

Package ‘smoothROC’

August 7, 2026

Type Package
Title Kernel-Smoothed ROC and AUC with Bandwidth Selection
Version 0.1.0
Author Ruhul Ali Khan [aut, cre] (ORCID:
 <https://orcid.org/0000-0003-1173-8345>),
 Musie Ghebremichael [aut] (Project PI)
Maintainer Ruhul Ali Khan <ruhulali.khan@gmail.com>
Description Computes kernel-smoothed receiver operating characteristic
 (ROC) curves and area under the curve (AUC), including confidence
 intervals, using several bandwidth selection methods. Also provides
 the Youden index with confidence intervals, the Youden point, and
 optimal diagnostic cutoff estimation. Supports Gaussian, Biweight, and
 Epanechnikov kernels.
License MIT + file LICENSE
Depends R (>= 3.5)
Imports ggplot2 (>= 3.4.0), rlang, stats
Suggests knitr, rmarkdown
VignetteBuilder knitr
Config/roxygen2/version 8.0.0
Encoding UTF-8
LazyData true
NeedsCompilation no
Repository CRAN
Date/Publication 2026-08-07 17:10:07 UTC

Contents

dystrophy	2
smoothROC	3
smoothROCboot	7
Index	10

dystrophy

Detection of Muscular Dystrophy Carriers

Description

This dataset contains measurements from 209 female subjects evaluated for carrier status of Duchenne Muscular Dystrophy (DMD). The dataset is a classic example used in diagnostic test evaluation and appears in Andrews and Herzberg (2012). It includes basic demographic variables and four serum biomarkers that may be elevated in carriers even when outward symptoms are absent.

Usage

dystrophy

Format

A data frame with 209 rows and 10 variables:

OBS Observation number.

HospID Hospital ID number.

AGE Age in years.

M Month of examination.

Y Year of examination.

CK Serum creatine kinase (primary biomarker).

H Serum hemopexin.

PK Serum pyruvate kinase.

LD Serum lactate dehydrogenase.

Class Factor with levels "carrier" (diseased/positive) and "normal" (non-carrier).

Source

Adapted from the dataset in Andrews and Herzberg (2012)

References

Andrews, D. F., & Herzberg, A. M. (2012). Data: a collection of problems from many fields for the student and research worker. Springer Science & Business Media.

Description

This function estimates a smooth receiver operating characteristic (ROC) curve for a continuous biomarker using kernel-smoothing of the cumulative distribution functions (CDFs) in the non-diseased and diseased groups. It returns a smooth ROC curve, a kernel-based estimator of the area under the curve (AUC) with confidence interval, and the kernel-smoothed Youden index summary with its optimal cutoff.

Usage

```
smoothROC(
  data,
  biomarker,
  status,
  diseased,
  kernel = c("gaussian", "biweight", "epanechnikov"),
  bw_method = c("pdf", "AL", "PB", "BHP", "AR"),
  alpha = 0.05,
  logtrans = FALSE,
  grid_n = 1000
)

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

## S3 method for class 'smoothROC'
plot(x, label = TRUE, youden = TRUE, ...)

## S3 method for class 'smoothROC'
summary(object, ...)
```

Arguments

<code>data</code>	A data frame containing the biomarker and status variables.
<code>biomarker</code>	Character string; name of the numeric column containing biomarker values.
<code>status</code>	Character string; name of the column containing binary disease status.
<code>diseased</code>	The value in <code>status</code> indicating the diseased class.
<code>kernel</code>	Character string; kernel function for smoothing. One of "gaussian", "biweight", or "epanechnikov".
<code>bw_method</code>	Character string; bandwidth selection method. One of "pdf" (Silverman rule for density), "AL" (Altman-Leger), "PB" (Polansky-Baker multistage), "BHP" (normal-reference), or "AR" (adjusted rule-of-thumb).

alpha	Numeric; significance level for (1 - alpha) confidence intervals (default: 0.05).
logtrans	Logical; apply log-transformation to biomarker values? (default: FALSE).
grid_n	Integer; number of grid points for evaluating the ROC curve (default: 1000).
x	An object of class "smoothROC".
...	Additional arguments passed to the S3 methods.
label	Logical; if TRUE (default), include the AUC/Youden annotation label on the plot.
youden	Logical; if TRUE (default), include the Youden point marker and its guide segments on the plot.
object	An object of class "smoothROC".

