J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 4, No.1, 1-11, Juli 2007
ISBN:1858-1382

Estimasi Spline Dan Mars Menggunakan Kuadrat Terkecil

Bambang Widjanarko Otok[*], M. Sjahid Akbar[**], Raupong[***]

Abstrak

Tujuan penelitian ini adalah mendapatkan estimasi dari estimator spline dan MARS, fungsi optimum menggunakan kurva *truncated spline* dan Multivariate Adaptive Regression Splines (MARS) melalui data simulasi dari fungsi tertentu dengan berbagai ukuran sampel. Selanjutnya pada data simulasi digunakan bootstrap untuk melihat kestabilan varians dari kedua kurva tersebut. Kriteria MSE untuk melihat fungsi yang optimum.
Hasil penelitian menunjukkan bahwa taksiran dari estimator:

Truncated Spline : $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = [X(\boldsymbol{\lambda})^T X(\boldsymbol{\lambda})]^{-1} X(\boldsymbol{\lambda})^T \tilde{y}$

MARS : $(B'B)a = (B'Y)$ dengan $B = [1, \prod_{i=1}^{K_m} \{(x_{v,(k,m)} - t_{km})\}_1^K]$

dimana $\hat{a}$ diperoleh dari

$$\frac{\partial ASR}{\partial \hat{a}} = 0$$

Dengan

$$ASR(a) = \frac{1}{N}\left[\sum_{k=1}^{N}(y_k - \bar{y})^2 - \sum_{i=1}^{M+1} a_i (c_i + \delta D_{ii} a_i)\right]$$

Pada kajian empirik, basis fungsi optimal piecewise linear berdasarkan MSE terjadi pada varians yang kecil untuk sembarang pengamatan (n). Kurva truncated spline pada piecewise linear lebih baik dibanding dengan MARS, karena hanya melibatkan data dimensi rendah. Sedangkan pembootstrapan pada kurva spline maupun MARS memberikan hasil yang lebih baik dibanding kurva asli.

Kata Kunci: *Bootstrap, MARS, MSE, regresi nonparametrik, Truncated Spline,*

1. Pendahuluan

Analisis regresi memperlihatkan hubungan dan pengaruh variabel prediktor terhadap variabel respon. Misalnya y adalah variabel respon dan t adalah variabel prediktor, untuk n buah pengamatan, secara umum hubungan antara y dan t dapat ditulis sebagai berikut :

$$y_i = g(t_i) + \varepsilon_i; \quad i = 1,2...n \qquad (1)$$

dengan ε adalah sesatan random dan $g(t_i)$ merupakan kurva regresi.

[*] *Mahasiswa S3 Matematika, UGM Yogyakarta,* Dosen Jurusan Statistika, ITS Surabaya, otok_bw@yahoo.com
[**] Dosen Jurusan Statistika, ITS Surabaya, sjahid_alakbar@yahoo.com
[***] *Staf pengajar pada Jurusan Matematika F.MIPA Universitas Hasanuddin Makassar* raupong@yahoo.com

Untuk mengestimasi kurva atau parameter regresi, ada dua pendekatan yaitu parametrik dan nonparametrik. Pendekatan parametrik digunakan jika ada informasi sebelumnya tentang bentuk kurva, yang diperoleh berdasarkan teori atau pengalaman masa lalu, sehingga dapat dikatakan bahwa mengestimasi kurva ekuivalen dengan mengestimasi parameter, dimana hasil estimasi mengikuti model tertentu. Sedang pendekatan nonparametrik digunakan jika tidak ada informasi tentang bentuk kurva $g(t_i)$, tidak tergantung pada asumsi bentuk kurva tertentu, sehingga memberikan fleksibilitas yang lebih besar (Eubank , 1988; Hardle, 1990; Emond and Steven, 1997)

Wahba (1990) menunjukkan bahwa spline memiliki sifat-sifat statistik yang berguna untuk menganalisis hubungan dalam regresi. Spline dalam regresi nonparametrik terus berkembang sampai pada model adaptive (Bilier and Fahrmeir, 2001) dan multivariate respon (Holmes dan Mallick, 2003). Untuk mengestimasi basis fungsi spline telah dikembangkan beberapa metode seperti *monotonicity* (He dan Shi, 1998) dan *penalised* (Hall dan Opsomer, 2005). Friedman (1991) memodifikasi keterbatasan yang dimiliki metode *recursive partitioning regression*, yang dikenal dengan MARS.

