PENAKSIRAN FUNGSI DENSITAS UNTUK SUATU DATA DENGAN PENAKSIR KERNEL

Netty Sunandi
R. Alam Malau

ABSTRACT

One of the estimating of density function which has been recognized is histogram. Histogram has some weaknesses, i.e. the different starting points and the width of class intervals. Different starting points or different class intervals result different histogram forms. This article is about the estimating of the density function by using kernel function. This method does not require the determination starting points and the interval class width. The obtained curve has a smooth density function, a small sampling variance, and the important information from data are still kept.

Keywords : kernel function, oversmoothing, cross validation, Schwartz Bayesian Criterion

PENDAHULUAN

Fungsi densitas merupakan suatu konsep dasar dalam statistika yaitu sebagai penentu besar probabilitas untuk suatu selang yang diberikan.

Misalnya: $P(a < X < b) = \int_a^b f(x)\,dx$, di mana $f(x)$ ialah fungsi densitas dari peubah acak X.

Dalam praktek fungsi densitas dari suatu peubah acak tidak diketahui, jadi perlu ditaksir.

Pada penaksiran fungsi densitas secara parametrik diperlukan asumsi mengenai distribusi suatu peubah acak (misalnya distribusi normal, gamma) dan yang ditaksir ialah parameter-parameter dari distribusi tersebut dengan menggunakan data tentang peubah acaknya. Sedangkan penaksiran fungsi densitas secara nonparametrik (yang akan dibahas di sini) tidak memerlukan asumsi mengenai distribusi, data diperbolehkan berbicara mengenai dirinya sendiri. Pendekatan ini digunakan jika tidak ada informasi yang tepat mengenai bentuk dari fungsi densitas yang sebenarnya. Taksiran fungsi densitas yang diperoleh diharapkan dapat menggambarkan keadaan data tersebut. Penaksir tersebut dapat memberikan petunjuk mengenai kemiringan, modus ganda, variansi, klasifikasi dari suatu peubah acak.[5]

Tujuan penaksiran fungsi densitas ini ialah untuk mendapatkan kurva fungsi densitas yang merupakan kurva mulus dengan variansi sampling tidak besar dan informasi penting dari data tidak hilang.

Salah satu penaksiran fungsi densitas yang sudah dikenal ialah histogram. Histogram mempunyai beberapa kelemahan yaitu bentuknya dipengaruhi oleh pemilihan titik awal dan lebar interval kelas. Dengan titik awal yang berbeda akan didapat bentuk histogram yang berbeda, begitu pula dengan lebar interval kelas yang berbeda. Juga karena histogram mempunyai bentuk yang tidak kontinu pada batas kelas, sedangkan fungsi densitas yang sebenarnya merupakan fungsi yang kontinu, maka di sini akan dibahas penaksiran fungsi densitas yang memberikan suatu fungsi yang kontinu. Salah satunya ialah penaksiran fungsi densitas dengan metode kernel.

Penaksir kernel untuk suatu fungsi densitas dari peubah acak X berbentuk $\hat{f}_n(x)=\frac{1}{nh}\sum_{i-1}^{n}K\left(\frac{x-X_i}{h}\right)$, dengan *x* = suatu nilai tertentu, X_i = peubah acak yang independen dan berdistribusi identik, *K(.)* = fungsi kernel, *n* = besar sampel (banyak data), *h* = lebar bandwith.

Penaksir ini tidak bergantung kepada pemilihan titik awal tetapi bergantung kepada h, sehingga supaya didapat penaksir fungsi densitas yang baik, maka pemilihan h menjadi penting.

Pemilihan h optimal akan ditentukan dengan memperhatikan fungsi validasi silang. [3] Setelah h optimal dirumuskan secara teoretis, penghitungan h optimal untuk suatu data yang diberikan didapat dengan bantuan komputer di mana di sini akan digunakan program S-plus. Kemudian dapat dibuat penaksiran fungsi densitasnya.([1],[2])

Penaksir Kernel

Jika peubah acak *X* mempunyai fungsi densitas *f,* maka $f(x)=\lim^{\frac{1}{2h}} P(x-h<X<x+h)$

$P(x-h < X < x+h)$ ditaksir dengan $\frac{f_i}{n}$, di mana *n* ialah besar sampel dengan peubah acak $X_1,X_2,\ldots,X_n$ dan f_i ialah banyak X_i yang ada dalam selang *(x-h,x+h).* Sehingga suatu penaksir untuk f dapat ditentukan sebagai $\hat{f}(x)=\frac{1}{2hn}\left[banyak\ X_{\hat{ij}}(x-h,x+h)\right]$ dengan memilih h yang kecil.

