Prosiding Seminar Nasional MIPA 2016
"Peran Penelitian Ilmu Dasar dalam Menunjang Pembangunan Berkelanjutan"
Jatinangor, 27-28 Oktober 2016
ISBN 978-602-72216-1-1

Menentukan Pusat Elips Pada Metode MVE Menggunakan Jarak Robust

Kankan Parmikanti*, I. Irianingsih[#], K. Joebaedi[^], Rusyaman

Departemen Matematika, FMIPA, Unpad
E-mail: **parmikanti@yahoo.co.id*
[#]*iin_mtk@yahoo.com*
[^]*khafsah.jbd@gmail.com*

Abstrak

Banyak metode yang dapat digunakan untuk mendeteksi pencilan data, hal ini dikarenakan data yang mengandung banyak pencilan akan mengganggu keseimbangan data sehingga sangat memungkinkan mengurangi kecocokan model yang dibuat. Kecocokan dan keakuratan suatu model regresi sangat tergantung pada data dan metode yang digunakan. Pada makalah ini akan diuraikan salah satu metode pendeteksi pencilan data multivariat yaitu metode *Minimum Volume Elipsoida* atau lebih dikenal dengan MVE. Tujuannya adalah untuk menganalisis hal-hal yang berkaitan dengan elipsoida seperti pusat elips, jari-jari, dan volume. Konsep jarak yang akan digunakan adalah jarak *robust*, di mana data yang terletak di luar ellips akan diidentifikasi sebagai pencilan data. Dari contoh kasus, metode MVE bisa sangat diandalkan untuk mendeteksi pencilan data dibandingkan dengan menggunakan metode lain atau konsep jarak yang lain seperti konsep jarak Mahalanobis.

Kata kunci: *MVE, elipsoida, pencilan data, robust, multivariat*

1. Pendahuluan

Metode *Minimum Volume Ellipsoid* adalah salah satu metode *robust* yang dapat digunakan untuk mendeteksi adanya pencilan data [4]. Pendeteksian pencilan merupakan langkah penting dalam analisis data, karena akan sangat berpengaruh terhadap taksiran parameter model. Terdapatnya satu atau dua pencilan saja pada data dapat mengaburkan pengaruh dalam pengambilan kesimpulan. Untuk mengatasi masalah pencilan ini Rousseeuw memperkenalkan metode *robust* yang resisten terhadap adanya pencilan, yaitu Metode *Minimum Volume Ellipsoid* (MVE) [3].

MVE menjadi populer berkat resistensi yang tinggi terhadap outlier pencilan data , yang membuatnya menjadi alat yang handal untuk mendeteksi pencilan. Dalam makalah ini, pertama akan mengulas definisi MVE dan algoritma resampling standar untuk menghitung perkiraan MVE dalam praktek dan memberikan referensi untuk algoritma alternatif yang sering melibatkan pendeteksi pencilan data multivariat . Dalam MVE, paling tidak ada dua konsep jarak yang dapat digunakan, yaitu konsep jarak *robust* dan konsep jarak Mahalanobis. Namun demikian dalam makalah ini akan di titik beratkan pada konsep jarak *robust*, sedangkan konsep jarak mahalanobis akan digunakan sebagai pembanding melalui bahasan yang sangat singkat.

Walaupun MVE juga sekaligus bisa digunakan untuk menaksir parameter regresi linear berganda, tapi dalam kesempatan ini manfaat tersebut tidak akan digunakan.

2. Tinjauan Pustaka

Metode *Minimum Volume Ellipsoid* (MVE) mulai diperkenalkan oleh Rosaeuw pada tahun 1985, namun demikian Prof. Byod dalam makalahnya [1] mengemukakan bahwa pada tahun tujuh puluhan Shor dan Yudin telah mengembangkan pendekatan lain yang mirip dengan MVE ini dengan menggunakan vektor gradien, bahkan tahun 1979 metode ini digunakan oleh Khachian dalam menyusun algoritma polinom waktu.

Sebagai metode penaksir parameter, MVE juga merupakan penaksir yang sangat penting dalam statistik *robust*, bahkan Jun-ya Gotoh and Aiko Takeda, 2006 [2] telah membuat karakterisasi persentase dari sejumlah titik di $\mathbf{R}^n$ dalam elips dengan 100β % , dengan $\beta \in$ [0,5 , 1,0). Makalahnya menyajikan formulasi baru dalam mengkonstruksi sebuah elips berbasis teknik minimisasi CvaR (Conditional value-at-Risk).

