MODEL SPLINE DENGAN ERROR BERKORELASI

Nalim, I Nyoman Budiantara dan Kartika Fitriasari
Jurusan Statistika ITS
Kampus ITS Sukolilo Surabaya 60111

Abstract. Spline smoothing is a popular method for estimating the function in nonparametric regression model. Its performance depends greatly on the choice of smoothing parameters. Many methods of selecting smoothing parameters such as GCV, GML and UBR are developed under the assumption of independent observations. They fail badly when data are correlated. In nonparametric regression, correlated error could be solved by finding weighted estimator and determine the correlation matrix from the error. Estimation of nonparametric function is obtained by minimizing the penalized weighted least-square (PWLS). In this paper, the extension of the GML method to estimate the smoothing parameters and correlation simulataneously is presented. Simulation was conducted to evaluate and to compare the performance of the original GML and the extended GML method. The extended GML is recommended since it works well in all simulation scheme. This method is also able to illustrate the data concentration data in a continous chemical process.

Keywords: Generalized cross validation, generalized maximum likelihood, nonparametric regression, penalized weighted least-square

1. PENDAHULUAN

Misalkan diberikan data berpasangan $(t_i, y_i),\ i = 1,2,...,n$ dan hubungan antara t_i dan y_i diasumsikan mengikuti model

$$y_i = f(t_i) + \varepsilon_i, \quad t_i \in [a,b],\ i = 1,2,...,n \tag{1.1}$$

dimana $f(t_i)$ adalah kurva regresi dan ε_i adalah sesatan random yang diasumsikan memiliki mean nol dan varian σ^2. Dalam pembahasan model regresi parametrik bentuk kurva regresi diketahui, sedangkan dalam pendekatan regresi nonparametrik tidak ada asumsi terhadap bentuk kurva $f(t_i)$. Kurva tersebut hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi ini dimotivasi oleh sifat kehalusan (*smoothness*) yang dimiliki oleh fungsi regresi tersebut. Adapun regresi semiparametrik merupakan gabungan dari pendekatan regresi parametrik dan nonparametrik.

Pada estimasi kurva regresi nonparametrik, spline memainkan peranan yang cukup penting. Spline dikembangkan dari potongan polinomial yang mempunyai sifat fleksibel dan efektif untuk menangani sifat lokal data [4]. Spline yang didasarkan pada suatu persoalan optimasi telah dikembangkan oleh [10].

Pada penelitian-penelitian sebelumnya, sebagian besar penulis membahas bagaimana mengestimasi kurva regresi nonparametrik berdasarkan asumsi bahwa sesatan random adalah independen. Tidak sedikit kasus-kasus yang ada dalam aplikasi, sering terjadi observasi yang diperoleh saling berkorelasi, misalnya data time series, data spasial. Banyak penulis yang mempelajari efek korelasi, diantaranya ada yang membahas bagaimana efek korelasi berpengaruh terhadap pemilihan *bandwidth* ([1],[5],[6],[7]). Estimasi *bandwidth* dari mean square error (MSE) telah dikaji oleh beberapa peneliti ([1],[5]). Data timeseries dengan metode *cross-validation* untuk mengetimasi *bandwidth* dan fungsi autokorelasi diteliti oleh [6]. Beberapa metode juga telah dikembangkan untuk penghalus spline. [3] memperluas metode GCV untuk mengestimasi parameter penghalus dan pa-

rameter autokorelasi. [9] merepresentasikan penghalus spline dengan suatu model *state-space* dan menggunakan CV, GCV dan GML untuk mengestimasi suatu barisan *autoregressive moving average*. [8] menggunakan suatu frekuensi domain *cross validation* untuk mengestimasi parameter penghalus.

Observasi berkorelasi sangat berpengaruh terhadap pemilihan parameter penghalus, yang merupakan hal penting pada pendekatan spline. Beberapa metode yang cukup populer dalam memilih parameter penghalus adalah *generalized maximum likelihood* (GML), *generalized cross validation* (GCV), dan *unbiased risk* (UBR) [10]. Metode-metode tersebut cenderung kurang mampu mengestimasi dengan baik terhadap parameter penghalus jika data yang digunakan memiliki korelasi positif dan korelasinya diabaikan. Untuk mengatasi masalah tersebut, pada penelitian ini akan diajukan suatu model penghalus spline yang mampu mengestimasi kurva regresi dan parameter penghalus jika data yang digunakan adalah observasi data yang terdapat korelasi. Untuk menunjukkan bagaimana perbandingan metode pemilihan parameter penghalus standar yang diakibatkan oleh error yang berkorelasi akan dilakukan simulasi data dari model $y_i = sin(2\pi i/100) + \varepsilon_i$, $i = 1,...,n$, n=50, 100, 150, dengan ε_i dibangkitkan dari proses *autoregressive* orde pertama (AR(1)), dengan mean nol, variansi $\sigma^2 = 0.1,\ 0.3$ dan korelasi orde pertama $\rho = 0.1,\ 0.5,\ 0.8$.

