J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 12, No. 2, 120-130, Januari 2016
ISBN:1858-1382

Reproducing Kernel dan Estimator Spline dalam Regresi Nonparametrik

Ruliana*

Abstrak

Spline merupakan polinomial yang memiliki sifat tersegmen yang memberikan fleksibel lebih dari polinomial biasa, yang bisa menyesuaikan diri secara efektif terhadap karakteristik suatu data. *Estimator Spline* didapatkan dari meminimalkan kriteria *penalized least square* yaitu kriteria kesesuaian kurva terhadap data (*goodness of fit*) dengan kekasaran kurva (*roughness penalty*), dimana keseimbangan antara kesesuaian kurva ini dikontrol oleh suatu parameter penghalus λ, dimana $\lambda > 0$. Jika suatu data berpasangan $\{(\boldsymbol{t}_i, \boldsymbol{y}_i), i = 1,2, ...,n\}$, $\boldsymbol{t}_i \in \mathrm{R}$, $\boldsymbol{y}_i \in \mathrm{R}$ dan hubungan antara $\boldsymbol{t}_i$ dan $\boldsymbol{y}_i$, diasumsikan mengikuti model $\boldsymbol{y}_i = L_i \boldsymbol{f} + \boldsymbol{\varepsilon}_i$, $i = 1,2, ...,n$, dengan $\boldsymbol{f} \in H_R$, dan L_i, $i = 1,2,\ldots,n$ adalah fungsional linear terbatas pada H_R dimana $H_R = H_0 \oplus H_1$, maka estimasi dari $\boldsymbol{f}$ adalah $\hat{\boldsymbol{f}} = \sum_{v=1}^{m} \boldsymbol{\alpha}_v \boldsymbol{\varphi}_v + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{\xi}_i \in \Pi^{2m-1}$. yang merupakan *polynomial natural spline*. Estimator ini dipengaruhi oleh pemilihan λ. Dengan menggunakan data simulasi dalam kasus ini diperoleh bahwa $I_{GCV} = \frac{L(\lambda_{GCV})}{L(\lambda_{opt})}$, lebih efisien dari $I_{CV} = \frac{L(\lambda_{CV})}{L(\lambda_{opt})}$ dalam mendapatkan parameter penghalus optimal.

Kata Kunci: Spline, Penalized Least Square, reproducing Kernel, parameter penghalus, GCV, CV, asimtotik efisien.

1. Pendahuluan

Analisis regresi merupakan salah satu alat statistik yang banyak digunakan untuk mengetahui hubungan antara sepasang variabel atau lebih. Misalkan diberikan data $\{(\boldsymbol{t}_i, \boldsymbol{y}_i), i = 1,2, ...,n\}$, $\boldsymbol{t}_i \in \mathrm{R}$, $\boldsymbol{y}_i \in \mathrm{R}$ dan hubungan antara $\boldsymbol{t}_i$ dan $\boldsymbol{y}_i$ diasumsikan mengikuti model regresi

$$\boldsymbol{y}_i = \boldsymbol{f}(\boldsymbol{t}_i) + \boldsymbol{\varepsilon}_i, \boldsymbol{t}_i \in [0,1], i = 1,2,\ldots,n, \quad (1)$$

dimana $\boldsymbol{f}(\boldsymbol{t}_i)$ adalah fungsi regresi dan $\boldsymbol{\varepsilon}_i$ adalah sesatan random yang diasumsikan berdistribusi normal independen dengan mean nol dan variansi σ^2.

Ada dua metode yang dapat digunakan untuk menaksir fungsi $\boldsymbol{f}$, yaitu metode regresi parametrik dan metode regresi nonparametrik. Pertimbangan untuk memilih metode mana yang akan digunakan ada kaitannya dengan asumsi dari fungsi $\boldsymbol{f}$ tersebut. Metode regresi parametrik akan sesuai jika bentuk fungsi $\boldsymbol{f}$ diketahui atau ada sumber lain yang bisa digunakan untuk

* *Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, email: ruliana@yahoo.co.id*

menentukan bentuk fungsi $\boldsymbol{f}$ tersebut. Tetapi jika fungsi $\boldsymbol{f}$ tersebut tidak diketahui bentuknya, maka metode regresi nonparametrik lebih sesuai digunakan (Eubank, 1988). Dalam hal ini fungsi $\boldsymbol{f}$ hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi tersebut biasanya dimotivasi oleh sifat kemulusan (*smoothness*) yang dimiliki oleh fungsi $\boldsymbol{f}$ tersebut.

Beberapa pendekatan nonparametrik yang cukup populer dalam mengestimasi fungsi $\boldsymbol{f}$ antara lain *Estimator Deret Orthogonal* (Eubank 1988), Spline (Wahba, 1990), *Estimator Kernel* (Hardle, 1990), Histogram (Green, 1994). Khusus untuk Spline, pendekatan ini merupakan *piecewise polynomial*, yaitu polinomial yang memiliki sifat tersegmen. Sifat inilah yang memberikan fleksibel lebih daripada polynomial biasa, sehingga memungkinkan untuk menyesuaikan diri secara efektif terhadap karakteristik lokal dari fungsi atau data.