Sekar wulandari tahun 2010 [5] pada penelitiannya membandingkan metode kombinasi *Minimum Covariant determinant* (*MCD*) dan *Minimum Volume Ellipsoid* (*MVE*) dengan Metode Regresi *Robust*: *Least Median Square* (*LMS*) dan *Least Trimmed Square* (LTS), dalam mengukur tingkat resistensi terhadap *outlier*. dengan membandingkan nilai Bias dan *MSE* (*Means Square Error)* pada beberapa ukuran sampel dan persentase *outlier* yang

ISBN 978-602-72216-1-1

berbeda. Hasil yang diperoleh menunjukkan bahwa metode *MCD-LMS* lebih baik dari pada metode yang lainnya karena memiliki nilai Bias dan *MSE* yang minimum.

.

3. Metode Volume Minimum Elipsoida

Prisip utama dari metode Volume Minimum Elipsoida (MVE) ini adalah bagaimana kita membuat/menentukan sebuah elipsoida yang volumenya minimum dari sekumpulan data, sedemikian sehingga titik-titik data yang berada di luar elips dikatakan sebagai pencilan (*outlier*). Jika n adalah banyaknya pengamatan dan p adalah banyaknya variabel bebas, maka terdapat sebanyak kombinasi (p+1) dari n atau $\binom{n}{p+1}$ himpunan bagian (sub sampel) yang memuat (p+1) pengamatan. Berangkat dari data sebanyak inilah, banyaknya elips yang harus dibuat, lalu ditentukan elips mana yang paling minimum volumenya. Dengan demikian bila banyaknya pengamatan terlalu besar, maka banyaknya elipsoida yang harus diperiksapun akan menjadi banyak, sehingga akan menjadi pekerjaan yang tidak praktis. Sebagai contoh, bila ada p = 3 buah variabel acak yaitu X_1 , X_2 , dan X_3 , dengan n = 40 pengamatan yaitu $\{x_{i1}$, x_{i2} , $x_{i3}\}$; i = 1, 2 , 3, . . . , 40, maka akan ada sebanyak $\binom{40}{4}$ = 91.390 subsampel yang harus diolah dan dihitung volume elipsnya. Namun demikian dengan adanya teknologi masalah tersebut bisa terpecahkan. Dengan demikian untuk penelitian yang bertujuan menganalisis Metode Minimum Volume Elipsoida akan lebih efisien diterapkan pada data yang hasil pengamatannya sedikit dengan dimensi yang kecil.

4. Algoritma

Misalkan diberikan data hasil n observasi (pengamatan) dengan p variabel acak sebagai berikut:

Pengamatan:	X_1	X_2	⋯	X_p
1	x_{11}	x_{12}	⋯	x_{13}
2	x_{21}	x_{22}	⋯	x_{23}
3	x_{31}	x_{32}	⋯	x_{33}
4	x_{41}	x_{42}	⋯	x_{43}
5	x_{51}	x_{52}	⋯	x_{53}
⋮	⋮	⋮	⋱	⋮
n	x_{n1}	x_{n2}		x_{n3}

Dari data dengan n pengamatan ini, akan terdapat subsampel sebanyak $\binom{n}{p+1}$ yang memuat (p+1) pengamatan.

Untuk mendapatkan elipsoida dengan volume minimum, dilakukan langkah-langkah sebagai berikut:

