PENDEKATAN REGRESI KUADRAT TERKECIL PARSIAL ROBUST MULTIRESPONS DALAM MODEL KALIBRASI

Ismah, Aji Hamim Wigena, Anik Djuraidah

Sekolah Pascasarjana, Program Studi Statistik, Institut Pertanian Bogor.
Ismah.fr@gmail.com

Abstrak

Regresi Kuadrat Terkecil Parsial (RKTP) merupakan sebuah tehnik prediktif yang mampu mengatasi peubah bebas yang berdimensi besar, khususnya ketika terdapat masalah multikolinearitas. Skor dalam RKTP dihitung dengan memaksimalkan kriteria koragam antara peubah *x* dan *y* sehingga dalam teknik ini respons telah dilibatkan dalam analisis sejak awal. SIMPLS merupakan salah satu algoritma RKTP yang dikenalkan oleh De Jong (1993). Karena SIMPLS didasari dari matriks koragam silang empirik antara peubah respon dan peubah bebas dan dalam regresi linier kuadrat terkecil, maka SIMPLS tidak resisten terhadap pengamatan pencilan (*outlier*). Untuk mengatasi masalah pencilan diperlukan suatu metode penduga yang tegar terhadap pencilan yang disebut sebagai metode *robust*. Dua metode RKTP *robust*, RSIMCD dan RSIMPLS, yang dibangun dari matriks koragam robust untuk data berdimensi besar dan regresi linier *robust*, mampu mengatasi pengaruh pengamatan pencilan. Selanjutnya nilai RMSECV *robust* diperoleh untuk membangun model kalibrasi dan RMSEP *robust* digunakan untuk validasi model. Diagnosa plot akan dibuat sebagai visualisasi dan klasifikasi pencilan.

Kata kunci : RKTP, SIMPLS, regresi *robust*, regresi MCD, regresi ROBPCA.

PENDAHULUAN

Regresi adalah suatu teknik statistika yang dapat digunakan untuk menggambarkan hubungan antara satu atau lebih peubah bebas (***X***) dengan satu atau lebih peubah respons (***Y***). Metode kuadrat terkecil dikenal sebagai metode penduga terbaik dalam analisis regresi, namun metode ini sangat peka terhadap adanya penyimpangan asumsi pada data. Jika terjadi pelanggaran asumsi yaitu terdapat kolerasi tinggi di antara peubah bebas (multikolinieritas) maka penduga yang dihasilkan masih tetap tak bias dan konsisten, tetapi tidak efisien sehingga ragam dari koefisien regresi menjadi tidak minimum (*over estimate*). Sedangkan jika banyaknya peubah bebas lebih besar dari pada banyaknya pengamatan, maka struktur matriks peubah bebas menjadi singular. Hal ini mengakibatkan matriks $\mathbf{X^T X}$ tidak mempunyai kebalikan unik (khas). Asumsi penting lainnya yang berkaitan dengan inferensia model adalah asumsi sebaran normal (normalitas). Apabila terdapat pencilan dalam data, maka bentuk sebaran data tidak lagi simetrik tetapi cenderung menjulur ke arah pencilan sehingga melanggar asumsi normalitas.

Regresi Kuadrat Terkecil Parsial (RKTP) merupakan sebuah tehnik prediktif yang mampu mengatasi peubah bebas yang berdimensi besar, khususnya ketika terdapat masalah multikolinearitas. Penerapan RKTP dapat digunakan dalam bidang *chemometry* khususnya pada model kalibrasi. Salah satu algoritma RKTP adalah SIMPLS yang dikenalkan oleh De Jong (1993). Namun, SIMPLS tidak dapat mendeteksi pencilan karena algoritma yang digunakan tidak resisten terhadap pengamatan pencilan. Skor dalam RKTP dihitung berdasarkan matriks koragam silang contoh antara peubah-peubah *x* dan *y* (S_{xy}), dan matriks koragam empirik peubah *x* (S_x) dimana besar kemungkinan terinfeksi oleh pencilan. Untuk mengatasi masalah pencilan diperlukan suatu metode penduga yang tegar terhadap pencilan yang disebut metode *robus*t. Metode *robust* bagi S_x yang cukup populer adalah metode *Minimum Covariance Determinant* (MCD). Penduga MCD

bagi S_x diperoleh dari subhimpunan data berukuran h yang memiliki nilai determinan matriks koragam terkecil. Namun metode tersebut tidak dapat diaplikasikan ketika banyaknya peubah bebas jauh lebih besar dari pada banyaknya pengamatan ($p >> n$), karena matriks koragam $h < p$ selalu singular. Metode *robust* bagi S_x lainnya yang dapat diaplikasikan ketika $p >> n$ yaitu ROBPCA. Metode ROBPCA mengkombinasikan dua pendekatan antara pursuit proyeksi dan penduga koragam *robust* dengan metode MCD. Pursuit proyeksi digunakan untuk mereduksi dimensi kemudian penduga MCD diaplikasikan kedalam ruang data yang telah diperkecil dimensinya.