Spline mempunyai kelemahan pada saat orde spline tinggi, knots yang banyak dan knot yang terlalu dekat akan membentuk matrik dalam perhitungan yang hampir singular, sehingga persamaan normal tidak dapat diselesaikan. Basis lain yang dapat mengatasi kelemahan ini adalah basis B-Spline.

Bootstrap adalah metode yang berbasis komputer yang digunakan untuk pengukuran akurasi dari taksiran statistik (pendugaan besaran statistik dan selang kepercayaan). Bootstrap merupakan salah satu metode intensif perhitungan yang terbaru untuk memperbaiki estimasi/klasifikasi yang tidak stabil. Secara ekstrim berguna untuk masalah rangkaian data berdimensi tinggi.

Spline dalam regresi nonparametrik, memuat parameter penghalus yang dapat dipilih dengan berbagai metode. Penelitian ini melakukan kajian teori, simulasi untuk membandingkan nilai MSE pada kurva spline *truncated* dan MARS sebelum dan sesudah dilakukan pembootstrapan untuk menentukan fungsi yang optimal dan terapan.

2. *Spline Truncated* dan MARS

Misalkan diberikan model regresi $y_i = g(t_i) + \varepsilon_i$; $i = 1,2...n$, dengan ε_i merupakan residual dan $g(t_i)$ kurva regresi. Pada persamaan (1) jika tidak ada informasi sebelumnya mengenai bentuk kurva regresi $g(t_i)$ maka digunakan regresi nonparametrik, dengan g kurva regresi yang tidak diketahui (akan diestimasi), ε_i residual random independen dengan mean nol dan variansi σ^2. Dalam regresi nonparameterik hanya diasumsikan bentuk kualitatif dari g yaitu kontinu dan diferensiabel.

Pada fungsi smooth, apabila secara geometrik gradiennya berubah tidak terlalu cepat, maka kita dapat menggunakan suatu titik disekitar titik tersebut sebagai estimasi titik knotnya. Berdasarkan konsep di atas, misalkan g fungsi smooth dalam arti g termuat dalam suatu ruang fungsi, khususnya ruang Sobolev $W_2^d[a,b] = \{f \mid f, f', f'', ..., f^{(d-1)}\}$ kontinu absolut pada [a,b], $f^d \in L_2[a,b]$, dengan f^d adalah fungsi turunan ke-d. $L_2[a,b]$ adalah himpunan fungsi yang kuadratnya terintegral pada interval [a,b], atau

$$L_2[a,b] = \{f \mid \int_a^b [f^d(t)]^2 dt < \infty\} \qquad \text{(Schumaker, 1981)}$$

Apabila digunakan pendekatan kurva spline *truncated*, maka kurva regresi g dapat ditulis menjadi : (Eubank, 1988)

$$g(t) = \sum_{i=1}^{m} \alpha_i t^{i-1} + \sum_{j=1}^{K} \beta_j (t-u_j)_+^{m-1} \quad (2)$$

dimana $u_j, j = 1,2,...,K$ dengan $u_1 < u_2 < ... < u_K$ adalah knot dan $m \in N_0$ (integer non negatif). Nilai m menunjukkan derajat spline *truncated.*

Recursive Partitioning Regression (RPR) merupakan pendekatan dari fungsi *f(t)* yang tidak diketahui dengan:

$$\hat{f}(t) = \sum_{j=1}^{S} c_j(t) B_j(t) \quad (3)$$

dimana, $B_j(t) = I[t \in R_j]$, I[.] menunjukkan fungsi indikator yang mempunyai nilai 1 (satu) jika pernyataan benar ($t \in R_j$) dan 0 (nol) jika salah, $c_j(t)$ merupakan koefisien (konstanta) yang ditentukan dalam subregion.

Penentuan knots pada regresi dummy dilakukan secara manual, karena memiliki dimensi data yang rendah dan hal ini tidak akan mengalami kesulitan, sedangkan untuk data yang berdimensi tinggi terdapat kesulitan. Untuk mengatasi hal tersebut digunakan model *Recursive Partition Regression* karena penentuan knots tergantung (otomatis) dari data. Tetapi model ini masih terdapat kelemahan yaitu model yang dihasilkan tidak kontinu pada knots

