Package ‘ICV’

January 15, 2017

Title Indirect cross-validation (ICV) for kernel density estimation

Version 0.1.2017

Author Olga Savchuk

Maintainer Olga Savchuk <olga.y.savchuk@gmail.com>

Description Functions for computing the global and local Gaussian density estimates based on the ICV bandwidth. See the article of Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. Journal of the American Statistical Association, 105(489), 415-423.

Depends R (>= 3.1.1)

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 5.0.1

R topics documented:

C_ICV *The ICV rescaling constant.*

Description

Computing the ICV rescaling constant defined by expression (3) of Savchuk, Hart, and Sheather (2010).

Usage

```
C_ICV(alpha, sigma)
```

Arguments

alpha	first parameter of the selection kernel,
sigma	second parameter of the selection kernel.

Details

Calculation of the ICV rescaling constant C defined by (3) in Savchuk, Hart, and Sheather (2010). The constant is a function of the parameters (α, σ) of the selection kernel L_ICV defined by expression (4) in the same article. The Gaussian kernel is to be used for computing the ultimate density estimate.

Value

The ICV rescaling constant C.

References

Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.

See Also

ICV, h_ICV, L_ICV, MISE_mixnorm, KDE_ICV, LocICV.

Examples

```
# ICV rescaling constant for the selection kernel with (alpha,sigma)=(2.42,5.06).
C_ICV(2.42,5.06)
```

h_ICV *The ICV bandwidth.*

Description

Calculation of the ICV bandwidth for the Gaussian density estimator corresponding to expression (12) of Savchuk, Hart, and Sheather (2010).

Usage

```
h_ICV(x)
```

Arguments

x	numerical vector of data.

Details

Computing the ICV bandwidth for a univariate numerical data set of size $n < 12,058$. The ICV bandwidth is consistent for the MISE optimal bandwidth (see Wand and Jones (1995)). The Gaussian kernel is used for computing the ultimate density estimate. The following values of the paramaters of the selection kernel L_ICV are used: $(\alpha, \sigma) = (2.42, 5.06)$. The ICV bandwidth does not exceed the oversmoothed bandwidth of Terrell (1990). See expression (12) of Savchuk et al. (2010).

Value

The ICV bandwidth.

References

- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.
- Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.
- Terrel, G. (1990). The maximum smoothing principle in density estimation. *Journal of the American Statistical Association*, 85, 470-477.

See Also

ICV, C_ICV, L_ICV, MISE_mixnorm, KDE_ICV, LocICV.

Examples

```
# ICV bandwidth for a random sample of size n=100 from a N(0,1) density.
h_ICV(rnorm(100))
```

h_isemixnorm *The ISE-optimal bandwidth in the case when the true density is the specified mixture of normal distributions.*

Description

Computing the ISE-optimal bandwidth in the case when the true density is the specified mixture of normal distributions and the Gaussian kernel is used to compute the ultimate density estimate.

Usage

```
h_isemixnorm(x, w, mu, sdev)
```

Arguments

`x`	numerical vector of data,
`w`	vector of weighs (positive numbers between 0 and 1 that add up to one),
`mu`	vector of means,
`sdev`	vector of standard deviations.

Details

Computing the ISE-optimal bandwidth (i.e. the minimizer of the ISE function) in the case when the true density is the mixture of normal distributions defined by the vector of weights w, the vector of means μ, and the vector of standard deviations σ. See expression (2.3) of Marron and Wand (1992). It is assumed that the normals are defined as parsimonious as possible. The normal distributions in the mixture should be ordered such that the means in μ are sorted in a nondecreasing order. The Gaussian kernel is used for computing the ultimate density estimate.

Value

The ISE-optimal bandwidth.

References

Marron, J.S., Wand, M.P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2), 712-736.

See Also

`mixnorm`, `ISE_mixnorm`, `MISE_mixnorm`.

