Estimasi Densitas Mulus dengan Metode Kernel

(Kernel Method in Smooth Density Estimation)

Oleh Suparti[1)] dan Sudargo[2)]

Abstract

Let $\{X_i\}_{i = 1,2,\ldots,n}$ be independent observation data from a distribution with an unknown density function f . The function f could be estimated by parametric and nonparametric approach.

In nonparametric approach, the function f is assumed to be a smooth function or quadratic integrable function, so the function f could be estimated by kernel estimator.

The smoothing level of kernel estimator depends to the smoothing parameter. The big smoothing parameter gives a estimation function which over smooth and the contrary.

Key words : smooth density, kernel estimator.

Dalam analisis data cenderung diartikan sebagai proses perhitungan dalam penerapan metode statistika, misalnya perhitungan mean, varian, koefisien regresi ataupun perhitungan jumlah kuadrat dalam analisa varian, sehingga peranan dan kegunaan sebenarnya menjadi sering terlupakan. Proses analisis data pada dasarnya meliputi upaya penelusuran dan pengungkapan informasi yang relevan yang terkandung dalam data seperti penelusuran dan pengungkapan struktur dan pola data, dan penyajian hasilnya dalam bentuk lebih ringkas dan sederhana, sehingga pada akhirnya mengarah kepada keperluan adanya penjelasan dan penafsiran. Penelusuran

[1)] Staf Pengajar Jur. Matematika ,FMIPA, Undip, Semarang
[2)] Staf Pengajar Jur. Pend. Matematika ,IKIP PGRI, Semarang

struktur data bertujuan memeriksa apakah suatu data dapat diwakili oleh suatu model tertentu, sedangkan dalam penelusuran pola data bertujuan untuk memeriksa apakah distribusi datanya cenderung mengumpul di satu nilai tertentu atau pada beberapa nilai.

Jika diberikan data pengamatan independen $\{X_i\}_{i = 1,2,\ldots,n}$, untuk menentukan distribusi dari X ekivalen dengan menentukan fungsi densitasnya. Untuk mengestimasi fungsi densitas f dapat dilakukan dengan dua pendekatan yaitu pendekatan parametrik dan nonparametrik. Pendekatan parametrik dilakukan jika asumsi bentuk f diketahui dan tergantung pada suatu parameter, sehingga mengestimasi f ekivalen dengan mengestimasi parameternya, sedangkan pendekatan nonparametrik dilakukan jika asumsi bentuk f tidak diketahui. Dalam hal ini diasumsikan bahwa fungsi f termuat dalam kelas fungsi mulus dalam arti mempunyai turunan kontinu atau terintegralkan secara kuadrat.

Salah satu teknik untuk mengestimasi fungsi mulus adalah teknik pemulus kernel (Hardle ,1990). Teknik pemulus kernel pada estimator densitas merupakan pengembangan dari estimator histogram (Odgen,1997). Dalam tulisan ini dibahas tentang pencarian estimator kernel dari densitas mulus, sifat-sifat dan contoh simulasinya dengan program S-Plus for Windows.

Jika diberikan $\{X_i\}_{i = 1,2,\ldots,n}$ data pengamatan independen dari suatu distribusi identik dengan densitas f yang tak diketahui, maka ada dua cara untuk membuat suatu keputusan tentang densitas f yaitu dengan pendekatan parametrik dan nonparametrik. Pendekatan parametrik dilakukan jika asumsi model distribusi X diketahui, misalnya data dari distribusi normal dengan mean μ dan varian σ^2 yang tak diketahui, maka

mengestimasi f ekivalen dengan mengestimasi parameter μ dan σ^2 dari data, sedangkan pendekatan nonparametrik dilakukan jika asumsi model distribusi X tak diketahui. Berikut metode nonparametrik untuk mengestimasi densitas f .

