15
estimate_eif_projection . . . . .	16
estimate_nuisances . . . . .	17
estimate_type_predictiveness . . . . .	18
est_predictiveness . . . . .	18
est_predictiveness_cv . . . . .	20
extract_sampled_split_predictions . . . . .	22
format.predictiveness_measure . . . . .	23
format.vim . . . . .	23
get_cv_sl_folds . . . . .	24
get_full_type . . . . .	24
get_test_set . . . . .	25
make_folds . . . . .	25
make_kfold . . . . .	26
measure_accuracy . . . . .	26
measure_anova . . . . .	28
measure_auc . . . . .	29
measure_average_value . . . . .	30
measure_cross_entropy . . . . .	32
measure_deviance . . . . .	33
measure_mse . . . . .	34
measure_r_squared . . . . .	36
merge_vim . . . . .	37
predictiveness_measure . . . . .	38
print.predictiveness_measure . . . . .	40
print.vim . . . . .	41
process_arg_lst . . . . .	41
run_sl . . . . .	42
sample_subsets . . . . .	43
scale_est . . . . .	44
spvim_ics . . . . .	44
spvim_se . . . . .	45
sp_vim . . . . .	46
vim . . . . .	50

<i>average_vim</i>	3
--------------------	---

<i>vimp_accuracy</i>	55
<i>vimp_anova</i>	60
<i>vimp_auc</i>	63
<i>vimp_ci</i>	68
<i>vimp_deviance</i>	69
<i>vimp_hypothesis_test</i>	74
<i>vimp_regression</i>	75
<i>vimp_rsquared</i>	78
<i>vimp_se</i>	83
<i>vrc01</i>	83

<b>Index</b>	<b>124</b>
--------------	------------

<i>average_vim</i>	<i>Average multiple independent importance estimates</i>
--------------------	--

**Description**

Average the output from multiple calls to *vimp\_regression*, for different independent groups, into a single estimate with a corresponding standard error and confidence interval.

**Usage**

```
average_vim(..., weights = rep(1/length(list(...)), length(list(...))))
```

**Arguments**

<i>...</i>	an arbitrary number of <i>vim</i> objects.
<i>weights</i>	how to average the vims together, and must sum to 1; defaults to 1/(number of vims) for each vim, corresponding to the arithmetic mean

**Value**

an object of class *vim* containing the (weighted) average of the individual importance estimates, as well as the appropriate standard error and confidence interval. This results in a list containing:

- *s* - a list of the column(s) to calculate variable importance for
- *SL.library* - a list of the libraries of learners passed to SuperLearner
- *full\_fit* - a list of the fitted values of the chosen method fit to the full data
- *red\_fit* - a list of the fitted values of the chosen method fit to the reduced data
- *est*- a vector with the corrected estimates
- *naive*- a vector with the naive estimates
- *update*- a list with the influence curve-based updates
- *mat* - a matrix with the estimated variable importance, the standard error, and the  $(1 - \alpha) \times 100\%$  confidence interval

- full\_mod - a list of the objects returned by the estimation procedure for the full data regression (if applicable)
- red\_mod - a list of the objects returned by the estimation procedure for the reduced data regression (if applicable)
- alpha - the level, for confidence interval calculation
- y - a list of the outcomes

### Examples

```
# generate the data
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- smooth + stats::rnorm(n, 0, 1)

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# get estimates on independent splits of the data
samp <- sample(1:n, n/2, replace = FALSE)

# using Super Learner (with a small number of folds, for illustration only)
est_2 <- vimp_regression(Y = y[samp], X = x[samp, ], indx = 2, V = 2,
  run_regression = TRUE, alpha = 0.05,
  SL.library = learners, cvControl = list(V = 2))

est_1 <- vimp_regression(Y = y[-samp], X = x[-samp, ], indx = 2, V = 2,
  run_regression = TRUE, alpha = 0.05,
  SL.library = learners, cvControl = list(V = 2))

ests <- average_vim(est_1, est_2, weights = c(1/2, 1/2))
```

---

bootstrap\_se

---

*Compute bootstrap-based standard error estimates for variable importance*


---

### Description

Compute bootstrap-based standard error estimates for variable importance

**Usage**

```
bootstrap_se(
  Y = NULL,
  f1 = NULL,
  f2 = NULL,
  cluster_id = NULL,
  clustered = FALSE,
  type = "r_squared",
  b = 1000,
  boot_interval_type = "perc",
  alpha = 0.05
)
```

**Arguments**

Y	the outcome.
f1	the fitted values from a flexible estimation technique regressing Y on X. A vector of the same length as Y; if sample-splitting is desired, then the value of f1 at each position should be the result of predicting from a model trained without that observation.
f2	the fitted values from a flexible estimation technique regressing either (a) f1 or (b) Y on X withholding the columns in <code>indx</code> . A vector of the same length as Y; if sample-splitting is desired, then the value of f2 at each position should be the result of predicting from a model trained without that observation.
cluster_id	vector of the same length as Y giving the cluster IDs used for the clustered bootstrap, if <code>clustered</code> is TRUE.
clustered	should the bootstrap resamples be performed on clusters rather than individual observations? Defaults to FALSE.
type	the type of importance to compute; defaults to <code>r_squared</code> , but other supported options are <code>auc</code> , <code>accuracy</code> , <code>deviance</code> , and <code>anova</code> .
b	the number of bootstrap replicates (only used if <code>bootstrap</code> = TRUE and <code>sample_splitting</code> = FALSE); defaults to 1000.
boot_interval_type	the type of bootstrap interval (one of "norm", "basic", "stud", "perc", or "bca", as in <code>boot{boot.ci}</code> ) if requested. Defaults to "perc".
alpha	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.

**Value**

a bootstrap-based standard error estimate

---

check\_fitted\_values      *Check pre-computed fitted values for call to vim, cv\_vim, or sp\_vim*

---

### Description

Check pre-computed fitted values for call to vim, cv\_vim, or sp\_vim

### Usage

```
check_fitted_values(
  Y = NULL,
  f1 = NULL,
  f2 = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  sample_splitting_folds = NULL,
  cross_fitting_folds = NULL,
  cross_fitted_se = TRUE,
  V = NULL,
  ss_V = NULL,
  cv = FALSE
)
```

### Arguments

Y	the outcome
f1	estimator of the population-optimal prediction function using all covariates
f2	estimator of the population-optimal prediction function using the reduced set of covariates
cross_fitted_f1	cross-fitted estimator of the population-optimal prediction function using all covariates
cross_fitted_f2	cross-fitted estimator of the population-optimal prediction function using the reduced set of covariates
sample_splitting_folds	the folds for sample-splitting (used for hypothesis testing)
cross_fitting_folds	the folds for cross-fitting (used for point estimates of variable importance in cv_vim and sp_vim)
cross_fitted_se	logical; should cross-fitting be used to estimate standard errors?
V	the number of cross-fitting folds
ss_V	the number of folds for CV (if sample_splitting is TRUE)
cv	a logical flag indicating whether or not to use cross-fitting

**Details**

Ensure that inputs to `vim`, `cv_vim`, and `sp_vim` follow the correct formats.

**Value**

None. Called for the side effect of stopping the algorithm if any inputs are in an unexpected format.

---

check_inputs	<i>Check inputs to a call to <code>vim</code>, <code>cv_vim</code>, or <code>sp_vim</code></i>
--------------	--

---

**Description**

Check inputs to a call to `vim`, `cv_vim`, or `sp_vim`

**Usage**

```
check_inputs(Y, X, f1, f2, indx)
```

**Arguments**

<code>Y</code>	the outcome
<code>X</code>	the covariates
<code>f1</code>	estimator of the population-optimal prediction function using all covariates
<code>f2</code>	estimator of the population-optimal prediction function using the reduced set of covariates
<code>indx</code>	the index or indices of the covariate(s) of interest

**Details**

Ensure that inputs to `vim`, `cv_vim`, and `sp_vim` follow the correct formats.

**Value**

None. Called for the side effect of stopping the algorithm if any inputs are in an unexpected format.

---

create_z	<i>Create complete-case outcome, weights, and Z</i>
----------	---

---

### Description

Create complete-case outcome, weights, and Z

### Usage

```
create_z(Y, C, Z, X, ipc_weights)
```

### Arguments

Y	the outcome
C	indicator of missing or observed
Z	the covariates observed in phase 1 and 2 data
X	all covariates
ipc_weights	the weights

### Value

a list, with the complete-case outcome, weights, and Z matrix

---

cv_vim	<i>Nonparametric Intrinsic Variable Importance Estimates and Inference using Cross-fitting</i>
--------	--

---

### Description

Compute estimates and confidence intervals using cross-fitting for nonparametric intrinsic variable importance based on the population-level contrast between the oracle predictiveness using the feature(s) of interest versus not.

### Usage

```
cv_vim(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  V = ifelse(is.null(cross_fitting_folds), 5, length(unique(cross_fitting_folds))),
```



```

sample_splitting = TRUE,
final_point_estimate = "split",
sample_splitting_folds = NULL,
cross_fitting_folds = NULL,
stratified = FALSE,
type = "r_squared",
run_regression = TRUE,
SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
alpha = 0.05,
delta = 0,
scale = "identity",
na.rm = FALSE,
C = rep(1, length(Y)),
Z = NULL,
ipc_scale = "identity",
ipc_weights = rep(1, length(Y)),
ipc_est_type = "aipw",
scale_est = TRUE,
nuisance_estimators_full = NULL,
nuisance_estimators_reduced = NULL,
exposure_name = NULL,
cross_fitted_se = TRUE,
bootstrap = FALSE,
b = 1000,
boot_interval_type = "perc",
clustered = FALSE,
cluster_id = rep(NA, length(Y)),
...
)

```

## Arguments

- Y** the outcome.
- X** the covariates. If `type = "average_value"`, then the exposure variable should be part of X, with its name provided in `exposure_name`.
- cross\_fitted\_f1** the predicted values on validation data from a flexible estimation technique regressing Y on X in the training data. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
- cross\_fitted\_f2** the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in `cross_fitted_f1`, or (b) Y, on X with-

	holding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length <code>V</code> , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as <code>Y</code> ; if using a list, then the summed length of each element across the list should be the same length as <code>Y</code> (i.e., each observation is included in the predictions).
<code>f1</code>	the fitted values from a flexible estimation technique regressing <code>Y</code> on <code>X</code> . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>f2</code>	the fitted values from a flexible estimation technique regressing either (a) <code>f1</code> or (b) <code>Y</code> on <code>X</code> withholding the columns in <code>indx</code> . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of <code>V</code> -fold cross-fitting is done. See Details for a more detailed explanation.
<code>sample_splitting</code>	should we use sample-splitting to estimate the full and reduced predictiveness? Defaults to <code>TRUE</code> , since inferences made using <code>sample_splitting = FALSE</code> will be invalid for variables with truly zero importance.
<code>final_point_estimate</code>	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference (" <code>split</code> ", the default), or should they instead be based on the full dataset (" <code>full</code> ") or the average across the point estimates from each sample split (" <code>average</code> ")? All three options result in valid point estimates – sample-splitting is only required for valid inference.
<code>sample_splitting_folds</code>	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if <code>run_regression = FALSE</code> .
<code>cross_fitting_folds</code>	the folds for cross-fitting. Only used if <code>run_regression = FALSE</code> .
<code>stratified</code>	if <code>run_regression = TRUE</code> , then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
<code>type</code>	the type of importance to compute; defaults to <code>r_squared</code> , but other supported options are <code>auc</code> , <code>accuracy</code> , <code>deviance</code> , and <code>anova</code> .
<code>run_regression</code>	if outcome <code>Y</code> and covariates <code>X</code> are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is <code>TRUE</code> , then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to SuperLearner, if <code>f1</code> and <code>f2</code> are <code>Y</code> and <code>X</code> , respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.

delta	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
na.rm	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_scale	what scale should the inverse probability weight correction be applied on (if any)? Defaults to "identity". (other options are "log" and "logit")
ipc_weights	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
nuisance_estimators_full	(only used if type = "average_value") a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.
nuisance_estimators_reduced	(only used if type = "average_value") a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.
exposure_name	(only used if type = "average_value") the name of the exposure of interest; binary, with 1 indicating presence of the exposure and 0 indicating absence of the exposure.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
bootstrap	should bootstrap-based standard error estimates be computed? Defaults to FALSE (and currently may only be used if sample_splitting = FALSE).
b	the number of bootstrap replicates (only used if bootstrap = TRUE and sample_splitting = FALSE); defaults to 1000.

boot_interval_type	the type of bootstrap interval (one of "norm", "basic", "stud", "perc", or "bca", as in <code>boot{boot.ci}</code> ) if requested. Defaults to "perc".
clustered	should the bootstrap resamples be performed on clusters rather than individual observations? Defaults to FALSE.
cluster_id	vector of the same length as <code>Y</code> giving the cluster IDs used for the clustered bootstrap, if <code>clustered</code> is TRUE.
...	other arguments to the estimation tool, see "See also".

### Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution.

Cross-fitted VIM estimates are computed differently if sample-splitting is requested versus if it is not. We recommend using sample-splitting in most cases, since only in this case will inferences be valid if the variable(s) of interest have truly zero population importance. The purpose of cross-fitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data from estimating  $P_0$ ; this can result in improved performance, especially when using flexible learning algorithms. The purpose of sample-splitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data; this allows valid inference under the null hypothesis of zero importance.

Without sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $P_{n,k}$  of  $P_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{V(f_{n,k}, P_{n,k}) - V(f_{n,k,s}, P_{n,k})\}.$$

With sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $2K$  folds. These folds are further divided into 2 groups of folds. Then, for each fold  $k$  in the first group, estimator  $f_{n,k}$  of  $f_0$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K-1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k} = V(f_{n,k}, P_{n,k}).$$

Similarly, for each fold  $k$  in the second group, estimator  $f_{n,k,s}$  of  $f_{0,s}$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K-1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k,s} = V(f_{n,k,s}, P_{n,k}).$$

Finally,

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{v_{n,k} - v_{n,k,s}\}.$$

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind the `cv_vim` function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

**s** the column(s) to calculate variable importance for

**SL.library** the library of learners passed to SuperLearner

**full\_fit** the fitted values of the chosen method fit to the full data (a list, for train and test data)

**red\_fit** the fitted values of the chosen method fit to the reduced data (a list, for train and test data)

**est** the estimated variable importance

**naive** the naive estimator of variable importance

**ef** the estimated efficient influence function

**ef\_full** the estimated efficient influence function for the full regression

**ef\_reduced** the estimated efficient influence function for the reduced regression

**se** the standard error for the estimated variable importance

**ci** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate

**test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test

**p\_value** a p-value based on the same test as `test`

**full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)

**red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)

**alpha** the level, for confidence interval calculation

**sample\_splitting\_folds** the folds used for hypothesis testing

**cross\_fitting\_folds** the folds used for cross-fitting

**y** the outcome

**ipc\_weights** the weights

**cluster\_id** the cluster IDs

**mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

## Value

An object of class `vim`. See Details for more information.

## See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package.

## Examples

```

n <- 100
p <- 2
# generate the data
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- as.matrix(smooth + stats::rnorm(n, 0, 1))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm")

# -----
# using Super Learner (with a small number of folds, for illustration only)
# -----
set.seed(4747)
est <- cv_vim(Y = y, X = x, indx = 2, V = 2,
type = "r_squared", run_regression = TRUE,
SL.library = learners, cvControl = list(V = 2), alpha = 0.05)

# -----
# doing things by hand, and plugging them in
# (with a small number of folds, for illustration only)
# -----
# set up the folds
indx <- 2
V <- 2
Y <- matrix(y)
set.seed(4747)
# Note that the CV.SuperLearner should be run with an outer layer
# of 2*V folds (for V-fold cross-fitted importance)
full_cv_fit <- suppressWarnings(SuperLearner::CV.SuperLearner(
Y = Y, X = x, SL.library = learners, cvControl = list(V = 2 * V),
innerCvControl = list(list(V = V))
))
full_cv_preds <- full_cv_fit$SL.predict
# use the same cross-fitting folds for reduced
reduced_cv_fit <- suppressWarnings(SuperLearner::CV.SuperLearner(
Y = Y, X = x[, -indx, drop = FALSE], SL.library = learners,
cvControl = SuperLearner::SuperLearner.CV.control(
V = 2 * V, validRows = full_cv_fit$folds
),
innerCvControl = list(list(V = V))
))
reduced_cv_preds <- reduced_cv_fit$SL.predict
# for hypothesis testing
cross_fitting_folds <- get_cv_sl_folds(full_cv_fit$folds)
set.seed(1234)

```

```

sample_splitting_folds <- make_folds(unique(cross_fitting_folds), V = 2)
set.seed(5678)
est <- cv_vim(Y = y, cross_fitted_f1 = full_cv_preds,
cross_fitted_f2 = reduced_cv_preds, indx = 2, delta = 0, V = V, type = "r_squared",
cross_fitting_folds = cross_fitting_folds,
sample_splitting_folds = sample_splitting_folds,
run_regression = FALSE, alpha = 0.05, na.rm = TRUE)

```

---

estimate

*Estimate a Predictiveness Measure*


---

### Description

Generic function for estimating a predictiveness measure (e.g., R-squared or classification accuracy).

### Usage

```
estimate(x, ...)
```

### Arguments

x	An R object. Currently, there are methods for predictiveness_measure objects only.
...	further arguments passed to or from other methods.

---

estimate.predictiveness\_measure

*Obtain a Point Estimate and Efficient Influence Function Estimate for a Given Predictiveness Measure*


---

### Description

Obtain a Point Estimate and Efficient Influence Function Estimate for a Given Predictiveness Measure

### Usage

```
## S3 method for class 'predictiveness_measure'
estimate(x, ...)
```

### Arguments

x	an object of class "predictiveness_measure"
...	other arguments to type-specific predictiveness measures (currently unused)

**Value**

A list with the point estimate, naive point estimate (for ANOVA only), estimated EIF, and the predictions for coarsened data EIF (for coarsened data settings only)

---

```
estimate_eif_projection
```

*Estimate projection of EIF on fully-observed variables*

---

**Description**

Estimate projection of EIF on fully-observed variables

**Usage**

```
estimate_eif_projection(
  obs_grad = NULL,
  C = NULL,
  Z = NULL,
  ipc_fit_type = NULL,
  ipc_eif_preds = NULL,
  ...
)
```

**Arguments**

<code>obs_grad</code>	the estimated (observed) EIF
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
<code>ipc_fit_type</code>	if "external", then use <code>ipc_eif_preds</code> ; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
<code>ipc_eif_preds</code>	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
<code>...</code>	other arguments to SuperLearner, if <code>ipc_fit_type = "SL"</code> .

**Value**

the projection of the EIF onto the fully-observed variables



---

estimate_nuisances	<i>Estimate nuisance functions for average value-based VIMs</i>
--------------------	---

---

## Description

Estimate nuisance functions for average value-based VIMs

## Usage

```
estimate_nuisances(
  fit,
  X,
  exposure_name,
  V = 1,
  SL.library,
  sample_splitting,
  sample_splitting_folds,
  verbose,
  weights,
  cross_fitted_se,
  split = 1,
  ...
)
```

## Arguments

<code>fit</code>	the fitted nuisance function estimator
<code>X</code>	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of <code>X</code> , with its name provided in <code>exposure_name</code> .
<code>exposure_name</code>	(only used if <code>type = "average_value"</code> ) the name of the exposure of interest; binary, with 1 indicating presence of the exposure and 0 indicating absence of the exposure.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of V-fold cross-fitting is done. See Details for a more detailed explanation.
<code>SL.library</code>	a character vector of learners to pass to SuperLearner, if <code>f1</code> and <code>f2</code> are Y and X, respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>sample_splitting</code>	should we use sample-splitting to estimate the full and reduced predictiveness? Defaults to <code>TRUE</code> , since inferences made using <code>sample_splitting = FALSE</code> will be invalid for variables with truly zero importance.
<code>sample_splitting_folds</code>	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if <code>run_regression = FALSE</code> .

verbose	should we print progress? defaults to FALSE
weights	weights to pass to estimation procedure
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
split	the sample split to use
...	other arguments to the estimation tool, see "See also".

**Value**

nuisance function estimators for use in the average value VIM: the treatment assignment based on the estimated optimal rule (based on the estimated outcome regression); the expected outcome under the estimated optimal rule; and the estimated propensity score.

---

estimate\_type\_predictiveness

*Estimate Predictiveness Given a Type*

---

**Description**

Estimate the specified type of predictiveness

**Usage**

```
estimate_type_predictiveness(arg_lst, type)
```

**Arguments**

arg_lst	a list of arguments; from, e.g., predictiveness_measure
type	the type of predictiveness, e.g., "r_squared"

---

est\_predictiveness

*Estimate a nonparametric predictiveness functional*

---

**Description**

Compute nonparametric estimates of the chosen measure of predictiveness.

**Usage**

```
est_predictiveness(
  fitted_values,
  y,
  a = NULL,
  full_y = NULL,
  type = "r_squared",
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(C)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(C)),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  ...
)
```

**Arguments**

<code>fitted_values</code>	fitted values from a regression function using the observed data.
<code>y</code>	the observed outcome.
<code>a</code>	the observed treatment assignment (may be within a specified fold, for cross-fitted estimates). Only used if <code>type = "average_value"</code> .
<code>full_y</code>	the observed outcome (from the entire dataset, for cross-fitted estimates).
<code>type</code>	which parameter are you estimating (defaults to <code>r_squared</code> , for R-squared-based variable importance)?
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either <code>NULL</code> (if no coarsening) or a matrix-like object containing the fully observed data.
<code>ipc_weights</code>	weights for inverse probability of coarsening (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
<code>ipc_fit_type</code>	if <code>"external"</code> , then use <code>ipc_eif_preds</code> ; if <code>"SL"</code> , fit a <code>SuperLearner</code> to determine the correction to the efficient influence function.
<code>ipc_eif_preds</code>	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
<code>ipc_est_type</code>	IPC correction, either <code>"ipw"</code> (for classical inverse probability weighting) or <code>"aipw"</code> (for augmented inverse probability weighting; the default).
<code>scale</code>	if doing an IPC correction, then the scale that the correction should be computed on (e.g., <code>"identity"</code> ; or <code>"logit"</code> to logit-transform, apply the correction, and back-transform).
<code>na.rm</code>	logical; should NA's be removed in computation? (defaults to <code>FALSE</code> )

nuisance\_estimators  
 (only used if type = "average\_value") a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.

... other arguments to SuperLearner, if ipc\_fit\_type = "SL".

### Details

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function and the definition of the parameter of interest.

### Value

A list, with: the estimated predictiveness; the estimated efficient influence function; and the predictions of the EIF based on inverse probability of censoring.

---

est\_predictiveness\_cv *Estimate a nonparametric predictiveness functional using cross-fitting*

---

### Description

Compute nonparametric estimates of the chosen measure of predictiveness.

### Usage

```
est_predictiveness_cv(
  fitted_values,
  y,
  full_y = NULL,
  folds,
  type = "r_squared",
  C = rep(1, length(y)),
  Z = NULL,
  folds_Z = folds,
  ipc_weights = rep(1, length(C)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(C)),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = FALSE,
  ...
)
```

**Arguments**

fitted_values	fitted values from a regression function using the observed data; a list of length $V$ , where each object is a set of predictions on the validation data, or a vector of the same length as $y$ .
y	the observed outcome.
full_y	the observed outcome (from the entire dataset, for cross-fitted estimates).
folds	the cross-validation folds for the observed data.
type	which parameter are you estimating (defaults to <code>r_squared</code> , for R-squared-based variable importance)?
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
folds_Z	either the cross-validation folds for the observed data (no coarsening) or a vector of folds for the fully observed data $Z$ .
ipc_weights	weights for inverse probability of coarsening (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
ipc_fit_type	if "external", then use <code>ipc_eif_preds</code> ; if "SL", fit a SuperLearner to determine the correction to the efficient influence function.
ipc_eif_preds	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NA's be removed in computation? (defaults to FALSE)
...	other arguments to SuperLearner, if <code>ipc_fit_type = "SL"</code> .

**Details**

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function and the definition of the parameter of interest. If sample-splitting is also requested (recommended, since in this case inferences will be valid even if the variable has zero true importance), then the prediction functions are trained as if  $2K$ -fold cross-validation were run, but are evaluated on only  $K$  sets (independent between the full and reduced nuisance regression).

**Value**

The estimated measure of predictiveness.