Dalam kajian ini, akan diberikan metode GML untuk observasi berkorelasi yang menggunakan pendekatan spline. Metode ini diaplikasikan pada data konsentrasi dari suatu proses kimia (Box and Jenkins, 1994).

2. ESTIMATOR MODEL SPLINE DENGAN ERROR BERKORELASI

Berikut diberikan beberapa definisi yang akan digunakan pada pembahasan ini.

Definisi 1. *Ruang inner product* adalah suatu ruang vektor *X* dengan *suatu inner product* yang didefinisikan pada *X*.
Inner product pada *X* adalah suatu pemetaan dari $X \times X$ ke skalar di *X*; yaitu, untuk setiap pasang vektor *x* dan *y*, terdapat suatu hubungan dengan sebuah skalar yang ditulis sebagai $\langle x, y \rangle$

Definisi 2. *Reproducing kernel* (r.k.) dari *H* adalah suatu fungsi *R* yang didefinisikan pada $[0,1]\times[0,1]$ sedemikian hingga untuk setiap titik tertentu $t \in [0,1]$ berlaku $R_t \in H$ dengan $R_t(s) = R(s,t)$ dan $f(t) = < R_t, f >, \quad f \in H$.

Definisi 3 *Reproducing kernel Hilbert space* (r.k.h.s.) *H* adalah suatu ruang Hilbert dari fungsi-fungsi bernilai real pada [0,1] dengan sifat bahwa untuk setiap $t \in [0,1]$, fungsional L_t yang menghubungkan f dengan $f(t)$, $L_t f \to f(t)$, merupakan fungsional linier terbatas, dalam arti terdapat bilangan real c sedemikian hingga $|L_t f| = |f(t)| \le c \| f \|$, untuk semua f di r.k.h.s, dengan $\|\cdot\|$ adalah norm di ruang Hilbert.

Asumsikan data mengikuti model (1.1) dengan $f \in W_m^2$ dan $\varepsilon_i \sim N(0, \sigma^2 W^{-1})$. Jika model data dihubungkan dengan permasalahan spline secara umum maka model (1.1) menjadi:

$$y_i = L_i f + \varepsilon_i, \qquad i = 1,2,...,n, \tag{2.1}$$

dengan $\varepsilon_i \sim N(0, \sigma^2 W^{-1})$, $f \in H$, dan $L_1,...,L_n$ merupakan fungsional linier terbatas di *H*. Misalkan *H* dapat didekomposisi menjadi
$H = H_0 \oplus H_1$,
dimana $\oplus$ adalah *direct sum*. Estimator f diperoleh dengan mencari $f \in H$ yang meminimumkan

$$n^{-1} \sum_{i=1}^{n} w_i (y_i - f(t_i))^2 + \lambda \| Pf \|^2 \tag{2.2}$$

dengan P adalah proyeksi ortogonal f pada H_1 di H. Bentuk estimator f diberikan melalui teorema berikut.

Teorema 2.1. Misalkan $\phi_1,...,\phi_m$ basis ruang H_0 dan $\xi_1,...,\xi_n$ basis H_1. Jika P adalah proyeksi ortogonal dari f ke H_1 dalam H dan matrik $\boldsymbol{T}_{n\times m}$ adalah matrik full rank yang diberikan oleh

$$\boldsymbol{T}_{n\times m} = \{ L_i\phi_v \}_{i=1\ \ v=1}^{n\ \ m} \qquad (2.3)$$

maka $\hat{f}$, yang meminimumkan (2.2) diberikan oleh:

$$\hat{f} = \sum_{v=1}^{m} d_v\phi_v + \sum_{i=1}^{n} c_i\xi_i, \qquad (2.4)$$

dengan

$$\boldsymbol{d} = (d_1,....,d_m)' = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}$$

$$\boldsymbol{c} = (c_1,....,c_n)' = \boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1})\boldsymbol{y}, \qquad (2.5)$$

$$\boldsymbol{M} = \boldsymbol{\Sigma} + n\lambda\boldsymbol{W}^{-1},$$

$$\boldsymbol{\Sigma} = \{ < \xi_i, \xi_j > \}, i, j = 1,2,...,n.$$

Bukti.

Karena $f \in H = H_0 \oplus H_1$, maka f selalu dapat ditulis menjadi

$f = \boldsymbol{\varphi}'\boldsymbol{d} + \boldsymbol{\xi}'\boldsymbol{c}$, dengan $\boldsymbol{\varphi}'\boldsymbol{d} \in H_0$,

$\boldsymbol{\xi}'\boldsymbol{c} \in H_1$,

dimana

$\boldsymbol{\varphi}'\boldsymbol{d} = \phi_1 d_1 + ... + \phi_m d_m$,

$\boldsymbol{\xi}'\boldsymbol{c} = \xi_1 c_1 + ... + \xi_n c_n$.