Pada model MARS selain penentuan knots yang dilakukan secara otomatis dari data, juga menghasilkan model yang kontinu pada knots. Model MARS dapat ditulis sebagai berikut:

$$\hat{f}(t) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} [s_{km}.(t_{v(k,m)} - u_{km})] \quad (4)$$

dimana:

a_0	=	basis fungsi induk
a_m	=	koefisien dari basis fungsi ke-*m*
M	=	maksimum basis fungsi (*nonconstant basis fungsi*)
K_m	=	derajat interaksi
S_{km}	=	nilainya ±1
$t_{v(k,m)}$	=	variabel independen
u_{km}	=	nilai knots dari variabel independen $t_{v(k,m)}$

Penjabaran dari Persamaan (4) dapat disajikan sebagai berikut:

$$\begin{aligned}\hat{f}(t) = a_0 &+ \sum_{m=1}^{M} a_m [s_{1m}.(t_{v(1,m)} - u_{1m})] \\ &+ \sum_{m=1}^{M} a_m [s_{1m}.(t_{v(1,m)} - u_{1m})][s_{2m}.(t_{v(2,m)} - u_{2m})] \\ &+ \sum_{m=1}^{M} a_m [s_{1m}.(t_{v(1,m)} - u_{1m})][s_{2m}.(t_{v(2,m)} - u_{2m})][s_{3m}.(t_{v(3,m)} - u_{3m})] \\ &+ ...\end{aligned} \quad (5)$$

dan secara umum Persamaan (4) dapat dituliskan sebagai berikut:

$$\hat{f}(t) = a_0 + \sum_{K_m=1} f_i(t_i) + \sum_{K_m=2} f_{ij}(t_i, t_j) + \sum_{K_m=3} f_{ijk}(t_i, t_j, t_k) + ... \quad (6)$$

Persamaan (6) menyatakan bahwa penjumlahan pertama meliputi semua basis fungsi untuk satu variabel, penjumlahan kedua meliputi semua basis fungsi untuk interaksi antara dua variabel, penjumlahan ketiga meliputi semua basis fungsi untuk interaksi antara tiga variabel dan seterusnya. Misalkan $V(m) = \{v(k,m)\}_1^{K_m}$ adalah himpunan dari variabel yang dihubungkan dengan basis fungsi B_m ke-m, maka setiap penjumlahan pertama pada Persamaan (6) dapat dinyatakan sebagai:

$$f_i(t_i) = \sum_{\substack{K_m=1 \\ i \in V(m)}} a_m B_m(t_i) \tag{7}$$

Jadi *fi(t$_i$)* merupakan penjumlahan semua basis fungsi untuk satu variabel x_i dan merupakan spline dengan derajat q=1 yang merepresentasikan fungsi univariat. Setiap fungsi bivariat pada Persamaan (5) dapat ditulis sebagai:

$$f_{ij}(t_i,t_j) = \sum_{\substack{Km=2 \\ (i,j) \in V(m)}} a_m B_m(t_i,t_j) \tag{8}$$

Persamaan tersebut merepresentasikan penjumlahan semua basis fungsi dua variabel t_i dan t_j. Penambahan ini untuk menghubungkan kontribusi univariat, yang dituliskan sebagai berikut:

$$f_{ij}^*(t_i,t_j) = f_i(t_i) + f_j(t_j) + f_{ij}(t_i,t_j) \tag{9}$$

Fungsi trivariat pada penjumlahan yang ketiga diperoleh dengan menjumlahkan semua basis fungsi untuk tiga variabel, yang dituliskan sebagai berikut:

$$f_{ijk}(t_i,t_j,t_k) = \sum_{\substack{Km=3 \\ (i,j,k) \in V(m)}} a_m B_m(t_i,t_j,t_k) \tag{10}$$

Penambahan fungsi univariate dan bivariate mempunyai kontribusi dalam bentuk:

$$\begin{aligned} f_{ijk}^*(t_i,t_j,t_k) = {} & f_i(t_i) + f_j(t_j) + f_k(t_k) + f_{ij}(t_i,t_j) + f_{ik}(t_i,t_k) \\ & + f_{jk}(t_j,t_k) + f_{ijk}(t_i,t_j,t_k) \end{aligned} \tag{11}$$

Pemilihan model MARS dapat dilakukan dengan metode stepwise (forward dan backward). Forward untuk mendapatkan fungsi dengan jumlah basis fungsi maksimum. Kriteria pemilihan basis fungsi pada forward adalah dengan meminimumkan *Average Sum Square Residual* (ASR) atau *Mean Square Error* (MSE). Backward untuk memilih basis fungsi yang dihasilkan dari forward dengan meminimumkan nilai Generalized Cross-Validation (GCV). (Friedman and Silverman, 1989).