---

```
extract_sampled_split_predictions
```

*Extract sampled-split predictions from a CV.SuperLearner object*

---

## Description

Use the cross-validated Super Learner and a set of specified sample-splitting folds to extract cross-fitted predictions on separate splits of the data. This is primarily for use in cases where you have already fit a CV.SuperLearner and want to use the fitted values to compute variable importance without having to re-fit. The number of folds used in the CV.SuperLearner must be even.

## Usage

```
extract_sampled_split_predictions(
  cvsl_obj = NULL,
  sample_splitting = TRUE,
  sample_splitting_folds = NULL,
  full = TRUE,
  preds = NULL,
  cross_fitting_folds = NULL,
  vector = TRUE
)
```

## Arguments

cvsl_obj	An object of class "CV.SuperLearner"; must be entered unless preds is specified.
sample_splitting	logical; should we use sample-splitting or not? Defaults to TRUE.
sample_splitting_folds	A vector of folds to use for sample splitting
full	logical; is this the fit to all covariates (TRUE) or not (FALSE)?
preds	a vector of predictions; must be entered unless cvsl_obj is specified.
cross_fitting_folds	a vector of folds that were used in cross-fitting.
vector	logical; should we return a vector (where each element is the prediction when the corresponding row is in the validation fold) or a list?

## Value

The predictions on validation data in each split-sample fold.

## See Also

[CV.SuperLearner](#) for usage of the CV.SuperLearner function.

---

`format.predictiveness_measure`*Format a predictiveness\_measure object*

---

### Description

Nicely formats the output from a predictiveness\_measure object for printing.

### Usage

```
## S3 method for class 'predictiveness_measure'  
format(x, ...)
```

### Arguments

<code>x</code>	the predictiveness_measure object of interest.
<code>...</code>	other options, see the generic format function.

---

`format.vim`*Format a vim object*

---

### Description

Nicely formats the output from a vim object for printing.

### Usage

```
## S3 method for class 'vim'  
format(x, ...)
```

### Arguments

<code>x</code>	the vim object of interest.
<code>...</code>	other options, see the generic format function.

---

get_cv_sl_folds	<i>Get a numeric vector with cross-validation fold IDs from CV.SuperLearner</i>
-----------------	---

---

**Description**

Get a numeric vector with cross-validation fold IDs from CV.SuperLearner

**Usage**

```
get_cv_sl_folds(cv_sl_folds)
```

**Arguments**

cv\_sl\_folds      The folds from a call to CV.SuperLearner; a list.

**Value**

A numeric vector with the fold IDs.

---

get_full_type	<i>Obtain the type of VIM to estimate using partial matching</i>
---------------	--

---

**Description**

Obtain the type of VIM to estimate using partial matching

**Usage**

```
get_full_type(type)
```

**Arguments**

type              the partial string indicating the type of VIM

**Value**

the full string indicating the type of VIM



---

get_test_set	<i>Return test-set only data</i>
--------------	----------------------------------

---

**Description**

Return test-set only data

**Usage**

```
get_test_set(arg_lst, k)
```

**Arguments**

arg_lst	a list of estimates, data, etc.
k	the index of interest

**Value**

the test-set only data

---

make_folds	<i>Create Folds for Cross-Fitting</i>
------------	---------------------------------------

---

**Description**

Create Folds for Cross-Fitting

**Usage**

```
make_folds(y, V = 2, stratified = FALSE, C = NULL, probs = rep(1/V, V))
```

**Arguments**

y	the outcome
V	the number of folds
stratified	should the folds be stratified based on the outcome?
C	a vector indicating whether or not the observation is fully observed; 1 denotes yes, 0 denotes no
probs	vector of proportions for each fold number

**Value**

a vector of folds

---

make_kfold	<i>Turn folds from 2K-fold cross-fitting into individual K-fold folds</i>
------------	---

---

**Description**

Turn folds from 2K-fold cross-fitting into individual K-fold folds

**Usage**

```
make_kfold(
  cross_fitting_folds,
  sample_splitting_folds = rep(1, length(unique(cross_fitting_folds))),
  C = rep(1, length(cross_fitting_folds))
)
```

**Arguments**

cross_fitting_folds	the vector of cross-fitting folds
sample_splitting_folds	the sample splitting folds
C	vector of whether or not we measured the observation in phase 2

**Value**

the two sets of testing folds for K-fold cross-fitting

---

measure_accuracy	<i>Estimate the classification accuracy</i>
------------------	---

---

**Description**

Compute nonparametric estimate of classification accuracy.

**Usage**

```
measure_accuracy(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
)
```

```

    ipc_est_type = "aipw",
    scale = "logit",
    na.rm = FALSE,
    nuisance_estimators = NULL,
    a = NULL,
    ...
)

```

## Arguments

<code>fitted_values</code>	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
<code>y</code>	the observed outcome (may be within a specified fold, for cross-fitted estimates).
<code>full_y</code>	the observed outcome (not used, defaults to NULL).
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
<code>ipc_weights</code>	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $ipc\_weights = 1 / [\text{estimated probability weights}]$ ).
<code>ipc_fit_type</code>	if "external", then use <code>ipc_eif_preds</code> ; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
<code>ipc_eif_preds</code>	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
<code>ipc_est_type</code>	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
<code>scale</code>	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
<code>na.rm</code>	logical; should NAs be removed in computation? (defaults to FALSE)
<code>nuisance_estimators</code>	not used; for compatibility with <code>measure_average_value</code> .
<code>a</code>	not used; for compatibility with <code>measure_average_value</code> .
<code>...</code>	other arguments to SuperLearner, if <code>ipc_fit_type = "SL"</code> .

## Value

A named list of: (1) the estimated classification accuracy of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

measure\_anova

*Estimate ANOVA decomposition-based variable importance.***Description**

Estimate ANOVA decomposition-based variable importance.

**Usage**

```
measure_anova(
  full,
  reduced,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "logit",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

full	fitted values from a regression function of the observed outcome on the full set of covariates.
reduced	fitted values from a regression on the reduced set of observed covariates.
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.

ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
nuisance_estimators	not used; for compatibility with measure_average_value.
a	not used; for compatibility with measure_average_value.
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

### Value

A named list of: (1) the estimated ANOVA (based on a one-step correction) of the fitted regression functions; (2) the estimated influence function; (3) the naive ANOVA estimate; and (4) the IPC EIF predictions.

---

measure_auc	<i>Estimate area under the receiver operating characteristic curve (AUC)</i>
-------------	--

---

### Description

Compute nonparametric estimate of AUC.

### Usage

```
measure_auc(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "logit",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

fitted_values	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
nuisance_estimators	not used; for compatibility with measure_average_value.
a	not used; for compatibility with measure_average_value.
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

**Value**

A named list of: (1) the estimated AUC of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

measure\_average\_value    *Estimate the average value under the optimal treatment rule*

---

**Description**

Compute nonparametric estimate of the average value under the optimal treatment rule.

**Usage**

```
measure_average_value(
  nuisance_estimators,
  y,
  a,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = FALSE,
  ...
)
```

**Arguments**

nuisance_estimators	a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
a	the observed treatment assignment (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

**Value**

A named list of: (1) the estimated classification accuracy of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

measure\_cross\_entropy *Estimate the cross-entropy*

---

**Description**

Compute nonparametric estimate of cross-entropy.

**Usage**

```
measure_cross_entropy(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

fitted_values	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.



ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
nuisance_estimators	not used; for compatibility with measure_average_value.
a	not used; for compatibility with measure_average_value.
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

**Value**

A named list of: (1) the estimated cross-entropy of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

measure_deviance	<i>Estimate the deviance</i>
------------------	------------------------------

---

**Description**

Compute nonparametric estimate of deviance.

**Usage**

```
measure_deviance(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "logit",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

fitted_values	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
nuisance_estimators	not used; for compatibility with measure_average_value.
a	not used; for compatibility with measure_average_value.
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

**Value**

A named list of: (1) the estimated deviance of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

measure\_mse

---

*Estimate mean squared error*


---

**Description**

Compute nonparametric estimate of mean squared error.

**Usage**

```
measure_mse(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

<code>fitted_values</code>	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
<code>y</code>	the observed outcome (may be within a specified fold, for cross-fitted estimates).
<code>full_y</code>	the observed outcome (not used, defaults to <code>NULL</code> ).
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either <code>NULL</code> (if no coarsening) or a matrix-like object containing the fully observed data.
<code>ipc_weights</code>	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
<code>ipc_fit_type</code>	if "external", then use <code>ipc_eif_preds</code> ; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
<code>ipc_eif_preds</code>	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
<code>ipc_est_type</code>	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
<code>scale</code>	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
<code>na.rm</code>	logical; should NAs be removed in computation? (defaults to <code>FALSE</code> )
<code>nuisance_estimators</code>	not used; for compatibility with <code>measure_average_value</code> .
<code>a</code>	not used; for compatibility with <code>measure_average_value</code> .
<code>...</code>	other arguments to SuperLearner, if <code>ipc_fit_type = "SL"</code> .

**Value**

A named list of: (1) the estimated mean squared error of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

measure_r_squared	<i>Estimate R-squared</i>
-------------------	---------------------------

---

**Description**

Estimate R-squared

**Usage**

```
measure_r_squared(
  fitted_values,
  y,
  full_y = NULL,
  C = rep(1, length(y)),
  Z = NULL,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "external",
  ipc_eif_preds = rep(1, length(y)),
  ipc_est_type = "aipw",
  scale = "logit",
  na.rm = FALSE,
  nuisance_estimators = NULL,
  a = NULL,
  ...
)
```

**Arguments**

fitted_values	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
y	the observed outcome (may be within a specified fold, for cross-fitted estimates).
full_y	the observed outcome (not used, defaults to NULL).
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either NULL (if no coarsening) or a matrix-like object containing the fully observed data.
ipc_weights	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_fit_type	if "external", then use ipc_eif_preds; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.

ipc_eif_preds	if ipc_fit_type = "external", the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
ipc_est_type	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
scale	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
na.rm	logical; should NAs be removed in computation? (defaults to FALSE)
nuisance_estimators	not used; for compatibility with measure_average_value.
a	not used; for compatibility with measure_average_value.
...	other arguments to SuperLearner, if ipc_fit_type = "SL".

**Value**

A named list of: (1) the estimated R-squared of the fitted regression function; (2) the estimated influence function; and (3) the IPC EIF predictions.

---

merge\_vim

---

Merge multiple vim objects into one

---

**Description**

Take the output from multiple different calls to vimp\_regression and merge into a single vim object; mostly used for plotting results.

**Usage**

```
merge_vim(...)
```

**Arguments**

... an arbitrary number of vim objects, separated by commas.

**Value**

an object of class vim containing all of the output from the individual vim objects. This results in a list containing:

- s - a list of the column(s) to calculate variable importance for
- SL.library - a list of the libraries of learners passed to SuperLearner
- full\_fit - a list of the fitted values of the chosen method fit to the full data
- red\_fit - a list of the fitted values of the chosen method fit to the reduced data
- est- a vector with the corrected estimates
- naive- a vector with the naive estimates

- eif- a list with the influence curve-based updates
- se- a vector with the standard errors
- ci- a matrix with the CIs
- mat - a tibble with the estimated variable importance, the standard errors, and the  $(1 - \alpha) \times 100\%$  confidence intervals
- full\_mod - a list of the objects returned by the estimation procedure for the full data regression (if applicable)
- red\_mod - a list of the objects returned by the estimation procedure for the reduced data regression (if applicable)
- alpha - a list of the levels, for confidence interval calculation

### Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- smooth + stats::rnorm(n, 0, 1)

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# using Super Learner (with a small number of folds, for illustration only)
est_2 <- vimp_regression(Y = y, X = x, indx = 2, V = 2,
  run_regression = TRUE, alpha = 0.05,
  SL.library = learners, cvControl = list(V = 2))

est_1 <- vimp_regression(Y = y, X = x, indx = 1, V = 2,
  run_regression = TRUE, alpha = 0.05,
  SL.library = learners, cvControl = list(V = 2))

ests <- merge_vim(est_1, est_2)
```

---

predictiveness\_measure

*Construct a Predictiveness Measure*

---

### Description

Construct a Predictiveness Measure

**Usage**

```

predictiveness_measure(
  type = character(),
  y = numeric(),
  a = numeric(),
  fitted_values = numeric(),
  cross_fitting_folds = rep(1, length(fitted_values)),
  full_y = NULL,
  nuisance_estimators = list(),
  C = rep(1, length(y)),
  Z = NULL,
  folds_Z = cross_fitting_folds,
  ipc_weights = rep(1, length(y)),
  ipc_fit_type = "SL",
  ipc_eif_preds = numeric(),
  ipc_est_type = "aipw",
  scale = "identity",
  na.rm = TRUE,
  ...
)

```

**Arguments**

<code>type</code>	the measure of interest (e.g., "accuracy", "auc", "r_squared")
<code>y</code>	the outcome of interest
<code>a</code>	the exposure of interest (only used if <code>type = "average_value"</code> )
<code>fitted_values</code>	fitted values from a regression function using the observed data (may be within a specified fold, for cross-fitted estimates).
<code>cross_fitting_folds</code>	<p>folds for cross-fitting, if used to obtain the fitted values. If not used, a vector of ones.</p>
<code>full_y</code>	the observed outcome (not used, defaults to <code>NULL</code> ).
<code>nuisance_estimators</code>	<p>a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). For the average value measure: an estimator of the optimal treatment rule (<code>f_n</code>); an estimator of the propensity score under the estimated optimal treatment rule (<code>g_n</code>); and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule (<code>q_n</code>).</p>
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either <code>NULL</code> (if no coarsening) or a matrix-like object containing the fully observed data.
<code>folds_Z</code>	either the cross-validation folds for the observed data (no coarsening) or a vector of folds for the fully observed data <code>Z</code> .

