ESTIMASI DENSITAS DENGAN METODE WAVELET THRESHOLDING

(Density Estimation by Wavelet Thresholding Method)

Suparti, Rukun Santoso dan Yulia Sugiyanti

Program Studi Statistika Jurusan Matematika
FMIPA Universitas Diponegoro Semarang

Abstract

Let X_1, X_2, … , X_n be independent observation data with unknown density function *f*. In nonparametric approach, the function *f* is assumed to be quadratic integrable and smooth function, so the function *f* could be estimated by orthogonal series estimator, especially by Fourier series estimator. Another orthogonal series estimator which could be used to estimate *f* is wavelet estimator. Wavelet estimator is divided by two methods, they are linear wavelet and wavelet thresholding methods. Wavelet thresholding estimator is an extention of linear wavelet estimator and Fourier series estimator. The major strength of the wavelet thresholding estimator is that they can capture very well local features they can estimate the smooth or unsmooth functions.

Keywords: density estimation, Fourier series estimator, wavelet thresholding estimator.

Pendahuluan

Fungsi densitas merupakan suatu konsep dasar dalam statistika sebagai penentu besar probabilitas untuk suatu selang yang diberikan. Misalnya $P(a<X<b)=\int_a^b f(x)dx$, dengan $f(x)$ adalah fungsi densitas dari variabel acak *X*. Jika diberikan data pengamatan independen $\{X_i\}_{i\ =\ 1,2,\ldots,n}$, untuk menentukan distribusi dari X ekivalen dengan menentukan fungsi densitasnya.Dalam praktek fungsi densitas dari suatu variabel acak sering tidak diketahui, sehingga perlu ditaksir (diestimasi).

Estimasi fungsi densitas *f* dapat dilakukan melalui dua pendekatan, yaitu pendekatan parametrik maupun nonparametrik. Pada estimasi fungsi densitas dengan pendekatan parametrik diperlukan asumsi mengenai bentuk distribusi suatu variabel acak (misalnya distribusi normal, gamma,dll) dan yang akan diestimasi adalah parameter-parameter dari distribusi tersebut dengan menggunakan data tentang variabel acaknya. Sedangkan estimasi fungsi densitas dengan pendekatan nonparametrik tidak memerlukan asumsi mengenai bentuk distribusi. Pendekatan ini digunakan jika tidak ada informasi yang tepat mengenai bentuk dari fungsi densitas yang sebenarnya. Dalam estimasi nonparametrik, sebelum muncul estimasi densitas dengan metode wavelet, telah dikenal dua metode standar yaitu estimasi densitas dengan histogram, kernel dan deret ortogonal.

Walaupun metode wavelet telah banyak dibicarakan oleh statistikawan pada dekade terakhir, bukanlah berarti bahwa metode tersebut sebagai pengganti/menggeser metode sebelumnya, namun lebih pada pengayaan metode. Namun kenyataan dalam praktek, keberadaan metode wavelet dengan kelebihannya dari metode standar sebelumnya, tentu akan banyak disenangi dan dipilih sebagai alternatif terbaik untuk menyelesaikan masalah. Menurut Ogden [1], metode wavelet mempunyai kelebihan, diantaranya akan diperoleh bentuk estimator yang lebih sederhana, lokalisasi ruang dan frekuensi yang sangat baik. Metode wavelet dapat mengestimasi baik fungsi mulus maupun tidak mulus, sedangkan metode standar sebelumnya tidak demikian.Metode wavelet dibedakan menjadi dua macam, yaitu metode wavelet linier dan metode wavelet nonlinier. Metode wavelet nonlinier yang disebut juga sebagai metode wavelet thresholding merupakan suatu metode wavelet yang didasarkan pada prinsip penyusutan koefisien-koefisien wavelet menuju nol, sehingga fungsi yang didekati dengan wavelet thresholding diharapkan bebas dari noise.

Deret Fourier

Jika diberikan data observasi independen X_1, X_2, … , X_n dari suatu distribusi dengan densitas f tak diketahui, maka dari data tersebut dapat digunakan untuk mengestimasi fungsi densitas.

