PEMILIHAN THRESHOLD OPTIMAL PADA ESTIMATOR REGRESI WAVELET THRESHOLDING DENGAN METODE CROSS VALIDASI

Suparti[1] , Tarno[2], Paula Meilina Dwi Hapsari[3]
[1,2] Staf Pengajar Program Studi Statistika FMIPA UNDIP
[3] Staf Biro Perasuransian, BAPEPAM-LK, Departemen Keuangan, Jakarta

Abstract

If x is a predictor variable and y is a response variable of the regression model y = f (x)+ $\in$ with f is a regression function which not yet been known and $\in$ is independent random variable with mean 0 and variance σ^2, hence function f can be estimated by parametric and nonparametric approach. In this paper function f is estimated with a nonparametric approach. Nonparametric approach that used is a wavelet shrinkage or a wavelet threshold method. In the function estimation with a wavelet threshold method, the value of threshold has the most important role to determine level of smoothing estimator. The small threshold give function estimation very no smoothly, while the big value of threshold give function estimation very smoothly. Therefore the optimal value of threshold should be selected to determine the optimal function estimation. One of the methods to determine the optimal value of threshold by minimize a cross validation function. The cross validation method that be used is two-fold cross validatiaon. In this cross validation, it compute the predicted value by using a half of data set. The original data set is split into two subsets of equal size : one containing only the even indexed data, and the other, the odd indexed data. The odd data will be used to predict the even data, and vice versa. Based on the result of data analysis, the optimal threshold with cross validation method is not uniq, but they give the uniq of wavelet thersholding regression estimation.

Keywords : Nonparametric Regression, Wavelet Threshold Estimator, Cross Validation.

1. Pendahuluan

Model regresi standar dari data observasi $\{(x_i, y_i)\}_{i=1}^n$ adalah:

$$y_i = f(x_i) + \varepsilon_i \quad (1)$$

dengan x_i variabel prediktor, y_i variabel respon, f fungsi regresi yang tidak diketahui, dan ε_i variabel random independen dengan mean 0 dan varian σ^2. Karena f fungsi regresi yang tidak diketahui maka f perlu diestimasi. Fungsi f dapat diestimasi, salah satunya dengan pendekatan non parametrik. Pendekatan non parametrik yang sudah populer adalah metode kernel dan metode Fourier. Pada pendekatan ini mengasumsikan bahwa fungsi f termuat dalam kelas fungsi mulus, artinya mempunyai turunan yang kontinu. Sehingga jika fungsinya tidak mulus metode ini kurang baik untuk digunakan. Pendekatan non parametrik yang dapat mengatasi kekurangan metode kernel dan deret Fourier dalah metode wavelet. Dalam metode wavelet diasumsikan fungsi yang akan diestimasi dapat diintegralkan secara kuadrat. Jadi dengan metode wavelet fungsi yang akan diestimasi dapat berupa fungsi mulus maupun tidak mulus.

Estimator wavelet dari regresi non parametrik adalah pengembangan dari estimator regresi deret Fourier dan estimator kernel. Estimator wavelet sendiri dibedakan menjadi dua macam, yaitu estimator wavelet linier dan estimator wavelet non linier. Estimator wavelet non linier memiliki nilai Error Kuadrat Rata-rata Terintegrasi / Integrated Mean Square Error (IMSE) yang merupakan salah satu ukuran kebaikan dari estimator, lebih cepat menuju nol daripada IMSE wavelet linier[10].

Prinsip dari estimator wavelet thresholding, mempertahankan koefisien wavelet yang nilainya lebih besar dari suatu nilai threshold tertentu dan mengabaikan koefisien wavelet yang kecil. Selanjutnya koefisien yang besar ini digunakan untuk merekonstruksi estimator fungsi yang dicari.