<code>ipc_weights</code>	weights for inverse probability of coarsening (IPC) (e.g., inverse weights from a two-phase sample) weighted estimation. Assumed to be already inverted. (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
<code>ipc_fit_type</code>	if "external", then use <code>ipc_eif_preds</code> ; if "SL", fit a SuperLearner to determine the IPC correction to the efficient influence function.
<code>ipc_eif_preds</code>	if <code>ipc_fit_type = "external"</code> , the fitted values from a regression of the full-data EIF on the fully observed covariates/outcome; otherwise, not used.
<code>ipc_est_type</code>	IPC correction, either "ipw" (for classical inverse probability weighting) or "aipw" (for augmented inverse probability weighting; the default).
<code>scale</code>	if doing an IPC correction, then the scale that the correction should be computed on (e.g., "identity"; or "logit" to logit-transform, apply the correction, and back-transform).
<code>na.rm</code>	logical; should NAs be removed in computation? (defaults to FALSE)
<code>...</code>	other arguments to SuperLearner, if <code>ipc_fit_type = "SL"</code> .

**Value**

An object of class "predictiveness\_measure", with the following attributes:

---

```
print.predictiveness_measure
```

*Print predictiveness\_measure objects*

---

**Description**

Prints out a table of the point estimate and standard error for a `predictiveness_measure` object.

**Usage**

```
## S3 method for class 'predictiveness_measure'
print(x, ...)
```

**Arguments**

`x` the `predictiveness_measure` object of interest.

`...` other options, see the generic `print` function.



---

print.vim	<i>Print vim objects</i>
-----------	--------------------------

---

**Description**

Prints out the table of estimates, confidence intervals, and standard errors for a vim object.

**Usage**

```
## S3 method for class 'vim'
print(x, ...)
```

**Arguments**

x	the vim object of interest.
...	other options, see the generic print function.

---

process_arg_lst	<i>Process argument list for Super Learner estimation of the EIF</i>
-----------------	--

---

**Description**

Process argument list for Super Learner estimation of the EIF

**Usage**

```
process_arg_lst(arg_lst)
```

**Arguments**

arg_lst	the list of arguments for Super Learner
---------	---

**Value**

a list of modified arguments for EIF estimation

run\_sl

*Run a Super Learner for the provided subset of features***Description**

Run a Super Learner for the provided subset of features

**Usage**

```
run_sl(
  Y = NULL,
  X = NULL,
  V = 5,
  SL.library = "SL.glm",
  univariate_SL.library = NULL,
  s = 1,
  cv_folds = NULL,
  sample_splitting = TRUE,
  ss_folds = NULL,
  split = 1,
  verbose = FALSE,
  progress_bar = NULL,
  indx = 1,
  weights = rep(1, nrow(X)),
  cross_fitted_se = TRUE,
  full = NULL,
  vector = TRUE,
  ...
)
```

**Arguments**

Y	the outcome
X	the covariates
V	the number of folds
SL.library	the library of candidate learners
univariate_SL.library	the library of candidate learners for single-covariate regressions
s	the subset of interest
cv_folds	the CV folds
sample_splitting	logical; should we use sample-splitting for predictiveness estimation?
ss_folds	the sample-splitting folds; only used if sample_splitting = TRUE
split	the split to use for sample-splitting; only used if sample_splitting = TRUE

verbose	should we print progress? defaults to FALSE
progress_bar	the progress bar to print to (only if verbose = TRUE)
indx	the index to pass to progress bar (only if verbose = TRUE)
weights	weights to pass to estimation procedure
cross_fitted_se	if TRUE, uses a cross-fitted estimator of the standard error; otherwise, uses the entire dataset
full	should this be considered a "full" or "reduced" regression? If NULL (the default), this is determined automatically; a full regression corresponds to s being equal to the full covariate vector. For SPVIMs, can be entered manually.
vector	should we return a vector (TRUE) or a list (FALSE)?
...	other arguments to Super Learner

**Value**

a list of length V, with the results of predicting on the hold-out data for each v in 1 through V

---

sample_subsets	<i>Create necessary objects for SPVIMs</i>
----------------	--

---

**Description**

Creates the Z and W matrices and a list of sampled subsets, S, for SPVIM estimation.

**Usage**

```
sample_subsets(p, gamma, n)
```

**Arguments**

p	the number of covariates
gamma	the fraction of the sample size to sample (e.g., gamma = 1 means sample n subsets)
n	the sample size

**Value**

a list, with elements Z (the matrix encoding presence/absence of each feature in the uniquely sampled subsets), S (the list of unique sampled subsets), W (the matrix of weights), and z\_counts (the number of times each subset was sampled)

**Examples**

```
p <- 10
gamma <- 1
n <- 100
set.seed(100)
subset_lst <- sample_subsets(p, gamma, n)
```

---

scale_est	<i>Return an estimator on a different scale</i>
-----------	---

---

### Description

Return an estimator on a different scale

### Usage

```
scale_est(obs_est = NULL, grad = NULL, scale = "identity")
```

### Arguments

obs_est	the observed VIM estimate
grad	the estimated efficient influence function
scale	the scale to compute on

### Details

It may be of interest to return an estimate (or confidence interval) on a different scale than originally measured. For example, computing a confidence interval (CI) for a VIM value that lies in (0,1) on the logit scale ensures that the CI also lies in (0, 1).

### Value

the scaled estimate

---

spvim_ics	<i>Influence function estimates for SPVIMs</i>
-----------	--

---

### Description

Compute the influence functions for the contribution from sampling observations and subsets.

### Usage

```
spvim_ics(Z, z_counts, W, v, psi, G, c_n, ics, measure)
```

Arguments

Z	the matrix of presence/absence of each feature (columns) in each sampled subset (rows)
z_counts	the number of times each unique subset was sampled
W	the matrix of weights
v	the estimated predictiveness measures
psi	the estimated SPVIM values
G	the constraint matrix
c_n	the constraint values
ics	a list of influence function values for each predictiveness measure
measure	the type of measure (e.g., "r_squared" or "auc")

Details

The processes for sampling observations and sampling subsets are independent. Thus, we can compute the influence function separately for each sampling process. For further details, see the paper by Williamson and Feng (2020).

Value

a named list of length 2; contrib\_v is the contribution from estimating V, while contrib\_s is the contribution from sampling subsets.

---

spvim_se	<i>Standard error estimate for SPVIM values</i>
----------	---

---

Description

Compute standard error estimates based on the estimated influence function for a SPVIM value of interest.

Usage

```
spvim_se(ics, idx = 1, gamma = 1, na_rm = FALSE)
```

Arguments

ics	the influence function estimates based on the contributions from sampling observations and sampling subsets: a list of length two resulting from a call to spvim_ics.
idx	the index of interest
gamma	the proportion of the sample size used when sampling subsets
na_rm	remove NAs?

**Details**

Since the processes for sampling observations and subsets are independent, the variance for a given SPVIM estimator is simply the sum of the variances based on sampling observations and on sampling subsets.

**Value**

The standard error estimate for the desired SPVIM value

**See Also**

[spvim\\_ics](#) for how the influence functions are estimated.

---

sp_vim	<i>Shapley Population Variable Importance Measure (SPVIM) Estimates and Inference</i>
--------	---

---

**Description**

Compute estimates and confidence intervals for the SPVIMs, using cross-fitting.

**Usage**

```
sp_vim(
  Y = NULL,
  X = NULL,
  V = 5,
  type = "r_squared",
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  univariate_SL.library = NULL,
  gamma = 1,
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  stratified = FALSE,
  verbose = FALSE,
  sample_splitting = TRUE,
  final_point_estimate = "split",
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_scale = "identity",
  ipc_weights = rep(1, length(Y)),
  ipc_est_type = "aipw",
  scale = "identity",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

**Arguments**

Y	the outcome.
X	the covariates. If type = "average_value", then the exposure variable should be part of X, with its name provided in exposure_name.
V	the number of folds for cross-fitting, defaults to 5. If sample_splitting = TRUE, then a special type of V-fold cross-fitting is done. See Details for a more detailed explanation.
type	the type of importance to compute; defaults to r_squared, but other supported options are auc, accuracy, deviance, and anova.
SL.library	a character vector of learners to pass to SuperLearner, if f1 and f2 are Y and X, respectively. Defaults to SL.glmnet, SL.xgboost, and SL.mean.
univariate_SL.library	(optional) a character vector of learners to pass to SuperLearner for estimating univariate regression functions. Defaults to SL.polymars
gamma	the fraction of the sample size to use when sampling subsets (e.g., gamma = 1 samples the same number of subsets as the sample size)
alpha	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
delta	the value of the $\delta$ -null (i.e., testing if importance < $\delta$ ); defaults to 0.
na.rm	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
stratified	if run_regression = TRUE, then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
verbose	should sp_vim and SuperLearner print out progress? (defaults to FALSE)
sample_splitting	should we use sample-splitting to estimate the full and reduced predictiveness? Defaults to TRUE, since inferences made using sample_splitting = FALSE will be invalid for variables with truly zero importance.
final_point_estimate	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference ("split", the default), or should they instead be based on the full dataset ("full") or the average across the point estimates from each sample split ("average")? All three options result in valid point estimates – sample-splitting is only required for valid inference.
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_scale	what scale should the inverse probability weight correction be applied on (if any)? Defaults to "identity". (other options are "log" and "logit")

<code>ipc_weights</code>	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
<code>ipc_est_type</code>	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if <code>C</code> is not all equal to 1.
<code>scale</code>	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
<code>scale_est</code>	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
<code>cross_fitted_se</code>	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
<code>...</code>	other arguments to the estimation tool, see "See also".

## Details

We define the SPVIM as the weighted average of the population difference in predictiveness over all subsets of features not containing feature  $j$ .

This is equivalent to finding the solution to a population weighted least squares problem. This key fact allows us to estimate the SPVIM using weighted least squares, where we first sample subsets from the power set of all possible features using the Shapley sampling distribution; then use cross-fitting to obtain estimators of the predictiveness of each sampled subset; and finally, solve the least squares problem given in Williamson and Feng (2020).

See the paper by Williamson and Feng (2020) for more details on the mathematics behind this function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list containing:

**SL.library** the library of learners passed to SuperLearner  
**v** the estimated predictiveness measure for each sampled subset  
**fit\_lst** the fitted values on the entire dataset from the chosen method for each sampled subset  
**preds\_lst** the cross-fitted predicted values from the chosen method for each sampled subset  
**est** the estimated SPVIM value for each feature  
**ics** the influence functions for each sampled subset  
**var\_v\_contribs** the contributions to the variance from estimating predictiveness  
**var\_s\_contribs** the contributions to the variance from sampling subsets  
**ic\_lst** a list of the SPVIM influence function contributions  
**se** the standard errors for the estimated variable importance  
**ci** the  $(1 - \alpha) \times 100\%$  confidence intervals based on the variable importance estimates  
**p\_value** p-values for the null hypothesis test of zero importance for each variable  
**test\_statistic** the test statistic for each null hypothesis test of zero importance



**test** a hypothesis testing decision for each null hypothesis test (for each variable having zero importance)

**gamma** the fraction of the sample size used when sampling subsets

**alpha** the level, for confidence interval calculation

**delta** the delta value used for hypothesis testing

**y** the outcome

**ipc\_weights** the weights

**scale** the scale on which CIs were computed

**mat** - a tibble with the estimates, SEs, CIs, hypothesis testing decisions, and p-values

### Value

An object of class `vim`. See Details for more information.

### See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package.

### Examples