Berdasarkan teorema representasi Riesz diperoleh

$$\{ < \eta_i, f > \} = \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c}. \qquad (2.6)$$

Karena P adalah proyeksi ortogonal dari f pada H_1 dan $\phi' \in H_0$, maka $P\boldsymbol{\varphi}'\boldsymbol{d} = 0$.

Sehingga diperoleh

$$\|Pf\|^2 = < \boldsymbol{\xi}'\boldsymbol{c}, \boldsymbol{\xi}'\boldsymbol{c} > = \boldsymbol{c}'\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}. \qquad (2.7)$$

Substitusi (2.6) dan (2.7) ke (2.2) diperoleh

$$n^{-1}(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c}) + \lambda\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}. \qquad (2.8)$$

Dengan memaksimumkan (12) diperoleh

$$\boldsymbol{c} = \boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1})\boldsymbol{y}. \qquad (2.9)$$

$$\boldsymbol{d} = (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}. \qquad (2.10)$$

■

3. PEMILIHAN PARAMETER PENGHALUS

Parameter penghalus merupakan pengontrol keseimbangan antara kesesuaian kurva terhadap data dan kemulusan kurva. Beberapa peneliti ([4],[10]) telah menunjukkan bahwa memasangkan parameter penghalus yang sangat kecil atau besar akan memberikan bentuk fungsi penyelesaian yang sangat kasar atau mulus. Dilain pihak diinginkan suatu bentuk estimator disamping mempunyai suatu derajat kemulusan, juga sesuai dengan datanya. Memilih parameter penghalus pada prinsipnya adalah sama dengan memilih titik knot optimal yang menghasilkan nilai GML minimum [2]. Berikut diberikan metode untuk memilih parameter penghalus dan parameter korelasi yang optimal dengan metode *Generalized Maximum Likelihood* (GML).

Diberikan dua buah vektor $\boldsymbol{z}$ dan $\boldsymbol{w}$ dengan dekomposisi sebagai berikut

$$\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}'_2 \\ \frac{1}{\sqrt{\eta}}\boldsymbol{T}' \end{pmatrix} \boldsymbol{y} \qquad (3.1)$$

yang memenuhi $\boldsymbol{Q}'_2\boldsymbol{Q}_2 = \boldsymbol{I}_{n-m}$, $\boldsymbol{Q}'_2\boldsymbol{T} = 0_{(n-m)\times m}$, $\eta = a/b$ dan $\lambda = \sigma^2/nb$

Berdasarkan dekomposisi ini, akan ditentukan distribusi $\boldsymbol{z}$ dan $\boldsymbol{w}$ melalui teorema berikut

Teorema 3.1. Misalkan diberikan variabel random $\boldsymbol{y} = (y_1,..., y_n)'$,

$\boldsymbol{f} = (f(t_1),..., f(t_n))'$ dan $\boldsymbol{\varepsilon} = (\varepsilon_1,..., \varepsilon_n)'$ berdistribusi normal dengan mean nol dan mengikuti model

$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$,

dengan $E(\boldsymbol{f}) = 0$, $E(\boldsymbol{f}\boldsymbol{f}') = b\boldsymbol{\Sigma}_f$,

$\boldsymbol{\Sigma}_f = \frac{a}{b}\boldsymbol{T}\boldsymbol{T}' + \boldsymbol{\Sigma}$, $E(\boldsymbol{\varepsilon}) = 0$,

$E(\boldsymbol{\varepsilon\varepsilon}') = \sigma^2\boldsymbol{W}^{-1}$, dan $E(\boldsymbol{f}\,\boldsymbol{\varepsilon}') = 0$. Jika $f(t)$, $t \in [0,1]$ mempunyai distribusi prior improper dengan fungsi prior adalah

$$f(t) = \sum_{v-1}^{m} \theta_v \phi_v(t) + b^{1/2} X(t), \quad t \in [0,1]$$

dengan $\theta = (\theta_1,...,\theta_m)' \sim$ N(0, a**I**), a dan b adalah konstanta positif dan **z,w** mempunyai dekomposisi (3.1), maka untuk $a \to \infty$

(i) $\boldsymbol{z} \sim \mathrm{N}(0, b\boldsymbol{Q}'_2(\boldsymbol{\Sigma} + n\lambda\boldsymbol{W}^{-1})\boldsymbol{Q}_2)$

(ii) $\boldsymbol{w} \sim \mathrm{N}(0, b(\boldsymbol{T}'\boldsymbol{T})(\boldsymbol{T}'\boldsymbol{T}))$