3. Sifat-Sifat Asimtotik Estimator

Misalkan $Y_1,...,Y_n$ adalah sampel random dengan fungsi kepadatan probabilitas $f(.,\beta),\ \beta \in \Omega$ maka sifat konsistensi barisan estimator didefinisikan sebagai berikut:

Definisi 1: (Rohatgi,1975)

Barisan estimator $T_n = T_n(\mathbf{Y})$ disebut konsisten dalam probabilitas untuk β bila untuk setiap $\varepsilon > 0$ dan setiap $\beta \in \Omega$

$$\lim_{n\to\infty} P(|T_n - \beta| < \varepsilon) = 1 \text{ atau } T_n \xrightarrow{p} \beta \text{ bila } n \to \infty$$

Definisi 2: (Rohatgi,1975)

Jika $\{X_n\}$ barisan variabel random, maka $\{X_n\}$ dikatakan konvergen ke X dalam probabilitas, ditulis dengan $X_n \xrightarrow{p} X$ jika untuk setiap $\varepsilon > 0, P\{|X_n - X| > \varepsilon\} \to 0$ untuk n $\to \infty$.

Definisi 3: (Rohatgi,1975)

Barisan $\{X_n\}$ dikatakan konvergen hampir pasti ke X dan dinyatakan sebagai $X_n \xrightarrow{a.s.} X$ jika $P\{\lim_{n\to\infty} X_n = X\} = 1$ untuk n $\to \infty$.

Definisi 4: (Rohatgi,1975)

Barisan $\{X_n\}$ dinamakan konvergen dalam distribusi ke X dan dinyatakan sebagai $X_n \xrightarrow{d} X$ jika $F_{Xn}(x) \to F_X(x), n \to \infty$ untuk semua x, dimana $F_X(x)$ kontinu.

Dari definisi konvergensi diperoleh hubungan kekonvergenan sebagai berikut:

- Jika suatu barisan konvergen hampir pasti, maka barisan tersebut juga konvergen dalam probabilitas.
- Jika suatu barisan konvergen dalam probabilitas maka barisan tersebut juga konvergen dalam distribusi. (Fergusson,1996)

Teorema 1: Teorema Limit Sentral Multivariat (Pranab K.sen,1993)

Jika $\{X_i\}$ merupakan vektor random i.i.d di R^p dengan mean μ_i dan matrik kovarian $\sum_i$ maka

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \to N^{d}(0,\Sigma) \quad \text{dengan } \sum = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\Sigma_i .$$

Definisi 5: (Hardle,1990)

Barisan $\{a_i\}$ dan $\{b_i\}$ dinyatakan sebagai $a_n = O(b_n)$ jika terdapat bilangan real *M* sedemikian hingga $\lim_{n\to\infty} \frac{a_n}{b_n} = M$.

Definisi 6: (Hardle,1990)

Barisan $\{a_i\}$ dan $\{b_i\}$ dinyatakan sebagai $a_n = o(b_n)$ jika $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

Selanjutnya 'O' mempunyai sifat-sifat sebagai berikut: (Rohattgi, 1975)

- Jika $a_{n1} = O(b_{n1}),\ a_{n2} = O(b_{n2})$ maka $a_{n1} + a_{n2} = O(b_{n1} + b_{n2})$.
- Jika $\alpha > 0$ adalah konstanta, $a_n = O(\alpha b_n)$ maka $a_n = O(b_n)$.
- Jika $a_{n1} = O(b_{n1}),\ a_{n2} = O(b_{n2})$ maka $a_{n1}a_{n2} = O(b_{n1}b_{n2})$.

Definisi 7. (Hardle,1990)

Barisan $\{a_i\}$ dan $\{b_i\}$ dinyatakan sebagai $a_n = O_p(b_n)$ jika terdapat bilangan real M dan N sedemikian hingga $P\left\{\left|\frac{a_n}{b_n}\right| > M\right\} \le \varepsilon,\ \forall n > N$.

Definisi 8. (Hardle,1990)

Barisan $\{a_i\}$ dan $\{b_i\}$ dinyatakan sebagai $a_n = o_p(b_n)$ jika $\forall \varepsilon > 0$ sedemikian hingga $\lim_{n\to\infty} P\left\{\left|\frac{a_n}{b_n}\right| \geq \varepsilon\right\} = 0$.