$\hat{f}(x)$ ini disebut penaksir naïf.[5]

Jika $\hat{f}(x)$ dinyatakan dengan fungsi bobot w, maka $w(u)=\begin{cases}\frac{1}{2}, & jika|u|<1\\ 0, & jika|u|\geq 1\end{cases}$

sehingga : $\hat{f}(x)=\frac{1}{n}\sum_{i-1}^{n}\frac{1}{h}w\left(\frac{x-X_i}{h}\right)$

Penaksir naif menghasilkan $\hat{f}$ y ang tidak kontinu (ada tangga-tangga). Kesulitan ini diatasi dengan mengganti fungsi w dengan fungsi K yang memenuhi:

$\int K(t)dt=1$ dan *K* simetris terhadap titik 0.

Untuk mendapat $\hat{f}$ yang kontinu dipilih K yang kontinu.

Contoh kernel adalah fungsi densitas Gaussian : $K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$, ¥ < U < ¥

Jadi didapat penaksir kernel untuk suatu fungsi densitas f :

$$\hat{f}(x)=\frac{1}{nh}\sum_{i-1}^{n}K\left(\frac{x-X_i}{h}\right)$$

dengan *x* = suatu nilai tertentu, X_i = peubah acak yang independen dan berdistribusi identik, *K(.)* = fungsi kernel, *n* = besar sampel (banyak data), *h* = lebar pita (*bandwith*).

Integral dari $\hat{f}(x)_h(x)$ *yaitu* $\int \hat{f}_n(x)dx=1$, sehingga $\hat{f}_h(x)$ merupakan suatu fungsi densitas.

Karena penaksir kernel dijabarkan dari penaksir naïf, maka penaksir kernel tidak bergantung kepada pemilihan titik awal. Dalam penaksir kernel terdapat dua parameter yaitu h yang juga disebut lebar pita (*bandwith*) dan fungsi kernel K. Secara teoritis [3] dapat dijelaskan bahwa dengan menggunakan fungsi kernel yang berbeda bentuk global dari taksiran fungsi densitas suatu data ialah sama. Sedangkan perubahan h mempengaruhi bentuk global dari taksiran fungsi densitas. Jadi penentuan h menjadi penting.

Validasi Silang (Cross Validation)

Akan ditentukan bagaimana menentukan h optimal yang dapat digunakan dalam praktek. Di sini akan dibahas dua macam validasi silang, yaitu :[3]

1. Validasi silang dengan memaksimumkan fungsi kemungkinan (validasi silang Kulback Liebler).
2. Validasi silang kuadrat terkecil.

Untuk yang pertama, mula-mula ingin diuji:

H_0 : $\hat{f}_h(x) = f(x)$ terhadap H_1 : $\hat{f}_h(x)^1 f(x)$ untuk suatu h tertentu.

Akan digunakan uji ratio kemungkinan (likelihood ratio test) $f(x)/\hat{f}_h(x)$. Untuk h yang baik, statistik ini dekat dengan satu. Jadi $E[\log\left(\frac{f}{\hat{f}_h}\right)(x)]$ dekat dengan 0.

$E[\log\left(\frac{f}{\hat{f}_h}\right)(x)] = \int \log\left(\frac{f}{\hat{f}_h}\right)(x)\, f(x)\, dx$ dikenal sebagai informasi Kulback-Liebler.

Informasi Kulback-Liebler ini memenuhi sifat-sifat suatu jarak sehingga dinotasikan dengan $d_{KL}(f, \hat{f}_h)$ dan $d_{KL}(f, \hat{f}_h)$ ini tidak dapat dihitung dari data, karena fungsi densitas $f(x)$ tak diketahui. Tetapi dapat disimpulkan bahwa jika $d_{KL}(f, \hat{f}_h)$ dekat nol, maka H_0 benar. Oleh karena itu akan dicari h yang meminimumkan $d_{KL}(f, \hat{f}_h)$.

Misalkan bahwa ada observasi-observasi tambahan $X_1,X_2,\ldots,X_n$ yang independen. Fungsi kemungkinan (*likelihood function*) untuk observasi-observasi tersebut ialah: $f(X_1).f(X_2)\ldots f(X_n)$.