Selanjutnya hasil skor-skor RKTP *robust* yang terbentuk diregresikan dengan peubah respon menggunakan metode *robust*. Sampai saat ini berbagai metode *robust* untuk analisis regresi terus berkembang dan digunakan dalam berbagai bidang, diantaranya adalah regresi MCD dan regresi ROBPCA (M.Hubert dan K.Vanden Branden, 2003). Kedua regresi *robust* tersebut dapat diaplikasikan ketika dimensi peubah respon lebih dari satu (multirespon).

Dalam tulisan ini akan dibandingkan tingkat ketegaran (resistensi) metode RSIMCD yang merupakan hasil dari metode regresi MCD dan RSIMPLS yang merupakan hasil dari metode regresi ROBPCA sebagai metode RKTP *robust* dengan menggunakan nilai bias dan MSE pada beberapa ukuran sampel dan prosentase pencilan.

Algoritma SIMPLS

Metode SIMPLS mengasumsikan peubah-peubah *x* dan *y* dihubungkan dalam model bilinier seperti berikut ini :

$$\mathbf{x}_i = \bar{\mathbf{x}} + P_{p,k}\tilde{\mathbf{t}}_i + \mathbf{g}_i \qquad (1)$$

$$\mathbf{y}_i = \bar{\mathbf{y}} + A'_{q,k}\tilde{\mathbf{t}}_i + \mathbf{f}_i \qquad (2)$$

Dalam model tersebut, $\bar{\mathbf{x}}$ dan $\bar{\mathbf{y}}$ merupakan rata-rata dari peubah *x* dan *y*. $\tilde{\mathbf{t}}_i$ adalah skor berdimensi *k*, dengan $k << p$. $P_{p,k}$ adalah matriks loading *x*, sedangkan sisaan dalam model ini dinotasikan dengan $\mathbf{g}_i$ dan $\mathbf{f}_i$. Matriks $A_{k,q}$ direpresentasikan sebagai matriks slope model regresi $\mathbf{y}_i$ dalam $\tilde{\mathbf{t}}_i$.

Struktur model bilinier (1) dan (2) mengimplikasikan sebuah algoritma 2 langkah. Setelah data dipusatkan, langkah yang pertama SIMPLS yaitu menentukan komponen *k* $\left(\tilde{T}_{n,k} = \left(\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_n\right)'\right)$ dan yang kedua peubah respon akan diregresikan kedalam komponen *k* yang telah ditentukan.

Langkah Pertama : Menentukan komponen *k*

Yang membedakan PLS dengan regresi komponen utama (RKU) komponen-komponen *k* tidak semata-mata ditentukan berdasarkan peubah *x*. Tetapi, dibentuk sebagai sebuah kombinasi linier peubah *x* yang memiliki nilai koragam maksimum dengan kombinasi linier peubah *y*.

Element-element skor $\tilde{\mathbf{t}}_i$ didefinisikan sebagai kombinasi linier rata-rata data pusat: $\tilde{t}_{ia} = \tilde{\mathbf{x}}'_i\mathbf{r}_a$ atau sama dengan $\tilde{T}_{n,k} = \tilde{X}_{n,p}R_{p,k}$ dengan $R_{p,k} = \left(\mathbf{r}_1, \ldots, \mathbf{r}_k\right)$.

De Jong (1993) menganjurkan untuk menetapkan bobot supaya memaksimumkan koragam vektor-vektor skor $\mathbf{t}_a$ dan $\mathbf{u}_a$ dengan beberapa kendala. Dia juga menentukan empat kondisi yang khusus untuk mengontrol solusi, yaitu :

1. Memaksimumkan koragam : $\mathbf{u}'_a\mathbf{t}_a = \mathbf{q}'_a(\mathbf{Y}'_a\mathbf{X}_a)\mathbf{r}_a = \max!$
2. Menormalisasi bobot $\mathbf{r}_a$: $\mathbf{r}'_a\mathbf{r}_a = 1$
3. Menormalisasi bobot $\mathbf{q}_a$: $\mathbf{q}'_a\mathbf{q}_a = 1$
4. orthogonal skor-skor $\mathbf{t}$: $\mathbf{t}'_b\mathbf{t}_a = 0$, untuk $a > b$

$\tilde{X}_{n,p}$ dan $\tilde{Y}_{n,q}$ merupakan matriks rata-rata data pusat, dengan $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ dan $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}}$. Komponen-komponen *k* adalah sebuah kombinasi linier peubah-peubah *x* yang memaksimumkan koragam dengan kombinasi linier peubah-peubah *y*, dan komponen-komponen *k* mengandung

normalisasi vektor bobot MKT $\mathbf{r}_a$ dan $\mathbf{q}_a$ untuk setiap $a = 1,\ldots,,k$, sebagai vektor yang memaksimumkan koragam antara komponen-komponen *x* dan *y*.

$$\max_{\|\mathbf{r}_a\|=1,\|\mathbf{q}_a\|=1} \operatorname{cov}\left(\tilde{Y}_{n,q}\mathbf{q}_a, \tilde{X}_{n,p}\mathbf{r}_a\right) = \max_{\|\mathbf{r}_a\|=1,\|\mathbf{q}_a\|=1} \mathbf{q}_a' \frac{\tilde{Y}'_{n,q}, \tilde{X}_{n,p}}{n-1}\mathbf{r}_a = \max_{\|\mathbf{r}_a\|=1,\|\mathbf{q}_a\|=1} \mathbf{q}_a' S_{yx}\mathbf{r}_a \tag{3}$$