Teorema 2:

Misalkan $Z \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ dan **A** suatu matrik, maka vektor $U = \mathbf{A}Z \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}')$.

Bukti:

$$\begin{aligned}
\mathrm{E}(U) &= \mathrm{E}(\mathbf{A}Z) = \mathbf{A}\mathrm{E}(Z) = \mathbf{A}\mu \\
Var(U) &= \mathrm{E}[(U - \mathrm{E}(U))(U - \mathrm{E}(U))'] \\
&= \mathrm{E}[(\mathbf{A}Z - \mathbf{A}\mu))(\mathbf{A}Z - \mathbf{A}\mu)'] \\
&= \mathrm{E}[(\mathbf{A}(Z - \mu))(Z - \mu)'\mathbf{A}'] \\
&= \mathbf{A}\,\mathrm{E}[((Z - \mu))(Z - \mu)']\,\mathbf{A}' \\
&= \mathbf{A}\,\mathrm{Var}(Z)\,\mathbf{A}' \\
&= \mathbf{A}\,\Omega\,\mathbf{A}'
\end{aligned}$$

Karena U fungsi linier dari $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ dan juga persamaan (2.1.16) dan (2.1.17) maka terbukti $\boldsymbol{U} = \mathbf{A}\boldsymbol{Z} \sim \boldsymbol{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}')$. ■

Definisi 9: (Royden,1968)

Suatu fungsi dikatakan kontinu di titik *x* jika untuk setiap $\varepsilon > 0$ ada $\delta > 0$, sehingga $|f(x) - f(y)| < \varepsilon$ dengan $|x - y| < \delta$

Definisi 10. (Bartle,1976)

Suatu fungsi $f(.)$ dikatakan memenuhi kondisi Lipschitz order 1 jika ada konstanata A > 0 sehingga $|f(t_i) - f(t_j)| \leq A\,|t_i - t_j|^1$.

4. Hasil Penelitian

Penelitian ini secara garis besar terbagi dalam dua kajian, yaitu kajian teori berkaitan dengan kajian parameter pada estimator spline dan MARS, kajian data empiris melalui data simulasi dengan fungsi tertentu, dan kajian terapan berkaitan dengan implementasi hasil kajian teori.

4.1. Estimasi Spline dan MARS

Spline least squares merupakan generalisasi regresi polinomial (Eubank,1988), dimana estimasi kurva regresi *f* diperoleh melalui fungsi yang didefinisikan sebagai berikut: $g(t) = \sum_{i=1}^{m} \alpha_i t^{i-1} + \sum_{j=1}^{K} \beta_j (t - u_j)_+^{m-1}$, dalam persamaan ini, *g* merupakan spline orde-m dengan knots $\xi_1, \ldots, \xi_k$. Himpunan dari semua fungsi ini, $S^m(\xi_1, \ldots, \xi_k))$ adalah suatu ruang vektor berdimensi m+k yang terdiri dari potongan polinomial orde-m yang memiliki m-2 turunan yang kontinyu dan diskontinyu pada turunan ke-(m-1) di titik ξ_j. Dengan memilih $\lambda = \{\xi_1, \ldots, \xi_k\}$,

maka f dapat diestimasikan dengan mengestimasi koefisien-koefisien dari persamaan (2.1.34). Salah satu metode untuk menyelesaikan hal tersebut adalah dengan menggunakan least-squares. Didefinisikan:

$$\begin{aligned} x_1(t) &= 1, \\ x_2(t) &= t, \\ &\vdots \\ x_m(t) &= t^{m-1}, \\ x_{m+1}(t) &= (t-\xi_1)_+^{m-1}, \\ &\vdots \\ x_{m+k}(t) &= (t-\xi_k)_+^{m-1} \end{aligned} \tag{12}$$

dengan $\tilde{\boldsymbol{\beta}} = (\theta_1, \cdots, \theta_m, \delta_1, \cdots, \delta_k)^T$ (13)