Details

Let X and Y denote biomarker values from non-diseased and diseased subjects with CDFs F and G , and survival functions $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$. The ROC curve is defined as

$$ROC(p) = \bar{G}\{\bar{F}^{-1}(p)\}, \quad p \in [0, 1],$$

that is, the true positive rate plotted against the false positive rate as the threshold varies over the real line. The AUC is the integral of the ROC curve over $[0, 1]$ and can be written as

$$AUC = \int_{-\infty}^{\infty} F(x) dG(x) = P(Y > X),$$

representing the probability that a randomly chosen diseased subject has a higher biomarker value than a randomly chosen non-diseased subject.

Nonparametric ROC estimation based on the empirical distributions of X and Y yields a stepwise ROC curve that may be jagged and sensitive to sampling variability, especially in small or moderate samples. To obtain a smooth ROC curve while avoiding parametric distributional assumptions, this function applies kernel-based CDF estimators of the form

$$\hat{F}(x) = \frac{1}{m} \sum_{i=1}^m K\left(\frac{x - X_i}{h_m}\right), \quad \hat{G}(x) = \frac{1}{n} \sum_{j=1}^n K\left(\frac{x - Y_j}{h_n}\right),$$

where K is the integrated kernel and h_m , h_n are bandwidths for the non-diseased and diseased groups. The smooth ROC curve is then obtained by plugging $\hat{F}$ and $\hat{G}$ into the ROC functional.

Bandwidth selection is critical for balancing bias and variance in the smoothed CDFs and the resulting ROC curve. The argument `bw_method` implements several rules:

- "pdf": a density-based rule-of-thumb (Silverman) that uses a kernel density bandwidth and is convenient but does not satisfy the usual asymptotic conditions for CDF estimation.
- "BHP": a CDF-based normal-reference bandwidth that minimizes an approximation to the integrated mean squared error of $\hat{F}$, with a robust scale estimate based on $\min(\text{SD}, \text{IQR}/1.34)$.
- "AR": an adjusted CDF reference bandwidth for the Gaussian kernel, obtained by shrinking the normal-reference constant to improve performance for non-Gaussian data while preserving the $m^{-1/3}$ CDF rate.

- "AL": a fully data-driven CDF-based bandwidth in the spirit of Altman and Leger, where the unknown roughness functional is estimated via an auxiliary kernel estimator.
- "PB": a multistage plug-in CDF-based bandwidth (two-stage version of Polansky and Baker) that uses a pilot normal-reference step followed by data-driven refinement.

These choices focus on optimal smoothing of the CDFs rather than the density, which is more directly aligned with ROC curve estimation.

The smooth ROC estimators in this package are based on kernel CDF estimators constructed from a univariate kernel function $k(u)$ and its integral $K(u) = \int_{-\infty}^u k(v) dv$. The following kernels are implemented:

- **Gaussian kernel:** $k(u) = (2\pi)^{-1/2} \exp(-u^2/2)$, with CDF $K(u)$ equal to the standard normal distribution function. This kernel has infinite support and is often used as a default choice in smooth distribution and ROC estimation.
- **Epanechnikov kernel:** $k(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{\{|u| \leq 1\}}$, with compact support on $[-1, 1]$ and optimal second-order efficiency under many mean squared error criteria.
- **Biweight kernel:** $k(u) = \frac{15}{16}(1 - u^2)^2\mathbf{1}_{\{|u| \leq 1\}}$, a smoother higher-order alternative with support on $[-1, 1]$ that produces more rounded ROC and CDF estimates near the boundaries.

The AUC is estimated from the kernel CDFs via

$$\hat{\delta} = \int_{-\infty}^{\infty} \hat{F}(x) d\hat{G}(x),$$

and is asymptotically equivalent to the empirical AUC based on the Mann–Whitney statistic. To quantify uncertainty, a kernel-smoothed analogue of DeLong’s variance estimator is used: empirical placement values are replaced by their kernel-smoothed counterparts, producing a more stable variance estimate in small samples while retaining the large-sample properties of the Mann–Whitney-based estimator.