Dalam tulisan ini pendekatan regresi yang digunakan adalah regresi nonparametrik spline. Dari bentuk (1), apabila fungsi $\boldsymbol{f} \in W_2^m[0,1] = \{\boldsymbol{f} : \boldsymbol{f}^{(k)}, k = 0,1,..., m-1,\ \boldsymbol{f}^{(m)} \in L_2[0,1]\}$, dimana $L_2[0,1]$ menyatakan himpunan fungsi-fungsi kuadrat terintegral pada interval [0,1], dengan kesesuaian kurva terhadap data adalah $n^{-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{t}_i))^2$ dan kekasaran kurva adalah $\int_0^1 (\boldsymbol{f}''(t))^2\,dt$ maka estimasi $\boldsymbol{f}$ diperoleh dengan meminimumkan *Penalized Least Square*,

$$n^{-1}\sum_{i=1}^{n}\{\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{t}_i)\}^2 + \lambda\int_0^1\{\boldsymbol{f}''(t)\}^2 dx,\ \lambda > 0. \qquad (2)$$

Permasalahan optimasi pada (2) dapat diselesesaikan dengan pendekatan *Reproducing Kernel* (Kemildrorf & Wahba, 1971). Sifat dari *Reproducing Kernel* adalah dapat ditentukan representasi dari suatu fungsional linear terbatas, sehingga kurva regresi $\boldsymbol{f} \in W_2^m[0,1]$ yang merupakan penyelesaian yang optimal dari persamaan (2) bisa diperoleh (Wahba, 2000).

Dalam menghasilkan estimasi kurva regresi yang baik, pemilihan λ yang optimal merupakan hal yang penting. Dengan menggunakan *Reproducing Kernel* untuk mengestimasi kurva regresi pada (1), selanjutnya akan dipilih suatu nilai λ yang optimal. Beberapa metode untuk memilih λ yaitu metode *Cross Validation* (CV), *Generalized Cross Validation* (GCV), metode *Generalized Maximum Likelihood* atau GML (Kou, 2003).

Dalam tulisan ini dibahas penyelesaian optimal dari (2) dengan menggunakan *Reproducing Kernel*, dan pemilihan parameter penghalus λ dengan menggunakan metode GCV dan CV serta melihat sifat efisien dari kedua metode tersebut dengan menggunakan data simulasi.

2. Spline dalam Regresi Nonparametrik

Ada beberapa pendekatan nonparametrik yang cukup populer dalam mengestimasi $\boldsymbol{f}$, salah satunya adalah spline. Spline adalah potongan polinomial order *r*. Titik bersama dari potongan-potongan tersebut disebut dengan knots. Spline order *r* dengan knots pada $\zeta_1,...,\zeta_k$ diberikan dalam fungsi *S* dengan bentuk:

$$S(t) = \sum_{i=0}^{r-1}\boldsymbol{\theta}_i t^i + \sum_{j=1}^{k}\boldsymbol{\delta}_j\,(t-\zeta_j)_+^{r-1} \qquad (3)$$

dengan $(t-\zeta_j)_+^{r-1} = \begin{cases} (t-\zeta_j) & , jika (t-\zeta_j) \geq 0 \\ 0 & , jika (t-\zeta_j) < 0 \end{cases}$ (Eubank, 1988).

Spline (3) mempunyai sifat sebagai berikut:

1. *S* merupakan potongan polinomial derajat *r*-1 pada setiap subinterval $[\zeta_j, \zeta_{j+1}]$.
2. *S* mempunyai turunan ke (*r*-2) yang kontinu.
3. *S* mempunyai turunan ke (*r*-1) yang merupakan fungsi tangga dengan titik-titik lompatan pada $(\zeta_1, ..., \zeta_k)$.

3. *Reproducing Kernel*

Definisi 1. (Wahba, 2000**)**

Reproducing Kernel dari H adalah suatu fungsi R yang didefinisikan pada [0,1]×[0,1] *sedemikian hingga untuk setiap titik tetap t*∈[0,1] *berlaku* $R_t \in H$ *dengan* $R_t(s) = R(s,t)$ *dan*

$$L_t \boldsymbol{f} = < R_t, \boldsymbol{f} > = \boldsymbol{f}(t). \qquad (4)$$

Didefinisikan ruang H_0 memuat koleksi fungsi $H_0 = \{ \boldsymbol{h} : \boldsymbol{h}^{(i-1)}(0) = (0)$, $i = 1,...,m$, $\boldsymbol{h}^{(i-1)}$ kontinu absolut pada [0,1] dan $\boldsymbol{h}^{\boldsymbol{m}} \in [0,1] \}$. H_0 adalah ruang inner product yang lengkap yang disebut *Hilbert space* . H_0 merupakan *Reproducing Kernel Hilbert Space* karena untuk setiap $\boldsymbol{h} \in H_0$ dan pemetaan $L_t : H_0 \to R$, $h \to L_t(h) = 0$ L_t memenuhi sifat linearitas, dengan domain [0,1] didalam ruang H_0, dengan *reproducing kernel* dari H_0 adalah $R_t^0 \in H_0$ yaitu $R_t(s) = \sum_{i=1}^{m} \phi_i(s)\phi_i(t)$, dan untuk titik tetap *t* berlaku $L_t \boldsymbol{h} = \left\langle R_t^0, h \right\rangle = h(t)$, $\boldsymbol{h} \in \boldsymbol{H}_0$.

