Package ‘JOPT’

May 23, 2025

Type Package

Title J-optimal Subdata Selection

Version 0.1.0

Description Implements J-optimal subsample selection for regression models. Provides efficient tools for selecting data subsets to optimize statistical efficiency in large-scale or computationally demanding analyses. Includes functions for model specification, subsample selection, and comparative benchmarking.
The methodology is based on Cia-Mina et al. (2025, <https://doi.org/10.1109/TBDATA.2025.3552343>).

URL https://github.com/alvarocia/JOPT

Author Alvaro Cia-Mina <aciamina@unav.es>

Maintainer Alvaro Cia-Mina <aciamina@unav.es>

License GPL

Encoding UTF-8

LazyData true

Imports Matrix,
latex2exp

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Contents

create_model_function 2
jseq 2
run_efficiency_comparison_example 4

Index **5**

create_model_function *Create a Model Function from Expressions*

Description

This function takes a vector of mathematical expressions (as character strings) and generates a function that, given an input vector x, computes the specified expressions and returns the results as a column matrix.

Usage

```
create_model_function(expressions)
```

Arguments

expressions	A character vector of mathematical expressions to define the model. Each expression should be valid R code and reference elements of x (e.g., "x[1]", "x[2]^2").

Value

A function that takes an input vector x and evaluates the functions in expressions, returning a column matrix of the results.

Examples

```
# Define the model expressions
expressions <- c("1", "x[1]", "x[1]*x[2]^2")

# Create the model function
model_function <- create_model_function(expressions)

# Test the model function with an input vector
input_vector <- c(2, 3) # x[1] = 2, x[2] = 3
result <- model_function(input_vector)
print(result)
```

jseq *J-Optimal Subsample Selection*

Description

This function implements the J-optimal subsample selection method, as described in Cia-Mina et al. (2025). It takes a dataset of covariates x, a subsample proportion alpha (between 0 and 1), and a vector defining a regression model. Additional parameters can be specified to control the selection process.

Usage

```
jseq(
  x,
  alpha,
  model_vec,
  k0 = 5 * length(model_vec),
  q = 5/8,
  gamma = 1/10,
  eps1 = 0
)
```

Arguments

x	A dataset (data frame) containing the covariates for the regression model.
alpha	A numeric value between 0 and 1 specifying the subsample proportion.
model_vec	A character vector defining the regression model. Each element should represent a term in the model, written as an expression involving x. For example, "1" for the intercept, "x[1]" for the first covariate, or "x[1]*x[2]^2" for an interaction term.
k0	An integer specifying the initial size of the subsample. Defaults to 5*length(model_vec).
q	A numeric value between 0.5 and 1. Defaults to 5/8.
gamma	A numeric value between 0 and q-0.5. Defaults to 1/10.
eps1	A small positive value. Defaults to 0.

Details

The J-optimal subsample selection algorithm selects a subset of observations from the dataset x that optimizes the statistical efficiency of the model defined by model_vec. For technical details, refer to Cia-Mina et al. (2025).

Value

A list with the following components:

x_j	A subsample of x containing the selected observations (rows) according to J-optimality.
idx	A vector of indices corresponding to the selected rows of x.

Examples

```
# Example 1: Bivariate regression
set.seed(123)
x1 <- runif(1e3, min = -1, max = 1)
x2 <- runif(1e3, min = -1, max = 1)
x <- data.frame(x1 = x1, x2 = x2)
model_vec <- c("1", "x[1]", "x[2]", "x[1]*x[2]", "x[1]^2", "x[2]^2")
result <- jseq(x, 0.3, model_vec)

# Plot the full dataset and the selected subsample
plot(x$x1, x$x2, col = "black", pch = 16, cex = 0.7, xlab = "x1", ylab = "x2")
points(result$x_j$x1, result$x_j$x2, col = "red", pch = 16, cex = 0.7)
title(main = "J-OPT", line = 1)
```

```
# Example 2: Univariate regression
set.seed(123)
x <- data.frame(x = rnorm(1e4))
model_vec <- c("1", "x[1]", "x[1]^2")
result <- jseq(x, 0.3, model_vec)

# Plot the density of the selected subsample
plot(density(result$x_j$x, bw = 2 / 100, kernel = "epanechnikov"),
     ylab = "", lwd = 1.7, xlim = c(-3.5, 3.5), main = "", xlab = "")
```

run_efficiency_comparison_example

Run Efficiency Comparison Example

Description

This function provides an example of running an efficiency comparison. It compares the efficiency of two subdata selection methods: J-optimal and D-optimal. It evaluates their performance based on predefined efficiency criteria and generates comparative plots to visualize the results. Theoretical J-optimal is included for comparison.

Usage

```
run_efficiency_comparison_example()
```

Examples

```
# To run the example:
run_efficiency_comparison_example()
```

Index

create_model_function, 2

jseq, 2

run_efficiency_comparison_example, 4