Estimator histogram

Metode klasik yang paling populer untuk mengetahui bentuk fungsi densitas adalah metode histogram. Suatu histogram disusun dengan meletakkan titik-titik data ke dalam suatu bin atau klas. Setiap bin dinyatakan secara grafik oleh segiempat dengan lebar sama dan tinggi proporsional dengan banyaknya titik-titik data yang terletak dalam bin terkait. Bin ditentukan dengan memilih titik awal x_0 dan lebar bin/pita (binwidth) h. Untuk sembarang integer *l*, suatu bin memuat interval setengah terbuka $[x_o+lh,\ x_o+(l+1)h)$. Nilai estimator densitas histogram di sembarang titik x dapat dinyatakan sebagai $\hat{f}(x) = \frac{1}{nh}\#X_i$ dalam bin yang sama dengan x.

Pemilihan lebar bin h kecil, histogram memuat banyak batang kecil-kecil, sedangkan untuk h besar histogram memuat sedikit batang besar-besar.

Estimator kernel

Suatu fungsi K(.) disebut fungsi kernel jika K fungsi kontinyu, berharga riil, simetris , terbatas dan $\int_{-\infty}^{\infty} K(y)dy = 1$. Jika K suatu kernel dengan sifat

1. $\int_{-\infty}^{\infty} x^j K(x)\,dx = 0$, untuk $j=1,2,\ldots,r-1$.

2. $\int_{-\infty}^{\infty} x^r K(x)\,dx \neq 0$ atau ∞ , maka K disebut kernel order r .

Beberapa contoh fungsi kernel diantaranya:

1. Seragam (Uniform)

$$K(x) = \begin{cases} 1/2, \text{ untuk } |x| \le 1 \\ 0, \quad \text{untuk x yang lain} \end{cases}$$

2. Segitiga

$$K(x) = \begin{cases} 1-|x|, \text{ untuk } |x| \le 1 \\ 0, \quad \text{untuk x yang lain} \end{cases}$$

3. Epanechnikov

$$K(x) = \begin{cases} 3/4(1-x^2), \text{ untuk } |x| \le 1 \\ 0, \quad \text{untuk x yang lain} \end{cases}$$

4. Gauss

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \text{ untuk } |x| < \infty$$

Estimator densitas kernel merupakan pengembangan dari estimator histogram. Jika $\{X_i\}_{i = 1,2,\ldots,n}$ data pengamatan independen dari suatu distribusi dengan densitas f (tak diketahui), maka estimator densitas kernel f dengan kernel K dan lebar jendela h didefinisikan sebagai $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$

Dalam estimator kernel, tingkat kemulusan $\hat{f}_h$ ditentukan oleh fungsi kernel K dan lebar jendela h yang disebut parameter pemulus, tetapi pengaruh kernel K tidak sedominan parameter pemulus h. Nilai h yang kecil memberikan grafik yang kurang mulus sedangkan nilai h yang besar memberikan grafik yang sangat mulus. Oleh karena itu, perlu dipilih nilai h optimal untuk mendapatkan grafik optimal. Salah

satu cara memilih parameter pemulus h optimal menurut Hardle (1990), dengan meminimalkan IMSE dari $\hat{f}_h$. Berikut besar IMSE dari estimator densitas kernel.

Lema

Jika diberikan pengamatan $\{Xi\}_{i=1,2,...,n}$ dari variabel random berdistribusi identik dan independen dengan densitas f, K suatu kernel order r dan f mempunyai derivatif paling sedikit tingkat r, maka $E(\hat{f}_h(x)) - f(x)) = \frac{h^r}{r!} f^{(r))}(x) \int_{-\infty}^{\infty} s^r K(s)ds$ untuk $h \to 0$.

Bukti :

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

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

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

$$= \frac{1}{h} E(K\left(\frac{x - z}{h}\right)) - f(x) \text{ , karena } X_i \text{ iid.}$$

$$= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x - z}{h}\right) f(z)dz - f(x)$$

$$= \frac{1}{h}\int_{-\infty}^{\infty} K(-s) f(x + hs)hds - f(x)$$

$$= \int_{-\infty}^{\infty} K(s)\left(f(x) + hsf'(x) + \frac{h^2 s^2}{2} f''(x) + ... + \frac{h^r s^r}{h^r} f^{(r)}(x) + o(h^r)\right)ds - f(x)$$

$$= \left(f(x)\int_{-\infty}^{\infty} K(s)ds + hf'(x)\int_{-\infty}^{\infty} sK(s)ds + \frac{h^2}{2} f''(x)\int_{-\infty}^{\infty} s^2 K(s)ds + ... + \right)$$

$$\left(\frac{h^r}{r!} f^{(r))}(x)\int_{-\infty}^{\infty} s^r K(s)ds + o(h^r)\right) - f(x)$$

$$= f(x) + \left(\frac{h^r}{r!} f^{(r))}(x)\int_{-\infty}^{\infty} s^r K(s)ds + o(h^r)\right) - f(x)$$

$$= \frac{h^r}{r!} f^{(r))}(x)\int_{-\infty}^{\infty} s^r K(s)ds \text{. Terbukti.}$$