Bukti:

(i) Karena $\mathrm{var}(\boldsymbol{y}) = b\boldsymbol{\Sigma}_f + \sigma^2\boldsymbol{W}^{-1}$ dan

$\boldsymbol{\Sigma}_f = \frac{a}{b}\boldsymbol{T}\boldsymbol{T}' + \boldsymbol{\Sigma}$ sehingga

$\mathrm{var}(\boldsymbol{y}) = b(\eta\boldsymbol{T}\boldsymbol{T}' + \boldsymbol{\Sigma} + n\lambda\boldsymbol{W}^{-1})$,

Sehingga berdasarkan dekomposisi (3.1) diperoleh

$$\mathrm{var}(\boldsymbol{z}) = \boldsymbol{Q}'_2\,\mathrm{var}(\boldsymbol{y})[\boldsymbol{Q}'_2]' = b(\boldsymbol{Q}'_2\boldsymbol{\Sigma}\boldsymbol{Q}_2 + n\lambda\boldsymbol{Q}'_2\boldsymbol{W}^{-1}\boldsymbol{Q}_2),$$

(ii) Dari dekomposisi (3.1) diperoleh $\boldsymbol{w} = \left(\frac{1}{\sqrt{\eta}}\boldsymbol{T}'\right)\boldsymbol{y}$, sehingga

$$var(\boldsymbol{w}) = \left(\tfrac{1}{\sqrt{\eta}}\boldsymbol{T}'\right)var(\boldsymbol{y})\left(\tfrac{1}{\sqrt{\eta}}\boldsymbol{T}'\right)$$

$$= \frac{b}{\eta}\boldsymbol{T}'(\eta\boldsymbol{T}\boldsymbol{T}' + \boldsymbol{\Sigma} + n\lambda\boldsymbol{W}^{-1})\boldsymbol{T}$$

untuk $\eta \to \infty$ diperoleh

$$\lim_{\eta\to\infty} \mathrm{var}(\boldsymbol{w}) = b(\boldsymbol{T}'\boldsymbol{T})(\boldsymbol{T}'\boldsymbol{T}).$$

■

Teorema diatas memperlihatkan bahwa hanya $\boldsymbol{z}$ yang tergantung pada λ, sehingga estimasi generalized maximum likelihood (GML) untuk λ dan ρ adalah memaksimumkan log likelihood berdasarkan $\boldsymbol{z}$: $l_1(\lambda, \rho, \mathrm{b}|\boldsymbol{z})$

$$= -\frac{n-m}{2}\log b - \frac{1}{2}\log\left|\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2\right| - \frac{1}{2b}\boldsymbol{z}'(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2)^{-1}\boldsymbol{z} + C_1$$

Memaksimumkan l_1 terhadap b diperoleh

$$\hat{b} = \frac{\boldsymbol{z}'(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2)^{-1}\boldsymbol{z}}{n-m}.$$

Estimasi GML terhadap λ dan ρ adalah dengan cara maksimumkan:

$$\mathrm{l}_2(\lambda,\rho \mid \hat{\mathrm{b}}) = -\frac{n-m}{2}\log\hat{b} - \frac{1}{2}\log\left|\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2\right| - \frac{1}{2\hat{b}}\boldsymbol{z}'(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2)^{-1}\boldsymbol{z} + \mathrm{C2}$$

$$= -\frac{n-m}{2}\log\frac{\boldsymbol{z}'\ \boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2\boldsymbol{z}}{[det(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q})^{-1}]^{1/(n-m)}} + C_3$$

dengan C_3 adalah suatu konstanta. Untuk memaksimumkan $l_2(\lambda,\tau \mid \hat{b})$ ekivalen dengan meminimumkan

$$M(\lambda,\tau) = \frac{\boldsymbol{z}'(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2)^{-1}\boldsymbol{z}}{[\det(\boldsymbol{Q}'_2\boldsymbol{B}(\lambda,\rho)\boldsymbol{Q}_2)^{-1}]^{1/(n-m)}} = \frac{\boldsymbol{y}'\ \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}}{[det^{+}(\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{A})]^{1/(n-m)}} \quad (3.2)$$

dimana $\det^+$ adalah hasil kali nilai-nilai eigen yang tak nol. Karena $\lambda = \sigma^2/nb$ maka estimasi varians σ^2 adalah

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}'\ \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}}{n-m}.$$