Examples

```
# ISE optimal bandwidth for a random sample of size n=100 generated from a normal mixture defined by
# w=c(1/5,1/5,3/5), mu=(0,1/2,13/12), sdev=c(1,2/3,5/9).
# This corresponds to the skewed unimodal density of Marron and Wand (1992).
h_isemixnorm(rnorm(100),c(1/5,1/5,3/5),c(0,1/2,13/12),c(1,2/3,5/9))
```

ICV *The ICV function.*

Description

Computing $ICV(h)$, the value of the ICV function, at a given bandwidth h (vector) for a data set x of size $n < 12,058$. See Savchuk, Hart, and Sheather (2010).

Usage

```
ICV(h, x)
```

Arguments

h	numerical vector of bandwidth values (in the final scale),
x	numerical vecror of data.

Details

Computation of $ICV(h)$ for given h (bandwidth vector) and x (data vector). The sample size $n < 12,058$. The Gaussian kernel is to be used for computing the ultimate kernel density estimate. The parameters of the selection kernel are $(\alpha, \sigma) = (2.42, 5.06)$. The ICV bandwidth h_ICV is the minimizer of the ICV function.

Value

The value of $ICV(h)$ for given h and data (x).

References

Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.

See Also

h_ICV, C_ICV, L_ICV, MISE_mixnorm, KDE_ICV, LocICV.

Examples

```
#Example 1. Computation of ICV(h) at h=0.4 for a random sample of size n=100 from a N(0,1) distribution.
ICV(0.4,rnorm(100))

#Example 2. (Calculations for a random sample of size n=250 from the separated bimodal density).
w=c(1/2,1/2)
mu=c(-3/2,3/2)
sdev=c(1/2,1/2)
# Generating a sample of size n=250 from the separated bimodal density of Marron and Wand (1992).
dat=mixnorm(250,w,mu,sdev)
h_OS=(243/(35*length(dat)*2*sqrt(pi)))^0.2*sd(dat)    # Computing the oversmoothed bandwidth.
h_opt=round(h_ICV(dat),digits=4)
harg=seq(0.1,3,len=100)
X11()
plot(harg,ICV(harg,x=dat),'l',lwd=3,xlab="h",ylab="ICV",cex.lab=1.7,cex.axis=1.7)
```

```
title(main="ICV(h)",cex.main=1.7)
lines(c(h_OS,h_OS),c(-0.5,0.5),lty="dashed",lwd=3)
legend(0.75,-0.05,legend="Vertical line shows the oversmothed bandwidth")
legend(1.35,0.1,legend=paste("h_ICV=",h_opt),cex=2,bty="n")
# Notice that the scale of the ICV function is such that its minimizer is the ICV bandwidth h_ICV.
# Thus, no additional rescaling of the ICV function's minimizer is needed to obtain the ICV bandwidth.
X11()
dens=density(dat,bw=h_opt)
plot(dens,main="",cex.lab=1.7,cex.axis=1.7,lwd=3,xlab=paste("n=250, ","h=h_ICV=",h_opt),ylim=c(0,0.45))
title(main="KDE based on h_ICV",cex.main=1.7)
arg=seq(-3.5,3.5,len=1000)
lines(arg,w[1]*dnorm(arg,mu[1],sd=sdev[1])+w[2]*dnorm(arg,mu[2],sd=sdev[2]),lwd=3,lty="dashed")
legend(-1,0.45,lty=c("solid","dashed"),lwd=c(3,3),legend=c("ICV estimate","True density"),bty="n")
```

ISE_mixnorm — *The ISE function in the case when the underlying density is the specified mixture of normal distributions.*

Description

Computing $ISE(h)$ for given h in the case when the underlying density is the specified mixture of normal distributions and the Gaussian kernel is used to compute the ultimate density estimate.

Usage

```
ISE_mixnorm(h, x, w, mu, sdev)
```

Arguments

h	numerical vector of bandwidth values,
x	numerical vector of data,
w	vector of weighs (positive numbers between 0 and 1 that add up to one),
mu	vector of means,
sdev	vector of standard deviations.