Dari lema di atas , dapat disimpulkan bahwa estimator densitas kernel $\hat{f}_h(x)$ merupakan estimator yang tak bias secara asimtotis dari f(x). Dengan menggunakan sifat tak bias asimtotis dari $(\hat{f}_h(x))$, akan dihitung var$(\hat{f}_h(x))$, MSE$(\hat{f}_h(x))$ dan IMSE$(\hat{f}_h(x))$.

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

$$= (nh)^{-2}\sum_{i=1}^{n} \text{var}\left(K\left(\frac{x - X_i}{h}\right)\right)$$

$$= n^{-1}h^{-2}\text{var}\left(K\left(\frac{x - X}{h}\right)\right)$$

$$= n^{-1}h^{-2}\left\{\int_{-\infty}^{\infty} K^2\left(\frac{x - u}{h}\right)f(u)du - \left(\int_{-\infty}^{\infty} K\left(\frac{x - u}{h}\right)f(u)du\right)^2\right\}$$

$$= n^{-1}h^{-1}\left\{\int_{-\infty}^{\infty} K^2(u)f(x + uh)du\right\}, \text{ untuk } h \to 0$$

$$= n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du\, f(x), \text{ untuk } h \to 0.$$

Karena $\text{var}(\hat{f}_h(x)) = n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du\, f(x)$ dan

bias $(\hat{f}_h(x)) = \left(\frac{h^r}{r!} f^{(r))}(x)\int_{-\infty}^{\infty} s^r K(s)ds\right)$, maka

$$\text{MSE}\,(\hat{f}_h(x)) = n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du\, f(x) + \left(\frac{h^r}{r!} f^{(r))}(x)\int_{-\infty}^{\infty} s^r K(s)ds\right)^2$$

$$= n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du\, f(x) + \left(\frac{h^{2r}}{(r!)^2}(f^{(r))}(x))^{2r}\, k_r{}^2\right)$$

dengan $k_r = \int_{-\infty}^{\infty} s^r K(s)ds$.

$$\text{IMSE}\,(\hat{f}_h(x)) = \int_{-\infty}^{\infty} \text{MSE}\,(\hat{f}_h(x))dx$$

$$= n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du \int_{-\infty}^{\infty} f(x)dx + \left(\frac{h^{2r}}{(r!)^2}\int_{-\infty}^{\infty} (f^{(r))}(x))^2 dx\, k_r{}^2\right)$$

$$= n^{-1}h^{-1}\int_{-\infty}^{\infty} K^2(u)du + \left(\frac{h^{2r}}{(r!)^2} k_r \int_{-\infty}^{\infty} (f^{(r))}(x))^2 dx\right) .$$

Akibat

Jika kernel K mempunyai order 2 dan $f \in C^2$ maka

$\text{MSE}(\hat{f}_h(x)) \approx (nh)^{-1} f(x)c_K + \frac{1}{4} h^4[f''(x)]^2 d_K{}^2$,