4. HASIL SIMULASI

Studi simulasi dilakukan untuk mengevaluasi dan membandingkan GML original dan GML dengan error berkorelasi berdasarkan kriteria MSE minimum. Model yang dipakai dalam simulasi adalah

$$y_i = sin(2\pi i/n) + \varepsilon_i, \quad i = 1,...,n.$$

dan membangkitkan error ε_i dari proses AR(1) dengan mean 0, dan menetapkan tiga ukuran sampel *n* = 50, 100, 150; dua variansi σ^2 = 0.1, 0.3 dan tiga korelasi yaitu ρ = 0.1, 0.5, 0.8. Dengan demikian akan terdapat 3 x 2 x 3 = 18 percobaan. Titik knot optimum dipilih berdasarkan nilai

GML minimum, selanjutnya model spline original dan spline dengan error berkorelasi dihitung nilai MSE-nya untuk berbagai ukuran sampel, koefisien korelasi dan variansi. Fungsi spline yang digunakan adalah *Cubic spline*, yaitu

$$\hat{f}(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t-k_1)_+^3 + \beta_5 (t-k_2)_+^3 + \ldots + \beta_{p+3} (t-k_p)_+^3 \quad (3.3)$$

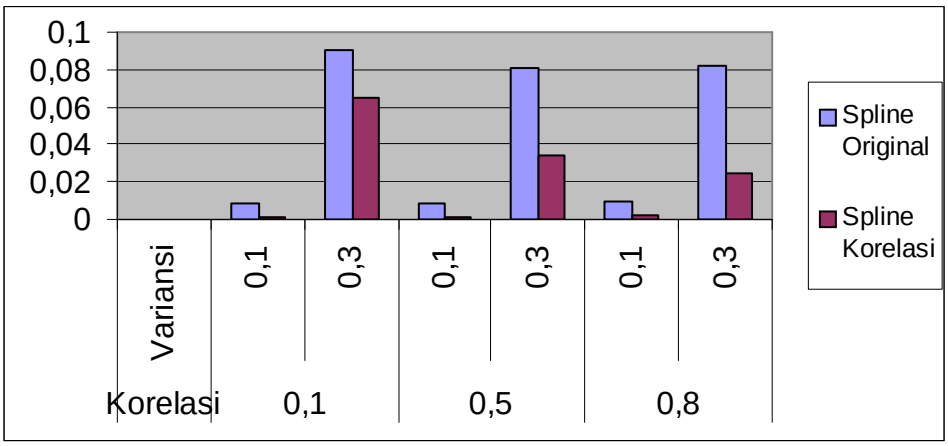
dengan $\beta_1, \ldots, \beta_{p+3}$ adalah parameter model dan $k_1, \ldots, k_p$ adalah titik-titik knot.

Pada simulasi dengan n = 50, terdapat enam percobaan yang berbeda, yang masing-masing akan dicobakan pada dua ukuran variansi dan tiga ukuran korelasi. Adapun *setting* pertama dari simulasi dengan n = 50 adalah dengan memasang korelasi sebesar 0,1 dan variansi sebesar 0.1. Pada semua replikasi, seluruh nilai MSE spline error berkorelasi lebih kecil daripada MSE spline original. Ini menunjukkan bahwa spline error berkorelasi memiliki kualitas yang lebih baik daripada spline original. Rata-rata semua nilai MSE pada simulasi untuk n = 50 dengan berbagai korelasi dan variansi diberikan pada Tabel 1. Terlihat dari Tabel 1, dari sebanyak duapuluh lima replikasi yang dilakukan, diperoleh rata-rata MSE spline original sebesar 0.0087 yang lebih besar dari rata-rata spline dengan error berkorelasi 0.0009. Gambar 1 berikut adalah histogram rata-rata MSE untuk n = 50.

Gambar 1 memperlihatkan, dengan meningkatnya variansi menyebabkan MSE untuk kedua model spline semakin besar dan jika koefisien korelasi meningkat, maka selisih MSE kedua model spline semakin besar. Ini menunjukkan bahwa, apabila koefisien korelasi besar maka pendekatan spline dengan error berkorelasi akan meng-

Tabel 1. Rata-rata nilai MSE untuk n = 50 dengan berbagai korelasi dan variansi

ρ	σ^2	**Spline Original**	**Spline dengan Error Berkorelasi**
		MSE	**MSE**
0.1	0.1	0.0087	0.0009
	0.3	0.0899	0.0652
0.5	0.1	0.0084	0.0012
	0.3	0.0804	0.0336
0.8	0.1	0.0097	0.0021
	0.3	0.0817	0.0244

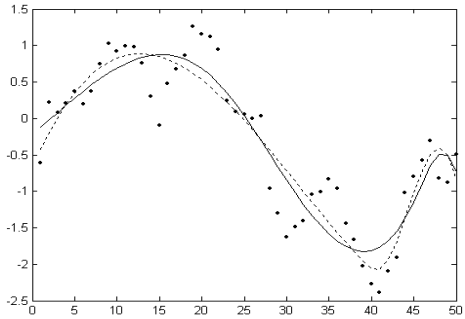


Gambar 1. Histogram rata-rata MSE untuk n = 50

hasilkan model yang lebih baik. Plot data simulasi untuk n = 50, ρ = 0.8 dan $\sigma^2 = 0.3$ beserta kurva estimasi spline dengan error berkorelasi dan spline original diberikan pada Gambar 2.