The Youden index is defined as

$$J = \max_t \{\text{sensitivity}(t) + \text{specificity}(t) - 1\} = \max_t \{F(t) - G(t)\},$$

with corresponding optimal cutoff $t_0 = \arg \max_t \{F(t) - G(t)\}$. Using the kernel CDFs, the function computes a smoothed Youden index $\hat{J}$ and its maximizing cutoff $\hat{t}_0$ on a search grid. When multiple cutoffs achieve the same maximum, secondary criteria can be applied (e.g., favoring higher sensitivity or higher specificity), or the median of all maximizers can be reported. A Delta-method approximation is used for the variance of $\hat{J}$, from which a Wald-type confidence interval is obtained.

Value

An object of class "smoothROC" with components that include:

- **curve:** data frame with columns FPR, TPR, threshold, and J (Youden index at that threshold).
- **AUC:** kernel-smoothed AUC estimate.
- **AUC_ci, AUC_ci_lo, AUC_ci_hi:** confidence interval for the AUC, based on the kernel-smoothed DeLong-type variance.

- J: kernel-smoothed Youden index estimate.
- J_ci, J_ci_lo, J_ci_hi: Delta-method confidence interval for the Youden index.
- t0: estimated optimal cutoff associated with the Youden index.
- sensitivity, specificity: estimated sensitivity and specificity at the Youden point.
- kernel, bandwidth_method: the chosen kernel and bandwidth selection method.
- hX, hY: selected CDF bandwidths for the non-diseased and diseased groups.
- plot: a ggplot2 object showing the smooth ROC curve, the Youden point, and the annotation label.

Print, plot, and summary methods are available for objects of class "smoothROC".

References

- Khan, R. A., & Ghebremichael, M. (2025). Smooth ROC Curve Estimation. *Journal Name*. (Preprint)
- Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16(19), 2143-2156.
- Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93(444), 1356-1364.
- Zhou, X.-H., & Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, 21(14), 2045-2055.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837-845.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32-35.
- Altman, N., and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46(2), 195-214.
- Bowman, A., Hall, P., and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85(4), 799-808.
- Polansky, A. M., and Baker, E. R. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, 65(1-4), 63-80.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
- Andrews, D. F., & Herzberg, A. M. (2012). *Data: a collection of problems from many fields for the student and research worker*. Springer Science & Business Media.
- The example dataset [dystrophy](#) contains serum biomarkers for Duchenne muscular dystrophy carriers and non-carriers and is used to illustrate kernel-smoothed ROC analysis.

Examples

```
data(dystrophy)

roc <- smoothROC(
  data      = dystrophy,
```

```

    biomarker = "CK",
    status     = "Class",
    diseased   = "carrier",
    kernel     = "biweight",
    bw_method  = "PB",
    alpha      = 0.05,
    logtrans   = TRUE,
    grid_n     = 1000
  )

## Basic use
print(roc)      # Summary print method
summary(roc)    # Summary method
plot(roc)       # ROC plot with annotation label
plot(roc, label = FALSE, youden = FALSE) # ROC plot without annotation label

## Optional: direct slot access
roc$plot        # ggplot object for the ROC curve
head(roc$curve) # FPR, TPR, threshold, and J, one row per unique FPR
roc$AUC         # Kernel-smoothed AUC estimate
roc$AUC_ci      # Confidence interval for AUC
roc$J           # Youden index estimate
roc$J_ci        # Confidence interval for the Youden index
roc$t0          # Estimated optimal cutoff (Youden point)
roc$sensitivity  # Sensitivity at the Youden point
roc$specificity  # Specificity at the Youden point
roc$hX          # Selected CDF bandwidth for the non-diseased group
roc$hY          # Selected CDF bandwidth for the diseased group

```

smoothROCboot	<i>Bootstrap Percentile Confidence Intervals for the Youden Index, Cutoff, and AUC</i>
---------------	--

Description

Computes bootstrap percentile confidence intervals for the kernel-smoothed Youden index, its optimal cutoff, and the AUC, by resampling the non-diseased and diseased groups independently. This is recommended as a companion to [smoothROC](#) in small samples.