1. Dipilih himpunan bagian yang memuat (p+1) pengamatan. Selanjutnya untuk setiap himpunan bagian berukuran (p+1) tersebut, sebutlah himpunan indeks
   $J = \{ i_1 , i_2 , \ldots , i_{p+1} \} \subset \{ 1 , 2 , \ldots , n \}$.
2. Menentukan rata-rata (mean) sampel dengan rumus:
   $$T_J = \bar{x}_J = \frac{1}{p+1} \sum_{i=1}^{p+1} x_{ij} \qquad (1)$$
   sehingga diperoleh pusat elips
   $T_J = (\bar{x}_1 , \bar{x}_2 , \bar{x}_3)$ di mana:
   $$\bar{x}_1 = \frac{x_{11} + x_{21} + x_{31} + \cdots + x_{(p+1)1}}{p+1}$$
   $$\bar{x}_2 = \frac{x_{12} + x_{22} + x_{32} + \cdots + x_{(p+1)2}}{p+1}$$
   $$\bar{x}_3 = \frac{x_{13} + x_{23} + x_{33} + \cdots + x_{(p+1)3}}{p+1} .$$
   Selanjutnya ditentukan matriks kovarian dari sampel dengan rumus
   $$S_J = \frac{1}{p} \sum_{j=1}^{p+1} (x_{ij} - \bar{x}_J)(x_{ij} - \bar{x}_J)^T , \qquad (2)$$
   yang secara rinci bisa dinyatakan dalam bentuk matriks:
   $$S_J = S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_p^2 \end{bmatrix}$$
   di mana $s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$
   Matriks kovarian S_J ini pada umumnya nonsingular, dan memang harus nonsingular. Bila tidak, banyaknya pengamatan harus ditambah sampai pemilihan (p+1) himpunan bagian dari sampel ini mengakibatkan S_J nonsingular.
3. Selanjutnya dihitung jarak kuadrat:
   $$D_J^2 = \left[(x_i - \bar{x}_J)^T (S_J)^{-1} (x_i - \bar{x}_J) \right]_{h:n} \qquad (3)$$
   di mana h:n menunjukkan jarak kuadrat terkecil ke-h diantara jarak kuadrat pada n pengamatan dalam X. Dengan menggunakan jarak kuadrat tersebut, dibuat faktor skala D_J^2 / c^2 di mana $c = \sqrt{\chi^2_{p,\alpha}}$.
   Nilai $\frac{D_J}{c}$ ini merupakan jari-jari elips ke arah sumbu-X_J .

ISBN 978-602-72216-1-1

4. Sebagai penentu yang sangat penting dalam MVE adalah volume elipsoida yang proporsional dengan nilai

$$V_J = \left(\frac{D_j}{c}\right)^p \sqrt{det(S_j)} \qquad (4)$$

Setelah selesai langkah ke-4 dengan memperoleh nilai volume elips untuk subsampel pertama, selanjutnya ulangi langkah 1 sampai 4 di atas untuk subsampel ke-2 berukuran sama yaitu (p+1) sehigga diperoleh volume elips ke-2. Proses terus diulang sampai sebanyak $\binom{n}{p+1}$ subsampel. Selanjutnya kita memilih subsampel yang elipsnya memiliki volume paling minimum.

Dari elips terpilih tersebut, berikutnya dihitung $T(X)$ dan $S(X)$ di mana

$$T(x) = T_J \quad \text{dan} \quad S(X) = \frac{c^2(n,p)}{\chi^2_{p,\alpha}} D_J^2 S_j, \qquad (5)$$

dengan

$c^2(n,p) = \left[1 + \frac{15}{n-p}\right]^2$ yang disebut dengan *correction term*.

Berdasarkan T(X) dan S(X) tersebut di atas, dihitung Jarak Robust dengan rumus

$$RD_i = \sqrt{(x_i - T(X))\, S(X)^{-1} (x_i - T(X))^T} \qquad (6)$$

untuk setiap pengamatan i .

Selanjutnya pencilan data ditentukan apabila

$$RDi > \sqrt{\chi^2_{p,\alpha}} \ .$$

5. Implementasi

Dalam tahap implementasi ini, akan diolah data tentang produktifitas primer fitoplankton Pada budi daya jala apung yang dipengaruhi oleh intensitas cahaya, pH air, dan kerapatan fitoplankton itu sendiri. Dengan banyaknya pengamatan n = 20, dan variabel bebasnya p = 3, maka ditentukan hal sebagai berikut:

Y = produktifitas primer fitoplankton sebagai variabel tak bebasnya, sedangkan variabel bebasnya terdiri dari

X_1 = Intensitas cahaya
X_2 = PH
X_3 = Kerapatan Fitoplankton

Sebagai langkah awal, diambil subsampel pertama berukuran (p+1) = 4 dari subsampel sebanyak $C\binom{20}{4}$ = 4.845. Sebutlah subsampel ini data ke-1 , ke-2, ke-3, dan ke-4 dengan data sepeti tampak pada tabel berikut

Tabel 1: Data subsampel pertama

X1	X2	X3
6485,2	7,8	148
7030,5	7,38	194
6551,3	7,48	180
6140,3	7,52	134

Dengan persamaan (1) diperoleh mean yang berperan sebagai pusat elips adalah

$T(X) = T_j = [6551{,}825 \quad 7{,}545 \quad 164]$.