Estimator Deret Fourier

Diasumsikan $f \in L^2(R)$ dengan $L^2(R)$ ruang fungsi yang kuadratnya terintegralkan, dengan kata lain

$$L^2(R) = \left\{ f : \int_{-\infty}^{\infty} f^2(x)dx < \infty \right\}.$$

$L^2(R)$ merupakan ruang Hilbert [2] dengan perkalian skalar dan norma yang didefinisikan sebagai

$$<f,g> = \int_{-\infty}^{\infty} f(x)g(x)dx$$

dan $\|f\| = \sqrt{\langle f,f \rangle} = \sqrt{\int_{-\infty}^{\infty} f^2(x)dx}$.

Andaikan $\{\varphi_j\}_{j=1,2,...}$ sistem ortonormal lengkap (CONS) dari $L^2(R)$, maka sembarang $f \in L^2(R)$ dapat dinyatakan sebagai $f = \sum_{j=1}^{\infty} \alpha_j \varphi_j$ dengan $\alpha_j = \langle f, \varphi_j \rangle$ dan memenuhi identitas Parseval $\|f\|^2 = \sum_{j=1}^{\infty} \alpha_j^2$.

Karena $\int_{-\infty}^{\infty} f(x)^2 dx < \infty$ maka $\sum_{j=1}^{n} \alpha_j^2 < \infty$ 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 bilangan bulat J cukup besar. Khususnya jika $f \in L^2[0,2\pi]$, maka f dapat didekati dengan deret Fourier,

$$f_J(x) = \frac{1}{2}a_0 + \sum_{j=1}^{J} \left(a_j \cos(jx) + b_j \sin(jx) \right)$$

dengan koefisien Fourier

$$a_j = \frac{1}{\pi}\langle f, \cos(j.) \rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(jx)dx$$

, j=0,1,…,J dan

$$b_j = \frac{1}{\pi}\langle f, \sin(j.) \rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(jx)dx$$

, j=1,2,…,J.

Jika diberikan data observasi independen X_1, X_2, … ,X_n dengan suatu fungsi densitas f tak diketahui, maka estimator deret Fourier dari densitas f adalah

$$\hat{f}_J(x) = \frac{1}{2}\hat{a}_0 + \sum_{j=1}^{J} \left(\hat{a}_j \cos(jx) + \hat{b}_j \sin(jx) \right)$$

dengan estimator koefisien Fourier :

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

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

Tingkat kemulusan estimator deret Fourier $\hat{f}_J$ ditentukan oleh pemilihan parameter pemulus J. Semakin kecil parameter pemulus J, semakin mulus estimasinya . Dengan kata lain semakin besar parameter pemulus J, semakin kurang mulus estimasi dari f.

Deret Wavelet

Fungsi wavelet adalah suatu fungsi matematika yang mempunyai sifat- sifat tertentu diantaranya berosilasi di sekitar nol (seperti fungsi sinus dan cosinus) dan terlokalisasi dalam domain waktu artinya pada saat nilai domain relatif besar, fungsi wavelet berharga nol. Fungsi wavelet dibedakan atas dua jenis, yaitu wavelet ayah (ϕ) dan wavelet ibu (ψ) yang mempunyai sifat:

$$\int_{-\infty}^{\infty} \phi(x)dx = 1 \quad \text{dan}$$

$$\int_{-\infty}^{\infty} \psi(x)dx = 0 .$$

Dengan dilatasi diadik dan translasi integer, wavelet ayah dan wavelet ibu melahirkan keluarga wavelet yaitu $\phi_{j,k}(x) = (p2^j)^{1/2}\phi(p2^j x - k)$ dan

$$\psi_{j,k}(x) = (p2^j)^{1/2}\psi(p2^j x - k)$$

untuk suatu skalar p>0, dan tanpa mengurangi keumuman dapat diambil p=1, sehingga $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ dan