Usage

```

smoothROCboot(
  data,
  biomarker,
  status,
  diseased,
  kernel = c("gaussian", "biweight", "epanechnikov"),
  bw_method = c("pdf", "AL", "PB", "BHP", "AR"),

```

```

    alpha = 0.05,
    logtrans = FALSE,
    B = 1000,
    grid_n = 1000,
    tie_tol = 1e-06,
    seed = NULL,
    verbose = TRUE
)

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

Arguments

<code>data</code>	A data frame containing the biomarker and status variables.
<code>biomarker</code>	Character string; name of the numeric column containing biomarker values.
<code>status</code>	Character string; name of the column containing binary disease status.
<code>diseased</code>	The value in <code>status</code> indicating the diseased class.
<code>kernel</code>	Character string; kernel function for smoothing. One of "gaussian", "biweight", or "epanechnikov".
<code>bw_method</code>	Character string; bandwidth selection method. One of "pdf", "AL", "PB", "BHP", or "AR".
<code>alpha</code>	Numeric; significance level for (1 - alpha) confidence intervals (default: 0.05).
<code>logtrans</code>	Logical; apply log-transformation to biomarker values? (default: FALSE).
<code>B</code>	Integer; number of bootstrap replicates (default: 1000).
<code>grid_n</code>	Integer; number of grid points for evaluating the smoothed ROC curve on each replicate (default: 1000).
<code>tie_tol</code>	Numeric; tolerance for identifying tied maximizers of the Youden index when locating the cutoff (default: 1e-6).
<code>seed</code>	Optional integer. If supplied, sets the random seed via <code>set.seed()</code> before re-sampling, so results are reproducible. If NULL (the default), the random number generator is left as is.
<code>verbose</code>	Logical; print bootstrap progress messages via <code>message()</code> ? (default: TRUE). Set to FALSE, or wrap the call in <code>suppressMessages()</code> , to silence.
<code>x</code>	An object of class "smoothROCboot".
<code>...</code>	Additional arguments passed to the S3 methods.

Details

On each bootstrap replicate, both groups are resampled with replacement (X_b of size m from the non-diseased group, Y_b of size n from the diseased group), bandwidths are recomputed on the resampled data, and the kernel-smoothed Youden index, its maximizing cutoff, and the AUC are recomputed numerically using trapezoidal rule. Percentile confidence intervals are then formed from the empirical quantiles of the B bootstrap replicates.

Value

An object of class "smoothROCboot" with components:

- J, J_ci: Youden index estimate and its bootstrap percentile CI.
- t0, t0_ci: optimal cutoff and its bootstrap percentile CI.
- sensitivity, specificity: sensitivity and specificity at the Youden point (observed data).
- AUC, AUC_ci: AUC computed numerically using the trapezoidal rule, and its bootstrap percentile CI.
- boot_J, boot_t0, boot_AUC: the raw bootstrap replicate vectors.
- kernel, bandwidth_method, B, alpha: settings used.

A print method is available for objects of class "smoothROCboot".

Examples

```
data(dystrophy)

## For a real analysis, use larger B (e.g. 1000) and grid_n (e.g. 1000).
boot <- smoothROCboot(
  data      = dystrophy,
  biomarker = "CK",
  status    = "Class",
  diseased  = "carrier",
  kernel    = "biweight",
  bw_method = "PB",
  alpha     = 0.05,
  logtrans  = TRUE,
  B         = 1000,
  grid_n    = 1000,
  seed      = 1691
)

print(boot)
boot$J
boot$J_ci
boot$t0
boot$t0_ci
boot$AUC
boot$AUC_ci
```

Index

* **datasets**

dystrophy, [2](#)

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

plot.smoothROC (smoothROC), [3](#)

print.smoothROC (smoothROC), [3](#)

print.smoothROCboot (smoothROCboot), [7](#)

smoothROC, [3](#), [7](#)

smoothROCboot, [7](#)

summary.smoothROC (smoothROC), [3](#)