Dimana $S'_{yx} = S_{xy} = \dfrac{\tilde{X}'_{p,n}\tilde{Y}_{n,q}}{n-1}$ adalah matriks koragam silang empirik antara peubah *x* dan *y*. Maksimisasi mempunyai retriksi tambahan bahwa komponen-komponen $\tilde{T}_a = \tilde{X}\mathbf{r}_a$ tidak berkorelasi (orthogonal),

$$\mathbf{r}_j'\tilde{X}'\tilde{X}\mathbf{r}_a = \tilde{T}_j'\tilde{T}_a = \sum_{i=1}^{n}\tilde{t}_{ij}\tilde{t}_{ia} = 0, \;\; a > j \tag{4}$$

Kendala ini ditentukan untuk memperoleh lebih dari satu solusi dan untuk menghindari multikolinearitas antara peubah-peubah bebas.

Loading-*x*, $\mathbf{p}_j$ merupakan hubungan linier antara peubah *x* dan komponen $\tilde{X}\mathbf{r}_j$ ke-*j*.

$$\begin{aligned}\mathbf{p}_j &= \left(\mathbf{r}_j'\tilde{X}'\tilde{X}\mathbf{r}_j\right)^{-1}\tilde{X}'\tilde{X}\mathbf{r}_j \\ &= \left(\mathbf{r}_j' S_x \mathbf{r}_j\right)^{-1} S_x\mathbf{r}_j\end{aligned} \tag{5}$$

Dengan S_x adalah matriks koragam empirik antara peubah *x*. Definisi ini mengimplikasikan bahwa persamaan (4) dapat diselesaikan ketika $\mathbf{p}'_j\mathbf{r}_a = 0$ untuk $a > j$.

Vektor-vektor bobot SIMPLS adalah sepasang $(\mathbf{r}_a, \mathbf{q}_a)$, pasangan yang pertama $(\mathbf{r}_1, \mathbf{q}_1)$ diperoleh dari vektor-vektor singular kiri dan kanan yang pertama dari S_{xy}, sehingga mengimplikasikan bahwa $\mathbf{q}_1$ adalah vektor ciri dari $S_{yx}S_{xy}$ dan $\mathbf{r}_1$ adalah vektor ciri dari $S_{xy}S_{yx}$ dimana $\left(S_{xy} = S'_{yx}\right)$. Selanjutnya sepasang vektor bobot SIMPLS $(\mathbf{r}_a, \mathbf{q}_a)$ dengan $2 \le a \le k$ adalah vector ciri $S^a_{yx}S^a_{xy}$ dan $S^a_{xy}S^a_{yx}$.

$$S^a_{xy} = S^{a-1}_{xy} - \mathbf{v}_a\left(\mathbf{v}_a' S^{a-1}_{xy}\right) = (I_p - \mathbf{v}_a\mathbf{v}_a')S^{a-1}_{xy} \tag{6}$$

dan $S^1_{xy} = S_{xy}$. $\{\mathbf{v}_1,\ldots,\mathbf{v}_{a-1}\}$ direpresentasikan sebagai sebuah basis orthonormal terhadap semua loading-*x* $P_{a-1} = [\mathbf{p}_1,\ldots,\mathbf{p}_{a-1}]$. Maka, algoritma iterative ini diawali dengan $S_{xy} = S^1_{xy}$ dan mengulang proses ini sampai komponen *k* ditetapkan.

Salah satu tehnik untuk menentukan banyaknya komponen *k* yaitu dengan menghitung nilai *Root Mean Squared Error* (RMSE).

$$RMSE_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,k}\right)^2} \tag{7}$$

Jumlah komponen yang optimal ditentukan dari komponen *k* yang memiliki nilai RMSE minimum.

Langkah Kedua : Meregresikan peubah respons kedalam komponen-komponen *k*

Langkah kedua dalam algoritma ini, peubah-peubah respon diregresikan kedalam komponen-komponen *k*. Model formal regresi diberikan dibawah ini :

$$\mathbf{y}_i = \boldsymbol{\alpha}_0 + \mathbf{A}'_{q,k}\tilde{\mathbf{t}}_i + \mathbf{f}_i \tag{8}$$

Dimana $E(\mathbf{f}_i) = 0$ dan $\operatorname{cov}(\mathbf{f}_i) = \sum_f$ yang merupakan performa dari regresi linier berganda.