Taksiran dari fungsi kemungkinan di atas ialah : $\hat{f}_h(X_1).\ \hat{f}_h(X_2)\ \ldots\ \hat{f}_h(X_n) = \prod_{i=1}^{n} \hat{f}_h(X_i)$

Tetapi biasanya tidak ada observasi-observasi tambahan dari suatu data yang diberikan. Cara mengatasi penentuan $\hat{f}_h(X_i)$ ialah dengan menggantinya dengan $\hat{f}_{h,,i}(X_i) = \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)$

yaitu *satu* observasi untuk prediksi dan *(n-1)*observasi sisanya untuk menaksir. Taksiran seperti ini disebut validasi silang (*cross validation* = *CV*) atau leave-one-out estimate.

Jadi fungsi kemungkinan ditaksir dengan : $\prod_{i=1}^{n} \hat{f}_{h,i}(X_i) = \frac{1}{(n-1)^n h^n} \prod_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)$

Jika statistik ini diambil logaritmanya, kemudian dikalikan dengan *1/n* didapat

$$CV_{KL}(h) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_{h,i}(X_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \{\log(\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right))\} - \{\log(n-1)h\}$$

h optimal ialah yang *h* yang memaksimumkan $CV_{KL}(h)$ dengan alasan sebagai berikut:

$E[CV_{KL}(h)] = E[\log \hat{f}_{h,i}(X_i)]$

$= \int [\log \hat{f}_{h,i}(X_i)]\ f(x)\ dx$

$= - d_{KL}(f, \hat{f}_h) + \int [\log f(x)] f(x)\ dx$

Suku ke dua tidak bergantung kepada h, sehingga nilai h yang membuat : Maks $CV_{KL}(h)$ = Maks $E[CV_{KL}(h)]$ = Min $d_{KL}(f, \hat{f}_h)$.

Untuk yang kedua, tinjau integrasi dari kuadrat kesalahan $\hat{f}_h$ terhadap f yaitu ISE (*Integrated Squared Error*) dari h :

$$ISE(h) = d_1(h) = \int (\hat{f}_h - f)^2(x)\ dx$$
$$= \int \hat{f}_h^2(x)\ dx - 2\int (\hat{f}_h f)(x)\ dx + \int f^2(x)\ dx$$

Suku pertama dihitung dari data, suku ketiga tidak bergantung kepada h, jadi suku kedua ialah satu-satunya suku yang harus ditaksir dari data. Nilai h yang optimal ialah nilai h yang meminimumkan ISE(h). Nilai h optimal tersebut membuat:

$Min\ ISE(h) = Min\ \{d_1(h) - \int f^2(x)\,dx\}$

$= \text{Min}\ \{ \int \hat{f}_h^2(x)\,dx - 2\int (\hat{f}_h f)(x)\,dx\}$

$\int (\hat{f}_h f)(x)\,dx$ dapat dinyatakan sebagai $E[\hat{f}_h(x)]$.

Sehingga sebagai taksiran dari $E[\hat{f}_h(x)]$ diambil $\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i)$.

Jadi nilai h yang optimal akan meminimumkan : $CV_{LS}(h) = \int \hat{f}_h^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i)$

Terlalu mulus (Oversmoothing) [4]

Suatu batas atas untuk h optimal dengan kernel Gaussian ialah : $h_{os} = 1.144\ s\ n^{-1/5}$

Lebar pita terbaik

Lebar bandwith h yang terbaik merupakan nilai minimum antara h_{opt} dan h_{os}. Lebar pita terbaik ini menjadi lebar bandwith untuk menaksir fungsi densitas suatu data.

Penggunaan

Penaksiran suatu fungsi densitas akan ditentukan yaitu dengan menggambarkan kurva penaksiran fungsi densitasnya dengan menggunakan program S-plus.([1], [2])

Untuk menaksir fungsi densitas dengan penaksir kernel perlu ditentukan lebar bandwith (=h) yang sesuai. Karena itu dibuat fungsi validasi silang dengan program S-plus. Dalam validasi silang Kulback-Liebler ditentukan h_{opt} yang memaksimumkan fungsi validasi silangnya. Sedangkan dalam validasi silang kuadrat terkecil tentukan h_{opt} yang meminimumkan validasi silangnya. Kemudian bandingkan h_{opt} tersebut dengan h_{os}. Pilih h yang merupakan minimum antara h_{opt} dan h_{os}. Lebar bandwith terbaik ini menjadi lebar pita untuk menaksir fungsi densitas suatu data.

Menentukan suatu fungsi densitas untuk suatu data

Setelah memperoleh taksiran fungsi densitas suatu data, langkah selanjutnya adalah menentukan fungsi densitasnya, dengan kata lain menentukan parameter-parameternya.