Estimator spline least-square dari f adalah:

$$f_{\boldsymbol{\lambda}} = \sum_{j=1}^{m+k} \boldsymbol{\beta}_{\boldsymbol{\lambda} j} x_j \tag{14}$$

dimana $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\boldsymbol{\beta}_{\boldsymbol{\lambda}1}, \ldots, \boldsymbol{\beta}_{\boldsymbol{\lambda},m+k})^T$ adalah suatu minimizer dari

$$MSE(\boldsymbol{\beta};\boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m+k} \boldsymbol{\beta}_j x_j(t_i))^2 \tag{15}$$

terhadap $\tilde{\boldsymbol{\beta}}$. Lebih jelasnya jika didefinisikan:

$$X(\boldsymbol{\lambda}) = \{x_j(t_i)\}, i=1,\ldots,n; j=1,\ldots,m+k. \tag{16}$$

maka $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ adalah suatu penyelesaian untuk persamaan normal:

$$X(\boldsymbol{\lambda})^T X(\boldsymbol{\lambda}) \tilde{\boldsymbol{\beta}} = X(\boldsymbol{\lambda})^T y, \tag{17}$$

dimana $y = (y_1, \ldots, y_n)^T$. Jika X(λ) mempunyai rank m +k, maka:

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = [X(\boldsymbol{\lambda})^T X(\boldsymbol{\lambda})]^{-1} X(\boldsymbol{\lambda})^T \tilde{y} \tag{18}$$

Dari Persamaan (14) dan (18) terlihat bahwa dengan $\boldsymbol{\lambda} = \{\xi_1, \ldots, \xi_k\}$ }, $f_{\boldsymbol{\lambda}}$ adalah estimator linier dari f. Model MARS dapat dinyatakan dalam persamaan berikut:

$$Y = \mathbf{Ba} + \boldsymbol{\varepsilon} \tag{19}$$

dengan,

Y = variabel respon,

$\mathbf{a}$ = koefisien dari *basis function*

$\boldsymbol{\varepsilon}$ = error.

$$\mathbf{B} = [1, \prod_{k=1}^{K_m} \{(x_{v,(k,m)} - t_{km})\}_1^K]$$

= *basis function*,

Estimasi parameter dengan metode kuadrat terkecil, yang pada prinsipnya meminumumkan jumlah kuadrat error, dinyatakan sebagai berikut:

$$\begin{aligned}\varepsilon'\varepsilon &= (Y-Ba)'(Y-Ba)\\ &= (Y'-a'B')(Y-Ba)\\ &= Y'Y-a'B'Y-Y'Ba+a'B'Ba\\ &= Y'Y-2a'B'Y+a'B'Ba\end{aligned} \tag{20}$$

Persamaan normal diperoleh dengan menurunkan (20) secara parsial terhadap *a* dengan hasil sebagai berikut:

$$(B'B)a = (B'Y) \tag{21}$$

Jumlah parameter MARS selain basis fungsi induk, dapat diperoleh dengan mensubstitusikan Persamaan (21) ke dalam Persamaan (19) sebagai berikut:

$$Y = B(B'B)^{-1}(B'Y) \Rightarrow I = B(B'B)^{-1}B' \tag{22}$$

Nilai *Trace* dari matriks *I* adalah jumlah parameter basis fungsi selain konstanta dan jumlah parameter yang diestimasi adalah $C(\tilde{M}) = Trace[B(B'B)^{-1}B'] + 1$.

Penyelesaian Persamaan (21) dengan analisis numeric tidak efisien, sehingga dilakukan dengan *QR decompotition* (Wahba, 1990) dalam (Schott, 1997). Jika basis fungsi dipusatkan ke nilai rata-rata nol maka *B'B* proporsional dengan kovarians matriks basis fungsi berikut:

$$Va = c \tag{23}$$

$$V_{ij} = \sum_{k=1}^{N} B_j(x_k)[B_i(x_k) - \overline{B}_i] \tag{24}$$

$$c_i = \sum_{k=1}^{N} (y_k - \bar{y})B_i(x_k) \tag{25}$$

di mana $\overline{B}_i$ dan $\bar{y}$ adalah rata-rata dari seluruh data. Persamaan (23) diselesaikan untuk setiap lokasi *knot t*, untuk setiap variabel *v*, untuk semua basis fungsi *m* dan untuk semua interaksi *M*. Untuk q=1, formula dengan estimasi least square adalah:

$$(x-t)_+ - (x-u)_+ = \begin{cases} 0 & , x \le t \\ x-t & , t < x < u \\ u-t & , x \ge u \end{cases}$$