Details

Computing $ISE(h)$ in the case when the true density is the mixture of normal distributions defined by the vector of weights w, the vector of means μ, and the vector of standard deviations σ. See expression (2.3) of Marron and Wand (1992). It is assumed that the normals are defined as parsimonious as possible. The normal distributions in the mixture should be ordered such that the means in μ are arranged in a nondecreasing order. The Gaussian kernel is to be used for computing the ultimate density estimate.

Value

The vector of ISE values corresponding to the specifies values of h.

References

Marron, J.S., Wand, M.P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2), 712-736.

See Also

mixnorm, h_isemixnorm, MISE_mixnorm.

Examples

```
harg=seq(0.01,1,len=100)
w=c(3/4,1/4)
mu=c(0,3/2)
sdev=c(1,1/3)
# The vectors w, mu, and sdev define the skewed bimodal density of Marron and Wand (1992).
dat=mixnorm(300,w,mu,sdev)
h_ISE=round(h_isemixnorm(dat,w,mu,sdev),digits=4)
ISEarray=ISE_mixnorm(harg,dat,w,mu,sdev)
X11()
plot(harg,ISEarray,'l',lwd=3,xlab="h, n=300",ylab="ISE",cex.lab=1.7,cex.axis=1.7,main="")
title(main="ISE(h)",cex.main=1.7)
legend(0.2,0.08,legend=paste("h_ISE=",h_ISE),cex=2)
```

KDE_ICV *Computing the kernel density estimate based on the ICV bandwidth.*

Description

Computing the Gaussian density estimate based on h_ICV.

Usage

```
KDE_ICV(x)
```

Arguments

x	numerical vector of data.

Details

Computing the Gaussian density estimate based on h_ICV. The ICV selection kernel L_ICV is based on $(\alpha, \sigma) = (2.42, 5.06)$.

Value

A list containing the following components:

arg vector of sorted values of the argument at which the density estmate is computed,

y vector of density estimates at the corresponding values of the argument.

The function also produces a graph of the resulting ICV kernel density estimate.

References

- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.
- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2009). An empirical study of indirect cross-validation. *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger.* World Scientific Publishing, 288-308.

See Also

ICV, h_ICV, L_ICV, LocICV, C_ICV.

Examples

```
#Example (Density estimate for eruption duration of the Old Faithful Geyser).
data=faithful[[1]]
dens=KDE_ICV(data)
```

LocICV *The local ICV function.*

Description

Computing the local ICV function at the given estimation point, as explained in Section 6 of Savchuk, Hart, and Sheather (2010).

Usage

```
LocICV(h, xest, x, eta, alpha, sigma)
```

Arguments

h	bandwidth (scalar) in the final scale,
xest	estimation point (scalar),
x	numerical vector of data,
eta	smoothing parameter,
alpha	first parameter of the selection kernel,
sigma	second parameter of the selection kernel.

Details

Calculation of the local ICV function at the given estimation point xest. The Gaussian kernel is used for local weighting. The ultimate kernel density estimate is computed based on the Gaussian kernel. The parameters of the selection kernel L_ICV are α and σ. The minimizer of the local ICV function is to be used in computing the ultimate density estimate without additional rescaling. Parameter η is a smoothing parameter that determines the degree to which the cross-validation is local. A suggested value of η is $\eta = R/20$, where R is the range of data.

Value

The value of the local ICV function at the fixed estimation point and for the specified value of the bandwidth (see Section 6 of Savchuk, Hart, and Sheather (2010)).

References

- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.
- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2009). An empirical study of indirect cross-validation. *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger.* World Scientific Publishing, 288-308.
- Hall, P., and Schukany, W. R. (1989). A local cross-validation algorithm. *Statistics and Probability Letters*, 8, 109-117.

See Also

h_ICV, C_ICV, L_ICV, MISE_mixnorm, ICV, KDE_ICV.