Dengan mengacu pada kurva taksiran fungsi densitas dan membandingkannya dengan kurva fungsi densitas (teoretis) dari distribusi yang sudah dikenal, dapat ditentukan kandidat-kandidat distribusi dari suatu data. Seleksilah kandidat-kandidat tersebut dengan PP dan QQ plot.

Untuk kandidat-kandidat yang masuk seleksi akan ditentukan taksiran parameternya secara parametrik. Taksiran titik dari parameter-parameter suatu fungsi densitas dapat ditentukan dengan metode momen dan metode maksimum likelihood.

Jadi akan didapat sekelompok fungsi densitas yang dapat dijadikan kandidat dari fungsi densitas yang akan dicari. Yang dibutuhkan sekarang ialah menentukan pilihan pada suatu fungsi densitas yang paling tepat. Salah satunya adalah menggunakan salah satu uji "*Goosness of fit*" yaitu uji Kolmogorov Smirnov.

Uji Kolmogorov-Smirnov

Misalkan akan diuji : H_0 : $F(x) = F(x; \hat{\theta})$

H_1 : tidak demikian

$D_n = \max_x |F_n(x) - F(x;\hat{\theta})|$, dengan F_n adalah fungsi tangga sehingga untuk $F(.;\hat{\theta})$kontinu, nilai maksimum akan diperoleh pada titik lompatan x_i *atau* $\overline{x_i}$. Maksimum D_n dibandingkan dengan nilai dari tabel sebagai berikut :

Tingkat Signifikansi	Nilai Kritis
0.20	$1.07/\sqrt{n}$
0.10	$1.22/\sqrt{n}$
0.05	$1.36/\sqrt{n}$
0.01	$1.63/\sqrt{n}$

Ho ditolak jika D_n > Nilai kritis

Jika fungsi densitas yang satu lebih sederhana dari yang lain yaitu jika yang satu mempunyai banyak parameter berbeda dengan yang lain maka digunakan Uji Ratio Likelihood kemudian NLL(*negative log likelihood*) dikoreksi dengan '*penalty*'. Uji likelihood ratio ialah suatu uji yang menguji

Ho : distribusi yang lebih sederhana lebih baik (I)

H_1 : distribusi yang lebih kompleks lebih baik (II)

Statistik Uji yang digunakan adalah :

$X^2 = 2\ |NLL\ I - NLL\ II| \sim \chi^2_{\Delta p}$

Dengan NLL = negative log likelihood yaitu negatif dari besarnya fungsi likelihood pada titik maksimum, Δ_p = beda banyaknya parameter pada I dan II.

Ho ditolak jika X^2 > nilai tabel.

Besarnya 'penalty' untuk NLL adalah r log (n / 2p) dengan r adalah banyaknya parameter dan n adalah ukuran sampel.

Metode ini dikenal dengan Schwartz Bayesian Criterion (SBC)

Contoh:

Dengan n = 217, suatu data mempunyai kandidat sebagai berikut:

Model	NLL	Penalty	Score
Inverse exponential	520,27	3,54	523,81
Lognormal	498,29	7,08	505,37
Burr	498,41	10,63	509,04

Dalam contoh ini dapat disimpulkan bahwa data berdistribusi lognormal.

KESIMPULAN

Berdasarkan uraian di atas maka dapat disimpulkan bahwa untuk mencari fungsi densitas yang terbaik perhatikan :

- Taksir terlebih dahulu fungsi densitas dan tentukan kandidat-kandidat fungsi densitas teoretis. Kemudian bandingkan taksiran fungsi densitas dengan kandidat.
- NLL (makin kecil makin bagus)
- Nilai dari Statistik Uji Kolmogorov Smirnov (makin kecil makin bagus)

REFERENSI

1. Becker, R. A., Chamber, J.M. 1988. *The New S Language.* Bell Telephone Laboratories, Inc., Murray Hill, New Yersey.
2. Elan Computer Group . 1993. *S-plus User's Manual Version 3.2.* Math. Soft. Inc., Seattle.
3. Hardle,W. 1991. *Smoothing Techniques.* Springer-Verlag, New York.
4. Scott, D.W. 1992. *Multivariate Density Estimation.* John Wiley, New York.
5. Silverman,B.W. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.
6. Hogg, R. V., and Klugman, S. A. 1984. *Loss Distributions*. Library of Congress Cataloging in Publications Data.