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) .$$

Fungsi $\phi_{j,k}(x)$ dan $\psi_{j,k}(x)$ mempunyai sifat

$$\int_{-\infty}^{\infty} \phi_{j,k}(x)\phi_{j,k'}(x)dx = \delta_{k,k'}$$

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\phi_{j,k'}(x)dx = 0$$

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\psi_{j',k'}(x)dx = \delta_{j,j'}\delta_{k,k'}$$

dengan $\delta_{i,j} = \begin{cases} 1 & jika\ i = j \\ 0 & jika\ i \neq j. \end{cases}$

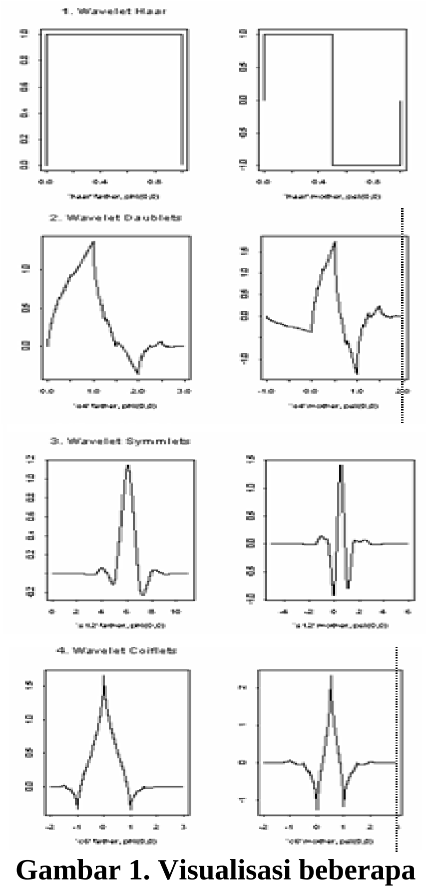
Contoh wavelet paling sederhana adalah wavelet Haar yang mempunyai rumus

$$\psi(x) = \begin{cases} 1 & ,0 \le x < 1/2 \\ -1 & ,1/2 \le x < 1 \\ 0 & ,x \quad \text{yang lain} \end{cases}$$

dan

$$\phi(x) = \begin{cases} 1 & ,0 \le x < 1 \\ 0 & ,x \quad \text{yang lain.} \end{cases}$$

Beberapa contoh wavelet selain wavelet haar diantaranya adalah wavelet Daubechies (Daublet), symmetris (Symmlet), dan Coifman (Coiflet). Visualisasi beberapa wavelet dapat ditunjukkan pada gambar 1 berikut:



Gambar 1. Visualisasi beberapa Wavelet

Analisis Multiresolusi. Analisis multiresolusi $L^2(R)$ adalah ruang bagian tertutup $\{V_j, j \in Z\}$ yang memenuhi

i) $\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots$

ii) $\bigcap_{j\in Z} V_j = \{0\}$, $\overline{\bigcup_{j\in Z} V_j} = L^2(R)$

iii) $f \in V_j \Leftrightarrow f(2.) \in V_{j+1}$

iv) $f \in V_0 \Rightarrow f(.-k) \in V_0, \forall k \in Z$

v) Terdapat sebuah fungsi $\phi \in V_0$ sehingga $\phi_{0,k} = \phi(.-k), k \in Z$ membentuk basis ortonormal untuk V_0 dimana untuk semua j,k $\in$ Z, $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$.

Jika $\{V_j,\ j \in Z\}$ analisis multiresolusi dari $L^2(R)$, maka ada basis ortonormal $\{\psi_{j,k}; j,k \in Z\}$ untuk $L^2(R)$: $\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$,

sehingga untuk sembarang f $\in L^2(R)$,

$P^j f = P^{j-1} f + \sum_{k \in Z} < f, \psi_{j-1,k} > \psi_{j-1,k}$.

,

yaitu $\psi(x)$ yang diturunkan dari

$\psi(x) = \sum_{k \in Z} (-1)^k c_{(-k+1)} \phi_{1,k}(x)$.

Untuk j $\to \infty$ maka $P^j f \to$ f.

Akibat.

