Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)
Volume 02, No. 3 (2013), hal 191 – 196.

GENERALIZED CROSS VALIDATION DALAM REGRESI *SMOOTHING SPLINE*

Andi Sayuti, Dadan Kusnandar, Muhlasah Novitasari Mara

INTISARI

Regresi nonparametrik adalah salah satu metode Statistika yang digunakan untuk mengetahui pola hubungan antara variabel independen dengan variabel dependen yang tidak diketahui bentuk fungsinya. Analisis regresi nonparametrik digunakan jika tidak ada informasi sebelumnya tentang bentuk kurva regresi. Estimasi fungsi regresi nonparametrik dilakukan berdasarkan data pengamatan dengan menggunakan teknik pemulusan (smoothing). Pendekatan yang digunakan untuk regresi nonparametrik dalam penelitian ini adalah pendekatan dengan regresi smoothing spline. Smoothing spline merupakan fungsi yang mampu memetakan data dengan baik serta mempunyai variansi error yang kecil. Regresi smoothing spline digunakan untuk mengetahui bentuk kurva f(x) pada regresi nonparametrik. Adapun metode yang digunakan dalam regresi smoothing spline adalah metode Generalized Cross Validation (GCV). Metode GCV adalah metode klasik yang digunakan untuk menentukan parameter pemulus pada regresi smoothing spline. Nilai dari parameter pemulus dipilih dari nilai GCV yang minimum. Hasil penelitian menunjukkan bahwa semakin besar nilai dari parameter pemulus maka kurva yang dihasilkan akan semakin mulus. Sebaliknya, semakin kecil nilai dari parameter pemulus maka kurva yang dihasilkan akan semakin kasar. Nilai optimal parameter pemulus pada penelitian ini adalah pada $\lambda = 4.3718 \times 10^{-4}$ *dimana nilai* $\mathrm{GCV}(\lambda) = 0.1519$.

Kata Kunci : Nonparametrik, *Smoothing Spline*, *Generalized Cross Validation*

PENDAHULUAN

Model regresi nonparametrik merupakan model regresi yang digunakan untuk mengestimasi kurva regresi yang hanya tergantung pada data amatan. Model regresi nonparametrik tidak memberikan asumsi terhadap bentuk kurva regresi. Kurva tersebut hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi ini biasanya dimotivasi oleh sifat kemulusan (*smoothness*) yang diasumsikan dimiliki oleh fungsi regresi tersebut. Ini memberikan fleksibilitas yang lebih besar didalam bentuk yang mungkin dari kurva regresi. Pada umumnya fungsi regresi hanya termuat dalam suatu ruang kurva yang berdimensi tak hingga. Untuk mengkonstruksi model regresinya dipilih ruang kurva yang sesuai, yang mana kurva regresi diyakini termasuk didalamnya [1].

Diberikan n pengamatan (x_i, y_i) dimana $i = 1,2,\cdots,n$ dengan x_i dan y_i dalam R. Variabel x_i merupakan vektor variabel independen pada pengamatan ke-i, variabel y_i merupakan variabel dependen pada pengamatan ke-i, dan R adalah bilangan riil. Hubungan antara x_i dan y_i diasumsikan mengikuti model regresi:

$$y_i = f(x_i) + \varepsilon_i \quad \text{dimana } i = 1,2,\cdots,n \tag{1}$$

fungsi f merupakan fungsi pemulus yang tidak spesifik dan ε_i adalah random *error* diasumsikan berdistribusi independen dengan rata-rata sama dengan nol dan variansi sama dengan σ^2 [2]. Pada persamaan (1) bentuk kurva regresi *f* belum diketahui.

Ada beberapa teknik untuk mengestimasi kurva regresi f dalam regresi nonparametrik, antara lain dengan menggunakan regresi kernel dan *smoothing spline* [1]. Metode *smoothing spline* memiliki hasil yang lebih baik daripada regresi kernel. Maka dalam penelitian ini akan membahas masalah regresi *smoothing spline* [3].