Lampiran: Contoh Penggunaan

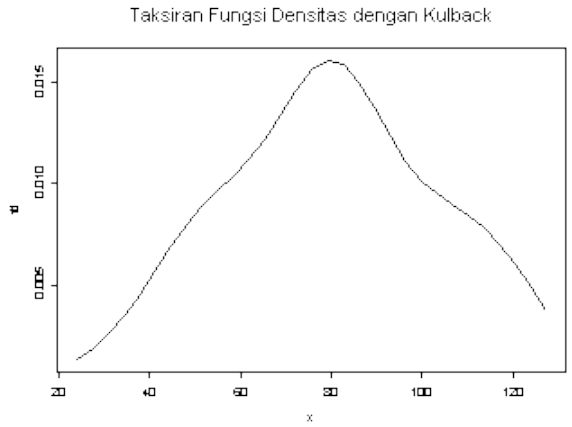
Suatu data berukuran n=35.

6766 7123 10562 14474 15351 16983 18383 19030 25304 29112 30146 33727
40596 41409 47905 49397 52600 59917 63123 77809 102942 103217 123680 140136
192013 198446 227338 329511 361200 421680 513586 545778 750389 863881 1638000

Penaksiran Fungsi Densitas dengan Program S-PLUS

```
#Masukkan data yang ada di "a:\data.txt".
> X <- scan("a:/data.txt")
#Untuk mengurutkan data :
> X <- sort(X)
#Menghitung nilai batas atas h (= hos) :
> s <- sqrt(var(X))
> hend <- 1.144*s*(length(X)^(-0.2))
> hend
#Masukkan function Kulback dan densitas :
> kulb<-function(X,h)
+ {
+ cvkl<-1:length(h)
+ jum<-1:length(X)
+ a<-1:length(X)
+ for(j in 1:length(h))
+ {
+ for(i in 1:length(X))
+ {
+ d<-(X[i]-X)/h[j]
+ K<-dnorm(d)
+ jum[i]<-sum(K)-dnorm(0)
+ a[i]<-log(jum[i])
+ }
+ n<-length(X)
+ cvkl[j]<-(sum(a)/n)-log((n-1)*h[j])
+ }
```

```
+ cvkl
+ }
> densitas<- function(x,h,X)
+ {K<-0
+ for(i in 1:length(X))
+ {K<-K+dnorm((X[i]-x)/h)}
+ fh<-K/(length(X)*h)
+ fh}
#Masukkan nilai nilai h :
> h <- 1:20
> h
#Mencari nilai h optimal dari CV_KL(h) yang dinamakan hmax :
> m <- kulb(X,h)
> cbind(m,h)
> max(m)
> hmax <- 10
#Karena nilai hmax lebih kecil dari pada hend maka untuk membuat fungsi #densitas digunakan nilai hmax.
#Menentukan nilai x yang akan dicari nilai fungsi densitasnya :
> lb <- (max(X)+1-min(X)+1)/30
> x <- min(X)-1+lb*c(0:30)
> x
#Bandingkan dengan nilai data :
> X
Menentukan nilai nilai fungsi densitas dari x :
> fd <- densitas(x,hmax,X)
Membuat grafik fungsi densitas dengan metode Kulback Liebler :
> win.graph()
> plot(x,fd,type="l",main="Taksiran Fungsi Densitas dengan Kulback")
>
```



#Masukkan function CVLS :

```
> cvls<-function(X,h)
+ {
+ m<-length(h)
+ cv<-rep(0,m)
+ n<-length(X)
+ n1<-n-1
+ nn1<-2+(2/n1)
+ for(i in 1:n1)
+ {
+ for(j in (i+1):n)
+ {
+ d<-X[j]-X[i]
+ for(k in 1:m)
+ {
+ dh<-d/h[k]
+ cv[k]<-cv[k]+dnorm(dh,0,1.414)-(dnorm(dh)*nn1)
+ } } }
+ for(k in 1:m)
+ {
+ cv[k]<-cv[k]+(dnorm(0,0,1.414)*n*0.5)
```

```
+ cv[k]<-cv[k]*2/(n*n*h[k])
+ }
+ cv }
```

#Mencari nilai h optimal dengan meminimumkan $CV_{LS}(h)$ yang dinamakan #hmin :

```
> p <-cvls(X,h)
> cbind(p,h)
> min(p)
> hmin <- 9
```

#Membuat grafik fungsi densitas dengan metode kuadrat terkecil :

```
> win.graph()
> pdf <- densitas(x,hmin,X)
> plot(x,pdf,main="Taksiran Fungsi Densitas dengan LS")
```