Berdasarkan hasil uji signifikansi model dan parameter yang meliputi pengujian hipotesis, analisis residual dan analisis variansi untuk spline error berkorelasi dan spline original maka diperoleh model pendekatan

$$\hat{f}(t) = -0.6796 + 0.2686\,t - 0.0132\,t^2 + 0.0001\,t^3 + 0.0135(t-39)_+^3 - 0.0222(t-40)_+^3$$



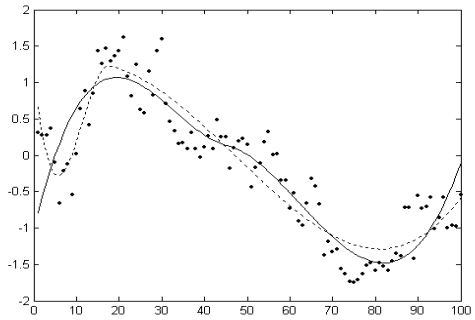
Gambar 2. Plot data simulasi (titik-titik), kurva pendekatan spline error berkorelasi (kurva solid) dan kurva pendekatan spline original (kurva putus-putus) untuk n = 50, ρ = 0.8 dan $\sigma^2 = 0.3$

dengan nilai MSE sebesar 0.1098. Sedangkan model kurva pendekatan spline dengan error berkorelasi adalah:

$$\hat{f}(t) = -0.0002t^3 + 0.007(t-21)_+^3 - 0.0121(t-45)_+^3$$

dengan nilai MSE untuk model ini adalah sebesar 0.0321. Berdasarkan nilai MSE, diperoleh kesimpulan bahwa model spline dengan error berkorelasi memiliki kualitas yang lebih baik.

Pada simulasi selanjutnya, dicobakan pada n = 100 dengan dua macam variansi dan tiga koefisien korelasi. Hasil dari semua simulasi untuk n = 100, ρ = 0.5 dan $\sigma^2 = 0.1$ dapat dilihat pada Tabel 2 (Lampiran). Nilai MSE pada spline dengan error berkorelasi seluruhnya lebih kecil daripada nilai MSE model spline original. Perbedaan rata-rata dua nilai MSE ini sangat kecil, ini disebabkan karena koefisien korelasi yang diambil sangat kecil. Selanjutnya dipasang variansi 0.3. Pemasangan variansi ini menyebabkan membesarnya nilai MSE dari kedua model spline. Percobaan terakhir dari simulasi dengan n = 100 adalah dengan memasang korelasi = 0.8 dan variansi = 0.3. Plot hasil simulasi pada percobaan ini dapat dilihat pada Gambar 3.



Gambar 3. Plot data simulasi untuk n =100, ρ = 0.8 dan σ^2 = 0.3

Data asal (titik-titik), estimasi spline dengan error berkorelasi (kurva solid), estimasi spline original(kurva titk-titik).

Berdasarkan hasil uji signifikansi model dan koefisien regresi untuk spline original diperoleh model pendekatan:

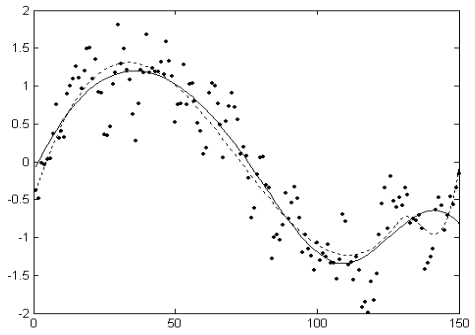
$$\hat{f}(t) = 1.1647 - 0.5687\,t + 0.06641\,t^2 - 0.0019\,t^3 + 0.0120(t-17)_+^3 - 0.0101(t-18)_+^3$$

dengan nilai MSE untuk model ini adalah 0.2627.

Sedangkan untuk spline dengan error berkorelasi diperoleh model pendekatan:

$$\hat{f}(t) = 01603 + 0.0254\,t - 0.0011\,t^2 + 0.0010\,t^3 - 0.0002(t-46)_+^3 + 0.0002(t-44)_+^3$$

dengan nilai MSE sebesar 0.1210. Terlihat dari nilai MSE, bahwa model pendekatan spline dengan error berkorelasi lebih baik daripada model spline original.