Bila ϕ adalah fungsi skala yang membangun analisis multiresolusi dan

$\psi(x) = \sum_{k \in Z} (-1)^k c_{(-k+1)} \phi_{1,k}(x)$

maka dekomposisi sembarang f$\in L^2(R)$ ke dalam deret wavelet ortonormal menjadi

$$f(x) = \sum_{k \in Z} c_{jo,k}\phi_{jo,k}(x) + \sum_{j \geq jo}\sum_{k \in Z} d_{j,k}\psi_{j,k}(x)$$

dengan $c_{jo,k} = < f, \phi_{jo,k} >$ dan $d_{j,k} = < f, \psi_{j,k} >$.

Estimator fungsi densitas dengan metode wavelet thresholding

Thresholding merupakan dasar pengerjaan statistik dengan wavelet yang didasarkan pada prinsip penyusutan koefisien-koefisien wavelet menuju nol. Dengan wavelet thresholding, diharapkan hasil fungsi yang diestimasi bebas dari noise. Dengan kata lain, wavelet thresholding dapat menghilangkan koefisien wavelet kecil yang dianggap menjadi noise. Dengan menghilangkan $d_{j,k}$ yang kecil dan menyimpan $d_{j,k}$ yang besar dapat digunakan untuk mengefisienkan estimasi. Gagasan ini dapat ditelusuri dari prinsip MRA; yaitu proyeksi (pendekatan) fungsi f pada ruang V_j, j $\in$ Z yang dinyatakan sebagai $P^j f = P^{j-1} f + g^{j-1}$ dengan g^{j-1} merupakan residu antara proyeksi $P^j f$ dan proyeksi sebelumnya $P^{j-1} f$. Fungsi "detail" $g^{j-1} = \sum_{k \in Z} \langle f, \psi_{j-1,k} \rangle \psi_{j-1,k}$; memuat koefisien-koefisien wavelet $d_{j-1,k} = \langle f, \psi_{j-1,k} \rangle$. Jika koefisien-koefisien wavelet tersebut dapat disusutkan menuju nol, maka akan diperoleh estimasi untuk f yang bebas dari noise.

Donoho dan Johnstone [3,4] memberikan metode yang menekankan rekonstruksi wavelet dalam mengestimasi suatu fungsi f tak diketahui dengan menggunakan sejumlah koefisien wavelet terbesar, yakni hanya koefisien yang lebih besar dari suatu nilai tertentu yang diambil, sedangkan koefisien selebihnya diabaikan, karena dianggap nol. Nilai tertentu tersebut dinamakan nilai threshold dan estimator waveletnya dinamakan estimator wavelet thresholding. Misal tersedia nilai threshold λ, maka estimator densitas f dengan metode wavelet thresholding dapat dinyatakan sebagai

$$\hat{f}_\lambda(x) = \sum_{k \in Z} \hat{c}_{jo,k}\phi_{jo,k}(x) + \sum_{j=jo}^{J-1}\sum_{k \in Z} \partial_\lambda\left(\hat{d}_{j,k}\right)\psi_{j,k}(x)$$

dengan : ∂ = fungsi thresholding

λ = parameter threshold

$$\hat{c}_{jo,k} = \frac{1}{n}\sum_{i=1}^{n} \phi_{jo,k}(X_i)$$

$$\hat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \Psi_{j,k}(X_i).$$

Tingkat kemulusan estimator ditentukan oleh pemilihan level resolusi J, fungsi thresholding ∂, dan parameter threshold λ. Namun pemilihan J dan ∂ tidak dominan, sedangkan λ dominan. Semakin besar λ yang digunakan menghasilkan estimasi densitas yang semakin mulus. Sebaliknya, semakin kecil λ yang digunakan menghasilkan estimasi densitas yang semakin kurang mulus.

Langkah-langkah Thresholding

Langkah-langkah thresholding pada estimasi densitas pada dasarnya analog dengan thresholding pada estimasi fungsi regresi, terdiri dari:

1. Pemilihan Fungsi Thresholding

Ada dua jenis fungsi thresholding ∂, yaitu:

Soft Thresholding,

$$\partial_\lambda^S(x) = \begin{cases} x-\lambda, & x>\lambda \\ 0, & |x|\le\lambda \\ x+\lambda, & x<-\lambda \end{cases}$$

Hard Thresholding,

$$\partial_\lambda^H(x) = \begin{cases} x, & |x|>\lambda \\ 0, & x \quad yang \quad lain \end{cases}$$

dengan λ merupakan parameter threshold.