Ruang H_1 memuat koleksi fungsi $H_1 = \{ g : g^{(k)}(0) = (0), k = 0,1,..., m-1$, $\boldsymbol{g}^{(k)}$ kontinu absolut pada [0,1] dan $\boldsymbol{g}^{\boldsymbol{m}} \in \boldsymbol{L}_2[0,1] \}$. H_1 adalah *Reproducing Kernel Hilbert Space* dengan *Inner Product*

$$\left\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \right\rangle_1 = \int_0^1 \boldsymbol{g}_1^{(\boldsymbol{m})}(\boldsymbol{t}) \boldsymbol{g}_2^{(\boldsymbol{m})}(\boldsymbol{t}) \boldsymbol{dt} \qquad (5)$$

dan *Reproducing Kernel* dari H_1 adalah

$$R_1(s,t) = \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{((m-1)!)^2} du \qquad (6)$$

H adalah ruang Hilbert dan terdapat dengan tunggal *reproducing kernel* R_t untuk titik $t \in [0,1]$ dalam *H* dan $L_t \boldsymbol{f} = \left\langle R_t, \boldsymbol{f} \right\rangle = \boldsymbol{f}(t)$, $\boldsymbol{f} \in H$, L_t adalah fungsional linear terbatas pada *H* yang memetakan fungsi $\boldsymbol{f}$ didalam ruang *H* kebilangan riil , $L_i : f \to f(t_i)$.

Dalam masalah regresi, untuk data berpasangan $\{(\boldsymbol{t}_i, \boldsymbol{y}_i)$, $i = 1,2,...,n\}$, $\boldsymbol{t}_i \in \mathrm{R}$, $\boldsymbol{y}_i \in \mathrm{R}$ dan hubungan antara t_i dan y_i diasumsikan mengikuti model regresi secara umum yaitu $\boldsymbol{y_i} = L_i \boldsymbol{f} + \boldsymbol{\varepsilon}_i$, $i = 1,2,...,n$, dengan $\boldsymbol{f}$ adalah fungsi regresi dan $\boldsymbol{\varepsilon}_i$ adalah sesatan random yang

diasumsikan berdistribusi independen dengan mean nol dan varians σ^2. Menaksir fungsi $\boldsymbol{f}$ adalah mencari $\boldsymbol{f}$ yang ada didalam ruang Hilbert yang meminimumkan $n^{-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - L_i\ \boldsymbol{f})^2 + \|P_1\ f\|_R^2$.

4. *Estimator Spline*

Misalkan diberikan model data

$$\boldsymbol{y}_i = L_i\boldsymbol{f} + \boldsymbol{\varepsilon}_i \; i = 1,2,\ldots,n, \tag{7}$$

dengan $\boldsymbol{f} \in H_R$ dan L_i i =1,2,…,n adalah fungsional linear terbatas pada H_R serta H_R mempunyai dekomposisi $H_R = H_0 \oplus H_1$, estimasi dari $\boldsymbol{f}$ adalah $\boldsymbol{f} \in H_R$ yang meminimumkan

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - L_i\boldsymbol{f})^2 + \lambda\|P_1\boldsymbol{f}\|_R^2 \text{ (Wahba, 1990).} \tag{8}$$

Teorema 1. (Wahba, 1990)

Apabila $H_R = H_0 \oplus H_1$ dan $\phi_1,...,\phi_m$ basis dari ruang H_0 serta $\boldsymbol{T_{nxm}}$ adalah matriks full rank kolom berorder nxm yang diberikan oleh:

$$\boldsymbol{T_{nxm}} = \{L_i\boldsymbol{\varphi}_v\}, i = 1,2,\ldots,n; \text{dan } v = 1,2,\ldots,m$$

maka $\hat{f}$ *yang meminimumkan* $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - <\eta_i, \boldsymbol{f}>)^2 + \lambda\|P_1\boldsymbol{f}\|_R^2$ (9)

adalah $\hat{\boldsymbol{f}} = \sum_{v=1}^{m}\boldsymbol{\alpha}_v\boldsymbol{\varphi}_v + \sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{\xi}_i$

dengan $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_m)' = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}$

$\boldsymbol{\beta} = (\beta_1,...,\beta_n)' = \boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1})\boldsymbol{y}$

$\boldsymbol{\xi}_i = P_1\eta_i$

$\boldsymbol{M} = \boldsymbol{\Sigma} + n\ \lambda\boldsymbol{I}$

$\boldsymbol{\Sigma} = \{<\boldsymbol{\xi}_i, \boldsymbol{\xi}_j>\}$, i , j = 1,2,3,…,n

Bukti:

$\hat{\boldsymbol{f}}$ dapat ditulis dalam bentuk

$$\hat{\boldsymbol{f}} = \sum_{v=1}^{m}\boldsymbol{\alpha}_v\boldsymbol{\varphi}_v + \sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{\xi}_i = \boldsymbol{\varphi}'\boldsymbol{\alpha} + \boldsymbol{\xi}'\boldsymbol{\beta} + \psi,\ \psi \in H_R$$

yang tegak lurus dengan $\phi_1,...,\phi_m$, $\xi_1,...,\xi_n$.

$$\hat{\boldsymbol{f}} = \sum_{v=1}^{m}\boldsymbol{\alpha}_v\boldsymbol{\varphi}_v + \sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{\xi}_i = \boldsymbol{\phi}'\boldsymbol{\alpha} + \boldsymbol{\xi}'\boldsymbol{\beta}$$

$$\{(L_i\boldsymbol{f})\} = \{\langle\eta_i, \hat{\boldsymbol{f}}\rangle\};\ \eta_i \in H_R = H_0 + H_1$$

$$= \{\langle\eta_i, \boldsymbol{\varphi}'\boldsymbol{\alpha} + \boldsymbol{\xi}'\boldsymbol{\beta}\rangle\} = \{\langle\eta_i, \boldsymbol{\varphi}'\boldsymbol{\alpha}\rangle\} + \{\langle\eta_i, \boldsymbol{\xi}'\boldsymbol{\beta}\rangle\},\ i = 1,\ldots,n$$

$$= \boldsymbol{T\alpha} + \boldsymbol{\Sigma\beta} \tag{10}$$

$$\|P\boldsymbol{f}\|_R^2 = \langle P\boldsymbol{f}, P\boldsymbol{f} \rangle$$
$$= \langle P(\boldsymbol{\phi}'\boldsymbol{\alpha} + \boldsymbol{\xi}'\boldsymbol{\beta}), P(\boldsymbol{\phi}'\boldsymbol{\alpha} + \boldsymbol{\xi}'\boldsymbol{\beta}) \rangle$$
$$= \boldsymbol{\beta}'\boldsymbol{\Sigma\beta} \tag{11}$$

$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - L_i\boldsymbol{f})^2 = \frac{1}{n}(\boldsymbol{y} - L\boldsymbol{f})'(\boldsymbol{y} - L\boldsymbol{f})$, dari persamaan (10) selanjutnya

$$= \|(\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{\Sigma\beta})\|^2 \tag{12}$$

Dari persaman (11) dan (12), maka persamaan (9) dapat ditulis sebagai

$$\Re(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n}\|(\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{\Sigma\beta})\|^2 + \lambda\boldsymbol{\beta}'\boldsymbol{\Sigma\beta} \tag{13}$$

Selanjutnya akan dicari α dan β yang meminimumkan persamaan (13)

$$\Re(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n}\|(\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{\Sigma\beta})\|^2 + \lambda\boldsymbol{\beta}'\boldsymbol{\Sigma\beta}$$
$$\Re(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|(\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{\Sigma\beta})\|^2 + n\lambda\boldsymbol{\beta}'\boldsymbol{\Sigma\beta} \tag{14}$$

Dengan menurunkan secara parsial persamaan (14) terhadap $\boldsymbol{\beta}$ dan hasilnya disamakan dengan nol

$$\frac{\partial\Re}{\partial\boldsymbol{\beta}} = -2\boldsymbol{\Sigma}'\boldsymbol{y} + 2\boldsymbol{\Sigma}'\boldsymbol{T\alpha} + 2(\boldsymbol{\Sigma}'\boldsymbol{\Sigma} + n\lambda\boldsymbol{\Sigma})\boldsymbol{\beta} = -\boldsymbol{\Sigma}'\boldsymbol{y} + \boldsymbol{\Sigma}'\boldsymbol{T\alpha} + (\boldsymbol{\Sigma}'\boldsymbol{\Sigma} + n\lambda\boldsymbol{\Sigma})\boldsymbol{\beta} = 0$$

$$\boldsymbol{\beta} = \boldsymbol{M}^{-1}(\boldsymbol{y} - \boldsymbol{T\alpha}) \tag{15}$$

Dengan cara yang sama persamaan (14) diturunkan terhadap $\boldsymbol{\alpha}$ di dapatkan

$$\frac{\partial\Re}{\partial\boldsymbol{\alpha}} = -2\boldsymbol{T}'\boldsymbol{y} + 2\boldsymbol{T}'\boldsymbol{T\alpha} + 2\boldsymbol{T}'\boldsymbol{\Sigma\beta} = 0$$
$$-\boldsymbol{T}'\boldsymbol{y} + \boldsymbol{T}'\boldsymbol{T\alpha} + \boldsymbol{T}'\boldsymbol{\Sigma\beta} = 0. \tag{16}$$

Sehingga

$$\boldsymbol{\alpha} = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y} \tag{17}$$
$$\boldsymbol{M} = \boldsymbol{\Sigma} + n\lambda\boldsymbol{I}.$$

Dari persamaan (17) disubtitusi pada persamaan (15)

$$\beta = M^{-1}(y - T\alpha)$$
$$\beta = M^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1})y. \tag{18}$$