Pada estimasi fungsi dengan metode wavelet thresholding, tingkat kemulusan estimator ditentukan oleh pemilihan fungsi wavelet, level resolusi, fungsi thresholding, dan parameter threshold. Namun yang paling dominan menentukan tingkat kemulusan estimator adalah parameter threshold. Nilai threshold yang kecil memberikan estimasi fungsi yang sangat tidak mulus (under smooth), sedangkan nilai threshold yang besar memberikan estimasi yang sangat mulus (over smoot). Oleh karena itu perlu dipilih nilai threshold yang optimal.

2. Bahan dan Metode

Penelitian ini merupakan kajian literatur yang kemudian dikembangkan dengan simulasi menggunakan software S-Plus For Windows. Dalam tulisan ini dilakukan pembahasan tentang estimator fungsi regresi non parametrik dengan wavelet thresholding dan prosedur mendapatkan parameter threshold optimal dengan metode cross validasi serta contoh penerapannya. Metode cross validasi yang digunakan adalah metode cross validasi dua lipatan (Twofold Cross Validation), yaitu mengeluarkan setengah dari data untuk menghasilkan nilai prediksi, seperti yang telah dikembangkan oleh Nason[6] dan Ogden[7].

3. Hasil dan Pembahasan

3.1. Estimator Fungsi Regresi Non Parametrik dengan Wavelet Thresholding

3.1.1. Fungsi 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 $\int_{-\infty}^{\infty}\phi(x)dx=1$ 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

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

Beberapa contoh wavelet selain wavelet Haar diantaranya wavelet Daubechies (Daublet), symmetris (Symmlet), dan Coifman (Coiflet)[2].

3.1.2. 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) $\cap_{j\in Z} V_j=\{0\},\ \overline{\cup}_{j\in Z} V_j = L^2(R)$

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 dan untuk semua $j,k\in Z$, $\phi_{j,k}(x)=2^{j/2}\phi\left(2^j x-k\right)$.

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

Akibat 1

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 ke dalam wavelet ortonormal untuk sembarang $f\in L^2(R)$ dapat dilakukan menjadi

$$f(x)=\sum_{k\in Z} c_{jo,k}\phi_{jo,k}(x)+\sum_{j\ge jo}\sum_{k\in Z} d_{j,k}\psi_{j,k}(x) \qquad (2)$$

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

3.1.3. Estimator Wavelet Linier

Jika terdapat sekumpulan data independen $\{(Xi,Yi)\}_{i=1}^n$ yang mempunyai model (1) dan $n=2^m$ dengan m bilangan bulat positip. Jika X_i rancangan titik reguler pada interval [0,1] dengan $X_i=i/n$, maka proyeksi f pada ruang V_J dapat ditulis menjadi $(P^J f)(x)=\sum_{k\in Z} c_{J,k}\phi_{J,k}(x)$ atau $f_J(x)=\sum_{k\in Z} c_{J,k}\phi_{J,k}(x)$

dengan $c_{J,k}=\langle f,\phi_{J,k}\rangle=\int_0^1 f(x)\phi_{J,k}(x)dx$.

Berdasarkan dekomposisi fungsi ke dalam wavelet ortonormal (2), maka untuk sembarang fungsi $f\in L^2(R)$ diperoleh

$$f_J(x)=\sum_{k\in Z} c_{jo,k}\phi_{jo,k}(x)+\sum_{j\ge jo}^{J-1}\sum_{k\in Z} d_{j,k}\psi_{j,k}(x) \qquad (3)$$

dengan $c_{jo,k} = < f, \phi_{jo,k} > = \int_0^1 f(x)\phi_{jo,k}(x)dx$ dan $d_{j,k} = < f, \psi_{j,k} > = \int_0^1 f(x)\psi_{j,k}(x)dx$. Karena fungsi regresi f tidak diketahui maka estimator f pada ruang V_J dapat ditulis sebagai