Penduga regresi linier berganda diperoleh sebagai berikut :

$$\hat{\mathbf{A}}_{k,q} = (S_t)^{-1}S_{ty} = \left(R'_{k,p}S_xR_{p,k}\right)^{-1}R'_{k,p}S_{xy}$$

$$\hat{\boldsymbol{\alpha}}_0 = \bar{\mathbf{y}} - \hat{\mathbf{A}}'_{q,k}\bar{\tilde{\mathbf{t}}}$$

$$S_f = S_y - \hat{\mathbf{A}}'_{q,k}S_t\hat{\mathbf{A}}_{k,q}$$

S_y dan S_t adalah matriks koragam empirik peubah-peubah *y* dan *t*. Karena $\overline{\tilde{\mathbf{t}}} = 0$ maka intersept $\boldsymbol{\alpha}_0$ diduga dengan $\overline{\mathbf{y}}$. Dengan $\tilde{\mathbf{t}}_i = R'_{k,p}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)$ dari persamaan (2), kita peroleh penduga parameter untuk model regresi linier original yaitu :

$$\hat{\mathbf{B}}_{p,q} = R_{p,k}\hat{\mathbf{A}}_{k,q}$$
$$\hat{\boldsymbol{\beta}}_0 = \overline{\mathbf{y}} - \hat{\mathbf{B}}'_{q,p}\overline{\mathbf{x}}$$

penduga $\sum_e$ yaitu S_e merupakan fungsi dalam parameter original :

$$S_e = S_y - \hat{\mathbf{B}}'_{q,p} S_x \hat{\mathbf{B}}$$

Sebagai catatan bahwa untuk peubah respons univariat $(q = 1)$, penduga parameter $\hat{\mathbf{B}}_{p,1}$ dapat ditulis sebagai vektor $\hat{\boldsymbol{\beta}}$ serta penduga ragam error $\hat{\sigma}_e^2 = s_e^2$.

Metode *Minimum Covariance Determinant* (MCD)

Misalkan $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ merupakan suatu contoh dari *n* pengamatan dalam $\mathbf{R}^k$ dan *h*, dengan $\frac{n}{2} < h < n$, cari subhimpunan $\mathbf{J}^*$ berukuran *h* sedemikian hingga :

$$\mathbf{J}^* = \min_{J \subset \{1,2,\ldots,n\} \# |J| = h} \det \hat{S}_J$$

Dimana $\hat{S}_J$ adalah matriks koragam berdasarkan pada pengamatan x_i dengan $i \in J$.
Penduga MCD diberikan sebagai berikut :

$$\overline{x}_{J^*} = \frac{i}{h}\sum_{i \in J^*} x_i$$
$$\hat{S}_{J^*} = \frac{1}{h}\sum_{i \in J^*}\left(x_i - \overline{x}_{J^*}\right)'\left(x_i - \overline{x}_{J^*}\right)$$

Regresi MCD

Penduga regresi *robust* diperoleh dengan menggantikan rataan dan matriks peragam klasik dengan penduga pusat dan sebaran bobot MCD.

$$\hat{\boldsymbol{\mu}}_R = \frac{\left(\sum_{i=1}^n w_i x_i\right)}{\left(\sum_{i=1}^n w_i\right)} \quad ; \quad \hat{\Sigma}_R = \frac{\left(\sum_{i=1}^n w_i (x_i - \hat{\boldsymbol{\mu}}_R)(x_i - \hat{\boldsymbol{\mu}}_R)'\right)}{\left(\sum_{i=1}^n w_i - 1\right)}$$

Ringkasnya, masing-masing x_i diberikan bobot w_i, $w_i = 1$ apabila $\left(x_i - \hat{\boldsymbol{\mu}}_0\right)' \hat{\Sigma}_0^{-1}\left(x_i - \hat{\boldsymbol{\mu}}_0\right) \le \chi^2_{q,0.975}$ dan $w_i = 0$ untuk lainnya. Penduga koefisien regresi diperoleh menggunakan metode OLS, perbedaannya hanya didasari dengan pemberian bobot terhadap pengamatan. Misal $\hat{\Sigma}_{\bar{f}}$ adalah penduga inisial untuk matriks peragam galat, maka parameter *robust* untuk model regresi linier original diberikan seperti dibawah ini :

$$\hat{\mathbf{B}}_{p,q} = R_{p,k}\hat{\mathbf{A}}_{k,q}$$
$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\alpha}}_0 - \hat{\mathbf{B}}'_{q,p}\hat{\boldsymbol{\mu}}_x$$
$$\hat{\Sigma}_e = \hat{\Sigma}_{\bar{f}}$$

Metode ini disebut dengan RSIMCD.

Regresi ROBPCA

Metode ROBPCA mengkombinasikan dua pendekatan, yaitu menggunakan *projection pursuit* yang dikembangkan oleh Donoho dan Stahel, dengan menentukan data pencilan untuk setiap pengamatan kemudian membentuk matriks peragam empirik titik-titik data *h* dengan pencilan yang paling kecil. Kemudian data di proyeksi kedalam subruang K_0 yang merentang dengan $k_0 << m$ vektor ciri dominan dari matriks peragam. Selanjutnya, metode MCD diaplikasikan untuk menduga pusat dan sebaran data dalam subruang yang telah diperkecil dimensinya. Dengan kata lain, pendugaan ini adalah *backtransformed* untuk ruang original dan penduga pusat robust $\hat{\mu}_z$ dari $Z_{n,m} = \left(X_{n,p}, Y_{n,q}\right)$ dan sebarannya $\hat{\Sigma}_z$. Matriks sebaran dapat didekomposisi sebagai berikut :

$$\hat{\Sigma}_z = \begin{pmatrix} \hat{\Sigma}_x & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y \end{pmatrix} = P^z L^z (P^z)'$$