Sehingga terbukti bahwa $\hat{\boldsymbol{f}}$ yang meminimumkan

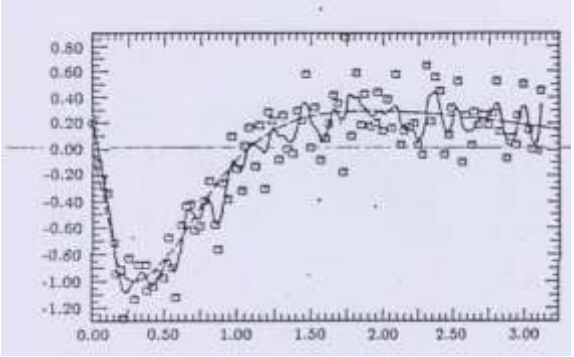
$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - <\eta_i, \boldsymbol{f}>)^2 + \lambda\|P_1\boldsymbol{f}\|_R^2 \text{ adalah } \boldsymbol{f}_\lambda = \sum_{v=1}^{m}\boldsymbol{\alpha}_v\boldsymbol{\varphi}_v + \sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{\xi}_i \,,$$

dengan $\boldsymbol{\alpha}$ dan $\boldsymbol{\beta}$ yaitu

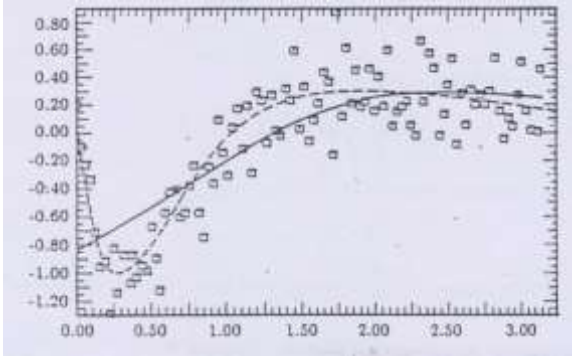
$$\boldsymbol{\alpha} = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}$$

$$\beta = M^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1})y.$$

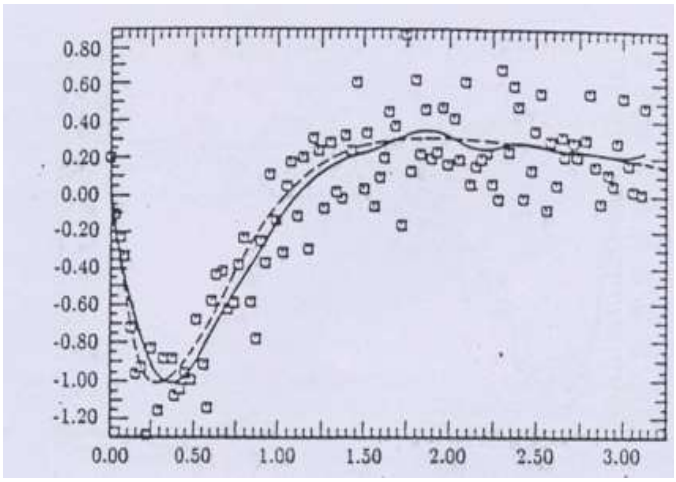
Bentuk estimator spline, dipengaruhi oleh parameter penghalus λ. nilai λ yang sangat kecil atau besar akan memberikan bentuk fungsi penyelesaian yang sangat kasar atau sangat mulus (Wahba 1990; Eubank, 1988).



Gambar 1. Estimasi Data dengan λ yang Kecil.



Gambar 2. Estimasi Data dengan λ yang Besar



Gambar 3. Estimasi Data dengan λ Optimal.

5. Pemilihan Parameter Penghalus

Dalam regresi nonparametrik, bentuk fungsi $\boldsymbol{f}$ tidak diketahui, fungsi $\boldsymbol{f}$ diasumsikan *smooth* dalam arti $\boldsymbol{f}$ merupakan anggota ruang Sobolev $W_2^m[0,1] = \{\boldsymbol{f} : \boldsymbol{f}^{(k)}, k = 0,1,...,m-1$, kontinu absolute pada [0,1] dan $\boldsymbol{f}^{(m)} \in L_2[0,1]\}$ dengan $L_2[0,1]$ menyatakan himpunan fungsi-fungsi kuadrat terintegral pada interval [0,1].

Idealnya akan dipilih suatu nilai λ yang meminimumkan fungsi kerugian $L(\lambda)$, akan tetapi dalam regresi nonparametrik tidak dapat dilakukan secara nyata sebab $L(\lambda)$ masih memuat

fungsi $\boldsymbol{f}$ yang tidak diketahui. Sehingga perlu mengestimasi data dan kemudian estimatornya diminimumkan terhadap λ untuk mendapat estimator $\boldsymbol{f}$ yang paling baik (Eubank, 1988). Salah satu pemilihan parameter penghalus λ adalah menggunakan metode *Generalized Cross Validation* (GCV). Ide dasar dari GCV adalah memodifikasi *Cross Validation* atau CV (Green & Silverman 1994).