$$\hat{f}_J(x) = \sum_{k\in Z} \hat{c}_{J,k}\phi_{J,k}(x) \text{ dengan } \hat{c}_{J,k} = \frac{1}{n}\sum_{i=1}^{n} y_i \phi_{J,k}(x_i) \text{ atau}$$

$$\hat{f}_J(x) = \sum_{k\in Z} \hat{c}_{jo,k}\phi_{jo,k}(x) + \sum_{j\geq jo}^{J-1}\sum_{k\in Z} \hat{d}_{j,k}\psi_{j,k}(x) \quad (4)$$

dengan $\hat{c}_{jo,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_{jo,k}(x_i)$ dan $\hat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{j,k}(x_i)$, yang merupakan estimator tak bias dari $c_{jo,k}$ dan $d_{j,k}$. Estimator wavelet (4) dinamakan estimator wavelet linier.

3.1.4. Estimator Wavelet Shrinkage.

Jika diberikan data $\{(x_i, y_i)\}_{i=1}^n$ dengan model (1), $n = 2^m$ dan $x_i = i/n$, maka $y_i \sim N(f(i/n), \sigma^2)$. Mean dan varian dari $\hat{d}_{j,k}$ adalah $E[\hat{d}_{j,k}] = d_{j,k}$ dan $\text{Var}(\hat{d}_{j,k}) = \frac{\sigma^2}{n}$ sehingga $\hat{d}_{j,k} \sim N(d_{j,k}, \frac{\sigma^2}{n})$. Jadi koefisien wavelet empiris $\hat{d}_{j,k}$ memuat sejumlah noise dan hanya relatif sedikit yang memuat sinyal signifikan sehingga dapat direkonstruksi estimator wavelet dengan menggunakan sejumlah koefisien terbesar. Oleh karena itu Hall&Patil[4] dan Ogden[7] memberikan metode yang menekankan rekonstruksi wavelet dengan menggunakan sejumlah koefisien wavelet terbesar, yakni hanya koefisien yang lebih besar dari suatu nilai tertentu yang diambil, sedangkan koefisien selebihnya diabaikan, karena dianggap 0. Nilai tertentu tersebut dinamakan nilai threshold (nilai ambang) dan estimatornya menghasilkan

$$\hat{f}_J(x) = \sum_{k\in Z} \hat{c}_{jo,k}\phi_{jo,k}(x) + \sum_{j\geq jo}^{J-1}\sum_{k\in Z} \hat{d}_{j,k}\psi_{j,k}(x) \quad (5)$$

dengan ∂_λ menyatakan fungsi thresholding atau fungsi ambang dengan nilai ambang atau threshold λ. Estimator (5) dinamakan estimator wavelet non linier, estimator wavelet shrinkage, atau estimator wavelet thresholding.

Prinsip dari estimator wavelet thresholding adalah mempertahankan koefisien wavelet yang nilainya lebih besar dari suatu nilai ambang atau nilai threshold tertentu dan mengabaikan koefisien wavelet yang kecil. Selanjutnya koefisien yang besar ini digunakan untuk merekonstruksi fungsi (estimator) yang dicari. Karena thresholding ini dirancang untuk membedakan antara koefisien wavelet empiris yang masuk dan yang keluar dari rekonstruksi wavelet, sedangkan untuk membuat keputusan ada 2 faktor yang mempengaruhi ketepatan estimator, yaitu ukuran sampel n dan tingkat noise σ^2, maka setiap koefisien merupakan calon kuat masuk didalam rekonstruksi wavelet jika ukuran sampel besar atau tingkat noise kecil. Karena $\sqrt{n}\,\hat{d}_{j,k}/\sigma$ berdistribusi normal dengan varian 1 untuk seluruh n dan σ, maka estimator thresholding dari $d_{j,k}$ adalah

$$\tilde{d}_{j,k} = \frac{\sigma}{\sqrt{n}}\partial_\lambda\left(\frac{\sqrt{n}\hat{d}_{j,k}}{\sigma}\right)$$