Selanjutnya dengan bantuan matriks kovarian, jarak kuadrat, dan konstanta

C = $\sqrt{\chi^2_{3,0.025}}$ = 3,0575 , maka diperoleh volume elips dari subsampel pertama yaitu:

$$V_J = 815{,}6704.$$

Setelah proses diulang untuk 4.845 subsampel, maka diperoleh volume minimum dengan pusat ellipsoida.

$T(X) = T_j = [2497, \quad 7{,}472 \quad 149{,}36]$.

Selanjutnya, setelah S(X) ditentukan, maka diperoleh jarak *robust* untuk setiap data seperti tampak pada Tabel-2. Data pada Tabel-2 ini dilengkapi dengan hasil perhitungan konsep jarak yang lain yaitu jarak Mahalanobis dengan rumus:

$$MD(x_i) = \sqrt{(x_i - \bar{x}_n)^t\, S_n^{-1}\, (x_i - \bar{x}_n)}$$

untuk i = 1, 2, 3, ... , n.

Tabel 2 Jarak Robust dan Mahalanobis

N	Mahalanobis Distances	Robust Distances
1	2.130549	5.115256
2	2.002981	2.223515
3	1.665619	1.478747
4	1.493529	1.313966
5	2.027032	1.987303
6	1.352333	5.056273
7	1.544290	1.185568
8	1.260015	1.069498
9	0.282805	0.865716
10	0.657289	1.138124
11	1.532475	4.807515
12	1.325619	1.146423
13	1.615588	1.343990
14	1.967271	3.057969
15	1.331119	1.619730
16	1.502753	5.159683
17	2.225202	1.810721
18	2.459557	8.140078
19	2.215150	3.689402
20	1.580432	1.795889

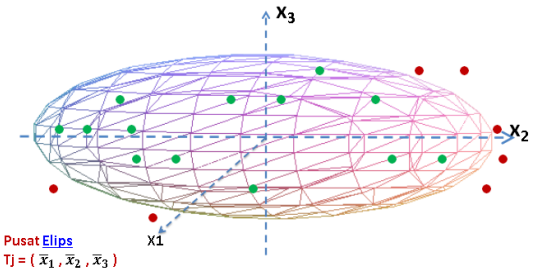
ISBN 978-602-72216-1-1

Dengan kriteria pencilan data:

$$Rdi > \sqrt{\chi^2_{3\ ,\ 0.025}} = 3{,}0575\ ,$$

maka terdeteksi data pencilan ada tujuh buah yaitu: data ke: 1, 6, 11, 14,16, 18, dan 19. Di lain fihak, jika digunakan jarak mahalanobis tidak terdeteksi adanya pencilan.

Sebagai ilustrasi, berikut adalah gambar elips berkaitan dengan pencilan data.



Gb.1 Pencilan Data di luar Elips

6. Simpulam

1. Mengingat volume elipsoida minimum harus dicari dari semua subsampel, maka metode MVE hanya cocok untuk ukuran sampel yang kecil.
2. Selain jarak *robust*, konsep jarak yang dapat digunakan dalam menentukan elipsoida adalah jarak Mahalanobis.
3. Metode MVE cukup akurat dalam mendeteksi pencilan data.

Daftar Pustaka

[1] Boyd S, Prof. , 2014 , Ellipsoid Method, note for EE364b, Stanford University, *stanford. edu/class/ee364b/lectures/**ellipsoid**_method_ notes.pdf* , unduh 28/04/16.

[2]Gotoh Jun-ya and Aiko Takeda, 2006, Conditional Minimum Volume Ellipsoid with Applications to Subset Selection for MVE Estimator and Multiclass Discrimination, Research Report on Mathematical and Computing Sciences ISSN 1342-2804

[3] Rousseeuw P, Va Aest S, 2009, Minimum Volume Ellipsoid, Reasearchgate, https://www.researchgate.net/publication/229803108, unduh 29/04/16.

[4] Rousseeuw and Pan Zomeren, 1991, Robust Distance: Simulations and Cutoff Values, England, John Willey & Sons

[5] Sekar Wulandari, Nur Salam, dan Dewi Anggraini, 2010, Pebandingan Metode Robust *Mcd- Lms*, *Mcd-Lts*, *Mve-Lms*, Dan *Mve-Lts* Dalam Analisis Regresi Komponen Utama, *Jurnal Matematika Murni dan Terapan Vol. 4 No.1 Juni 2010: 57 – 64.*