```
n <- 100
p <- 2
# generate the data
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- as.matrix(smooth + stats::rnorm(n, 0, 1))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm")

# -----
# using Super Learner (with a small number of CV folds,
# for illustration only)
# -----
set.seed(4747)
est <- sp_vim(Y = y, X = x, V = 2, type = "r_squared",
  SL.library = learners, alpha = 0.05)
```

**Description**

Compute estimates of and confidence intervals for nonparametric intrinsic variable importance based on the population-level contrast between the oracle predictiveness using the feature(s) of interest versus not.

**Usage**

```
vim(
  Y = NULL,
  X = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  type = "r_squared",
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  scale = "identity",
  na.rm = FALSE,
  sample_splitting = TRUE,
  sample_splitting_folds = NULL,
  final_point_estimate = "split",
  stratified = FALSE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_scale = "identity",
  ipc_weights = rep(1, length(Y)),
  ipc_est_type = "aipw",
  scale_est = TRUE,
  nuisance_estimators_full = NULL,
  nuisance_estimators_reduced = NULL,
  exposure_name = NULL,
  bootstrap = FALSE,
  b = 1000,
  boot_interval_type = "perc",
  clustered = FALSE,
  cluster_id = rep(NA, length(Y)),
  ...
)
```

**Arguments**

Y                      the outcome.

X	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of X, with its name provided in <code>exposure_name</code> .
f1	the fitted values from a flexible estimation technique regressing Y on X. A vector of the same length as Y; if sample-splitting is desired, then the value of f1 at each position should be the result of predicting from a model trained without that observation.
f2	the fitted values from a flexible estimation technique regressing either (a) f1 or (b) Y on X withholding the columns in <code>indx</code> . A vector of the same length as Y; if sample-splitting is desired, then the value of f2 at each position should be the result of predicting from a model trained without that observation.
indx	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
type	the type of importance to compute; defaults to <code>r_squared</code> , but other supported options are <code>auc</code> , <code>accuracy</code> , <code>deviance</code> , and <code>anova</code> .
run_regression	if outcome Y and covariates X are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is TRUE, then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
SL.library	a character vector of learners to pass to SuperLearner, if f1 and f2 are Y and X, respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
alpha	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
delta	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
na.rm	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
sample_splitting	should we use sample-splitting to estimate the full and reduced predictiveness? Defaults to TRUE, since inferences made using <code>sample_splitting = FALSE</code> will be invalid for variables with truly zero importance.
sample_splitting_folds	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if <code>run_regression = FALSE</code> .
final_point_estimate	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference ("split", the default), or should they instead be based on the full dataset ("full") or the average across the point estimates from each sample split ("average")? All three options result in valid point estimates – sample-splitting is only required for valid inference.
stratified	if <code>run_regression = TRUE</code> , then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are

	thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_scale	what scale should the inverse probability weight correction be applied on (if any)? Defaults to "identity". (other options are "log" and "logit")
ipc_weights	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., $\text{ipc\_weights} = 1 / [\text{estimated probability weights}]$ ).
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
nuisance_estimators_full	(only used if <code>type = "average_value"</code> ) a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.
nuisance_estimators_reduced	(only used if <code>type = "average_value"</code> ) a list of nuisance function estimators on the observed data (may be within a specified fold, for cross-fitted estimates). Specifically: an estimator of the optimal treatment rule; an estimator of the propensity score under the estimated optimal treatment rule; and an estimator of the outcome regression when treatment is assigned according to the estimated optimal rule.
exposure_name	(only used if <code>type = "average_value"</code> ) the name of the exposure of interest; binary, with 1 indicating presence of the exposure and 0 indicating absence of the exposure.
bootstrap	should bootstrap-based standard error estimates be computed? Defaults to FALSE (and currently may only be used if <code>sample_splitting = FALSE</code> ).
b	the number of bootstrap replicates (only used if <code>bootstrap = TRUE</code> and <code>sample_splitting = FALSE</code> ); defaults to 1000.
boot_interval_type	the type of bootstrap interval (one of "norm", "basic", "stud", "perc", or "bca", as in <code>boot{boot.ci}</code> ) if requested. Defaults to "perc".
clustered	should the bootstrap resamples be performed on clusters rather than individual observations? Defaults to FALSE.
cluster_id	vector of the same length as Y giving the cluster IDs used for the clustered bootstrap, if <code>clustered</code> is TRUE.
...	other arguments to the estimation tool, see "See also".

## Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution. VIM estimates are obtained by obtaining estimators  $f_n$  and  $f_{n,s}$  of  $f_0$  and  $f_{0,s}$ , respectively; obtaining an estimator  $P_n$  of  $P_0$ ; and finally, setting  $\psi_{n,s} := V(f_n, P_n) - V(f_{n,s}, P_n)$ .

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

- s** the column(s) to calculate variable importance for
- SL.library** the library of learners passed to SuperLearner
- type** the type of risk-based variable importance measured
- full\_fit** the fitted values of the chosen method fit to the full data
- red\_fit** the fitted values of the chosen method fit to the reduced data
- est** the estimated variable importance
- naive** the naive estimator of variable importance (only used if `type = "anova"`)
- eif** the estimated efficient influence function
- eif\_full** the estimated efficient influence function for the full regression
- eif\_reduced** the estimated efficient influence function for the reduced regression
- se** the standard error for the estimated variable importance
- ci** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate
- test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test
- p\_value** a p-value based on the same test as `test`
- full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)
- red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)
- alpha** the level, for confidence interval calculation
- sample\_splitting\_folds** the folds used for sample-splitting (used for hypothesis testing)
- y** the outcome
- ipc\_weights** the weights
- cluster\_id** the cluster IDs
- mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

## Value

An object of classes `vim` and the type of risk-based measure. See Details for more information.

**See Also**

[SuperLearner](#) for specific usage of the SuperLearner function and package.

**Examples**

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -1, 1)))

# apply the function to the x's
f <- function(x) 0.5 + 0.3*x[1] + 0.2*x[2]
smooth <- apply(x, 1, function(z) f(z))

# generate Y ~ Bernoulli (smooth)
y <- matrix(rbinom(n, size = 1, prob = smooth))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm")

# using Y and X; use class-balanced folds
est_1 <- vim(y, x, indx = 2, type = "accuracy",
            alpha = 0.05, run_regression = TRUE,
            SL.library = learners, cvControl = list(V = 2),
            stratified = TRUE)

# using pre-computed fitted values
set.seed(4747)
V <- 2
full_fit <- SuperLearner::CV.SuperLearner(Y = y, X = x,
                                       SL.library = learners,
                                       cvControl = list(V = 2),
                                       innerCvControl = list(list(V = V)))
full_fitted <- SuperLearner::predict.SuperLearner(full_fit)$pred
# fit the data with only X1
reduced_fit <- SuperLearner::CV.SuperLearner(Y = full_fitted,
                                       X = x[, -2, drop = FALSE],
                                       SL.library = learners,
                                       cvControl = list(V = 2, validRows = full_fit$folds),
                                       innerCvControl = list(list(V = V)))
reduced_fitted <- SuperLearner::predict.SuperLearner(reduced_fit)$pred

est_2 <- vim(Y = y, f1 = full_fitted, f2 = reduced_fitted,
            indx = 2, run_regression = FALSE, alpha = 0.05,
            stratified = TRUE, type = "accuracy",
            sample_splitting_folds = get_cv_sl_folds(full_fit$folds))
```

---

vimp_accuracy	<i>Nonparametric Intrinsic Variable Importance Estimates: Classification accuracy</i>
---------------	---

---

## Description

Compute estimates of and confidence intervals for nonparametric difference in classification accuracy-based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "accuracy"`.

## Usage

```
vimp_accuracy(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  final_point_estimate = "split",
  cross_fitting_folds = NULL,
  sample_splitting_folds = NULL,
  stratified = TRUE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "logit",
  ipc_est_type = "aipw",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

## Arguments

Y	the outcome.
X	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of X, with its name provided in <code>exposure_name</code> .
cross_fitted_f1	the predicted values on validation data from a flexible estimation technique regressing Y on X in the training data. Provided as either (a) a vector, where each

element is the predicted value when that observation is part of the validation fold; or (b) a list of length  $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as  $Y$ ; if using a list, then the summed length of each element across the list should be the same length as  $Y$  (i.e., each observation is included in the predictions).

<code>cross_fitted_f2</code>	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in <code>cross_fitted_f1</code> , or (b) $Y$ , on $X$ withholding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as $Y$ ; if using a list, then the summed length of each element across the list should be the same length as $Y$ (i.e., each observation is included in the predictions).
<code>f1</code>	the fitted values from a flexible estimation technique regressing $Y$ on $X$ . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>f2</code>	the fitted values from a flexible estimation technique regressing either (a) <code>f1</code> or (b) $Y$ on $X$ withholding the columns in <code>indx</code> . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of $V$ -fold cross-fitting is done. See Details for a more detailed explanation.
<code>run_regression</code>	if outcome $Y$ and covariates $X$ are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is <code>TRUE</code> , then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to <code>SuperLearner</code> , if <code>f1</code> and <code>f2</code> are $Y$ and $X$ , respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
<code>na.rm</code>	should we remove NAs in the outcome and fitted values in computation? (defaults to <code>FALSE</code> )
<code>final_point_estimate</code>	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference (" <code>split</code> ", the default), or should they instead be based on the full dataset (" <code>full</code> ") or the average across the point estimates from each sample split (" <code>average</code> ")? All three options result in valid point estimates – sample-splitting is only required for valid inference.



cross_fitting_folds	the folds for cross-fitting. Only used if run_regression = FALSE.
sample_splitting_folds	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if run_regression = FALSE.
stratified	if run_regression = TRUE, then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_weights	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., ipc_weights = 1 / [estimated probability weights]).
scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
...	other arguments to the estimation tool, see "See also".

## Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution.

Cross-fitted VIM estimates are computed differently if sample-splitting is requested versus if it is not. We recommend using sample-splitting in most cases, since only in this case will inferences be valid if the variable(s) of interest have truly zero population importance. The purpose of cross-fitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data from estimating  $P_0$ ; this can result in improved performance, especially when using flexible learning algorithms. The purpose of sample-splitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data; this allows valid inference under the null hypothesis of zero importance.

Without sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $P_{n,k}$  of  $P_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{V(f_{n,k}, P_{n,k}) - V(f_{n,k,s}, P_{n,k})\}.$$

With sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $2K$  folds. These folds are further divided into 2 groups of folds. Then, for each fold  $k$  in the first group, estimator  $f_{n,k}$  of  $f_0$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k} = V(f_{n,k}, P_{n,k}).$$

Similarly, for each fold  $k$  in the second group, estimator  $f_{n,k,s}$  of  $f_{0,s}$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k,s} = V(f_{n,k,s}, P_{n,k}).$$

Finally,

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{v_{n,k} - v_{n,k,s}\}.$$

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind the `cv_vim` function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

**s** the column(s) to calculate variable importance for

**SL.library** the library of learners passed to SuperLearner

**full\_fit** the fitted values of the chosen method fit to the full data (a list, for train and test data)

**red\_fit** the fitted values of the chosen method fit to the reduced data (a list, for train and test data)

**est** the estimated variable importance

**naive** the naive estimator of variable importance

**eif** the estimated efficient influence function

**eif\_full** the estimated efficient influence function for the full regression

**eif\_reduced** the estimated efficient influence function for the reduced regression

**se** the standard error for the estimated variable importance

**ci** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate

**test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test

**p\_value** a p-value based on the same test as `test`

**full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)

**red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)

**alpha** the level, for confidence interval calculation

**sample\_splitting\_folds** the folds used for hypothesis testing

**cross\_fitting\_folds** the folds used for cross-fitting

**y** the outcome

**ipc\_weights** the weights

**cluster\_id** the cluster IDs

**mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

## Value

An object of classes `vim` and `vim_accuracy`. See Details for more information.

## See Also

[SuperLearner](#) for specific usage of the `SuperLearner` function and package.

## Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -1, 1)))

# apply the function to the x's
f <- function(x) 0.5 + 0.3*x[1] + 0.2*x[2]
smooth <- apply(x, 1, function(z) f(z))

# generate Y ~ Normal (smooth, 1)
y <- matrix(rbinom(n, size = 1, prob = smooth))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# estimate (with a small number of folds, for illustration only)
est <- vimp_accuracy(y, x, indx = 2,
  alpha = 0.05, run_regression = TRUE,
  SL.library = learners, V = 2, cvControl = list(V = 2))
```

vimp\_anova

*Nonparametric Intrinsic Variable Importance Estimates: ANOVA***Description**

Compute estimates of and confidence intervals for nonparametric ANOVA-based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "anova"`. This type has limited functionality compared to other types; in particular, null hypothesis tests are not possible using `type = "anova"`. If you want to do null hypothesis testing on an equivalent population parameter, use `vimp_rsquared` instead.