sehingga estimator wavelet thresholding adalah

$$\hat{f}_{\lambda}(x)=\sum_{k}\hat{c}_{jo,k}\phi_{jo,k}(x)+\sum_{j\geq jo}^{J-1}\sum_{k=0}^{2^{j}-1}\frac{\sigma}{\sqrt{n}}\partial_{\lambda}\left(\frac{\sqrt{n}\hat{d}_{j,k}}{\sigma}\right)\psi_{j,k}(x) \qquad (6)$$

dengan

$\hat{c}_{jo,k}$: penduga koefisien fungsi skala $c_{jo,k}$

$\hat{\mathrm{d}}_{\mathrm{j,k}}$: penduga koefisien wavelet $\mathrm{d}_{\mathrm{j,k}}$

λ : parameter nilai threshold

∂_{λ} : fungsi threshold

3.1.5. Langkah-langkah Thresholding

Langkah-langkah thresholding terdiri dari:

i. Pemilihan Fungsi Thresholding

Ada dua jenis fungsi thresholding ∂_{λ}, yaitu:

Hard Thresholding,

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

Soft Thresholding,

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

dengan λ merupakan parameter thresholding.

Fungsi Hard thresholding lebih dikenal karena terdapat diskontinyu dalam fungsi thresholding sehingga nilai x yang berada diatas threshold λ tidak disentuh. Sebaliknya, fungsi soft thresholding kontinyu yaitu sejak nilai x berada diatas threshold λ. Motivasi penggunaan soft thresholding berasal dari prinsip bahwa noise mempengaruhi seluruh koefisien wavelet. Juga kekontinyuan dari fungsi soft thresholding membuat kondisi yang lebih baik untuk alasan statistik.

ii. Estimasi σ

Dalam merekonstruksi fungsi wavelet biasanya nilai σ tidak diketahui. Oleh karena itu, σ harus diestimasi dari data. Ogden[7] memberikan estimasi σ berdasarkan koefisien wavelet empiris pada level resolusi tertinggi dengan fungsi Median Deviasi Absolut (MAD), yaitu:

$$\hat{\sigma}=\frac{\mathrm{median}\left(\left|\hat{\mathrm{d}}_{\mathrm{J-1,k}}-\mathrm{median}\left(\hat{\mathrm{d}}_{\mathrm{J-1,k}}\right)\right|\right)}{0{,}6745}$$

iii. Pemilihan Parameter Thresholding

Pada estimasi wavelet thresholding, tingkat kemulusan estimator paling dominan ditentukan parameter threshold λ. Nilai λ yang terlalu kecil memberikan estimasi fungsi yang sangat tidak mulus (under smooth) sedangkan nilai λ yang terlalu besar memberikan estimasi yang sangat mulus (over smooth). Oleh karena itu perlu dipilih parameter threshold yang optimal untuk mendapatkan estimasi fungsi yang optimal. Untuk memilih nilai threshold optimal, ada dua kategori pemilihan yaitu memilih satu harga threshold untuk seluruh level resolusi (pemilihan secara global) dan pemilihan threshold yang tergantung pada level resolusi (dependent level thresholding).

Untuk pemilihan threshold global, Ogden[7] memberikan 2 pemilihan threshold yang hanya bergantung pada banyaknya data pengamatan n yaitu threshold universal ($\lambda = \sqrt{2\log n}$) dan threshold minimax yang telah ditabelkan oleh Donoho dan Johnstone[3] (Tabel 1). Nilai-nilai threshold minimax selalu lebih kecil dibandingkan dengan nilai threshold universal untuk ukuran sampel yang sama.