Dengan vektor ciri *Z* robust P^z_{m,k_0} dan akar ciri *Z* adalah $\left(L_{k_0,k_0}\right)$. Untuk menghitung skor *robust* yaitu tentukan vektor bobot $\mathbf{r}_a$ menggunakan algoritma SIMPLS sebagai tahap awal, tetapi matriks koragam S_{xy} diganti dengan $\hat{\Sigma}_{xy}$. Sedangkan vektor loading *x* didefinisikan $\mathbf{p}_j = \left(\mathbf{r}_j' \hat{\Sigma}_x \mathbf{r}_j\right)^{-1} \hat{\Sigma}_x \mathbf{r}_j$ kemudian performa $\hat{\Sigma}^a_{xy}$ sama seperti pada tahap SIMPLS. Dan pada masing-masing tahapan skor *robust* dihitung $t_{ia} = \breve{x}_i' \mathbf{r}_a = \left(x_i - \hat{\boldsymbol{\mu}}_x\right)' \mathbf{r}_a$.

Selanjutnya skor-skor *robust* diregresikan kedalam peubah respon, penduga pusat $\boldsymbol{\mu}$ dan sebaran $\sum$ dari (t, y) yaitu rataan dan matriks koragam terboboti.

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_t \\ \hat{\boldsymbol{\mu}}_y \end{pmatrix} = \frac{\sum_{i=1}^{n} w_i \begin{pmatrix} \mathbf{t}_i \\ \mathbf{y}_i \end{pmatrix}}{\sum_{i=1}^{n} w_i} \tag{9}$$

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_t & \hat{\Sigma}_{ty} \\ \hat{\Sigma}_{yt} & \hat{\Sigma}_y \end{pmatrix} = \frac{\sum_{i=1}^{n} w_i \begin{pmatrix} \mathbf{t}_i \\ \mathbf{y}_i \end{pmatrix} \begin{pmatrix} \mathbf{t}_i' & \mathbf{y}_i' \end{pmatrix}}{\sum_{i=1}^{n} w_i - 1} \tag{10}$$

Dengan $w_i = 1$ apabila pengamtan ke-*i* tidak diidentifikasi sebagai pencilan dengan metode ROBPCA dalam (x, y) dan $w_i = 0$ untuk lainnya.

Setelah $\hat{\boldsymbol{\mu}}$ dan $\hat{\Sigma}$ diperoleh, proses selanjutnya sama seperti konsep metode regresi MCD yaitu penduga koefisien regresi diperoleh menggunakan metode OLS. Metode ini disebut dengan RSIMPLS.

Model Kalibrasi dan Validasi

Untuk membangun model RKTP yaitu dengan memilih jumlah komponen yang optimal (k_{opt}). k_{opt} diperoleh dari nilai RMSECV *Robust* (R-RMSECV $_k$) minimum dari setiap *k*.

$$\mathbf{R-RMSECV}_k = \sqrt{\frac{i}{n_c q} \sum_{i \in G_c} \sum_{j=1}^{q} \left(y_{ij} - \hat{y}_{-ij(k)}\right)^2}$$

Masing-masing pengamatan pencilan dihilangkan $\left(c_{-i} = \min_K c_{-i(k)}\right)$ dan G_c merupakan subset pengamatan dimana $c_{-i} = 1$ dengan $|G_c| = n_c$.

Salah satu jenis pengujian untuk validasi model yaitu dengan menghitung nilai RMSEP *robust* (R-RMSEP $_{k_{opt}}$).

$$\mathbf{R-RMSEP}_{k(opt)} = \sqrt{\frac{i}{n_p q}\sum_{i\in G_p}\sum_{j=1}^{q}\left(y_{ij}-\hat{y}_{-ij(k)}\right)^2}$$

BAHAN DAN METODE PENELITIAN

Sumber Data

Banyaknya pengamatan yang digunakan untuk membangun model kalibrasi adalah 20 rimpang temulawak yang diukur menggunakan metode HPLC (*High Performance Liquid Chromatography*), mengenai konsentrasi senyawa aktif dalam rimpang temulawak yang disebut kurkuminoid sebagai peubah respon (***Y***). Dan data mengenai persen transmitan yang dihasilkan metode FTIR (*Fourier Transform Infrared*) pada 1866 titik di sepanjang kisaran bilangan gelombang 4000-400 cm^{-1} sebagai peubah bebas (***X***).

Metode Penelitian

1. Hitung matriks $\breve{X}_{n,p}$ dan $\breve{Y}_{n,q}$

$\breve{\mathbf{x}}_i = x_i - \hat{\boldsymbol{\mu}}_x$

$\breve{\mathbf{y}}_i = y_i - \hat{\boldsymbol{\mu}}_y$

2. Hitung sepasang vektor bobot RSIMPLS yang pertama, $\mathbf{r}_1$ dan $\mathbf{q}_1$.

$\mathbf{q}_1$ adalah vektor ciri dari $\hat{\Sigma}_{yx}\,\hat{\Sigma}_{xy}$

$\mathbf{r}_1 = \hat{\Sigma}_{xy}\,\mathbf{q}_1$; dimana $\hat{\Sigma}_z = \begin{pmatrix}\hat{\Sigma}_x & \hat{\Sigma}_{xy}\\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y\end{pmatrix} = P^z L^z \left(P^z\right)'$; $Z_{n,m} = \left(X_{n,p}, Y_{n,q}\right)$, dengan vektor ciri *Z robust* (P^z_{m,k_0}) dan akar ciri *Z*, $\mathbf{diag}(L_{k_0,k_0})$.