Gambar 4. Plot data simulasi (titik-titik), Estimasi dengan spline original (kurva putus-putus), estimasi dengan spline terbobot (kurva solid)

Pada bagian simulasi berikutnya dicobakan pada n = 150 dengan dua macam variansi dan tiga macam korelasi. Selanjutnya akan ditampilkan uji signifikansi model dan parameter untuk n= 150, $\rho = 0.8$ dan $\sigma^2 = 0.3$. Plot data simulasi beserta kurva pendekatan spline dengan error berkorelasi dan spline original bisa diberikan pada Gambar 4.

Berdasarkan hasil uji signifikansi model parameter, diberikan model pendekatan kurva regresi spline original adalah:

$$\hat{f}(t) = -0.4950 + 0.1204t - 0.0023t^2 + 0.001t^3 + 0.0047(t-129)_+^3 - 0.0042(t-128)_+^3 ,$$

dengan MSE sebesar 0.2213. Sedangkan model pendekatan kurva regresi dengan spline error berkorelasi adalah:

$$f(t) = -0.1452 + 0.0804t - 0.0013t^2 + 0.0182t^3 + 0.0567(t-74)_+^3 - 0.0883(t-110)_+^3 .$$

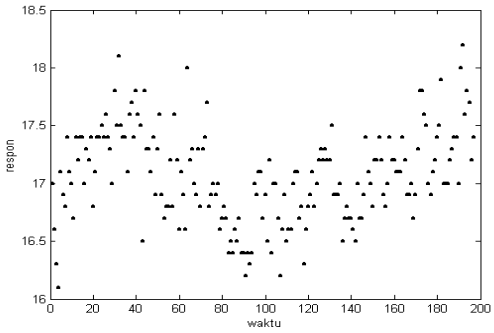
dengan nilai MSE untuk model ini adalah 0.1331.

Dari hasil keseluruhan simulasi diperoleh kesimpulan bahwa model spline dengan error berkorelasi lebih baik daripada model spline original untuk kasus data yang memiliki korelasi antar error, karena semua nilai MSE untuk model spline dengan error berkorelasi lebih kecil daripada spline original.

5. APLIKASI

Pada bagian ini, diberikan penerapan model spline dengan error berkorelasi pada konsentrasi dari suatu proses kimia. Data ini memuat 197 pengukuran yang dicatat setiap dua jam (Box and Jenkins, 1994). Permasalahan yang muncul adalah bagaimana menentukan estimasi untuk model data tersebut. Untuk mengestimasi kurva regresi pada data konsentrasi kimia digunakan *cubic splin*e. Estimasi kurva regresi pada data ini digunakan *cubic spline*.

[3] dan [12] telah menggunakan data ini untuk mengestimasi parameter penghalus dan parameter korelasi. Dari hasil yang diperoleh oleh [12] didapatkan error yang mengikuti model AR(1) dengan nilai parameter korelasi sebesar 0.305 dan variansi sebesar 0.098. Pada penelitian ini menggunakan hasil estimasi parameter yang diperoleh oleh [12].



Gambar 5. Plot data konsentrasi kimia yang dicatat setiap dua jam

Pertama digunakan pendekatan spline original tanpa mempertimbangkan adanya korelasi antar error. Pemilihan titik knot optimal didasarkan pada kriteria GML minimum. Dengan metode GML diperoleh titik knot optimal pada t = 125 dan t = 1 dengan nilai GML minimum 6.9143. Hasil uji signifikansi model dan koefisien regresi memberikan model kurva pendekatan untuk spline original adalah

$$\hat{f}(t) = 16.2171 + 2.1876t - 2.1649t^2 + 0.7215(t-125)_+^3 - 0.7215(t-1)_+^3$$

dengan nilai MSE sebesar 0.1093**.**

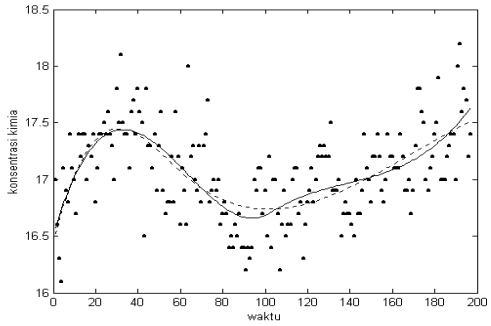
Selanjutnya, karena errornya saling berkorelasi dan mengikuti model AR(1),

maka dilakukan pendekatan spline dengan menggunakan bobot (Wang,1998). Matriks pembobot yang digunakan adalah matriks kovarian dari AR(1). Berdasarkan hasil uji signifikansi model dan koefisien regresi diperoleh model pendekatan spline dengan error berkorelasi, yaitu

$$f(t) = 15.6638 + 2.3394\,t - 2.2596\,t^2 + 0.7526\,t^3 + 0.03(t-57)_+^3 - 0.7435(t-1)_+^3,$$

dengan nilai MSE sebesar 0.0125. Berdasarkan tingkat signifikansi 5% diperoleh kesimpulan bahwa model spline dengan menggunakan bobot lebih baik dibandingkan spline original jika ditinjau dari nilai MSE. Plot data konsentrasi kimia beserta model pendekatan spline original dan spline dengan error berkorelasi diberikan pada Gambar 6.