**Usage**

```
vimp_anova(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  cross_fitting_folds = NULL,
  stratified = FALSE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "logit",
  ipc_est_type = "aipw",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

**Arguments**

<code>Y</code>	the outcome.
<code>X</code>	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of <code>X</code> , with its name provided in <code>exposure_name</code> .
<code>cross_fitted_f1</code>	the predicted values on validation data from a flexible estimation technique regressing <code>Y</code> on <code>X</code> in the training data. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation

fold; or (b) a list of length  $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as  $Y$ ; if using a list, then the summed length of each element across the list should be the same length as  $Y$  (i.e., each observation is included in the predictions).

<code>cross_fitted_f2</code>	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in <code>cross_fitted_f1</code> , or (b) $Y$ , on $X$ withholding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as $Y$ ; if using a list, then the summed length of each element across the list should be the same length as $Y$ (i.e., each observation is included in the predictions).
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of $V$ -fold cross-fitting is done. See Details for a more detailed explanation.
<code>run_regression</code>	if outcome $Y$ and covariates $X$ are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is <code>TRUE</code> , then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to SuperLearner, if <code>f1</code> and <code>f2</code> are $Y$ and $X$ , respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
<code>na.rm</code>	should we remove NAs in the outcome and fitted values in computation? (defaults to <code>FALSE</code> )
<code>cross_fitting_folds</code>	the folds for cross-fitting. Only used if <code>run_regression = FALSE</code> .
<code>stratified</code>	if <code>run_regression = TRUE</code> , then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either (i) <code>NULL</code> (the default, in which case the argument <code>C</code> above must be all ones), or (ii) a character vector specifying the variable(s) among $Y$ and $X$ that are thought to play a role in the coarsening mechanism. To specify the outcome, use " $Y$ "; to specify covariates, use a character number corresponding to the desired position in $X$ (e.g., " $1$ ").
<code>ipc_weights</code>	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).

scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
...	other arguments to the estimation tool, see "See also".

### Details

We define the population ANOVA parameter for the group of features (or single feature)  $s$  by

$$\psi_{0,s} := E_0\{f_0(X) - f_{0,s}(X)\}^2 / \text{var}_0(Y),$$

where  $f_0$  is the population conditional mean using all features,  $f_{0,s}$  is the population conditional mean using the features with index not in  $s$ , and  $E_0$  and  $\text{var}_0$  denote expectation and variance under the true data-generating distribution, respectively.

Cross-fitted ANOVA estimates are computed by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $E_{n,k}$  of  $E_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K E_{n,k}\{f_{n,k}(X) - f_{n,k,s}(X)\}^2 / \text{var}_n(Y),$$

where  $\text{var}_n$  is the empirical variance. See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function.

### Value

An object of classes `vim` and `vimp_anova`. See Details for more information.

### See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package.

### Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2
```

```
# generate Y ~ Normal (smooth, 1)
y <- smooth + stats::rnorm(n, 0, 1)

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# estimate (with a small number of folds, for illustration only)
est <- vimp_anova(y, x, indx = 2,
  alpha = 0.05, run_regression = TRUE,
  SL.library = learners, V = 2, cvControl = list(V = 2))
```

vimp\_auc

*Nonparametric Intrinsic Variable Importance Estimates: AUC***Description**

Compute estimates of and confidence intervals for nonparametric difference in \$AUC\$-based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "auc"`.

**Usage**

```
vimp_auc(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  final_point_estimate = "split",
  cross_fitting_folds = NULL,
  sample_splitting_folds = NULL,
  stratified = TRUE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "logit",
  ipc_est_type = "aipw",
  scale_est = TRUE,
```

```

    cross_fitted_se = TRUE,
    ...
)

```

## Arguments

Y	the outcome.
X	the covariates. If type = "average_value", then the exposure variable should be part of X, with its name provided in exposure_name.
cross_fitted_f1	the predicted values on validation data from a flexible estimation technique regressing Y on X in the training data. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
cross_fitted_f2	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in cross_fitted_f1, or (b) Y, on X withholding the columns in indx. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
f1	the fitted values from a flexible estimation technique regressing Y on X. If sample-splitting is requested, then these must be estimated specially; see Details. If cross_fitted_se = TRUE, then this argument is not used.
f2	the fitted values from a flexible estimation technique regressing either (a) f1 or (b) Y on X withholding the columns in indx. If sample-splitting is requested, then these must be estimated specially; see Details. If cross_fitted_se = TRUE, then this argument is not used.
indx	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
V	the number of folds for cross-fitting, defaults to 5. If sample_splitting = TRUE, then a special type of V-fold cross-fitting is done. See Details for a more detailed explanation.
run_regression	if outcome Y and covariates X are passed to vimp_accuracy, and run_regression is TRUE, then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
SL.library	a character vector of learners to pass to SuperLearner, if f1 and f2 are Y and X, respectively. Defaults to SL.glmnet, SL.xgboost, and SL.mean.



alpha	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
delta	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
na.rm	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
final_point_estimate	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference ("split", the default), or should they instead be based on the full dataset ("full") or the average across the point estimates from each sample split ("average")? All three options result in valid point estimates – sample-splitting is only required for valid inference.
cross_fitting_folds	the folds for cross-fitting. Only used if run_regression = FALSE.
sample_splitting_folds	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if run_regression = FALSE.
stratified	if run_regression = TRUE, then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_weights	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., ipc_weights = 1 / [estimated probability weights]).
scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
...	other arguments to the estimation tool, see "See also".

## Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution.

Cross-fitted VIM estimates are computed differently if sample-splitting is requested versus if it is not. We recommend using sample-splitting in most cases, since only in this case will inferences be valid if the variable(s) of interest have truly zero population importance. The purpose of cross-fitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data from estimating  $P_0$ ; this can result in improved performance, especially when using flexible learning algorithms. The purpose of sample-splitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data; this allows valid inference under the null hypothesis of zero importance.

Without sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $P_{n,k}$  of  $P_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{V(f_{n,k}, P_{n,k}) - V(f_{n,k,s}, P_{n,k})\}.$$

With sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $2K$  folds. These folds are further divided into 2 groups of folds. Then, for each fold  $k$  in the first group, estimator  $f_{n,k}$  of  $f_0$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k} = V(f_{n,k}, P_{n,k}).$$

Similarly, for each fold  $k$  in the second group, estimator  $f_{n,k,s}$  of  $f_{0,s}$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k,s} = V(f_{n,k,s}, P_{n,k}).$$

Finally,

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{v_{n,k} - v_{n,k,s}\}.$$

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind the `cv_vim` function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

**s** the column(s) to calculate variable importance for

**SL.library** the library of learners passed to SuperLearner

**full\_fit** the fitted values of the chosen method fit to the full data (a list, for train and test data)

**red\_fit** the fitted values of the chosen method fit to the reduced data (a list, for train and test data)

**est** the estimated variable importance

**naive** the naive estimator of variable importance

**EIF** the estimated efficient influence function

**EIF\_full** the estimated efficient influence function for the full regression

**EIF\_reduced** the estimated efficient influence function for the reduced regression

**SE** the standard error for the estimated variable importance

**CI** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate

**test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test

**p\_value** a p-value based on the same test as test

**full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)

**red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)

**alpha** the level, for confidence interval calculation

**sample\_splitting\_folds** the folds used for hypothesis testing

**cross\_fitting\_folds** the folds used for cross-fitting

**y** the outcome

**ipc\_weights** the weights

**cluster\_id** the cluster IDs

**mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

## Value

An object of classes `vim` and `vim_auc`. See Details for more information.

## See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package, and [performance](#) for specific usage of the ROCR package.

## Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -1, 1)))

# apply the function to the x's
f <- function(x) 0.5 + 0.3*x[1] + 0.2*x[2]
smooth <- apply(x, 1, function(z) f(z))

# generate Y ~ Normal (smooth, 1)
y <- matrix(rbinom(n, size = 1, prob = smooth))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")
```

```
# estimate (with a small number of folds, for illustration only)
est <- vimp_auc(y, x, indx = 2,
               alpha = 0.05, run_regression = TRUE,
               SL.library = learners, V = 2, cvControl = list(V = 2))
```

---

vimp_ci	<i>Confidence intervals for variable importance</i>
---------	---

---

## Description

Compute confidence intervals for the true variable importance parameter.

## Usage

```
vimp_ci(est, se, scale = "identity", level = 0.95, truncate = TRUE)
```

## Arguments

est	estimate of variable importance, e.g., from a call to <code>vimp_point_est</code> .
se	estimate of the standard error of est, e.g., from a call to <code>vimp_se</code> .
scale	scale to compute interval estimate on (defaults to "identity": compute Wald-type CI).
level	confidence interval type (defaults to 0.95).
truncate	truncate CIs to have lower limit at (or above) zero?

## Details

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function and the definition of the parameter of interest.

## Value

The Wald-based confidence interval for the true importance of the given group of left-out covariates.

vimp\_deviance

*Nonparametric Intrinsic Variable Importance Estimates: Deviance***Description**

Compute estimates of and confidence intervals for nonparametric deviance-based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "deviance"`.

**Usage**

```
vimp_deviance(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  final_point_estimate = "split",
  cross_fitting_folds = NULL,
  sample_splitting_folds = NULL,
  stratified = TRUE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "logit",
  ipc_est_type = "aipw",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

**Arguments**

**Y** the outcome.

**X** the covariates. If `type = "average_value"`, then the exposure variable should be part of X, with its name provided in `exposure_name`.

**cross\_fitted\_f1** the predicted values on validation data from a flexible estimation technique regressing Y on X in the training data. Provided as either (a) a vector, where each

element is the predicted value when that observation is part of the validation fold; or (b) a list of length  $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as  $Y$ ; if using a list, then the summed length of each element across the list should be the same length as  $Y$  (i.e., each observation is included in the predictions).

<code>cross_fitted_f2</code>	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in <code>cross_fitted_f1</code> , or (b) $Y$ , on $X$ withholding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length $V$ , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as $Y$ ; if using a list, then the summed length of each element across the list should be the same length as $Y$ (i.e., each observation is included in the predictions).
<code>f1</code>	the fitted values from a flexible estimation technique regressing $Y$ on $X$ . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>f2</code>	the fitted values from a flexible estimation technique regressing either (a) <code>f1</code> or (b) $Y$ on $X$ withholding the columns in <code>indx</code> . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of $V$ -fold cross-fitting is done. See Details for a more detailed explanation.
<code>run_regression</code>	if outcome $Y$ and covariates $X$ are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is <code>TRUE</code> , then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to <code>SuperLearner</code> , if <code>f1</code> and <code>f2</code> are $Y$ and $X$ , respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
<code>na.rm</code>	should we remove NAs in the outcome and fitted values in computation? (defaults to <code>FALSE</code> )
<code>final_point_estimate</code>	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference (" <code>split</code> ", the default), or should they instead be based on the full dataset (" <code>full</code> ") or the average across the point estimates from each sample split (" <code>average</code> ")? All three options result in valid point estimates – sample-splitting is only required for valid inference.

cross_fitting_folds	the folds for cross-fitting. Only used if run_regression = FALSE.
sample_splitting_folds	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if run_regression = FALSE.
stratified	if run_regression = TRUE, then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
C	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
Z	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
ipc_weights	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., ipc_weights = 1 / [estimated probability weights]).
scale	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
...	other arguments to the estimation tool, see "See also".

## Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution.

Cross-fitted VIM estimates are computed differently if sample-splitting is requested versus if it is not. We recommend using sample-splitting in most cases, since only in this case will inferences be valid if the variable(s) of interest have truly zero population importance. The purpose of cross-fitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data from estimating  $P_0$ ; this can result in improved performance, especially when using flexible learning algorithms. The purpose of sample-splitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data; this allows valid inference under the null hypothesis of zero importance.

Without sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $P_{n,k}$  of  $P_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{V(f_{n,k}, P_{n,k}) - V(f_{n,k,s}, P_{n,k})\}.$$

With sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $2K$  folds. These folds are further divided into 2 groups of folds. Then, for each fold  $k$  in the first group, estimator  $f_{n,k}$  of  $f_0$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k} = V(f_{n,k}, P_{n,k}).$$

Similarly, for each fold  $k$  in the second group, estimator  $f_{n,k,s}$  of  $f_{0,s}$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k,s} = V(f_{n,k,s}, P_{n,k}).$$

Finally,

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{v_{n,k} - v_{n,k,s}\}.$$

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind the `cv_vim` function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

**s** the column(s) to calculate variable importance for

**SL.library** the library of learners passed to SuperLearner

**full\_fit** the fitted values of the chosen method fit to the full data (a list, for train and test data)

**red\_fit** the fitted values of the chosen method fit to the reduced data (a list, for train and test data)

**est** the estimated variable importance

**naive** the naive estimator of variable importance

**eif** the estimated efficient influence function

**eif\_full** the estimated efficient influence function for the full regression

**eif\_reduced** the estimated efficient influence function for the reduced regression

**se** the standard error for the estimated variable importance

**ci** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate

**test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test

**p\_value** a p-value based on the same test as `test`

**full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)



**red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)

**alpha** the level, for confidence interval calculation

**sample\_splitting\_folds** the folds used for hypothesis testing

**cross\_fitting\_folds** the folds used for cross-fitting

**y** the outcome

**ipc\_weights** the weights

**cluster\_id** the cluster IDs

**mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

## Value

An object of classes `vim` and `vim_deviance`. See Details for more information.

## See Also

[SuperLearner](#) for specific usage of the `SuperLearner` function and package.

## Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -1, 1)))

# apply the function to the x's
f <- function(x) 0.5 + 0.3*x[1] + 0.2*x[2]
smooth <- apply(x, 1, function(z) f(z))

# generate Y ~ Normal (smooth, 1)
y <- matrix(stats::rbinom(n, size = 1, prob = smooth))

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# estimate (with a small number of folds, for illustration only)
est <- vimp_deviance(y, x, indx = 2,
  alpha = 0.05, run_regression = TRUE,
  SL.library = learners, V = 2, cvControl = list(V = 2))
```

---

`vimp_hypothesis_test`    *Perform a hypothesis test against the null hypothesis of  $\delta$  importance*

---

## Description

Perform a hypothesis test against the null hypothesis of zero importance by: (i) for a user-specified level  $\alpha$ , compute a  $(1 - \alpha) \times 100\%$  confidence interval around the predictiveness for both the full and reduced regression functions (these must be estimated on independent splits of the data); (ii) if the intervals do not overlap, reject the null hypothesis.

## Usage

```
vimp_hypothesis_test(
  predictiveness_full,
  predictiveness_reduced,
  se,
  delta = 0,
  alpha = 0.05
)
```

## Arguments

<code>predictiveness_full</code>	the estimated predictiveness of the regression including the covariate(s) of interest.
<code>predictiveness_reduced</code>	the estimated predictiveness of the regression excluding the covariate(s) of interest.
<code>se</code>	the estimated standard error of the variable importance estimator
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
<code>alpha</code>	the desired type I error rate (defaults to 0.05).

## Details

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function and the definition of the parameter of interest.

## Value

a list, with: the hypothesis testing decision (TRUE if the null hypothesis is rejected, FALSE otherwise); the p-value from the hypothesis test; and the test statistic from the hypothesis test.

vimp\_regression

*Nonparametric Intrinsic Variable Importance Estimates: ANOVA***Description**

Compute estimates of and confidence intervals for nonparametric ANOVA-based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "anova"`. This function is deprecated in `vimp` version 2.0.0.

**Usage**

```
vimp_regression(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  cross_fitting_folds = NULL,
  stratified = FALSE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "identity",
  ipc_est_type = "aipw",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

**Arguments**

<code>Y</code>	the outcome.
<code>X</code>	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of <code>X</code> , with its name provided in <code>exposure_name</code> .
<code>cross_fitted_f1</code>	the predicted values on validation data from a flexible estimation technique regressing <code>Y</code> on <code>X</code> in the training data. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length <code>V</code> , where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested,

then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).

<code>cross_fitted_f2</code>	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in <code>cross_fitted_f1</code> , or (b) Y, on X withholding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
<code>V</code>	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of V-fold cross-fitting is done. See Details for a more detailed explanation.
<code>run_regression</code>	if outcome Y and covariates X are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is TRUE, then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to SuperLearner, if <code>f1</code> and <code>f2</code> are Y and X, respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.
<code>na.rm</code>	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
<code>cross_fitting_folds</code>	the folds for cross-fitting. Only used if <code>run_regression = FALSE</code> .
<code>stratified</code>	if <code>run_regression = TRUE</code> , then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either (i) NULL (the default, in which case the argument C above must be all ones), or (ii) a character vector specifying the variable(s) among Y and X that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in X (e.g., "1").
<code>ipc_weights</code>	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
<code>scale</code>	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")

ipc_est_type	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if C is not all equal to 1.
scale_est	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
cross_fitted_se	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
...	other arguments to the estimation tool, see "See also".

### Details

We define the population ANOVA parameter for the group of features (or single feature)  $s$  by

$$\psi_{0,s} := E_0\{f_0(X) - f_{0,s}(X)\}^2 / \text{var}_0(Y),$$

where  $f_0$  is the population conditional mean using all features,  $f_{0,s}$  is the population conditional mean using the features with index not in  $s$ , and  $E_0$  and  $\text{var}_0$  denote expectation and variance under the true data-generating distribution, respectively.

Cross-fitted ANOVA estimates are computed by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $E_{n,k}$  of  $E_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K E_{n,k}\{f_{n,k}(X) - f_{n,k,s}(X)\}^2 / \text{var}_n(Y),$$

where  $\text{var}_n$  is the empirical variance. See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function.

### Value

An object of classes `vim` and `vim_regression`. See Details for more information.

### See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package.

### Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- smooth + stats::rnorm(n, 0, 1)
```

```
# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# estimate (with a small number of folds, for illustration only)
est <- vimp_regression(y, x, indx = 2,
  alpha = 0.05, run_regression = TRUE,
  SL.library = learners, V = 2, cvControl = list(V = 2))
```

vimp\_rsquared

*Nonparametric Intrinsic Variable Importance Estimates: R-squared*

### Description

Compute estimates of and confidence intervals for nonparametric  $R^2$ -based intrinsic variable importance. This is a wrapper function for `cv_vim`, with `type = "r_squared"`.

### Usage

```
vimp_rsquared(
  Y = NULL,
  X = NULL,
  cross_fitted_f1 = NULL,
  cross_fitted_f2 = NULL,
  f1 = NULL,
  f2 = NULL,
  indx = 1,
  V = 10,
  run_regression = TRUE,
  SL.library = c("SL.glmnet", "SL.xgboost", "SL.mean"),
  alpha = 0.05,
  delta = 0,
  na.rm = FALSE,
  final_point_estimate = "split",
  cross_fitting_folds = NULL,
  sample_splitting_folds = NULL,
  stratified = FALSE,
  C = rep(1, length(Y)),
  Z = NULL,
  ipc_weights = rep(1, length(Y)),
  scale = "logit",
  ipc_est_type = "aipw",
  scale_est = TRUE,
  cross_fitted_se = TRUE,
  ...
)
```

**Arguments**

Y	the outcome.
X	the covariates. If <code>type = "average_value"</code> , then the exposure variable should be part of X, with its name provided in <code>exposure_name</code> .
<code>cross_fitted_f1</code>	the predicted values on validation data from a flexible estimation technique regressing Y on X in the training data. Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
<code>cross_fitted_f2</code>	the predicted values on validation data from a flexible estimation technique regressing either (a) the fitted values in <code>cross_fitted_f1</code> , or (b) Y, on X withholding the columns in <code>indx</code> . Provided as either (a) a vector, where each element is the predicted value when that observation is part of the validation fold; or (b) a list of length V, where each element in the list is a set of predictions on the corresponding validation data fold. If sample-splitting is requested, then these must be estimated specially; see Details. However, the resulting vector should be the same length as Y; if using a list, then the summed length of each element across the list should be the same length as Y (i.e., each observation is included in the predictions).
f1	the fitted values from a flexible estimation technique regressing Y on X. If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
f2	the fitted values from a flexible estimation technique regressing either (a) f1 or (b) Y on X withholding the columns in <code>indx</code> . If sample-splitting is requested, then these must be estimated specially; see Details. If <code>cross_fitted_se = TRUE</code> , then this argument is not used.
<code>indx</code>	the indices of the covariate(s) to calculate variable importance for; defaults to 1.
V	the number of folds for cross-fitting, defaults to 5. If <code>sample_splitting = TRUE</code> , then a special type of V-fold cross-fitting is done. See Details for a more detailed explanation.
<code>run_regression</code>	if outcome Y and covariates X are passed to <code>vimp_accuracy</code> , and <code>run_regression</code> is TRUE, then Super Learner will be used; otherwise, variable importance will be computed using the inputted fitted values.
<code>SL.library</code>	a character vector of learners to pass to SuperLearner, if f1 and f2 are Y and X, respectively. Defaults to <code>SL.glmnet</code> , <code>SL.xgboost</code> , and <code>SL.mean</code> .
<code>alpha</code>	the level to compute the confidence interval at. Defaults to 0.05, corresponding to a 95% confidence interval.
<code>delta</code>	the value of the $\delta$ -null (i.e., testing if importance $< \delta$ ); defaults to 0.

<code>na.rm</code>	should we remove NAs in the outcome and fitted values in computation? (defaults to FALSE)
<code>final_point_estimate</code>	if sample splitting is used, should the final point estimates be based on only the sample-split folds used for inference ("split", the default), or should they instead be based on the full dataset ("full") or the average across the point estimates from each sample split ("average")? All three options result in valid point estimates – sample-splitting is only required for valid inference.
<code>cross_fitting_folds</code>	the folds for cross-fitting. Only used if <code>run_regression = FALSE</code> .
<code>sample_splitting_folds</code>	the folds used for sample-splitting; these identify the observations that should be used to evaluate predictiveness based on the full and reduced sets of covariates, respectively. Only used if <code>run_regression = FALSE</code> .
<code>stratified</code>	if <code>run_regression = TRUE</code> , then should the generated folds be stratified based on the outcome (helps to ensure class balance across cross-validation folds)
<code>C</code>	the indicator of coarsening (1 denotes observed, 0 denotes unobserved).
<code>Z</code>	either (i) NULL (the default, in which case the argument <code>C</code> above must be all ones), or (ii) a character vector specifying the variable(s) among <code>Y</code> and <code>X</code> that are thought to play a role in the coarsening mechanism. To specify the outcome, use "Y"; to specify covariates, use a character number corresponding to the desired position in <code>X</code> (e.g., "1").
<code>ipc_weights</code>	weights for the computed influence curve (i.e., inverse probability weights for coarsened-at-random settings). Assumed to be already inverted (i.e., <code>ipc_weights = 1 / [estimated probability weights]</code> ).
<code>scale</code>	should CIs be computed on original ("identity") or another scale? (options are "log" and "logit")
<code>ipc_est_type</code>	the type of procedure used for coarsened-at-random settings; options are "ipw" (for inverse probability weighting) or "aipw" (for augmented inverse probability weighting). Only used if <code>C</code> is not all equal to 1.
<code>scale_est</code>	should the point estimate be scaled to be greater than or equal to 0? Defaults to TRUE.
<code>cross_fitted_se</code>	should we use cross-fitting to estimate the standard errors (TRUE, the default) or not (FALSE)?
<code>...</code>	other arguments to the estimation tool, see "See also".

## Details

We define the population variable importance measure (VIM) for the group of features (or single feature)  $s$  with respect to the predictiveness measure  $V$  by

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,s}, P_0),$$

where  $f_0$  is the population predictiveness maximizing function,  $f_{0,s}$  is the population predictiveness maximizing function that is only allowed to access the features with index not in  $s$ , and  $P_0$  is the true data-generating distribution.



Cross-fitted VIM estimates are computed differently if sample-splitting is requested versus if it is not. We recommend using sample-splitting in most cases, since only in this case will inferences be valid if the variable(s) of interest have truly zero population importance. The purpose of cross-fitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data from estimating  $P_0$ ; this can result in improved performance, especially when using flexible learning algorithms. The purpose of sample-splitting is to estimate  $f_0$  and  $f_{0,s}$  on independent data; this allows valid inference under the null hypothesis of zero importance.

Without sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $K$  folds; then using each fold in turn as a hold-out set, constructing estimators  $f_{n,k}$  and  $f_{n,k,s}$  of  $f_0$  and  $f_{0,s}$ , respectively on the training data and estimator  $P_{n,k}$  of  $P_0$  using the test data; and finally, computing

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{V(f_{n,k}, P_{n,k}) - V(f_{n,k,s}, P_{n,k})\}.$$

With sample-splitting, cross-fitted VIM estimates are obtained by first splitting the data into  $2K$  folds. These folds are further divided into 2 groups of folds. Then, for each fold  $k$  in the first group, estimator  $f_{n,k}$  of  $f_0$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k} = V(f_{n,k}, P_{n,k}).$$

Similarly, for each fold  $k$  in the second group, estimator  $f_{n,k,s}$  of  $f_{0,s}$  is constructed using all data besides the  $k$ th fold in the group (i.e.,  $(2K - 1)/(2K)$  of the data) and estimator  $P_{n,k}$  of  $P_0$  is constructed using the held-out data (i.e.,  $1/2K$  of the data); then, computing

$$v_{n,k,s} = V(f_{n,k,s}, P_{n,k}).$$

Finally,

$$\psi_{n,s} := K^{(-1)} \sum_{k=1}^K \{v_{n,k} - v_{n,k,s}\}.$$

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind the `cv_vim` function, and the validity of the confidence intervals.

In the interest of transparency, we return most of the calculations within the `vim` object. This results in a list including:

**s** the column(s) to calculate variable importance for

**SL.library** the library of learners passed to SuperLearner

**full\_fit** the fitted values of the chosen method fit to the full data (a list, for train and test data)

**red\_fit** the fitted values of the chosen method fit to the reduced data (a list, for train and test data)

**est** the estimated variable importance

**naive** the naive estimator of variable importance

**eif** the estimated efficient influence function

**eif\_full** the estimated efficient influence function for the full regression

**eif\_reduced** the estimated efficient influence function for the reduced regression

**se** the standard error for the estimated variable importance

**ci** the  $(1 - \alpha) \times 100\%$  confidence interval for the variable importance estimate

**test** a decision to either reject (TRUE) or not reject (FALSE) the null hypothesis, based on a conservative test

**p\_value** a p-value based on the same test as test

**full\_mod** the object returned by the estimation procedure for the full data regression (if applicable)

**red\_mod** the object returned by the estimation procedure for the reduced data regression (if applicable)

**alpha** the level, for confidence interval calculation

**sample\_splitting\_folds** the folds used for hypothesis testing

**cross\_fitting\_folds** the folds used for cross-fitting

**y** the outcome

**ipc\_weights** the weights

**cluster\_id** the cluster IDs

**mat** a tibble with the estimate, SE, CI, hypothesis testing decision, and p-value

### Value

An object of classes `vim` and `vim_rsquared`. See Details for more information.

### See Also

[SuperLearner](#) for specific usage of the SuperLearner function and package.

### Examples