$$\begin{aligned} c_{M+1}(t) &= c_{M+1}(u) + \sum_{t \le x_{vk} \le u} (y_k - \bar{y})B_{mk}(x_{vk} - t) + (u-t)\sum_{x_{vk} \ge u} (y_k - \bar{y})B_{mk} \\ V_{i,M+1}(t) &= V_{i,M+1}(u) + \sum_{t \le x_{vk} < u} (B_{ik} - \overline{B}_i)B_{mk}(x_{vk} - t) + (u-t)\sum_{x_{vk} \ge u} (B_{ik} - \overline{B}_i)B_{mk} \\ V_{M+1,M+1}(t) &= V_{M+1,M+1}(u) + \sum_{t \le x_{vk} < u} B_{mk}^2 (x_{vk} - t)^2 + (u-t)\sum_{x_{vk} \ge u} B_{mk}^2 (2x_{vk} - t - u) \\ &\quad + (s^2(u) - s^2(t))/N \end{aligned} \tag{27}$$

dimana,

$s(t) = \sum_{x_{vk} \ge t} B_{mk}(x_{vk} - t)$, B_{ik} dan B_{mk} adalah elemen dari matriks data basis fungsi, x_k adalah elemen dari matriks data asli, dan y_k adalah data respon.

Persamaan normal dari pemodelan MARS dapat diselesaikan dengan menggunakan Cholesky Decompotition (Dongarra, Moler, Bunch and Stewart, 1979), (Scott, 1997), dengan modifikasi berikut,

$$(V + \varepsilon D)a = c \tag{28}$$

dimana *D* adalah matriks diagonal berukuran (M+1)(M+1) dari elemen matriks diagonal matriks *V*. Koefisien dari basis fungsi *a* diperoleh dengan menurunkan formula berikut,

$$ASR(a) = \frac{1}{N}\left[\sum_{k=1}^{N}(y_k - \bar{y})^2 - \sum_{i=1}^{M+1} a_i (c_i + \delta D_{ii} a_i)\right] \quad (29)$$

4.2. Kajian Empiris

Setelah membangun model $y_i = g(t_i) + \varepsilon_i$, maka langkah pertama didekati dengan kurva spline truncated dan MARS. Nilai MSE dengan berbagai variasi pada fungsi, pengamatan (n) secara lengkap tersaji pada Tabel 1 berikut.

Tabel 1. Nilai MSE Pada Spline Truncated dan MARS

Fungsi	N	σ^2	MSE		$MSE_{Bootstrap}$	
			Spline	MARS	Spline	MARS
$5e^{-5t}$	50	0,025	0.00058	0.0013	0.00054	0.00119
		0,5	0.13982	0.1636	0.13007	0.15513
		1	0.50940	0.7686	0.49840	0.74380
	100	0,025	0.00072	0.0011	0.00066	0.00890
		0,5	0.17331	0.2311	0.15684	0.20567
		1	0.67544	0.8576	0.61675	0.81307
	250	0,025	0.00073	0.0009	0.00066	0.00087
		0,5	0.20838	0.2550	0.18718	0.22019
		1	0.77183	0.9650	0.76443	0.94437
$\sin(2\pi t)$	50	0,025	0.00127	0.0011	0.00124	0.00098
		0,5	0.15802	0.1688	0.13578	0.14980
		1	0.45707	0.7298	0.44408	0.67151
	100	0,025	0.00125	0.0011	0.00120	0.00109
		0,5	0.18962	0.1997	0.16563	0.19021
		1	0.63835	0.9620	0.57318	0.92341
	250	0,025	0.00169	0.0009	0.00158	0.00087
		0,5	0.19760	0.2309	0.18147	0.20022
		1	0.78974	1.0030	0.77430	0.98990

Sumber: Data Diolah

Berdasarkan Tabel 1, ternyata nilai MSE dengan banyaknya pengamatan yang semakin besar dan varians σ^2 konstan memberikan hasil yang semakin besar pada fungsi $g(t_i) = 5e^{-5t}$. Sedang pada fungsi $g(t_i) = \sin(2\pi t)$, nilai MSE dengan banyaknya pengamatan yang semakin besar dan varians σ^2 konstan memberikan hasil yang semakin kecil. Secara keseluruhan fungsi optimal terjadi pada varians σ^2 kecil dengan n sembarang. Sedangkan pada MARS, nilai MSE dengan banyaknya pengamatan yang semakin besar dan varians σ^2 konstan memberikan hasil yang semakin kecil pada fungsi $g(t_i) = 5e^{-5t}$. Sedang pada fungsi $g(t_i) = \sin(2\pi t)$, nilai MSE dengan banyaknya pengamatan yang semakin besar pada varians σ^2 = 0,025 memberikan hasil yang semakin kecil, tetapi pada varians σ^2 = 0,5 dan 1 nilai MSE bervariasi. Secara keseluruhan fungsi optimal terjadi pada varians σ^2 kecil dengan n sembarang.

5. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan sebagai berikut:

Truncated Spline: $\mathbf{X}(\lambda)^T \mathbf{X}(\lambda)\boldsymbol{\beta} = \mathbf{X}(\lambda)^T \mathbf{y}$

MARS : $(B'B)a = (B'Y)$ dengan $B = [1, \prod_{i=1}^{K_m} \{(x_{v,(k,m)} - t_{km})\}_1^K]$

dimana $\hat{a}$ diperoleh dari

$$\frac{\partial ASR}{\partial \hat{a}} = 0 \text{ dengan } ASR(a) = \frac{1}{N}\left[\sum_{k=1}^{N}(y_k - \bar{y})^2 - \sum_{i=1}^{M+1} a_i(c_i + \delta D_{ii} a_i)\right].$$

Fungsi optimal terjadi pada varians yang kecil untuk sembarang pengamatan. Nilai MSE pada kurva truncated spline linier lebih kecil dibanding dengan MARS linier pada semua fungsi. Hal ini dapat diartikan bahwa kurva truncated spline lebih baik dibanding dengan MARS, karena masih menggunakan data dimensi rendah. Sementara pada kurva MARS menekankan pada interaksi variabel prediktor. Oleh karena perlu kajian lebih lanjut pada data dimensi tinggi (melibatkan banyak prediktor dan interaksinya).

Daftar Pustaka

[1] C. Bilier, dan L. Fahrmeir,2001, "*Bayesian varying-coefficient models using adaptive regression spline*", *Statistical Modelling.*

[2] I.N. Budiantara, S. Fredi, B.W. Otok dan S. Guritno, 2006, "*Pemodelan BSpline and MARS pada nilai ujian masuk terhadap IPK Mahasiswa Jurusan Disain Komunikasi Visual UK-PETRA Surabaya*", Jurnal TeknikIndustri UK PETRA SURABAYA

[3] B. Efron dan R.J. Tibshirani, 1993, "*An Introduction to the Bootstrap",* Chapman and Hall, Inc.

[4] R.L. Eubank, 1988, "*Spline Smooting and Nonparametric Regression"*, Marcel Deker: New York.

[5] J.H. Friedman dan B.W. Silverman, 1989, "*Flexible parsimony smoothing and additive modeling*", Technometrics, 31, 3 – 39.

[6] J.H. Friedman, 1991, "*Multivariate Adaptive Regression Splines* (With Discussion)", The Annal of Statistics. 19. 1 – 141.

[7] P. Hall dan J.D. Opsomer, 2005, "*Theory for penalised spline regression",* Biometrika, 92,1. page.105

[8] W. Hardle, 1990, "*Applied Nonparametric Regression"*, Cambrige University Press: New York.

[9] X. He dan P. Shi, 1998, "*Monotone B-Spline Smoothing",* Journal of the American Statistical Association; 93,442

[10] C.C. Holmes dan B.K. Mallick B.K, 2003, "*Generalized Nonlinear Modelling With Multivariate Free-Knot Regression Spline*", Journal of the American Statistical Association; 98,462.

[11] B.W. Otok, S. Guritno dan Subanar, 2004, "*Misclassified With Approach Nonparametric*", UNISBA, Bandung.

[12] B.W. Otok, S. Guritno dan Subanar, 2004, "*Analisis Diskriminan dan MARS untuk Klasifikasi Perbankan di Indonesia",* Seminar FKMS3MI ke II, UGM, Yogyakarta.

[13] B.W. Otok, 2005, "*Klasifikasi Perbankan dengan Pendekatan CART dan MARS*", Jurnal Widya Manajemen & Akuntansi, UWM Surabaya.

[14] B.W. Otok, S. Guritno dan Subanar, 2006, "*Optimize Knot and Basis Function at Truncated Spline and Multivariate Adaptive Regression Splines*", ICOMS, UNISBA, Bandung.

[15] L.L. Schumaker, 1981, "*Spline Funtions: Basic Theory",* John Wiley & Sons, Inc: Canada.

[16] J. Shao dan D. Tu, 1995, "*The Jacknife and Bootstrap",* Springer-Verlag New York, Inc.

[17] G. Wahba, 1990, "*Spline Models for Observational Data",* SIAM, CBMS-NSF Regional Conference Series in Applied mathematics, Philadelphia.