2. Pemilihan Parameter Threshold Optimal

Tingkat kemulusan estimator wavelet thresholding ditentukan oleh nilai threshold λ. Memilih nilai threshold yang terlalu besar akan menghasilkan estimasi terlalu mulus (oversmooth) karena koefisien yang masuk dalam rekonstruksi terlalu sedikit. Sebaliknya, jika memilih nilai threshold yang terlalu kecil akan menghasilkan estimasi kurang mulus (undersmooth) karena terlalu banyak koefisien yang dimasukkan dalam rekonstruksi. Oleh karena itu perlu dipilih λ optimal [5] . Ada beberapa cara untuk menentukan nilai threshold λ optimal pada fungsi thresholding yang dikelompokkan dalam dua kategori yaitu *global thresholding* dan *level-dependent thresholding* [1,5] .

(a.) Global thresholding

Global thresholding berarti memilih satu nilai λ_j yang digunakan untuk seluruh level resolusi j pada koefisien-koefisien wavelet empirik $\hat{d}_{j,k}$. Donoho dan Johnstone [4] memberikan dua cara global thresholding, yaitu:

1. Threshold Minimax

   Nilai-nilai threshold minimax ditentukan berdasarkan banyaknya data observasi n, yang dapat dilihat pada tabel berikut.

n	λ	n	λ
64	1,474	2048	2,414
128	1,669	4096	2,594
256	1,860	8192	2,773
512	2,074	16384	2,952
1024	2,232	32768	3,131

2. Threshold Universal

   Threshold universal didefinisikan dengan:

$$\lambda_j = \sqrt{2\log n}$$

   dengan n adalah banyaknya data observasi. Nilai-nilai threshold universal lebih besar dari threshold minimax, untuk nilai n yang sama. Sehingga threshold universal menghasilkan estimasi yang lebih mulus daripada estimasi minimax.

(b.) Level-Dependent Thresholding

Level-dependent thresholding berarti memilih λ_j bergantung pada level resolusi j. Dengan demikian ada kemungkinan perbedaan nilai threshold λ_j yang dipilih untuk tiap level wavelet j. Ada beberapa cara level-dependent thresholding, diantaranya yaitu:

1. Threshold Adapt

Threshold adapt adalah nilai threshold yang ditentukan berdasarkan pada level resolusi j. Pemilihan threshold ini didasarkan pada prinsip untuk meminimalkan *Stein Unbiased Risk Estimator* (SURE) pada suatu level resolusi. Threshold adapt untuk himpunan koefisien detail d_j yang beranggotakan K koefisien didefinisikan sebagai

$$\lambda_j = \arg\min_{t \geq 0} SURE(d_j, t)$$

dengan

$$SURE(d_j, t) = K - 2\sum_{k=1}^{K} 1_{[|d_{j,k}| \leq t\sigma_j]} + \sum_{k=1}^{K} \min\left\{(d_{j,k}/\sigma_j)^2, t^2\right\}$$

Threshold adapt akan memberikan hasil yang kurang baik jika koefisien-koefisiennya sangat jarang (sebagian besar koefisien pada level tersebut mendekati nol). Oleh karena itu, himpunan koefisien ini diuji dengan persamaan berikut:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{d_{j,k}}{\sigma_j} - 1 \leq \frac{\left({}^2\log K\right)^{3/2}}{\sqrt{K}}$$

Jika persamaan tersebut terpenuhi maka threshold yang digunakan pada level resolusi j adalah threshold universal, sedangkan jika tidak maka threshold adaptlah yang digunakan [6].

