J. Pilar Sains 2 (2): 1-5, 2003

ISSN 1412-5595

Parameter Ambang Dalam Regresi Wavelet

Elfis Suanto*, Sakur

Laboratorium Pendidikan Matematika, Jurusan Pendidikan MIPA FKIP
Universitas Riau Pekanbaru 28293 Riau

Diterima 7 Mei 2003 Disetujui 20 Juni 2003

Abstract

Considered a regression model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, 2, 3, \ldots, n$$

where ε_i assumed iid N(0,σ^2). To found the best estimator of fungtion f in regression wavelets will be associated with chosen threshold parameter t. The Cross-Validation method is a good to choose the parameter t.

Keywords: Threshold parameter, Cross-Validation method, Wavelet Regretion

Pendahuluan

Diberikan model regresi berikut:

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1,2, \ldots, n \qquad (1)$$

Sesatan random ε_i diasumsikan iid N(0,σ^2) dan $f(t_i)$ merupakan kurva regresi. Masalah yang selalu aktual dalam regresi adalah bagaimana bentuk estimasi terbaik dari kurva regresi f tersebut. Dalam hal ini ada dua model regresi yang dapat digunakan yaitu *model regresi parametrik* dan *model regresi non-parametrik.*

Dalam regresi parametrik diasumsikan bentuk f diketahui. Pembuatan asumsi tersebut didasarkan pada teori, pengalaman masa lalu atau tersedianya sumber-sumber lain yang dapat memberi pengetahuan/informasi yang terperinci. Pendekatan model regresi parametrik banyak ditulis antara lain oleh Draper dan Smith (1966), Seber (1977) dan Montgomery dan Peck (1982). Dalam kasus tertentu, apabila informasi yang tersedia tentang bentuk f sangat terbatas atau sedikit sekali sehingga kesulitan untuk menentukan bentuk fungsi f maka bagian terbesar informasi tentang f terletak pada data. Dalam hal seperti ini model regresi non-parametrik akan lebih cocok untuk digunakan.

Dalam konsep regersi non-parametrik, bentuk fungsi f tidak perlu diketahui. Tetapi fungsi f cukup diasumsikan termuat dalam suatu ruang fungsi, sehingga lebih fleksibel.

Dalam tulisan ini diambil pendekatan model regresi non-parametrik dengan mengambil f termuat dalam ruang $L^2(\mathbf{R})$.

$L^2(\mathbf{R})$ menyatakan himpunan semua fungsi yang kuadratnya terintegralkan pada **R**. Estimator wavelet dari f diperoleh dengan mencari f anggota $L^2(\mathbf{R})$ yang meminimumkan fungsi :

$$M(t) = E\left[\int \left(\hat{f}_t(x) - f(x)\right)^2\right] dx \qquad (2)$$

dengan t merupakan parameter ambang.

Untuk meminimumkan (2) akan berhubungan dengan pemilihan parameter ambang t. Berbagai cara sudah dilakukan untuk memilih parameter ambang t tersebut, seperti yang dilakukan oleh Donoho dan Johnstone (1994) memberikan dua cara yaitu cara optimal dan universal.

Berdasarkan cara kerja Donoho dan Johnstone tersebut dalam tulisan ini dikembangkan cara pemilihan parameter ambang dengan meggunakan metoda Cross-Validasi (CV).

Metode Penelitian

Penelitian ini merupakan studi leteratur dengan

menelusuri dan menganalisa beberapa jurnal dan buku-buku teks yang terkait. Langkah-langkah pendekantannya sebagai berikut:

1. Pertama dibahas terlebih dahulu tentang Cross-Validasi standar.
2. Kemudian baru dibahas penerapan ide-ide metoda Cross-Validasi standar tersebut ke regresi menggunakan konsep wavelet. Hasilnya dituangkan dalam bentuk prosedur (algoritma) kerjanya.

Hasil dan Pembahasan

Diperhatikan model regresi

$$\tilde{y} = \tilde{f} + \tilde{\varepsilon} \qquad (3)$$

dengan $\tilde{y} = (y_1, y_2, \ldots, y_n)'$, $\tilde{f} = (f_1, f_2, \ldots, f_n)'$, dan $\tilde{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ serta diasumsikan $\tilde{\varepsilon} \approx N(0, \sigma^2)_n$.