3. Untuk setiap $a = 1,2,\ldots,k$ normalisasi vektor bobot RSIMPLS $\mathbf{r}_a$ dan $\mathbf{q}_a$, ($\|\mathbf{r}_1\| = \|\mathbf{q}_1\| = 1$) didefinisikan sebagai vektor-vektor maksimum.

$$\mathrm{cov}(\breve{Y}_{n,q}\mathbf{q}_a, \breve{X}_{n,p}\mathbf{r}_a) = \mathbf{q}'_a \frac{\breve{X}_{n,p}\breve{Y}_{n,q}}{n-1}\mathbf{r}_a = \mathbf{q}'_a\,\hat{\Sigma}_{yx}\,\mathbf{r}_a$$

4. Hitung skor RSIMPLS

dimana,

$$T_{n,k} = \breve{X}_{n,p} R_{p,k}$$

dengan $R_{p,k} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_k)$

skor pertama, $\mathbf{t}_1$: $\mathbf{t}_1 = \breve{x}'_i\,\mathbf{r}_1$

5. periksa restriksi :

$$\mathbf{r}'_j\,\breve{X}'\breve{X}\mathbf{r}_a = \sum_{i=1}^{n}\tilde{t}_{ij}\tilde{t}_{ia} = 0$$

$$T_a' T_j = 0 \qquad , a > j$$

dimana komponen $\breve{X}\mathbf{r}_j$ diharapkan orthogonal guna memperoleh lebih dari satu solusi.

6. Hitung x-*loading*, $\mathbf{p}_j$ yang menggambarkan hubungan linier antara peubah-peubah *x* dan komponen $\breve{X}\mathbf{r}_j$ ke-*j*

$$\mathbf{p}_j = \left(\mathbf{r}_j'\,\hat{\Sigma}_x\,\mathbf{r}_j\right)^{-1}\hat{\Sigma}_x\,\mathbf{r}_j$$

7. langkah 5 dipenuhi apabila $\mathbf{p}'_j\mathbf{r}_a = 0$ untuk $a > j$.

8. Hitung sebuah basis ortonormal $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{a-1}\}$ loading x $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{a-1}\}$ untuk $(2 \le a \le k)$

Basis,

$$\mathbf{v}_1 = \mathbf{p}_1$$

$$\mathbf{v}_i = \mathbf{p}_i - \frac{\mathbf{v}_1'\mathbf{p}_i}{\mathbf{v}_1'\mathbf{v}_1}\mathbf{v}_1 - \ldots - \frac{\mathbf{v}_{i-1}'\mathbf{p}_i}{\mathbf{v}_{i-1}'\mathbf{v}_{i-1}}\mathbf{v}_{i-1}$$

Basis orthonormal,

$$\mathbf{v}_i'\mathbf{v}_j = \begin{cases} 0, \ i = j & (\textit{orthogonal}) \\ 1, \ i \neq j & (\textit{Normalisasi}) \end{cases}$$

9. Hitung matriks peragam silang, $\hat{\sum}_{xy}^{a}$.

$$\hat{\sum}_{xy}^{a} = \hat{\sum}_{xy}^{a-1} - \mathbf{v}_{a-1}(\mathbf{v}_{a-1}' \hat{\sum}_{xy}^{a})$$

10. Hitung vektor bobot RSIMPLS $\mathbf{r}_a$ dan $\mathbf{q}_a$ $(2 \le a \le k)$ sebagai vektor-vektor singular kiri dan kanan yang pertama $\hat{\sum}_{xy}^{a}$

11. Hitung skor selanjutnya untuk $2 \le a \le k$

$$T_a = \breve{X}_{n,p}\mathbf{r}_a$$

12. Ulangi langkah 4 untuk $2 \le a \le k$
13. Hitung penduga algoritma RSIMPLS

$$\hat{A}_{k,q} = (\hat{\sum}_t)^{-1}\hat{\sum}_{ty} = (R_{k,p}'\hat{\sum}_x R_{p,k})^{-1} R_{k,p}'\hat{\sum}_{xy}$$

$$\hat{\boldsymbol{\alpha}}_0 = \bar{\mathbf{y}} - \hat{A}_{q,k}'\bar{\mathbf{t}}$$

$$\hat{\sum}_{\bar{f}} = \hat{\sum}_y - \hat{A}_{q,k}'\hat{\sum}_t \hat{A}_{k,q}$$

$\hat{\sum}_y$ dan $\hat{\sum}_t$ adalah matriks peragam peubah-peubah y dan t.