Smoothing spline merupakan fungsi yang mampu memetakan data dengan baik dan mempunyai variansi *error* yang kecil. Berdasarkan model regresi pada persamaan (1) dimana f adalah fungsi pemulus yang tidak spesifik dan $\mathrm{E}(\varepsilon) = 0$, menduga kurva pemulus $\hat{f}x_i$ dapat diperoleh berdasarkan data amatan, yakni variabel dependen dan variabel independen [4]. Oleh karena itu, dengan menggunakan data amatan sebanyak *n*, maka $f(x_i)$ diperoleh dengan meminimumkan persamaan berikut:

$$PLS = \sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \lambda \int_{-\infty}^{\infty}\left(\frac{d^2 f(x)}{dx^2}\right)^2 dx \qquad (2)$$

Masalah yang dihadapi dalam meminimumkan persamaan (2) adalah bagaimana menentukan parameter pemulus (λ) pada persamaan (2). Dalam menentukan parameter pemulus pada regresi *smoothing spline* tersebut dapat menggunakan metode CV (*Cross Validation*), GCV (*Generalized Cross Validation*), Cp *Criterion*, AIC (*Akaike Information Criterion*), RECP (*Risk Estimation Using Classical Pilots*) and EDS (*Exact Double Smoothing*). Metode GCV merupakan metode unggulan dari beberapa metode tersebut. Untuk itu, dalam penelitian ini hanya menggunakan Metode GCV untuk mendapatkan hasil yang maksimal pada parameter pemulus tersebut. Nilai λ dipilih dari nilai GCV(λ) yang minimum [5]. Tujuan dari penelitian ini adalah menentukan parameter pemulus pada regresi *smoothing spline* dan menganalisis penggunaan regresi *smoothing spline* dalam menentukan bentuk kurva regresi.

Penelitian ini berupa studi literatur. Dalam penelitian ini menggunakan data hasil simulasi seperti yang dilakukan oleh [6], yaitu $y = 3^x - 2^x + e^{-5x} + e^{-20\left[x-\frac{1}{2}\right]^2}$ dimana $\varepsilon \sim N(0; 0.1)$. Dari data simulasi tersebut nilai λ (parameter pemulus) pada persamaan (2) ditentukan dengan menggunakan metode GCV (*Generalized Cross Validation*) agar mendapatkan bentuk kurva yang mulus.

FUNGSI SPLINE DALAM REGRESI NONPARAMETRIK

Regresi nonparametrik adalah salah satu metode yang digunakan untuk mengetahui pola hubungan antara variabel independen dan variabel dependen yang tidak diketahui bentuk fungsinya, hal ini dikarenakan tidak ada informasi sebelumnya tentang bentuk kurva regresi $f(x)$. Beberapa model pendekatan regresi nonparametrik yang telah dikembangkan, misalnya spline.

Secara umum fungsi spline orde ke-*m* dapat disajikan dalam bentuk:

$$f(x) = \beta_0 + \sum_{j=1}^{m}\beta_j x^j + \sum_{j=1}^{k}\theta_j\left(x - X_j\right)_+^m \qquad (3)$$

dengan fungsi terpotong sebagai berikut:

$$\left(x - X_j\right)_+^m = \begin{cases}\left(x - X_j\right)_+^m & ; x \geq X_j \\ 0 & ; x < X_j\end{cases}$$

dimana β_0 merupakan konstanta, β_j merupakan koefisien variabel x_j, θ_j merupakan koefisien pada variabel x_j pemotongan knot ke-*k*, x^j merupakan variabel independen orde ke-*j*, X_j merupakan knot ke-*j* pada variabel x^j, nilai $j = 1,2,\cdots,m$, *m* merupakan orde spline dan *k* adalah banyak knot.

GCV (*GENERALIZED CROSS VALIDATION*)

Metode GCV digunakan untuk menentukan parameter pemulus dalam regresi *smoothing spline*. Bentuk umum metode GCV adalah sebagai berikut [5]:

$$\text{GCV}(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - f_\lambda(x_i)\}^2}{\{1 - n^{-1}tr(\boldsymbol{S}_\lambda)\}^2} \tag{4}$$

dengan f_λ adalah estimator dari *smoothing spline* dan $tr(\mathbf{S}_\lambda) \leq n$. Nilai λ dipilih dari nilai GCV(λ) yang minimum.