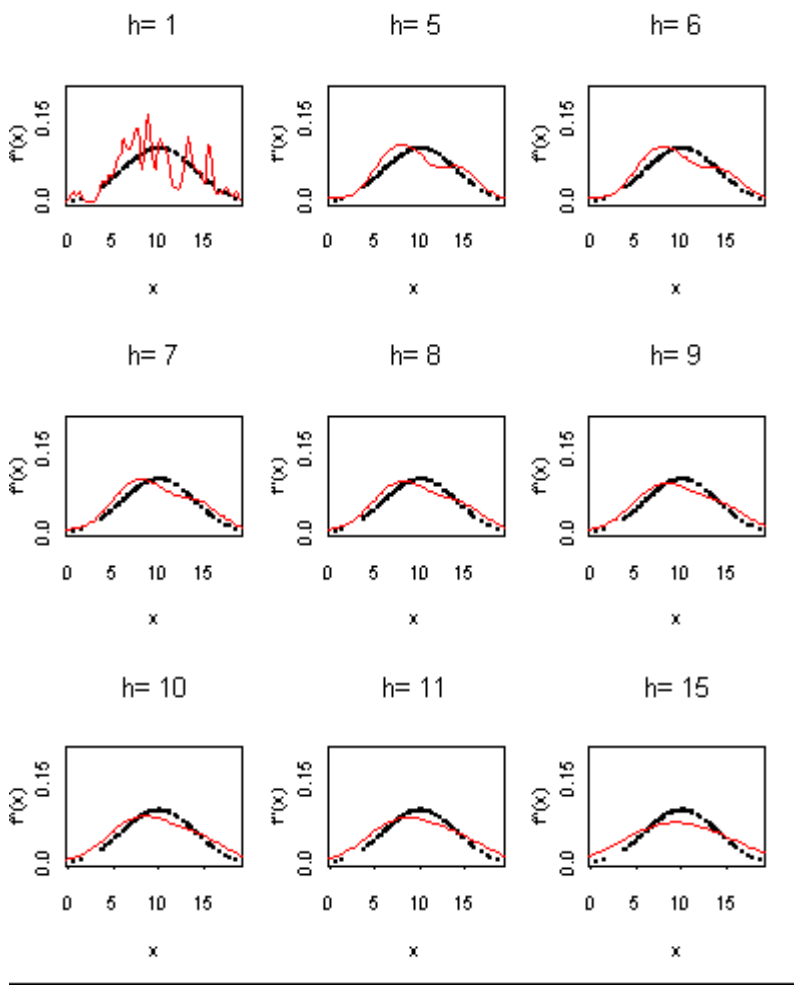
$\text{IMSE}(\hat{f}_h(x)) \approx \{(nh)^{-1}c_K + 1/4\,[h^4 d_K{}^2 \int_{-\infty}^{\infty} [f''(x)]^2]\, dx$

Dengan cara meminimalkan IMSE $((\hat{f}_h(x))$ diperoleh $h_{opt} \propto n^{-1/5}$ dan IMSE opt $\propto n^{-4/5}$. Jika $f \in C^r$, maka h opt $\propto n^{-1/(2r+1)}$ dan IMSE opt $\propto n^{-2r/(2r+1)}$.

Contoh simulasi estimasi densitas kernel

Diberikan 100 data X , yaitu X_i , i = 1,2,...,100 yang dibangkitkan dari bilangan random normal dengan mean 10 , sd = 4, maka estimasi densitas nonparametrik dari X dengan menggunakan kernel Gauss ditunjukkan pada gambar

1 berikut ini. Dari tampilan gambar di bawah ini, terlihat bahwa semakin besar h, semakin mulus estimasi densitasnya. Pada h = 9 terlihat bahwa estimasi densitasnya mendekati densitas data sebenarnya.



Gb.1. Estimasi densitas dengan kernel Gauss

..... : densitas sebenarnya (Normal mean = 10, sd = 4)
_____: estimasi densitas dari X

Kesimpulan

Dari uraian di atas , dapat disimpulkan bahwa untuk mengestimasi fungsi densitas f , jika informasi model distribusi X tak diketahui maka f dapat diestimasi dengan menggunakan pendekatan nonparametrik. Salah satu pendekatan nonparametrik dengan menggunakan teknik pemulus kernel. Tingkat kemulusan

fungsi estimasi ditentukan oleh parameter pemulus. Semakin besar parameter pemulusnya semakin mulus fungsi estimasinya dan sebaliknya.

DAFTAR PUSTAKA

Hardle,W.1990. *Smoothing Techniques With Implementation in S,* Springer-Verlag. New York.

Hardle,W.1990. *Smoothing Techniques With Implementation in S,* Springer-Verlag. New York

Odgen, R.T.1997. *Essential Wavelets for Statistical Applications and Data Analysi.,* Birkhauser, Boston.