14. Tentukan jumlah komponen k, pilih k_{opt} sebagai nilai k yang memberikan nilai $\mathbf{R} - \mathbf{RMSECV}_k$ minimum.
15. Hitung koefisien regresi RSIMPLS untuk peubah-peubah asal.

$$\mathbf{B}_{p,q} = R_{p,k}\hat{A}_{k,q}$$

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\alpha}}_0 - \mathbf{B}_{q,p}'\hat{\boldsymbol{\mu}}_x$$

$$\hat{\sum}_e = \hat{\sum}_{\bar{f}}$$

HASIL DAN PEMBAHASAN

Pada tahun 2003 Hubert dan Vanden Branden membandingkan tiga metode : SIMPLS, RSIMCD dan RSIMPLS menggunakan simulasi data dengan memilih n, p, q, k dan $\sum_t$ yang berbeda. Untuk setiap kondisi, data dibangkitkan sebanyak 1000 sampel. Kondisi yang pertama yaitu data yang tidak terkontaminasi, dimana data dibangkitkan berdasarkan model bilinier dibawah ini :

$T \sim N_k(\mathbf{0}_k, \sum_t)$; dengan $k < p$

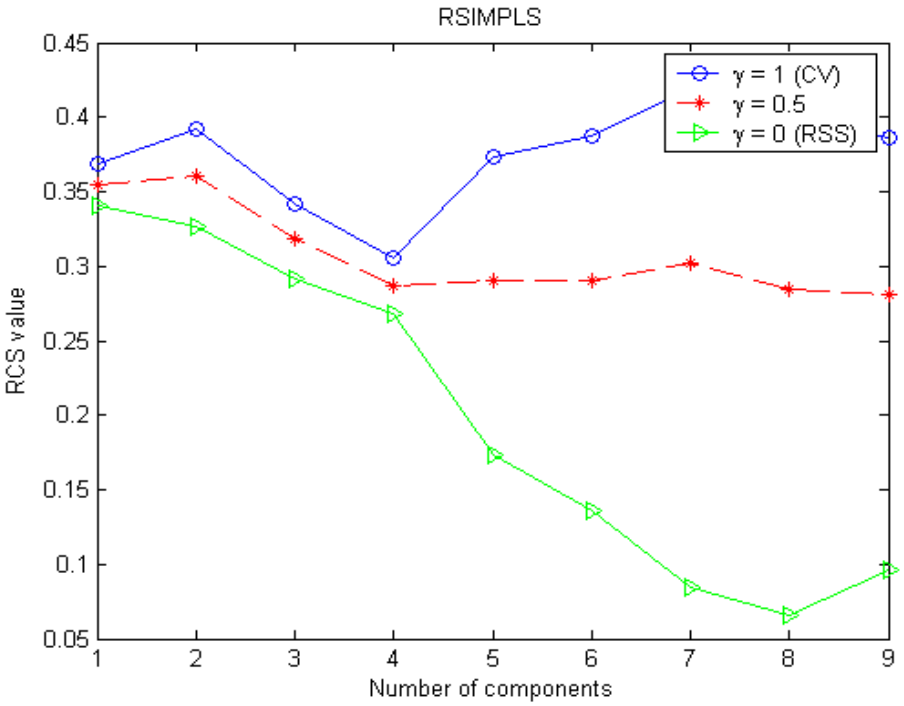
$X = TI_{k,p} + N_p(\mathbf{0}_p, 0.1I_p)$; I adalah matriks identitas

$Y = TA + N_q(\mathbf{0}_q, I_q)$; dengan $A \sim N_q(\mathbf{0}_q, I_q)$

Kondisi yang kedua yaitu data yang terkontaminsi dengan jenis-jenis pencilan yang berbeda, 10% orthogonal outlier, 10% bad leverage points dan 10% vertical outlier. Dari hasil simulasi data

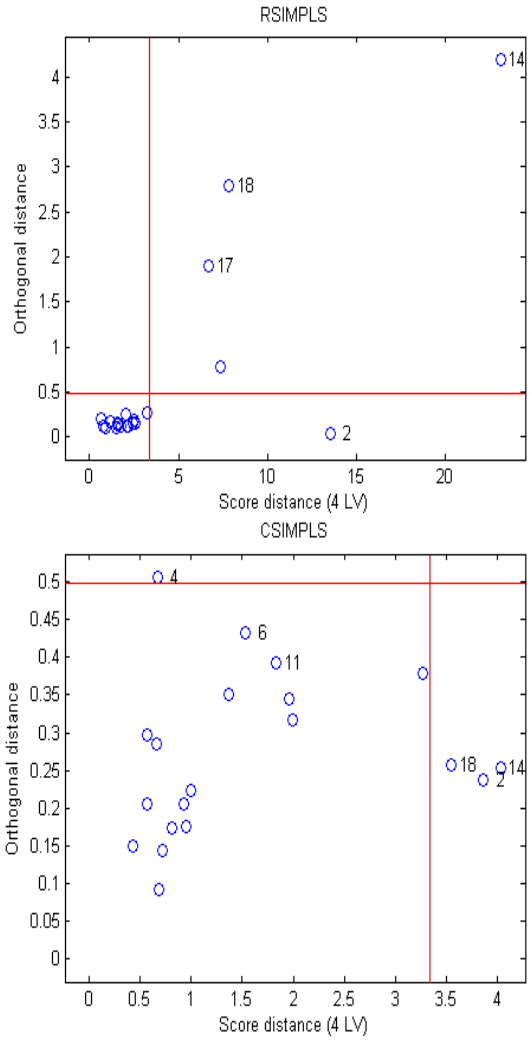
diperoleh, ketika data tidak terkontaminasi semua metode menunjukkan performa yang baik. SIMPLS menghasilkan nilai MSE paling rendah untuk $q = 1$ dan peubah bebas yang berdimensi besar, begitu juga RSMICD dan RSIMPLS memberikan hasil yang cukup baik. Sedangkan untuk data yang terkontaminasi, hasil SIMPLS menjadi terganggu, dimana nilai MSE untuk semua jenis pencilan menjadi meningkat. Sedangkan nilai MSE yang diperoleh RSIMCD dan RSIMPLS tidak mengalami peningkatan yang besar. Perbedaan RSIMCD dan RSIMPLS sangat kecil, tetapi karena komputasi RSIMPLS dua kali lebih cepat dari RSIMCD maka Hubert dkk menetapkan RSIMPLS merupakan metode terbaik.