Adapun langkah-langkah yang dilakukan untuk menentukan parameter pemulus pada regresi *smoothing spline* adalah:

1. Input data (x_i, y_i).
2. Hitung matrik **T** dan **H** kemudian matrik **L**.

$$\mathbf{T} = \left[\begin{pmatrix} \mathbf{R} + \boldsymbol{\lambda}\mathbf{I_n} & \mathbf{Q^t} \\ \mathbf{Q} & 0 \end{pmatrix}^{-1}\right]_{(n\times n)}$$

dimana notasi $[.]_{n\times n}$ menunjukkan submatrik berukuran $n \times n$ yang dibentuk dari bagian kiri atas matrik utama. Matrik $[.]_{n\times n}$ dituliskan sebagai berikut:

$$\left[\begin{array}{c:c} \mathbf{T} \; n\,x\,n & \mathbf{T_1} \; n\,x\,2 \\ \hdashline \mathbf{T_2} \; 2\,x\,n & \mathbf{T_3} \; 2\,x\,2 \end{array}\right]$$

$$\mathbf{H} = \left[(\mathbf{R} \quad \mathbf{Q^t})\begin{pmatrix} \mathbf{R} + \boldsymbol{\lambda}\mathbf{I_n} & \mathbf{Q^t} \\ \mathbf{Q} & 0 \end{pmatrix}^{-1}\right]_{(n\times n)}$$

dimana notasi $[.]_{n\times n}$ menunjukkan submatrik berukuran $n \times n$ yang dibentuk dari bagian kiri matrik utama. Matrik $[.]_{n\times n}$ dituliskan sebagai berikut:

$$\left[\begin{array}{c:c} \mathbf{H} \; n\,x\,n & \mathbf{H_1} \; n\,x\,2 \end{array}\right]$$

$$\mathbf{L} = \left(\mathbf{TH^{-1}}\right)^{\mathbf{t}}\mathbf{RTH^{-1}}$$

dimana

$$\mathbf{R} = \begin{pmatrix} \mathbf{0} & \frac{|X_1 - X_2|^3}{12} & \frac{|X_1 - X_3|^3}{12} & \cdots & \frac{|X_1 - X_n|^3}{12} \\ \frac{|X_2 - X_1|^3}{12} & \mathbf{0} & \frac{|X_2 - X_3|^3}{12} & \cdots & \frac{|X_2 - X_n|^3}{12} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{|X_n - X_1|^3}{12} & \frac{|X_n - X_2|^3}{12} & \frac{|X_n - X_3|^3}{12} & \cdots & \mathbf{0} \end{pmatrix} \text{ dan } \mathbf{Q} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ X_1 & X_2 & X_3 & \cdots & X_n \end{pmatrix}$$

3. Tentukan nilai $\boldsymbol{\lambda}$.

$$\lambda = \frac{1-q}{q} \qquad , 0 < q < 1$$

4. Hitung matrik $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{L})^{-1}$.

5. Hitung f_λ untuk berbagai nilai $\boldsymbol{\lambda}$.

$$f_\lambda = (\mathbf{I} + \lambda \mathbf{L})^{-1} y$$

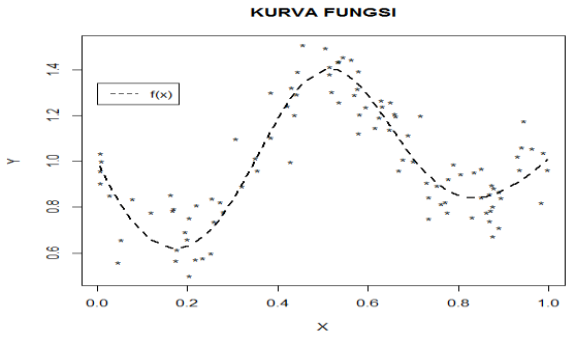
6. Pilih f_λ yang meminimumkan GCV($\boldsymbol{\lambda}$).