Examples

```
# Local ICV function for a random sample of size n=150 from the kurtotic density of Marron and Wand (1992).
dat=mixnorm(150,c(2/3,1/3),c(0,0),c(1,1/10))
a=2.42   # alpha
s=5.06   # sigma
harg=seq(0.025,1,len=100)
Xest=0.1    # estimation point
LocICV_Xest=numeric(length(harg))
for(i in 1:length(harg))
  LocICV_Xest[i]=LocICV(harg[i],Xest,dat,0.2,a,s)
h_Xest=optimize(LocICV,c(0.001,0.2),tol=0.001,xest=Xest,eta=0.2,x=dat,alpha=a,sigma=s)$minimum
h_Xest=round(h_Xest,digits=4)
X11()
plot(harg,LocICV_Xest,'l',lwd=3,xlab="harg",ylab="LocICV_Xest",main="",cex.lab=1.7, cex.axis=1.7)
title(main=paste("Local ICV function at x=",Xest),cex.main=1.7)
legend(0.1,max(LocICV_Xest),legend=paste("h_x=",h_Xest),cex=1.7)
legend(0.2,max(LocICV_Xest)-0.15,legend="Note: first local minimizer is used", cex=1.5,bty="n")
```

L_ICV — *The ICV selection kernel.*

Description

The ICV selection kernel L defined by expression (4) of Savchuk, Hart, and Sheather (2010).

Usage

```
L_ICV(u, alpha, sigma)
```

Arguments

u	numerical argument of the selection kernel,
alpha	first parameter of the selection kernel,
sigma	second parameter of the selection kernel.

Details

The ICV selection kernel $L(u;\alpha,\sigma) = (1+\alpha)\phi(u) - \alpha\phi(u/\sigma)/\sigma$, where ϕ denotes the Gaussian kernel.

Value

The value of $L(u;\alpha,\sigma)$.

References

Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.

See Also

ICV, h_ICV, C_ICV, MISE_mixnorm, KDE_ICV, LocICV.

Examples

```
# Graph of the ICV selection kernel with (alpha,sigma)=(2.42,5.06).
u=seq(-10,10,len=1000)
kern=L_ICV(u,2.42,5.06)
X11()
plot(u,kern,'l',lwd=2,ylim=c(-0.2,1.2),ylab="kernel",cex.lab=1.7,cex.axis=1.7,main="")
lines(u,dnorm(u),lwd=3,lty="dashed")
title(main="Selection kernel with (alpha,sigma)=(2.42,5.06)",cex.main=1.6)
legend(-11, 1.2, legend=c("ICV kernel","Gaussian kernel"),lwd=c(3,3),lty=c(1,2),bty="n",cex=1.3)
```

MISE_mixnorm

The MISE function in the case when the true density is the specified mixture of normal distributions and the selection kernel L_ICV *is used in the cross-validation stage.*

Description

Computing $MISE(h)$ for given h in the case when the true density is the specified mixture of normal distributions and the kernel is L_ICV defined by expression (4) of Savchuk, Hart, and Sheather (2010).

Usage

```
MISE_mixnorm(h, n, alpha, sigma, w, mu, sdev)
```

Arguments

h	numerical vector of bandwidth values,
n	sample size,
alpha	first parameter of the selection kernel,
sigma	second parameter of the selection kernel,
w	vector of weighs (positive numbers between 0 and 1 that add up to one),
mu	vector of means,
sdev	vector of standard deviations.

Details

Calculation of $MISE(h)$ in the case when the true density is the mixture of normal distributions defined by the vector of weights w, the vector of means μ, and the vector of standard deviations σ. See expression (2.3) of Marron and Wand (1992). It is assumed that the normals are defined as parsimonious as possible. The normal distributions in the mixture should be ordered such that the means in μ are arranged in a nondecreasing order. The MISE function is based on the selection kernel L_ICV defined by expression (4) of Savchuk, Hart, and Sheather (2010). Notice that the Gaussian kernel ϕ is the special case of L_ICV given that **(Case 1)** $\alpha = 0$, $\sigma > 0$ or **(Case 2)** $\sigma = 1$, $-\infty < \alpha < \infty$.