```
# generate the data
# generate X
p <- 2
n <- 100
x <- data.frame(replicate(p, stats::runif(n, -5, 5)))

# apply the function to the x's
smooth <- (x[,1]/5)^2*(x[,1]+7)/5 + (x[,2]/3)^2

# generate Y ~ Normal (smooth, 1)
y <- smooth + stats::rnorm(n, 0, 1)

# set up a library for SuperLearner; note simple library for speed
library("SuperLearner")
learners <- c("SL.glm", "SL.mean")

# estimate (with a small number of folds, for illustration only)
est <- vimp_rsquared(y, x, indx = 2,
  alpha = 0.05, run_regression = TRUE,
  SL.library = learners, V = 2, cvControl = list(V = 2))
```

---

vimp_se	<i>Estimate variable importance standard errors</i>
---------	---

---

**Description**

Compute standard error estimates for estimates of variable importance.

**Usage**

```
vimp_se(
  eif_full,
  eif_reduced,
  cross_fit = TRUE,
  sample_split = TRUE,
  na.rm = FALSE
)
```

**Arguments**

eif_full	the estimated efficient influence function (EIF) based on the full set of covariates.
eif_reduced	the estimated EIF based on the reduced set of covariates.
cross_fit	logical; was cross-fitting used to compute the EIFs? (defaults to TRUE)
sample_split	logical; was sample-splitting used? (defaults to TRUE)
na.rm	logical; should NA's be removed in computation? (defaults to FALSE).

**Details**

See the paper by Williamson, Gilbert, Simon, and Carone for more details on the mathematics behind this function and the definition of the parameter of interest.

**Value**

The standard error for the estimated variable importance for the given group of left-out covariates.

---

vrc01	<i>Neutralization sensitivity of HIV viruses to antibody VRC01</i>
-------	--

---

**Description**

A dataset containing neutralization sensitivity – measured using inhibitory concentration, the quantity of antibody necessary to neutralize a fraction of viruses in a given sample – and viral features including: amino acid sequence features (measured using HXB2 coordinates), geographic region of origin, subtype, and viral geometry. Accessed from the Los Alamos National Laboratory's (LANL's) Compile, Analyze, and tally Neutralizing Antibody Panels (CATNAP) database.

**Usage**

```
data("vrc01")
```

**Format**

A data frame with 611 rows and 837 variables:

**seqname** Viral sequence identifiers

**subtype.is.01\_AE** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.02\_AG** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.07\_BC** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.A1** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.A1C** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.A1D** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.B** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.C** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.D** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.O** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**subtype.is.Other** Dummy variables encoding the viral subtype as 0/1. Possible subtypes are 01\_AE, 02\_AG, 07\_BC, A1, A1C, A1D, B, C, D, O, Other.

**geographic.region.of.origin.is.Asia** Dummy variables encoding the geographic region of origin as 0/1. Regions are Asia, Europe/Americas, North Africa, and Southern Africa.

**geographic.region.of.origin.is.Europe.Americas** Dummy variables encoding the geographic region of origin as 0/1. Regions are Asia, Europe/Americas, North Africa, and Southern Africa.

**geographic.region.of.origin.is.N.Africa** Dummy variables encoding the geographic region of origin as 0/1. Regions are Asia, Europe/Americas, North Africa, and Southern Africa.

**geographic.region.of.origin.is.S.Africa** Dummy variables encoding the geographic region of origin as 0/1. Regions are Asia, Europe/Americas, North Africa, and Southern Africa.

**ic50.censored** A binary indicator of whether or not the IC-50 (the concentration at which 50 Right-censoring is a proxy for a resistant virus.

**ic80.censored** A binary indicator of whether or not the IC-80 (the concentration at which 80 Right-censoring is a proxy for a resistant virus.

- ic50.geometric.mean.imputed** Continuous IC-50. If neutralization sensitivity for the virus was assessed in multiple studies, the geometric mean was taken.
- ic80.geometric.mean.imputed** Continuous IC-90. If neutralization sensitivity for the virus was assessed in multiple studies, the geometric mean was taken.
- hxb2.46.E.1mer** Amino acid sequence features denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site. For example, hxb2.46.E.1mer records the presence of an E at HXB2-referenced site 46.
- hxb2.46.K.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.46.Q.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.46.R.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.46.T.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.61.H.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.61.Q.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.61.T.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.61.Y.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.97.E.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.97.K.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.97.N.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.97.R.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.124.F.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.124.P.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.125.I.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.125.L.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.127.I.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.127.V.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.

























































- [illegible]

























- hxb2.446.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.448.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.460.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.461.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.462.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.463.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.465.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.611.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.616.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.618.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.619.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.624.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.625.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.637.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.674.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.743.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.750.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.787.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.816.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- hxb2.824.sequon\_actual.1mer** Amino acid sequence feature denoting the presence (1) or absence (0) of a residue at the given HXB2-referenced site.
- sequons.total.env** The total number of sequons in various areas of the HIV viral envelope protein.
- sequons.total.gp120** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.v5** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.loop.d** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.loop.e** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.vrc01** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.cd4** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.sj.fence** The total number of sequons in various areas of the HIV viral envelope protein.

**sequons.total.sj.trimer** The total number of sequons in various areas of the HIV viral envelope protein.

**cysteines.total.env** The number of cysteines in various areas of the HIV viral envelope protein.

**cysteines.total.gp120** The number of cysteines in various areas of the HIV viral envelope protein.

**cysteines.total.v5** The number of cysteines in various areas of the HIV viral envelope protein.

**cysteines.total.vrc01** The number of cysteines in various areas of the HIV viral envelope protein.

**length.env** The length of various areas of the HIV viral envelope protein.

**length.gp120** The length of various areas of the HIV viral envelope protein.

**length.v5** The length of various areas of the HIV viral envelope protein.

**length.v5.outliers** The length of various areas of the HIV viral envelope protein.

**length.loop.e** The length of various areas of the HIV viral envelope protein.

**length.loop.e.outliers** The length of various areas of the HIV viral envelope protein.

**taylor.small.total.v5** The steric bulk of residues at critical locations.

**taylor.small.total.loop.d** The steric bulk of residues at critical locations.

**taylor.small.total.cd4** The steric bulk of residues at critical locations.

#### Source

<https://github.com/benkeser/vrc01/blob/master/data/fulldata.csv>

# Index

- \* **datasets**
  - vrc01, [83](#)
- average\_vim, [3](#)
- boot, [5](#), [12](#), [52](#)
- bootstrap\_se, [4](#)
- check\_fitted\_values, [6](#)
- check\_inputs, [7](#)
- create\_z, [8](#)
- CV.SuperLearner, [22](#)
- cv\_vim, [8](#)
- est\_predictiveness, [18](#)
- est\_predictiveness\_cv, [20](#)
- estimate, [15](#)
- estimate.predictiveness\_measure, [15](#)
- estimate\_eif\_projection, [16](#)
- estimate\_nuisances, [17](#)
- estimate\_type\_predictiveness, [18](#)
- extract\_sampled\_split\_predictions, [22](#)
- format.predictiveness\_measure, [23](#)
- format.vim, [23](#)
- get\_cv\_sl\_folds, [24](#)
- get\_full\_type, [24](#)
- get\_test\_set, [25](#)
- make\_folds, [25](#)
- make\_kfold, [26](#)
- measure\_accuracy, [26](#)
- measure\_anova, [28](#)
- measure\_auc, [29](#)
- measure\_average\_value, [30](#)
- measure\_cross\_entropy, [32](#)
- measure\_deviance, [33](#)
- measure\_mse, [34](#)
- measure\_r\_squared, [36](#)
- merge\_vim, [37](#)
- performance, [67](#)
- predictiveness\_measure, [38](#)
- print.predictiveness\_measure, [40](#)
- print.vim, [41](#)
- process\_arg\_lst, [41](#)
- run\_sl, [42](#)
- sample\_subsets, [43](#)
- scale\_est, [44](#)
- sp\_vim, [46](#)
- spvim\_ics, [44](#), [46](#)
- spvim\_se, [45](#)
- SuperLearner, [13](#), [49](#), [54](#), [59](#), [62](#), [67](#), [73](#), [77](#), [82](#)
- vim, [50](#)
- vimp\_accuracy, [55](#)
- vimp\_anova, [60](#)
- vimp\_auc, [63](#)
- vimp\_ci, [68](#)
- vimp\_deviance, [69](#)
- vimp\_hypothesis\_test, [74](#)
- vimp\_regression, [75](#)
- vimp\_rsquared, [78](#)
- vimp\_se, [83](#)
- vrc01, [83](#)