Estimasi Densitas Mulus dengan Metode Wavelet

(Wavelet Method in Smooth Density Estimation)

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

Abstract

Let $\{X_i\}_{i = 1,2,...,n}$ 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 or orthogonal series estimator, especially by Fourier series estimator.

Another orthogonal series estimator which could be use to estimate f is wavelet estimator. Wavelet estimator is an extention of Fourier series estimator but it has caracteristics like the kernel estimator.

Key words : smooth density, kernel estimator, Fourier series estimator, wavelet estimator.

1.PENDAHULUAN

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

[1)] Staf Pengajar FMIPA, Undip, Semarang

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 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.

Untuk mengestimasi fungsi mulus , teknik pemulusan yang banyak dibahas adalah teknik pemulus kernel dan deret ortogonal, khususnya deret Fourier. Estimator deret Fourier banyak dibahas oleh Eubank (1988), sedangkan estimator kernel banyak dibahas oleh Hardle (1990). Selanjutnya, para ilmuwan diantaranya Daubechies (1992), Vetterli dan Kovacevic (1995), Hall dan Patil (1995, 1996), Odgen(1997) mengembangkan dalam estimator wavelet. Dalam tulisan ini akan dibahas tentang pencarian estimator wavelet dari densitas mulus, sifat-sifat dan contoh simulasinya dengan program S+Wavelets for Windows.

[2)] Staf Pengajar FMIPA, UGM, Yogyakarta

2. TEORI DASAR

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 kontinu, 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 .

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

Lema

Diberikan $\{X_i\}_{i\,=\,1,2,\ldots,n}$ data pengamatan independen dari suatu distribusi dengan densitas f dan diasumsikan $f \in C^2(R)$, $c_K = \int_{-\infty}^{\infty} K^2(u)du$, $d_K = \int_{-\infty}^{\infty} u^2K^2(u)du$. Jika $n\to\infty$, $h\to 0$ dan $nh\to\infty$ maka Bias $(\hat{f}_h(x)) = (h^2/2)f''(x)d_K + o(h^2)$ dan $Var(\hat{f}_h(x)) = (nh)^{-1}c_K\, f(x) + o((nh)^{-1})$.

Akibat

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

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

Estimator kernel teritlak

K(.,.) disebut kernel teritlak jika K(.,.) merupakan fungsi simetris yang memenuhi:

1. $|K(x,y)| \le c_1 (1+|x-y|)^{-(1+c_2)}$, $\forall x, y \in R$, dengan c_1, c_2 suatu konstanta positip.

2. $\int_{-\infty}^{\infty} K(x,y)\,dy = 1, \forall x \in R$.

Dalam kasus kernel biasa K(x,y) $\equiv$ K_1(x - y), untuk suatu fungsi univariat K_1.

Jika diberikan ϕ suatu fungsi univariat dengan sifat :

1. $|\phi(x)| \le c_3 \left(1+|x|\right)^{-(1+c_2)}$, $\forall x \in R$

2. $\sum_k \phi(x-k) = 1, \forall x \in R$,

maka $K(x,y) = \sum_k \phi(x-k)\phi(y-k)$ merupakan suatu kernel teritlak. Dengan menggunakan kernel teritlak K(.,.) dan lebar jendela h, maka estimator densitas f adalah

$\hat{f}_h(x) = (nh)^{-1} K(h^{-1}x, h^{-1}X_i)$

Lema

Diberikan $\{X_i\}_{i = 1,2,\ldots,n}$ data pengamatan independen dari suatu distribusi dengan densitas f, $f \in C^2(R)$ dan didefinisikan $\lambda_2(x) = \int_{-\infty}^{\infty} y^2 K(x, x+y)dy$, $\kappa(x) = \int_{-\infty}^{\infty} K^2(x , x+y)dy$. Jika $n \to \infty$, $h \to 0$ dan $nh \to \infty$ maka

Bias ($\hat{f}_h(x)$) = $(h^2/2)f''(x)\lambda_2(h^{-1}x)+o(h^2)$ dan Var ($\hat{f}_h(x)$) = $(nh)^{-1}\kappa(h^{-1}x)f(x)+o((nh)^{-1})$.