$$\text{GCV}(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} \{y_i - f_\lambda(x_i)\}^2}{\{1 - n^{-1} tr(\boldsymbol{S}_\lambda)\}^2}$$

HASIL DAN PEMBAHASAN

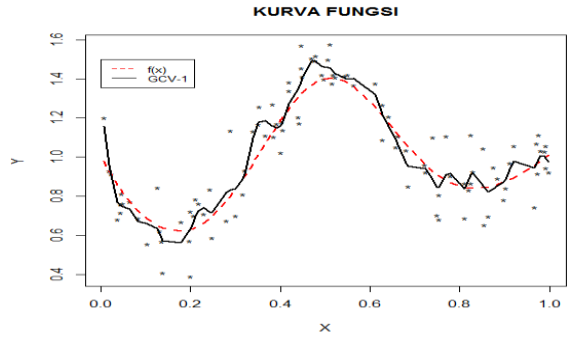
1. Deskripsi Data

Untuk menguji parameter pemulus dalam pendugaan kurva regresi *smoothing spline*, digunakan simulasi membangkitkan data seperti yang dilakukan oleh [6], yaitu $y = 3^x - 2^x + e^{-5x} + e^{-20\left[x-\frac{1}{2}\right]^2}$ dimana $\varepsilon \sim N(0; 0.1)$. Bentuk kurva regresi yang dihasilkan oleh fungsi tersebut ditunjukkan pada Gambar 1.

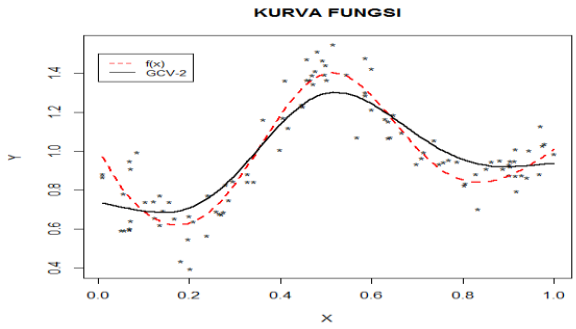


Gambar 1. Bentuk Kurva Fungsi $y = 3^x - 2^x + e^{-5x} + e^{-20\left[x-\frac{1}{2}\right]^2}$
Dimana $\varepsilon \sim N(0; 0.1)$

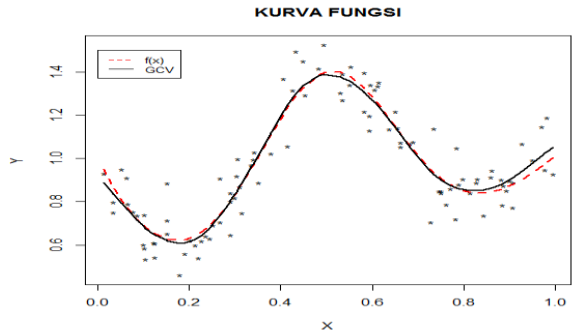
Pada gambar (2) sampai (4) dapat dilihat bahwa bentuk kurva mulus sangat dipengaruhi oleh nilai dari parameter pemulus λ. Semakin kecil nilai dari parameter pemulus maka bentuk kurva yang dihasilkan akan kasar (Gambar 2). Sebaliknya, apabila nilai dari parameter pemulus semakin besar maka bentuk kurva regresi yang dihasilkan akan mulus, namun hasil yang diperoleh belum tentu baik (Gambar 3). Kurva regresi dikatakan baik dan mulus apabila kurva regresi tersebut mendekati bentuk dari kurva regresi aslinya pada fungsi *f(x)* (Gambar 4).