Dari plot tersebut (gambar 6), terlihat bahwa model pendekatan spline dengan error berkorelasi lebih sesuai untuk mengestimasi kurva regresi pada data ini. Ini juga terbukti dengan lebih kecilnya nilai MSE spline dengan error berkorelasi daripada MSE spline original.



Gambar 6. Plot data konsentrasi kimia, kurva estimasi model spline dengan error berkorelasi (kurva solid), kurva estimasi spline original (kurva putus-putus)

6. DAFTAR PUSTAKA

[1] Altman, N.S. (1990), *Kernel Smoothing of Data With Correlated Errors*, Journal of the American Statistical Association, **85**: 749-759.

[2] Budiantara, I.N., dan Subanar (1999), *Estimator Spline Terbobot dalam Regresi Semiparametrik*, Majalah Ilmu Pengetahuan dan Teknologi, **10**: 103-109.

[3] Diggle, P.J., and Huthcinson, M.F. (1989), *On Spline Smoothing With Autocorrelated Errors*, The Australian Journal of Statistics, **16**: 113-119.

[4] Eubank, R. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

[5] Hart, J.D. (1991), *Kernel Regression Estimation with Tme Series Errors*, Journal of the Royal Statistical Society, **53**(B): 173-187.

[6] Hart, J.D. (1994), *Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-Validation*, Journal of the Royal Statistical Society, **56**(B): 173-187.

[7] Herrmann, E., Gasser, T., and Kneip, A. (1992), *Choice of Bandwith for Kernel Regression When Residuals are Correlated*,Biometrics, **79**: 783-795.

[8] Hurvich, C. M., and Zeger, S.L. (1990), *A Frequency Domain Criterion for Regression With Autocorrelated Errors*, Journal of the American Statistical Association, **85**: 705-714.

[9] Kohn, R., Ansley, C.F., and Wong, C. (1992), *Nonparametric Spline Regression With Autoregressive Moving Average Errors*, Biometrics, **79**: 335-346.

[10] Wahba, G., (1985), "*A Comparison of GCV dan GML for Choosing the Smoothing Parameters in the Generalized Spline Smoothing Problem*", The Annals of Statistics, **4**: 1378-1402.

[11] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM, **59**.

[12] Wang, Y. (1998), *Smoothing Spline Models with Correlated Random Errors*, Journal of The Royal Statistical Society, seri B.

LAMPIRAN

Tabel 2. Hasil simulasi untuk $n = 100$, $\rho = 0.5$ dan $\sigma^2 = 0.1$

Replikasi	Original Spline				Correlated Error Spline			
	Knot1	**Knot2**	**GML**	**MSE**	**Knot1**	**Knot2**	**GML**	**MSE**
1	32	11	0.4309	0.0098	11	30	0.0645	0.0015
2	37	12	0.3457	0.0079	37	12	0.0435	0.001
3	5	41	0.4295	0.0098	34	5	0.0614	0.0014
4	17	41	0.3197	0.0073	16	41	0.0358	0.0008
5	37	16	0.4375	0.0099	36	16	0.0509	0.0012
6	35	34	0.289	0.0066	34	35	0.0381	0.0009
7	14	40	0.504	0.0115	14	41	0.0615	0.0014
8	18	34	0.4753	0.0108	18	34	0.0582	0.0013
9	40	12	0.5142	0.0117	39	11	0.0631	0.0014
10	32	12	0.3829	0.0087	31	11	0.0414	0.0009
11	36	37	0.3617	0.0082	36	37	0.0554	0.0013
12	7	29	0.4984	0.0113	29	7	0.06	0.0014
13	28	27	0.5078	0.0115	26	28	0.0674	0.0015
14	16	17	0.3476	0.0079	17	15	0.0481	0.0011
15	38	10	0.4491	0.0102	9	38	0.0691	0.0016
16	9	45	0.3439	0.0078	9	45	0.0477	0.0011
17	34	10	0.5342	0.0121	9	34	0.0856	0.0019
18	27	28	0.4223	0.0096	29	27	0.0378	0.0009
19	15	34	0.4446	0.0101	35	15	0.0575	0.0013
20	35	33	0.5584	0.0127	13	40	0.087	0.002
21	18	19	0.3592	0.0082	18	20	0.043	0.001
22	14	29	0.2744	0.0062	14	28	0.0341	0.0008
23	32	33	0.3484	0.0079	32	33	0.0498	0.0011
24	25	31	0.442	0.01	25	32	0.0544	0.0012
25	32	15	0.3747	0.0085	31	15	0.0473	0.0011