Akibat

MSE($\hat{f}_h(x)$) $\approx$ $(nh)^{-1}$ $\kappa(h^{-1}x)$ $f(x)$ + ¼ $h^4[f''(x)\lambda_2(h^{-1}x)]^2$

IMSE($\hat{f}_h(x)$) $\approx$ $(nh)^{-1}$ $\int_{-\infty}^{\infty}$ $[\kappa(h^{-1}x)\ f(x)\ ]dx$ + 1/4 $[h^4 \int_{-\infty}^{\infty}$ $[f''(x)\ )\lambda_2(h^{-1}x)]^2]$ dx

Dalam estimator kernel / kernel teritlak , 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 $\hat{f}_h$. Dengan cara ini didapat $h_{opt} \approx n^{-1/5}$ dan IMSE opt $\approx n^{-4/5}$. Jika $f \in C^r$, maka h opt $\approx n^{-1/(2r+1)}$ dan IMSE opt $\approx n^{-2r/(2r+1)}$.

Estimator deret ortogonal

Diasumsikan $f \in L^2(R)$ dengan $L^2(R)$ ruang fungsi yang kuadratnya terintegralkan, dengan kata lain $L^2(R) = \{f : \int_{-\infty}^{\infty} f(x)^2 dx < \infty\}$. Menurut Vetterli dan Kovacecic (1995), $L^2(R)$ merupakan ruang Hilbert dengan perkalian skalar dan norma yang didefinisikan sebagai $< f, g >= \int_{-\infty}^{\infty} f(x)g(x)dx$ dan $\|f\| = \sqrt{< f, f >} = \sqrt{\int_{-\infty}^{\infty} f(x)^2 dx}$. Karena $L^2(R)$ merupakan ruang Hilbert dengan sendirinya merupakan ruang vektor (berdimensi tak hingga).

Jika $\{\varphi_j\}_{j=1,2,...}$ sistem ortonormal lengkap dari $L^2(R)$, maka sembarang $f \in L^2(R)$ dapat dinyatakan sebagai $f = \sum_{j=1}^{\infty} \alpha_j \varphi_j$ dengan α_j suatu skalar yang ditentukan dengan rumus

$\alpha_j = \langle f,\varphi_j \rangle$ dan memenuhi identitas Parseval $\|f\|^2 = \sum_{j=1}^{\infty} \alpha_j^2$. Karena $\int_{-\infty}^{\infty} f(x)^2 dx < \infty$, berakibat $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$, sehingga $\alpha_j \to 0$, untuk $j \to \infty$. Oleh karena itu, f dapat didekati oleh $f = \sum_{j=1}^{J} \alpha_j \varphi_j$, untuk suatu bilangan bulat J cukup besar.

Jika $\{X_i\}_{i = 1,2,\ldots,n}$ data pengamatan independen dari suatu distribusi dengan fungsi densitas f tak diketahui, maka estimator dari f adalah $\hat{f} = \sum_{j=1}^{J} \hat{\alpha}_j \varphi_j$ dengan $\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i)$. Khususnya jika $f \in L^2[0,2\pi]$, maka f dapat didekati oleh deret Fourier ,

$f_J(x) = \frac{1}{2} a_o + \sum_{j=1}^{J} \left(a_j \cos(jx) + b_j \sin(jx) \right)$, dengan koefisien Fourier

$a_j = 1/\pi \langle f,\cos(j.) \rangle$, j = 0,1,2,...,J

$b_j = 1/\pi \langle f,\sin(j.) \rangle$, j = 1,2,3,...,J

Estimator deret Fourier dari densitas f adalah $\hat{f}_J(x) = \frac{1}{2} \hat{a}_o + \sum_{j=1}^{J} \left(\hat{a}_j \cos(jx) + \hat{b}_j \sin(jx) \right)$,

dengan estimator koefisien Fourier :

$$\hat{a}_j = \frac{1}{n\pi} \sum_{i=1}^{n} \cos(jX_i) dx \quad , j = 0,1,2,...,J$$

$$\hat{b}_j = \frac{1}{n\pi} \sum_{i=1}^{n} \sin(jX_i) \, , \quad j = 1,2,3,...,J.$$