Value

The vector of MISE values corresponding to the specified values of h.

References

- Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489), 415-423.
- Marron, J.S., Wand, M.P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2), 712-736.

See Also

mixnorm, ISE_mixnorm, h_isemixnorm, L_ICV, ICV, h_ICV, C_ICV.

Examples

```
# Example 1. MISE for the separated bimodal density of Marron and Wand (1992).
# in the case (alpha,sigma)=(2.42,5.06), n=100.
harray=seq(0.05,1,len=1000)
w=c(1/2,1/2)
m=c(-3/2,3/2)
s=c(1/2,1/2)
MISEarray=MISE_mixnorm(harray,100,2.42,5.06,w,m,s)
hopt=round(optimize(MISE_mixnorm,c(0.01,1),n=100,alpha=2.42,sigma=5.06,w=w,mu=m,sdev=s)$minimum,digits=4)
X11()
plot(harray,MISEarray,'l',lwd=3,xlab="h",ylab="MISE",cex.lab=1.7,cex.axis=1.7,main="")
title(main="MISE(h) for the separated bimodal density",cex.main=1.5)
legend(0.45,0.45,legend=c(paste("h_MISE=",hopt),"n=100"),bty="n",cex=1.7)

# Example 2. MISE for the N(0,1) density in the case of the Gaussian kernel and n=500.
harray=seq(0.03,1,len=1000)
MISEarray=MISE_mixnorm(harray,500,1,1,1,0,1)
hopt=round(optimize(MISE_mixnorm,c(0.01,1),n=500,alpha=1,sigma=1,w=1,mu=0,sdev=1)$minimum,digits=4)
X11()
plot(harray,MISEarray,'l',lwd=3,xlab="h",ylab="MISE",cex.lab=1.7,cex.axis=1.7,main="")
title(main="MISE(h) for the standard normal density",cex.main=1.7)
legend(0.2,0.02,legend=c(paste("h_MISE=",hopt),"n=500"),bty="n",cex=1.7)
```

`mixnorm` *Generating a random sample from the specified mixture of normal distributions.*

Description

Generating a random sample of size n from the normal mixture defined by expression (2.3) of Marron and Wand (1992).

Usage

```
mixnorm(n, w, mu, sdev)
```

Arguments

`n`	desired sample size,
`w`	vector of weighs (positive numbers between 0 and 1 that add up to one),
`mu`	vector of means,
`sdev`	vector of standard deviations.

Details

Producing a random sample of size n from the normal mixture defined by the vector of weights w, the vector of means μ, and the vector of standard deviations σ. See Marron and Wand (1992). It is assumed that the normals are defined as parsimonious as possible. The normal distributions in the mixture should be ordered such that the means in μ are arranged in a nondecreasing order.

Value

A random sample of size n from the specified mixture of normals.

References

Marron, J.S., Wand, M.P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2), 712-736.

See Also

`ISE_mixnorm`, `h_isemixnorm`, `MISE_mixnorm`.

Examples

```
# Generating a sample of size n=300 from the separated bimodal density of Marron and Wand (1992).
w=c(0.5,0.5)
mu=c(-3/2,3/2)
sdev=c(1/2,1/2)
dat=mixnorm(300,w,mu,sdev) # generated data vector
arg=seq(-4,4,len=1000)  # argument
f=w[1]*dnorm(arg,mu[1],sd=sdev[1])+w[2]*dnorm(arg,mu[2],sd=sdev[2])     # true density
X11()
hist(dat,freq=F,ylab="",main="",cex.lab=1.7,cex.axis=1.7,xlim=c(-4,4),lwd=2,ylim=c(0,0.45),col='grey')
title(main="Separated bimodal density",cex.main=1.7)
```

```
legend(-5,0.4,legend="n=300",cex=2,bty="n")
lines(arg,f,lwd=3,'l')
```

Index