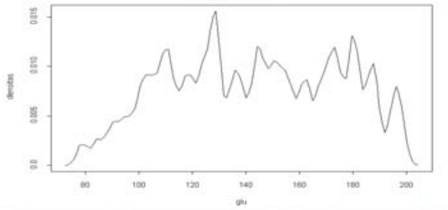
2. Threshold Top

Nilai threshold ini ditentukan dengan menentukan besarnya jumlah koefisien yang digunakan dari keseluruhan koefisien wavelet dalam merekonstruksi fungsi. Sehingga dapat ditentukan berapa jumlah koefisien wavelet yang akan digunakan untuk mengestimasi suatu fungsi [6]

Contoh estimasi densitas dengan metode wavelet thresholding

Berikut ini data pengukuran glukosa (diukur dalam milligram per deciliter) dari wanita berumur 21 tahun ke atas, berasal dari suku Indiana Pima dan tinggal di dekat Phoenix, Arizona. yang terkena diabetes [7]. Datanya sebagai berikut :
195,97,128,137,189,92,143,149,164,140,121,105,176,171,199,154,167,184,139, 134,131,158,112,181,168,144,107,125, 125,115,150,140,148,117,80,124,103, 124,112,148,145,151,144,187,129,167, 180,177,152,198,188,168,197,158,130, 151,115,194,184,95,100,138,100,175, 133,128,129,155,148,78,197,166,118, 119,102,90,111,171,180,109,100,136, 122,160,162,88,117,173,170,156,152, 163,104,179,129,128,109,109,196,109, 85,162,134,181, 179,119,184,113,155, 101,106,119,107,146,144,161,128,124, 155,109,152,122, 102,125, 196,189,173, 116,105,193,136,172,173,144,129,151,181,95,189,180,104,158,135,125,84, 163, 145,128,90,186,187,176,111,181,174, 138,112,97,179,136,155,145,111,162, 142, 169, 93,129,187,173,174,120,147, 187,181,128,170.

Data tersebut bersifat acak dan saling bebas. Dari data tersebut, dicari estimasi densitasnya menggunakan metode wavelet thresholding dengan fungsi hard thresholding dan threshold optimal yang digunakan yakni threshold minimax. Berikut ini hasil estimasi densitas dengan menggunakan program S+Wavelets.



Gambar 2: Estimasi densitas dengan metode wavelet hresholding

Kesimpulan

Estimator wavelet thresholding mempunyai kelebihan dalam menganalisis fungsi mulus maupun fungsi tidak mulus dengan kemampuannya mengadaptasi secara lokal. Tingkat kemulusan estimator sangat ditentukan oleh parameter threshold λ. Semakin besar λ yang digunakan, menghasilkan estimasi densitas yang semakin mulus dan sebaliknya. Oleh karena itu perlu dipilih λ optimal.

Ada beberapa cara untuk menentukan nilai threshold λ optimal yang dikelompokkan dalam dua kategori yaitu *global thresholding* dan *level-dependent thresholding*. Threshold minimax dan threshold universal merupakan cara pemilihan threshold optimal pada global thresholding sedangkan threshold adapt dan threshold top merupakan cara. pemilihan threshold optimal pada level-dependent thresholding.

Daftar Pustaka

[1]. Ogden, R.T. 1997. *Essential Wavelets for Statistical Applications and Data Analysis.* Birkhauser. Boston.

[2]. Vetterli, M. and Kovacevic, J. 1995. *Wavelets And Subband Coding*. Prentice Hall PTR. New Jersey.

[3] Hall, P and Patil. P. 1995. *On Wavelet Methods for Estimating Smooth Function*. Bernoulli 1(1/2). 041-058.

[4]. Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. 1996. *Density Estimation by Wavelet Thersholding.* The Annals of Statistics. Vol. 24. No. 2. 508-539.

[5]. Hall,P. and Patil,P.(1996), *On the Choice of Smoothing Parameter, Threshold and Truncation in Nonparametrik Regression by non-linier Wavelet Methods*,J.R.Statist.Soc.B (1996) 58, No.2, 361-377.

[6]. Bruce, A. and Gao, H Y. 1996. *Applied Wavelet Analysis with S - PLUS*. Springer-Verlag. New York.

[7].Http://www.en.wikipedia.org/wiki/Illustration_of_density_estimation