Definisi 2. (Eubank, 1988)
Jika $\boldsymbol{f}_\lambda = (f_{\lambda 1},..., f_{\lambda n})^T$ *adalah estimator untuk* $\boldsymbol{f}$ *maka fungsi kerugian (loss function) didefenisikan sebagai* $L(\lambda) = n^{-1}\sum_{i=1}^{n}(\boldsymbol{f}_i - \boldsymbol{f}_{\lambda i})^2$.

Teorema 2. (Craven dan Wahba, 1979**)**
Jika $\boldsymbol{f} \in W_2^m[0,1]$ *dan* λ_{GCV} , λ_{res} *menyatakan nilai* λ *yang meminimumkan EGCV(*λ*) dan EL(*λ*), maka terdapat suatu barisan* λ_{GCV} *sedemikian sehingga*

$$\frac{EL(\lambda_{GCV})}{EL(\lambda)} \to 1,\ n \to \infty.$$

Jika σ^2 diketahui maka λ optimal bisa didapat langsung dari kriteria prediksi *mean square error* atau fungsi kerugian yang didefenisikan oleh

$$L(\lambda) = n^{-1}\sum_{i=1}^{n}(\boldsymbol{f}_i - \boldsymbol{f}_{\lambda i})^2 \qquad (19)$$

fungsi dari persamaan (19) memberikan penaksiran dari kebaikan suatu estimator. Dalam hal σ^2 tidak diketahui maka dapat digunakan metode *Cross Validation* (CV) dan *Generalized Cross Validation* (GCV), untuk mendapatkan nilai λ optimal. *Cross validation* (CV) adalah metode untuk memilih λ yang meminimumkan

$$\mathrm{CV}(\lambda) = n^{-1}\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{f}_\lambda^{[j]}(t))^2 \qquad (20)$$

$\boldsymbol{f}_\lambda^{[j]}(t)$ adalah $\boldsymbol{f} \in W_2^m[0,1]$ yang meminimumkan

$$n^{-1}\sum_{i=1}^{n}\{\boldsymbol{y}_i - \boldsymbol{f}(t_i)\}^2 + \lambda\int_0^1\{\boldsymbol{f}''(t)\}^2 dx.$$

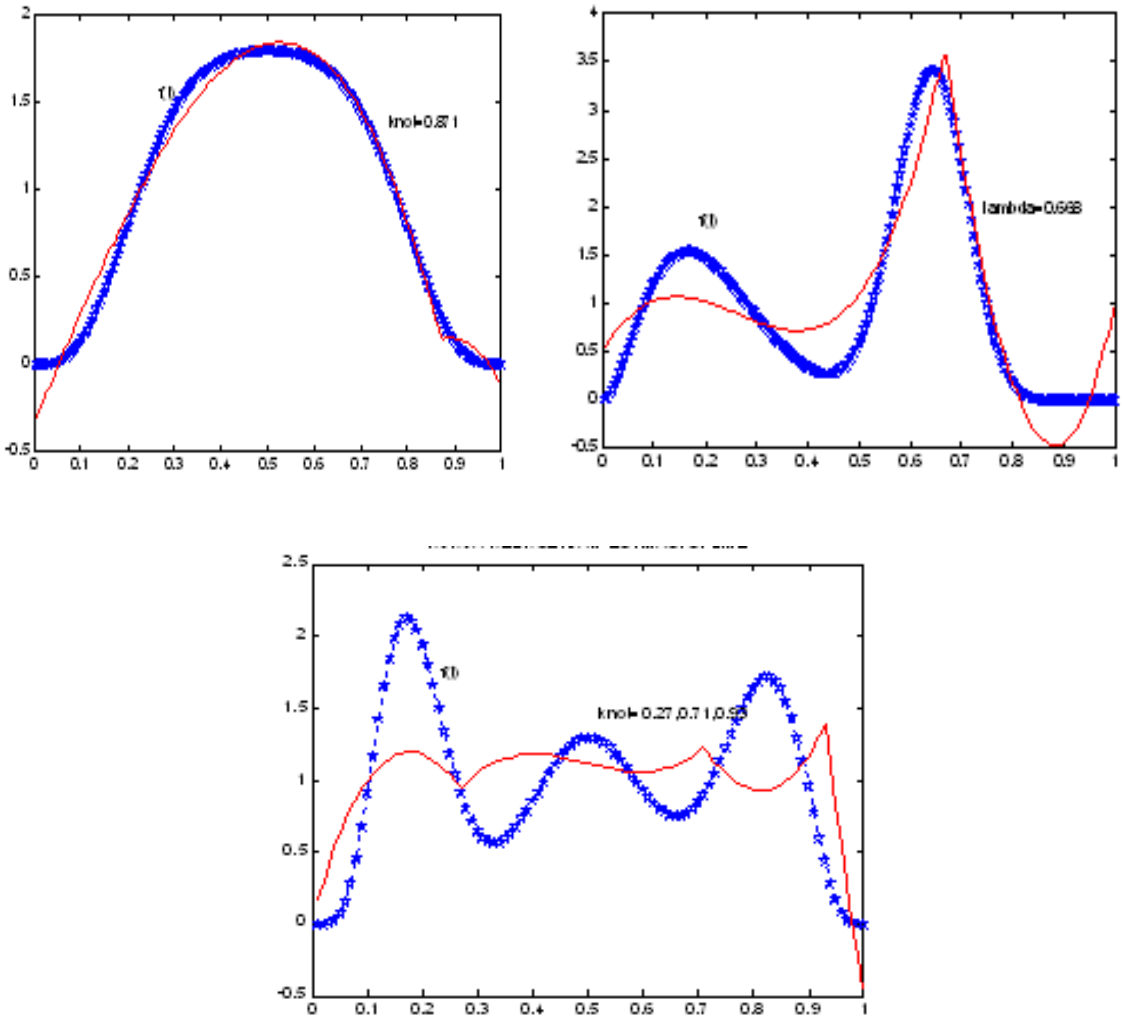
Sedangkan *Generalized Cross Validation* (GCV), adalah metode untuk memilih λ yang meminimumkan