Pemilihan threshold yang tergantung pada level resolusi berarti memilih λ_j bergantung level resolusi j. Dengan demikian ada kemungkinan perbedaan nilai threshold λ_j yang dipilh untuk tiap level j. Ada beberapa cara level-dependent thresholding diantaranya yaitu threshold adapt dan threshold top. Threshold adapt 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 $\lambda_j = \arg\min_{t\geq 0} \mathrm{SURE}(d_j, t)$ dengan $SURE(d_j,t) = K - 2\sum_{k=1}^{K} 1_{[|d_{j,k}|\leq t\sigma_j]} + \sum_{k=1}^{K} \min\{(d_{j,k}/\sigma_j)^2, t^2\}$ Sedangkan nilai threshold top ditentukan berdasarkan besar prosentase koefisien yang akan digunakan dari keseluruhan koefisien wavelet yang ada[1].

Tabel 1. Nilai Threshold Minimax berdasarkan Ukuran Sampel

n	λ	n	λ
2	0	512	2,074
4	0	1024	2,232
8	0	2048	2,414
16	1,200	4096	2,594
32	1,270	8192	2,773
64	1,474	16384	2,952
128	1,669	32768	3,131
256	1,860	65536	3,310

Pemilihan parameter threshold optimal universal, minimax, adapt, dan top merupakan pemilihan parameter threshold optimal standar dalam estimasi dengan wavelet thresholding dan telah tersedia dalam software S+Wavelet. Selain cara tersebut masih ada beberapa metode atau prosedur lain untuk mendapatkan threshold optimal dalam estimasi fungsi wavelet thresholding, diantaranya dengan prosedur uji hipotesa multipel[9], prosedur False Discovery Rate (FDR)[8] dan prosedur cross validasi.

3.2. Prosedur Penentuan Parameter Threshold Optimal dengan Metode Cross Validasi

Menentukan threshold λ optimal dengan metode cross validasi adalah mencari λ yang meminimalkan MISE (Mean Integrated Squared Error) antara estimator wavelet thresholding $\hat{f}_\lambda$ dan fungsi sebenarnya f yaitu mencari λ yang meminimalkan

$$M(\lambda) = E\int \left\{\hat{f}_\lambda(x) - f(x)\right\}^2 dx \tag{7}$$

Pada data diskret, persamaan (7) menjadi

$$M(\lambda) = E\left[\sum_{i=1}^{n}\left(\hat{f}_\lambda(x_i) - f(x_i)\right)^2\right] \tag{8}$$

Dalam prakteknya, fungsi sebenarnya f tidak diketahui (disini yang akan dicari estimatornya) sehingga M harus diestimasi. Selanjutnya yang diminimalkan adalah estimasi dari MISE.

Prosedur kerja metode Cross Validasi dimulai dengan mengeluarkan suatu titik data dari suatu himpunan data. Untuk setiap nilai parameter threshold λ, dikeluarkan sebuah titik data dari n buah titik-titik data dan sisanya n-1 titik-titik data digunakan untuk mendapatkan estimasi dari fungsi regresi dan kemudian dilakukan prediksi dari titik data yang dikeluarkan itu. Ukuran kebaikan di dalam prediksi ini diberikan oleh

$$CV(\lambda)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{9}$$

dengan $\hat{y}_i$ menyatakan prediksi dari y_i dengan menggunakan n-1 titik-titik data dengan meninggalkan data ke-i. Pemilihan threshold λ optimal yang meminimalkan (9) disebut prosedur cross validasi untuk n-1 titik data. Sedangkan persamaan (9) disebut Cross Validasi MSE (Mean Squared Error) yang merupakan estimasi dari $M(\lambda)$ pada (8). Memilih λ optimal yang meminimalkan (9) ekuivalen dengan memilih λ optimal yang meminimalkan

$$nCV(\lambda)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10}$$

Persamaan (10) disebut Cross Validasi Squared Error yang selanjutnya disebut Cross Validasi saja. Pemilihan λ optimal yang meminimalkan cross validasi memberikan nilai prediksi yang terdekat pada nilai data observasi yang sebenarnya[4].