Misal $C(\Lambda) = \{\tilde{f}_\lambda : \lambda \in \Lambda\}$ adalah suatu kelas estimator-estimator untuk f dengan asumsi elemen dari $C(\Lambda)$ merupakan estimator–estimator linier dengan Λ merupakan himpunan indeks. Berarti untuk setiap λ ada matriks $H(\lambda)$ yang berukuran n x n sedemikian sehingga

$$\tilde{f}_\lambda = H(\lambda)\tilde{y}$$

Untuk memilih estimator terbaik dari f yang merupakan elemen-elemen dari $C(\Lambda)$ maka diperhatikan tiga criteria yaitu fungsi kerugian, fungsi resiko dan fungsi resiko prediksi. Secara berturut-turut didefinisikan sebagai berikut:

$$L(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(f_j - f_{\lambda j}\right)^2 \qquad (4)$$

dengan $f_{\lambda j}$ merupakan elemen ke-j dari $\tilde{f}_\lambda$

$$R(\lambda) = E\left(\frac{1}{n}\sum_{j=1}^{n}\left(f_j - f_{\lambda j}\right)^2\right) \qquad (5)$$

$$P(\lambda) = E\left(\frac{1}{n}\sum_{j=1}^{n}\left(y_j^* - f_{\lambda j}\right)^2\right) \qquad (6)$$

degan y_j^* merupakan observasi baru.

Resiko prediksi (6) berhubungan erat dengan resiko (5), hal ini ditunjukkan oleh lemma berikut.

Lemma 1 : Jika $P(\lambda)$ adalah resiko prediksi seperti (6) dan $R(\lambda)$ adalah fungsi resiko seperti (5) maka berlaku hubungan

$$P(\lambda) = \sigma^2 + R(\lambda).$$

Bukti:

$$\begin{aligned} P(\lambda) &= \frac{1}{n}\sum_{j=1}^{n} E\left(y_j^* - f_{\lambda j}\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} E\left(f_j + \varepsilon_j^* - f_{\lambda j}\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} E\left((f_j - f_{\lambda j}) + E(\varepsilon_j^*)^2\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} E\left((f_j - f_{\lambda j}) + \sigma^2\right) \\ &= R(\lambda) + \sigma^2 \end{aligned}$$

Jadi suatu estimator dari f yang meminimumkan $P(\lambda)$ juga meminimumkan $R(\lambda)$ dan sebaliknya atau dengan kata lain meminimumkan $R(\lambda)$ ekivalen dengan meminimumkan $P(\lambda)$.

Disamping ketiga criteria diatas terdapat juga criteria lain, yaitu kerugian integrasi (integrated loss) dan resiko integrasi (integrated risk) (Eubank, 1988:18). Secara berturut turut didefinisikan sebagai berikut:

$$IL(\lambda) = \int_a^b \left(f(t) - f_\lambda(t)\right)^2 dt \qquad (7)$$

$$IR(\lambda) = \int_a^b E\left(f(t) - f_\lambda(t)\right)^2 dt \qquad (8)$$

Jadi untuk memperoleh estimator terbaik dari fungsi regresi f , idealnya dipilih suatu parameter smoothing λ yang meminimumkan $L(\lambda)$ atau $R(\lambda)$ atau $P(\lambda)$ atau $IL(\lambda)$ atau $IR(\lambda)$. Tetapi tidak satupun dari criteria tersebut yang dapat dihitung secara langsung tanpa diketahui terlebih dahulu fungsi regresi f. Dengan demikian dalam praktek, kriteria tersebut harus diestimasi dari data dan estimatornya kemudian diminimumkan terhadap λ.

Akibatnya diperlukan suatu metoda untuk mengestimasi $P(\lambda)$, $R(\lambda)$, $IL(\lambda)$ atau $IR(\lambda)$.

Untuk inilah digunakan/dibahahas metoda Cross-Validasi.

Cross-Validasi adalah suatu teknik yang mengandalkan pada usaha meminimumkan sesatan prediksi yang diakibatkan oleh pembandingan pada sebagian data ke sisa data. Jadi ide utama metoda cross-validasi adalah bagaimana memilih parameter smoothing sehingga memberikan estimator terbaik berdasarkan kemampuan prediksi.

Misal diberikan n titik data $y_1, y_2, y_3, \ldots, y_n$. Kemudian keluarkan satu titik data, katakan data ke-j maka diperoleh sub-data $y_1, y_2, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n$ yang berukuran n-1 titik data. Selanjutnya setiap sub data itu digunakan untuk menghampiri estimator $f_{\lambda j}$ Ambil $f_{\lambda(j)}$estimator yang menyerupai $f_{\lambda j}$ yang dihitung dari n-1 titik data yang tidak menyertakan observasi ke-j itu. Sehingga dapat dibuat konstruksi dasarnya sebagai berikut. Jika $X_{\lambda(j)}$ menyatakan matriks dengan elemen-elemennya

x_{kr} dengan $k = 1, 2, \ldots, j-1, j+1, \ldots, n$ dan $r \in \lambda$ serta

$\tilde{y}_{(j)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)'$

maka estimasi untuk f_j adalah

$$f_{\lambda(j)} = \tilde{x}'_{\lambda j}\beta_{\lambda(j)} \qquad (9)$$

dengan $\tilde{x}'_{\lambda j}$ merupakan vector nilai-nilai variable X yang diindeks didalam λ dan berhubungan dengan respon ke-j dan

$$\tilde{\beta}_{\lambda(j)} = \left(X'_{\lambda(j)}X_{\lambda(j)}\right)^{-1}X'_{\lambda(j)}y_j$$

merupakan estimasi kuadrat terkecil dari koefisien-koefisien variable $X_r, \in \lambda$ yang dihitung tanpa menggunakan $(\tilde{x}'_j, y_j)$