Berdasarkan hasil simulasi data, maka RSIMPLS diaplikasikan dalam data real rimpang temulawak menggunakan MATLAB 6.5.

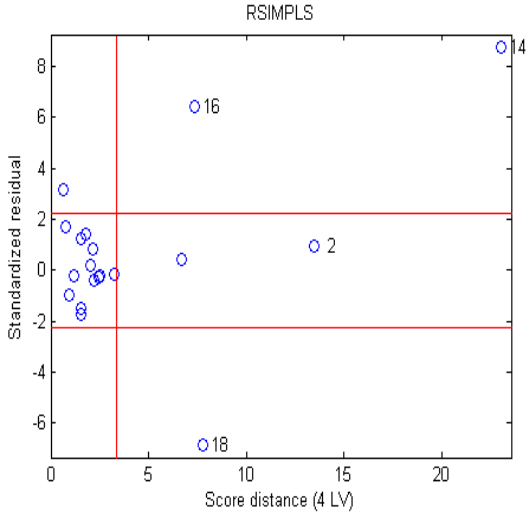


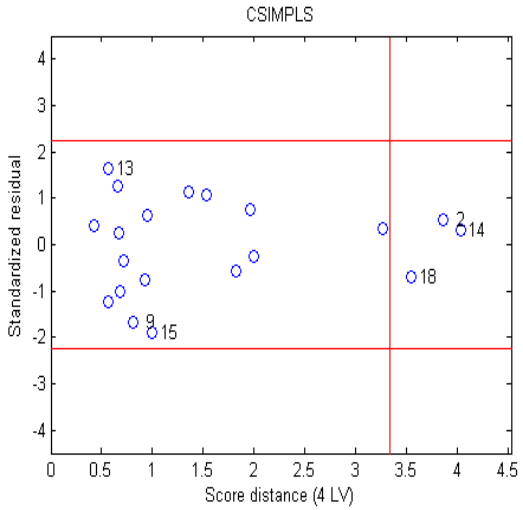
k	**R-RMSECV$_k$**
1	0.36868
2	0.392
3	0.34199
4	0.30526
5	0.37329
6	0.38787
7	0.41955
8	0.39771
9	0.386

Nilai R-RMSECV$_k$ minimum ketika k = 4, sehingga dipilih sebanyak 4 komponen dan diperoleh $h = 17$, dengan R^2 = 0.7954 dan untuk validasi model diperoleh nilai RMSEP = 0.2831.


RSIMPLS
Orthogonal distance
Score distance (4 LV)
14
18
17
2
CSIMPLS
Orthogonal distance
Score distance (4 LV)
4
6
11
18
14
2

Gambar 1


RSIMPLS
Standardized residual
Score distance (4 LV)
14
16
2
18



Gambar 2

Gambar 1 menunjukkan *score diagnostic plot* dengan RSIMPLS pengamtan 14, 17 dan 18 dideteksi sebagai titik bad PCA-leverage, dan pengamatan 2 sebagai titik good PCA-leverage. Namun, dengan SIMPLS mengindikasikan pengamatan 2, 14 dan 18 sebagai titik good PCA-leverage.

Gambar 2 menunjukkan *regression diagnostic plot* dengan RSIMPLS terdapat tiga titik bad leverage (14, 16, 18), dan satu titik good leverage (2). Sedangkan dengan SIMPLS semua titik bad leverage di masukkan kedalam titik good leverage.

DAFTAR PUSTAKA

Hubert, M., Rousseeuw, Peter J., dan Branden, Karlien V. (2004). ROBCA: A New Approach to Robust Principal Component Analysis. *Technometric*s. 47, No. 1. 64-79.

Verboven, S. dan Hubert, M. (2004). LIBRA: a MATLAB Library for Robust Analysis. http://www.wis.kuleuven.ac.be/stat/robust.html.

Hubert, M., Rousseeuw, P.J., Verboven, S. (2002), A fast robust method for principal componenet with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60, 101-111.

Hubert, M., dan Branden .K.V., (2003). Robust methods for Partial Least Squares Regression, *Journal of Chemometric*s. **17** : 537-549.

Debruyne, M., Engelen, S., Hubert, M., dan Rousseeuw, Peter J. (2006). Robustness and Outlier Detection in Chemometrics.

Rousseeuw ,P.J., Van Aelst, S., dan Van Driessen, K. (2004). Robust multivariate regression. *Technometrics*, 46: 293-305.