Metode cross validasi klasik menghitung nilai prediksi berdasarkan satu data dikeluarkan. Namun pada wavelet, karena data disarankan sebanyak 2^j dengan j integer sehingga jika satu data dikeluarkan, data yang tersisa menjadi (2^j-1). Ini bukan merupakan data pangkat dari 2. Oleh karena itu metode cross validasi klasik tidak dapat langsung diterapkan. Nason[6] dan Ogden[7] mengembangkan metode cross validasi dua lipatan (Twofold Cross Validation), yaitu menghitung nilai prediksi dengan mengeluarkan setengah dari data.

Prosedur cross validasi dua lipatan (Twofold Cross Validation) secara otomatis memilih atau menyeleksi suatu threshold λ untuk mendapatkan estimator wavelet yang bekerja pada himpunan data yang berisi sebanyak 2^M titik-titik data. Prosedur cross validasi dua lipatan bekerja dengan mengeluarkan setengah dari 2^M titik-titik data. Jadi tinggal sisanya 2^{M-1} titik-titik data yang kemudian digunakan untuk membentuk estimator wavelet menggunakan suatu threshold tertentu. Untuk jelasnya, berikut diberikan algoritmanya.

Diberikan data $y_1, y_2, \ldots, y_n$ dengan $n = 2^M$ yang diasumsikan mempunyai Model (1). Kemudian data ini berdasarkan indeksnya dikelompokkan menjadi 2 bagian yang berukuran sama. Satu kelompok berisi data yang semuanya berindeks genap dan yang lainnya kelompok data yang semuanya berindeks ganjil. Data yang berindeks genap akan digunakan untuk prediksi data yang berindeks ganjil dan sebaliknya. Kemudian dikeluarkan semua data y_i yang berindeks ganjil dari himpunan data tersebut. Jadi tinggal sisanya $\frac{n}{2}=2^{M-1}$ titik-titik y_i yang semuanya berindeks genap. Himpunan titik-titik data y_i yang berindeks genap ini, diindeks kembali berdasar $j=1,...,\frac{n}{2}$, dinyatakan dengan

$y_1^E, y_2^E, ..., y_{n/2}^E$ dengan $\frac{n}{2} = 2^{M-1}$. Selanjutnya $y_1^o, y_2^o, ..., y_{n/2}^o$ menyatakan indeks ulang untuk data yang berindeks ganjil. Berdasarkan data $y_1^E, y_2^E, ..., y_{n/2}^E$ itu dan menggunakan suatu threshold λ tertentu dikonstruksi estimator wavelet, misal $\hat{f}_\lambda^E$. Jadi $\hat{f}_\lambda^E$ merupakan nilai-nilai estimasi dari fungsi regresi f di titik ke 2,4,…,n-2,n. Sedangkan $\hat{f}_\lambda^o$ merupakan nilai-nilai estimasi dari fungsi regresi f di titik ke 1,3,…,n-1 menggunakan threshold λ.

Menggunakan data yang dikeluarkan itu, maka suatu versi interpolasi dari data tersebut diberikan oleh

$$\tilde{f}^{o}_{\lambda,j} = \begin{cases} \frac{1}{2}(\hat{f}^{o}{}_{\lambda,j+1} + \hat{f}^{o}{}_{\lambda,j}), & j = 1,2,3,...,\frac{n}{2}-1 \\ \frac{1}{2}(\hat{f}^{o}{}_{\lambda,1} + \hat{f}^{o}{}_{\lambda,j}), & j = \frac{n}{2} \end{cases}$$, untuk data berindeks ganjil dan

$$\tilde{f}^{E}_{\lambda,j} = \begin{cases} \frac{1}{2}(\hat{f}^{E}{}_{\lambda,j} + \hat{f}^{E}{}_{\lambda,n/2}), & j = 1 \\ \frac{1}{2}(\hat{f}^{E}{}_{\lambda,j-1} + \hat{f}^{E}{}_{\lambda,j}), & j = 2,3,4,...\frac{n}{2} \end{cases}$$, untuk data berindeks genap.