$$GCV(\lambda) = n^{-1}\sum_{j=1}^{n}\frac{\left(\boldsymbol{y}_j - \boldsymbol{f}_\lambda(\boldsymbol{t}_j)\right)^2}{\left(1 - n^{-1}\sum_{i=1}^{n}\boldsymbol{a}_{jj}(\lambda)\right)^2} \qquad (5.3)$$

6. Studi Simulasi Metode GCV dan CV dalam Mendapatkan λ Optimal

Dalam penelitian ini, diselidiki prilaku optimal estimator spline berdasarkan data simulasi, dengan metode GCV dan CV. Data dibangkitkan dari distribusi Beta dengan berbagai nilai *a* dan *b*:

$$B(a,b) = \frac{\tau(a+b)}{\tau(a)\tau(b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x).$$



Gambar 4. Plot Tiga Fungsi Beta dengan *a* dan *b* yang Berbeda.

Dalam menentukan λ optimal dengan metode *Generalized Cross Validation* (GCV) dan *Cross Validation* CV, dimana $I_{\text{GCV}} = \frac{L(\lambda_{GCV})}{L(\lambda_{opt})}$, $I_{CV} = \frac{L(\lambda_{CV})}{L(\lambda_{opt})}$ didapatkan hasil yang dapat dilihat dalam Tabel 1.

Tabel 1. Hasil Simulasi Lambda Optimal.

σ^2=0,0125	λ Optimal	n=25		n=50		n=100		n=200	
		λ	nilai	λ	Nilai	λ	Nilai	λ	nilai
	λ(GCV)	0,715	0,0077	0,720	0,0074	0,714	0,0069	0,715	0,0068
	λ(CV)	0,701	0,0106	0,702	0,0081	0,704	0,0071	0,712	0,0069
	λ(LOSS)	0,730	0,0047	0,723	0,0057	0,715	0,0061	0,714	0,0063
σ^2=0,025	λ(GCV)	0,734	0,0078	0,721	0,0079	0,718	0,0075	0,712	0,0072
	λ(CV)	0,718	0,0105	0,709	0,0086	0,712	0,0078	0,709	0,0073
	λ(LOSS)	0,73	0,0048	0,718	0,0058	0,721	0,0062	0,715	0,0064
σ^2=0,05	λ(GCV)	0,723	0,0101	0,732	0,0097	0,714	0,0091	0,717	0,0093
	λ(CV)	0,696	0,0127	0,714	0,0107	0,702	0,0091	0,708	0,0094
	λ(LOSS)	0,729	0,0052	0,717	0,0061	0,72	0,0062	0,714	0,0064
σ^2=0,1	λ(GCV)	0,687	0,0201	0,696	0,0194	0,702	0,0172	0,717	0,0163
	λ(CV)	0,675	0,0213	0,678	0,0208	0,693	0,0173	0,714	0,0164
	λ(LOSS)	0,729	0,0074	0,72	0,0063	0,717	0,0064	0,72	0,0066

Tabel 2. Ratio GCV dan Ratio CV.

σ^2	Ratio	n=25	n=50	n=100	n=200
0,0125	W_r(GCV)	1,562904	1,236415	1,10849220	1,05192572
	W_r(CV)	1,648624	1,220889	1,05407279	1,04367263
0,025	W_r(GCV)	1,564278	1,235004	1,10814558	1,05192572
	W_r(CV)	1,649797	1,219301	1,09358752	1,04367263
0,05	W_r(GCV)	1,572061	1,234827	1,10814558	1,05570839
	W_r(CV)	1,655471	1,219125	1,09341421	1,04745530
0,1	W_r(GCV)	1,587077	1,253831	1,12623377	1,06497100
	W_r(CV)	1,678571	1,239211	1,11168831	1,05672100

Pada Tabel 2 terlihat bahwa, dengan bertambahnya ukuran sampel, nilai ratio dari GCV semakin mendekati 1, begitu juga untuk nilai ratio dari CV, namun dalam hal ini ratio GCV selalu lebih kecil dari ratio CV untuk semua ukuran sampel n dan σ^2 yang diberikan.

Tabel 3. Nilai I_{GCV}, dan I_{CV}.