Estimator $f_{\lambda(j)}$berkaitan erat dengan $f_{\lambda j}$ dalam arti kedua-duanya merupakan estimator kuadrat terkecil dari f_jPerbedaannya hanya $f_{\lambda j}$ menggunakan $(\tilde{x}'_j, y_j)$ sedangkan $f_{\lambda(j)}$ tidak. Hubungan antaraf_{λ} dan $f_{\lambda(j)}$diberikan oleh lemma berikut.

Lemma 2: Jika $f_{\lambda(j)}$merupakan estimator dari f_j yang dihitung tanpa menggunakan $(\tilde{x}'_j, y_j)$ dan $f_{\lambda j}$ juga merupakan estimator dar f_j yang menggunakan $(\tilde{x}'_j, y_j)$ maka diperoleh

$$f_{\lambda(j)} = f_{\lambda j} - \frac{h_{jj}(\lambda)(y_j - f_{\lambda j})}{I - h_{jj}(\lambda)}$$

dengan $h_{jj}(\lambda)$ merupakan elemen diagonal ke-j dari $H(\lambda)$.

Bukti:

$$\begin{aligned} f_{\lambda(j)} &= \frac{f_{\lambda j} - h_{jj}(\lambda)y_j}{I - h_{jj}(\lambda)} \\ &= \frac{f_{\lambda i} - h_{jj}(\lambda)y_j - h_{jj}(\lambda)f_{\lambda j} + h_{jj}(\lambda)f_{\lambda j}}{I - h_{jj}(\lambda)} \\ &= \frac{(1 - h_{jj}(\lambda))f_{\lambda j} - h_{jj}(\lambda)(y_j - f_{\lambda j})}{I - h_{jj}(\lambda)} \\ &= f_{\lambda j} - \frac{h_{jj}(\lambda)(y_j - f_{\lambda j})}{I - h_{jj}(\lambda)} \end{aligned}$$

Selanjutnya $P(\lambda)$ atau criteria yang lainnya dapat diestimasi dengan

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - f_{\lambda(j)}\right)^2 \qquad (10)$$

dengan $f_{\lambda(j)}$ menurut (9). Metoda pemilihan λ yang meminimumkan $CV(\lambda)$ (10) yang disebut metoda Cross-Validasi.

Selanjutnya pembahasan secara mendalam tentang penerapan ide-ide Cross-Validasi ini untuk regresi menggunakan konsep Wavelet. Prosedur kerja Cross-Validasi standar seperti diatas tidak dapat diterapkan secara lansung ke regresi wavelet, karena adanya transformasi wavelet diskrit. Untuk itu dalam tulisan ini dibahas prosedur CV yang dimodifikasi dan disesuaikan sebagai berikut.

Diberikan data $y_1, y_2, \ldots, y_n$ dengan $n = 2^M$ M menyatakan tingkat resolusi (level) dan diasumsikan mempunyai model regresi (1). Kemudian data ini berdasarkan indeksnya dikelompokkan menjadi dua bagian yang berukuran sama. Satu kelompok berisi data yang semuanya berindeks genap dan yang lainnya berindeks ganjil. Sekarang keluarkan semua data y_i yang berindeks ganjil dari himpunan data

tersebut, jadi tinggal sisa data sebanyak $n/2=2^{M-1}$ titik-titik data yang berindeks genap. Selanjutnya, himpunan titik data yang berindeks genap ini diindeks kan kembali menjadi j = 1, 2, ... , n/2. Misalnya dinyatakan dengan $y_1^E, y_2^E, \dots, y_{n/2}^E$, $n/2=2^{M-1}$, sedangkan untuk himpunan titik data yang berindeks ganjil dinyatakan dengan $y_1^O, y_2^O, \dots, y_{n/2}^O$. Kemudian berdasarkan data $y_1^E, y_2^E, \dots, y_{n/2}^E$ dengan menggunakan parameter ambang t tertentu dikonstruksi estimator wavelet, misalkan $\hat{f}_t^E$. Jadi menggunakan data berindeks ganjil yang dikeluarkan itu maka diperoleh versi interpolasi dari data tersebut yang diberikan oleh:

$$\tilde{y}_j^O = \begin{cases} \frac{1}{2}(y_{2j-1} + y_{2j+1}), & \text{jika } j = 1,3,\dots,\frac{n}{2}-1 \\ \frac{1}{2}(y_1 + y_{n-1}), & \text{jika } j = \frac{n}{2} \end{cases} \quad (11)$$

dan

$$\tilde{y}_j^E = \begin{cases} \frac{1}{2}(y_n + y_2), & \text{jika } j = 1 \\ \frac{1}{2}(y_{2j+2} + y_{2j}), & \text{jika } j = 2,4,\dots,\frac{n}{2} \end{cases} \quad (12)$$

untuk data yang berindeks genap. Perhatikan bahwa indeks-indeks pada versi interpolasi dari data itu bersesuaian dengan pengindeks-an himpunan-himpunan bagian dari data tersebut, sehingga diperoleh estimator dari subset data tersebut, yaitu:

$$\tilde{f}_{t,j}^E = \begin{cases} \frac{1}{2}\left(\hat{f}_{t,j+1}^E + \hat{f}_{t,j}^E\right), & \text{jika } j = 1,3,\dots,\frac{n}{2}-1 \\ \frac{1}{2}\left(\hat{f}_{t,1}^E + \hat{f}_{t,n-1}^E\right), & \text{jika } j = \frac{n}{2} \end{cases} \quad (13)$$

untuk data berindeks genap, dan

$$\tilde{f}_{t,j}^O = \begin{cases} \frac{1}{2}\left(\hat{f}_{t,n}^O + \hat{f}_{t,2}^O\right), & \text{jika } j = 1 \\ \frac{1}{2}\left(\hat{f}_{t,2j-2}^O + \hat{f}_{t,2j}^O\right), & \text{jika } j = 2,4,\dots,\frac{n}{2} \end{cases} \quad (14)$$

untuk data yang berindeks ganjil.

Jadi estimasi penuh untuk fungsi M(t) diperoleh dengan membandingkan estimator wavelet interpolasi dan titik data yang dikeluarkan, yang diberikan oleh:

$$\hat{M}(t) = \sum_{j=1}^{\frac{n}{2}} \left\{ \left(\tilde{f}_{t,j}^E - \tilde{y}_j^O\right)^2 + \left(\tilde{f}_{t,j}^O - \tilde{y}_j^E\right)^2 \right\} \quad (15)$$

Jelas bahwa estimator $\hat{M}(t)$ tersebut bergantung pada dua estimasi dari f_t. Berdasarkan kerja Donoho dan Johnstone (1994) diketahui bahwa tersedia parameter ambang t yang bergantung pada n titik data dan asimtotik, yaitu

$$t_{UV}(n) =. \sqrt{2\log(n)}\hat{\sigma}_n \quad (16)$$

Besaran ini memberikan suatu metoda heuristic untuk memperoleh parameter ambang cross-validasi yang cocok untuk n titik data. Jika parameter ambang untuk n titik data adalah seperti (16) maka parameter ambang untuk n/2 titik data diberikan oleh

$$t_{UV}\left(\frac{n}{2}\right) = \sqrt{2\log\left(\frac{n}{2}\right)}\hat{\sigma}_{\frac{n}{2}} \quad (17)$$

Setelah estimasi diminimumkan, maka (17) digunakan untuk memperoleh parameter ambang final. Dengan kata lain bahwa parameter ambang final untuk himpunan data penuh diperoleh dari penerapan koreksi (17) ke parameter ambang cross-validasi untuk separoh titik data.

Kesimpulan

Jadi dapat disimpulkan metoda Cross-Validasi ini secara otomatis memilih dan menyeleksi suatu parameter ambang t untuk mendapatkan estimator wavelet terbaik yang bekerja pada himpunan data yang berisi sebanyak 2^M titik-titik data dengan menerapkan koreksi (17).

Daftar Pustaka

Donoho, D.L. and Jonhstone, I.M. 1994. Ideal Spatial Adaption by Wavelet Shrinkage. *Biometrika*, 81, 425-455.

_____ ,1995. Adapting to Unknown Smoothing via Wavelet Shrinkage. *Journal American Stat.Ass.* 90, No 432.

Eubank, R.T.1988. *Spiline Smoothing and Non parametric Regresion.* Marcel Dekker Inc., New York.

Gaust, R.F. and Mason, R.L. 1980. *Regression Analysis and Aplications: A Data-Oriented Approach.* Marcel Dekker, New York.

Mallat, S.G. 1989. A Theory for Multiresolution Signal Decomposition: The Wavelet Repesentation. *IEEE Trans. On Patt. Analysis and Machine Intell.*, 11,674-693.

Nason, G.P. and Silverman, B.W. 1994. The Discrate Wavelet Transform in S. *J.Comput. Graph Statist.*,3,281-299.

Stone, M. 1978. Cross-Validation: A Review, *Statistics,* 9,127-140.