Sehingga estimasi penuh (full estimate) untuk $M(\lambda)$ adalah membandingkan estimator wavelet yang diinterpolasi dan titik-titik data yang dikeluarkan itu, yang diberikan oleh

$$\hat{M}(\lambda) = \sum_{j=1}^{n/2} \left\{ \left(\tilde{f}^{E}{}_{\lambda,j} - y_{2j-1} \right)^2 + \left(\tilde{f}^{o}{}_{\lambda,j} - y_{2j} \right)^2 \right\} \quad (11)$$

Persamaan (11) merupakan fungsi Cross Validasi menggunakan n/2 data.. Estimasi $\hat{M}(\lambda)$ bergantung pada dua estimasi untuk $\hat{f}_\lambda^E$ dan $\hat{f}_\lambda^o$ berdasarkan pada $\frac{n}{2}$ titik-titik data. Sehingga λ yang diperoleh dengan meminimalkan (11) merupakan λ optimal untuk n/2 data, ditulis $\lambda(n/2)$. Pada threshold universal, besar threshold λ optimal untuk n data adalah $\lambda(n) = \sqrt{2\log(n)}$. Besaran ini memberikan suatu metode heuristic untuk memperoleh threshold cross validasi yang cocok untuk n titik-titik data. Karena threshold optimal untuk $n/2$ titik data adalah $\lambda\left(\frac{n}{2}\right) = \sqrt{2\log\left(\frac{n}{2}\right)}$ maka threshold optimal untuk n data adalah

$$\lambda(n) = \left(1 - \frac{\log 2}{\log n}\right)^{-1/2} \lambda\left(\frac{n}{2}\right) \quad (12)$$

3.3. Contoh Penerapan

Untuk menerapkan metode wavelet dengan metode Cross Validasi, diambil data simulasi tubrukan sepeda motor pada suatu PMTO (*Post Mortem Human Test Object*/ obyek uji pemeriksaan mayat manusia)[5]. Dalam hal ini variabel-variabelnya adalah sebagai berikut:

- Sebagai variabel respon, Y (percepatan dalam g) menyatakan percepatan setelah tubrukan yang disimulasikan.

- Sebagai variabel prediktor, X (waktu dalam milisekon) menyatakan waktu setelah simulasi tubrukan.

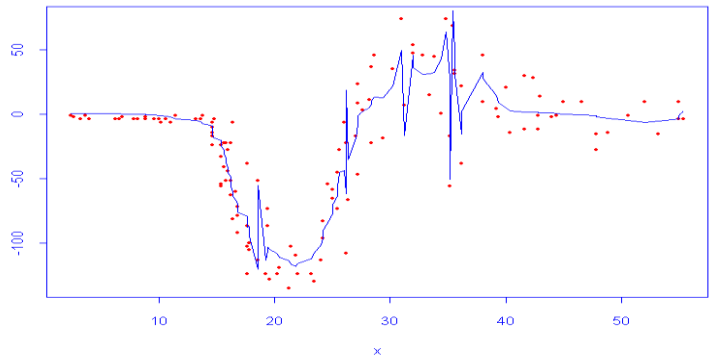
Dari data tersebut, dicari estimasi regresi wavelet thresholding dengan pemilihan threshold optimal menggunakan metode Cross Validasi. Dengan menggunakan program S+Wavelets diperoleh nilai dari fungsi Cross Validasi dan Threshold (λ) untuk setengah data yang dapat dilihat pada Tabel 2.

Berdasarkan Tabel 2, tampak bahwa nilai Cross Validasi yang terkecil (optimal) sebesar 97650,74 dicapai pada nilai threshold berkisar antara λ = 45,26 sampai dengan λ = 46,47 sehingga threshold optimal dapat dipilih salah satu diantaranya yaitu λ = 46,47. Karena dengan pemilihan λ optimal berkisar antara 45,26 sampai dengan 46,47 menghasilkan nilai cross validasi yang sama maka kurva estimasinya pun akan sama.