σ^2	efisiensi	n=25	n=50	n=100	n=200
0,0125	I_{GCV}	1,021277	1,000000	1,000000	1,000000
	I_{CV}	1,042553	1,017544	1,016393	1,000000
0,025	I_{GCV}	1,000000	1,000000	1,000000	1,000000
	I_{CV}	1,020833	1,000000	1,000000	1,000000
0,005	I_{GCV}	1,019231	1,000000	1,000000	1,000000
	I_{CV}	1,000000	1,000000	1,016129	1,016129
0,1	I_{GCV}	1,040541	1,000000	1,000000	1,000000
	I_{CV}	1,054054	1,01587	1,000000	1,000000

Dari Tabel 3 terlihat bahwa nilai I_{GCV} untuk ukuran sampel yang besar lebih banyak yang mendekati 1 dibandingkan dengan nilai I_{CV}.

Hasil simulasi ini adalah dengan melihat ratio dan efisiensi maka menggunakan metode *Generalized Cross Validation* (GCV) didalam mencari λ yang paling optimal adalah lebih efisien dari metode *Cross Validation* (CV).

6. Kesimpulan

Kesimpulan yang diperoleh dari hasil penelitan ini, yaitu bahwa apabila $H_R = H_0 \oplus H_1$ dan $\phi_1, \ldots \phi_m$ span dari ruang H_0 serta $\boldsymbol{T}_{nxm}$ adalah matriks full rank kolom berorder *nxm* yang diberikan oleh $\boldsymbol{T}_{nxm} = \{\boldsymbol{L}_i \phi_v\}$,*i* = 1,2,…,*n* dan *v* = 1,2,…,*m*, maka $\hat{\boldsymbol{f}}$ yang meminimumkan $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - <\eta_i, \boldsymbol{f}>)^2 + \lambda\|P_1\boldsymbol{f}\|_R^2$ adalah $\hat{\boldsymbol{f}} = \sum_{v=1}^{m}\boldsymbol{\alpha}_v\boldsymbol{\varphi}_v + \sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{\xi}_i$, dengan $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)' = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}$ dan $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)' = M^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1})y$. Dengan melihat ratio dan efisiensi maka menggunakan metode *Generalized Cross Validation* (GCV) didalam mencari λ yang paling optimal adalah lebih efisien dari metode *Cross Validation* (CV).

Daftar Pustaka

[1] Berberian, K. S., 1961, *Introduction to Hilbert Space*, New York.

[2] Budiantara, I. N. dan Subanar, 1997, Pemilihan parameter penghalus dalam regresi spline terbobot, *Majalah Ilmiah Matematika dan IPA-UGM,* Yogyakarta.

[3] Cox, D., 1981, Asymptotics for M-Type smoothing splines, *Technical Report*, 654, University of Wisconsin.

[4] Craven, P. dan Wahba, G., 1979, Smooting noisy data with spline functions estimating the correct degree of smoothing by method of Generalized Cross-Validation, *Journal Numerische Mathematics*, 31, 377-403.

[5] Eubank, R. L., 1988, *Spline smoothing and Nonparametrik Regression*, Marcel Dekker, New York.

[6] Green, P. J. dan Silverman, B.W., 1994, *Nonparametrik Regression and Generalized Linear Models (a Roughness Penalty Approach*), Chapman & Hall, New York.

[7] Hardle, G., 1990, *Applied Nonparametrik Regression,* Cambridge University Press, New York.

[8] Kimeldrof, G. dan Wahba, G., 1971, Some result on Tchebycheffian spline function, *Journal of Mathematical Analysis and Application*, 33, 82-95.

[9] Kou, S.C., 2003, On efficiency of selection criteria in spline regression, *Probab, Theory Relat. Fields*, 127, 153-176.

[10] Kreyszig, E., 1978, *Introductory Functional Analysis with Applications*, University of Windsor, John Wiley & Sons, New York.

[11] Mallows, C., 1973, Some comments on Cp, *Journal Technometrics*, 15, 661-675.

[12] Oeherlet, G. W., 1992, Relaxed boundary smoothing splines, *Journal the Annals of Statistics*, 20, 146-160.

[13] Rohatgi, V. K., 1976, *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley & Sons, New York.

[14] Silverman, B., 1984, A fast and efficient cross-validation method for smoothing parameter choice in spline regression, *Journal of the American Statistical Association*, 79,584-589.

[15] Silverman, B., 1985, some aspects of the spline smoothing approach to non parametric regression curve fitting (With discussion), *Journal Royal Statistical Society*, 47, 1-52.

[16] Speckman, P., 1980, Minimax estimates of linear functionals in a Hilbert space, *Unpublished Manuscript*.

[17] Speckman, P., 1985, Spline smoothing and optimal rates of convergence in nonparametric regression models, *Journal Ann. Statist*, 13, 970-983.

[18] Wahba, G., 1975, Smoothing noisy data by spline functions, *Journal Numer. Math*, 24, 383-393.

[19] Wahba, G., 1985, A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, *Journal the Annals of Statistics*, 13, 1378-1402.

[20] Wahba, G., 1990, Spline model for observational data, *Society For Industrial and Applied Mathematics*, Philadelphia.

[21] Wahba, G., 2000, An introduction to model building with reproducing kernel Hilbert spaces, *Technical Report* , 1020, University of Wisconsin.