Gambar 2. Kurva Fungsi *f(x)* dan Taksiran Kurva Regresi dengan *Smoothing Spline* dengan Parameter Pemulus $\lambda = 1.2789 \times 10^{-6}$



Gambar 3. Kurva Fungsi *f(x)* dan Taksiran Kurva Regresi dengan *Smoothing Spline* dengan Parameter Pemulus $\lambda = 5.3263 \times 10^{-3}$



Gambar 4. Kurva Fungsi *f(x)* dan Taksiran Kurva Regresi dengan *Smoothing Spline* dengan Parameter Pemulus $\lambda = 4.3718 \times 10^{-4}$

2. Pemilihan Parameter Pemulus

Pada penelitian ini nilai parameter pemulus optimal yang diperoleh untuk menduga kurva regresi adalah $\lambda = 4.3718 \times 10^{-4}$ dengan nilai GCV(λ)=0.1519. Kurva dugaan regresi dengan menggunakan *smooth.spline* pada nilai parameter pemulus optimal disajikan pada Gambar (4). Kurva dugaan regresi yang dihasilkan sangat baik karena mendekati fungsi *f(x)* yang sesungguhnya. Jika nilai dari parameter pemulus diubah maka kurva dugaan regresi spline semakin menjauh dari fungsi yang sebenarnya. Hal ini juga ditunjukkan oleh nilai GCV yang semakin besar dibanding nilai GCV pada saat parameter pemulus yang optimal sebesar 0.1519 (Tabel 1).

Tabel 1. Nilai-Nilai Parameter Pemulus yang Dicobakan

No.	Parameter Pemulus (λ)	GCV (λ)
1	5.3263×10^{-3}	0.1603
2	4.3718×10^{-4}	0.1519
3	1.2789×10^{-6}	0.3356

PENUTUP

Model regresi *smoothing spline* sangat berpengaruh pada nilai dari parameter pemulus. Nilai parameter pemulus memegang peranan penting dalam menentukan baik dan tidaknya kurva dugaan regresi yang dihasilkan. Metode yang digunakan dalam menentukan parameter pemulus pada regresi *smoothing spline* adalah metode *Generalized Cross Validation* (GCV). Nilai dari parameter pemulus dipilih dari nilai GCV yang minimum.

Untuk data simulasi yang dicobakan, Program R digunakan untuk mendapatkan nilai parameter pemulus yang optimal dan kurva dugaan yang dihasilkan sangat baik yaitu pada nilai $\lambda = 4.3718 \times 10^{-4}$ dengan nilai GCV(λ) = 0.1519.

Ada beberapa metode yang dapat digunakan dalam menentukan parameter pemulus pada regresi *smoothing spline* diantaranya, Cp *Criterion*, AIC (*Akaike Information Criterion*), RECP (*Risk Estimation Using Classical Pilots*) and EDS (*Exact Double Smoothing*). Untuk itu, pada penelitian selanjutnya dapat dicoba menggunakan metode tersebut agar dapat membandingkan metode mana yang lebih baik dari metode GCV (*Generalized Cross Validation*).

DAFTAR PUSTAKA

[1] Eubank, R. *Nonparametric Regression and Spline Smoothing*. *Second Edition*. New York: Marcel Dekker. 1999.

[2] Fox, J. *Nonparametric Regression* [Internet]. 2002 [cited 2009 Jan 24]. Available from: http://cran.r-project.org/doc/contrib/Fox-Companion/ appendix-nonparametric-regression.pdf

[3] Aydin, D. A Comparison of the Nonparametric Regression Models Using Smoothing Spline and Kernel Regression. *World Academy Science, Enginering and Technology*. 2007; (36): 253-257.

[4] Takezawa, K. *Introduction to Nonparametric Regression*. New York: John Wiley and Sons, inc. 2006.

[5] Lee, T. C. M. Smoothing Parameter Selection for Smoothing Splines: a Simulation Study. *Computational Statistic & Data Analysis*. 2003; (42): 139-148.

[6] Breaz, N. The Cross Validation Method in Smoothing Spline Regression. Romania: *Acta Universitatis Apulensis.* 2004; (7): 77-84.

ANDI SAYUTI : Jurusan Matematika FMIPA UNTAN, Pontianak, andisay86@yahoo.com

DADAN KUSNANDAR : Jurusan Matematika FMIPA UNTAN, Pontianak, dkusnand@yahoo.com

MUHLASAH NOVITASARI MARA : Jurusan Matematika FMIPA UNTAN, Pontianak, noveemara@gmail.com