Tabel 2. Cross Validasi pada Data Simulasi

Nilai dari Fungsi Cross Validasi ($\hat{M}(\lambda)$)	Nilai Threshold (λ)
97719,78	45,25
97650,74	45,26
97650,74	45,27
97650,74	45,28
97650,74	45,30
97650,74	45,35
97650,74	45,40
97650,74	45,45
97650,74	45,50
97650,74	45,55
97650,74	45,65
97650,74	45,75
97650,74	45,85
97650,74	45,95
97650,74	46,05
97650,74	46,25
97650,74	46,35
97650,74	46,45
97650,74	46,46
97650,74	46,47
98585,78	46,48
98585,78	46,49
98585,78	46,50

Karena threshold optimal dari setengah data sebesar 46,47 maka dengan menggunakan Persamaan (12) diperoleh threshold optimal untuk seluruh data sebesar 50,17. Plot data dan plot estimasi regresi wavelet thresholding pada threshold optimal λ = 50,17 dapat dilihat pada Gambar 1.



Gambar 1. Kurva Estimasi Wavelet Thresholding

Keterangan :

..... : plot data

——— : estimasi regresi wavelet thresholding

4. Kesimpulan

Metode cross validasi klasik dengan mengeluarkan satu data untuk menghitung nilai prediksi tidak dapat diterapkan pada estimator regresi wavelet, karena pada regresi wavelet data disarankan sebanyak 2^j dengan j integer sehingga jika satu data dikeluarkan, data yang tersisa bukan data pangkat dari 2. Oleh karena itu metode cross validasi yang digunakan adalah cross validasi dua lipatan (Twofold Cross Validation), yaitu menghitung nilai prediksi dengan mengeluarkan setengah dari data. Dari hasil simulasi, threshold optimal yang diperoleh tidak tunggal . Meskipun threshold optimal yang diperoleh tidak tunggal, tetapi setelah dimasukkan pada estimasi regresi wavelet thresholding menghasilkan estimasi optimal yang tunggal.

DAFTAR PUSTAKA

1. Bruce, A. and Hong-Ye, G., *Applied Wavelet Analysis with S-PLUS*, Springer, New York, 1996.
2. Daubechies, I., *Ten Lectures on Wavelets*, Capital City Press, Philadelpia, 1992.
3. Donoho, D.L and Johnstone, I.M., *Ideal Spatial Adaptation by Wavelet Shrinkage*, Biometrika, 1994, Vol. 81, No. 3: 425-455.
4. Hall, P. and Patil, P., On *Wavelet Methods for Estimating Smooth Functions*, Bernoulli 1995, Vol. 1, No. 2: 041-058.
5. Hardle, W., *Applied Nonparametric Regression*, Cambridge University Press, New York, 1993.
6. Nason, G.P., *Choice of the threshold parameter in wavelet function estimation. In Wavelets and Statistics.* Antoniadis, A. and Oppenheim, G.(eds.), Springer-Verlag, New York, 1995.
7. Ogden, R.T., *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser, Boston, 1997.
8. Suparti, Tarno dan Haryono, Y., *Pemilihan Parameter Threshold Optimal dalam Estimator Regresi Wavelet Thresholding dengan Prosedur False Discovery Rate (FDR)*, Media Statistika, 2008, Vol. 1, No. 1: 1-8.

9. Suparti, Santoso, R. dan Putra, S.W, *Pemilihan Threshold Optimal pada Estimator Regresi Wavelet thresholding dengan Prosedur Uji Hipotesis Multipel*, Jurnal Sains & Matematika, 2007, Vol. 15, No. 4: 187-197.
10. Suparti dan Subanar, H., *Estimasi Regresi dengan Metode Wavelet Shrinkage*, Jurnal Sains & Matematika, 2000, Vol. 8, No. 3: 105-113.