J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 14, No. 1, 46-53, Juli 2017

p-ISSN: 1858-1382
e-ISSN: 2614-8811

Modifikasi Penaksir Robust dalam Pelabelan Outlier Multivariat

Erna Tri Herdiani *

Abstrak

Outlier adalah suatu observasi yang polanya tidak mengikuti mayoritas data. Outlier dalam kasus multivariat sangat sulit dideteksi, khususnya ketika dimensi lebih dari 2. Kesulitan ini meningkat ketika data set berukuran besar, yakni jumlah variabel menjadi besar. Metode-metode pendeteksian outlier telah lama berkembang dan beberapa digunakan untuk pelabelan outlier sehingga data dapat dipisahkan antara data yang dicurigai sebagai outlier dan data set pada umumnya. Metode-metode tersebut adalah *minimum volume ellipsoid* disingkat MVE, *minimun covariance determinant* disingkat MCD, dan *minimum vector variance* disingkat MVV. Dari ketiga metode tersebut MVV memiliki waktu perhitungan yang paling cepat. Berdasarkan algoritma MVV, kriteria mengurutkan data menggunakan jarak mahalanobis, maka pada paper ini akan dimodifikasi kriteria pengurutan data dengan menghindari penulisan dalam bentuk invers dari matriks variansi kovariansi. Hasil yang diperoleh adalah metode MVV menjadi lebih cepat dengan menggunakan kriteria baru dengan kecermatan yang sama dengan MVV sebelumnya serta akan diaplikan untuk data real dan data simulasi.

Kata Kunci: Minimum Vector Variance (MVV), jarak Mahalanobis, invers matriks variansi kovariansi, trace matriks.

1. Pendahuluan

Pendeteksian outlier multivariat menjadi suatu ilmu yang menantang bagi para peneliti. Hal ini dikarenakan pendeteksian outlier dalam data multivariat tidaklah mudah. Masalah menjadi kompleks ketika terdapat dua atau lebih outlier yang berasal dari beberapa variabel yang lebih dari dua (Rousseeuw & Van Zomeren, 1990). Kesulitan ini meningkat ketika data set berukuran besar dengan ditandai besarnya jumlah variabel. Untuk masalah seperti ini, Angiulli & Pizzuti (2005) telah menyatakan bahwa efisiensi perhitungan merupakan hal yang penting. Seperti halnya efektifitas dari proses pendeteksian. Senada dengan peneliti ini, pada tahun 2001, Ye & Chen pernah mengungkapkan bahwa di dalam sistem informasi diperlukan suatu proses pendeteksian outlier yang cepat untuk data set yang besar, karena variabel yang terdapat dalam sistem informasi bisa terdiri dari ratusan bahkan sampai ribuan.

Metode-metode pendeteksian outlier sudah lama berkembang dan beberapa digunakan untuk pelabelan outlier yaitu memisahkan beberapa data yang dicurigai sebagai outlier dengan data set pada umumnya. Metode-metode tersebut adalah metode proyeksi (Pena & Prieto, 2001), *minimum volume ellipsoid* atau MVE (Werner, 2003), *minimum covariance determinant* atau MCD (Rousseeuw & Van Driessen,1999), dan *minimum vector variance* atau MVV (Herwindiati, 2006). Dari beberapa metode tersebut MVV merupakan metode yang lebih efisien.

Pada paper ini akan diusulkan suatu modifikasi dari metode MVV. Hal yang melatarbelakanginya adalah pada tahapan algoritma MVV terdapat suatu tahapan mengurutkan

* *Jurusan Matematika FMIPA Universitas Hasanuddin Makassar*

data dengan menggunakan jarak mahalanobis, dimana jarak mahalanobis dapat dilihat sebagai berikut:

$$d = sqrt\left(\left(\vec{X}_i - \bar{\vec{X}}\right)^t S^{-1}\left(\vec{X}_i - \bar{\vec{X}}\right)\right),\ i = 1,2,...,n. \qquad (1)$$

dimana $X_1, X_2, \cdots, X_n$ sampel acak dari suatu variabel acak $X \sim N_p\left(\vec{\mu}, \Sigma\right)$, dan secara berturut-turut $\bar{\vec{X}}$ dan S adalah taksiran dari vektor mean sampel dan matriks variansi kovariansi sampel. Menurut Ye & Chen (2001), perhitungan komputasi dari persamaan (1) akan mengalami kendala jika jumlah variabel dalam jumlah besar yaitu ratusan atau ribuan, hal itu diakibatkan karena adanya perhitungan invers dari matriks variansi kovariansi. Oleh karena itu, Djauhari (2003) dalam tulisannya telah menyatakan bahwa kuadrat jarak mahalanobis dapat dinyatakan bukan dalam bentuk invers matriks variansi kovariansi, seperti di bawah ini:

$$d^2 = 1 - \frac{|M_i|}{|S|},\ M_i = \begin{pmatrix} 1 & \left(\vec{X}_i - \bar{\vec{X}}\right)^t \\ \left(\vec{X}_i - \bar{\vec{X}}\right) & S \end{pmatrix},\ i = 1,2,...,n. \qquad (2)$$

Kriteria pengurutan data yang akan digunakan dalam paper ini adalah persamaan (1), (2) dan

$$|M_i|,\ i = 1,2,...,n, \qquad (3)$$

dimana kriteria ini telah digunakan oleh Djauhari & Umbara (2006) untuk data depth.

Adapun pembahasan dalam paper ini adalah bagian 1 berisikan pendahuluan yang berisikan latar belakang penulisan dan tujuan dari penulisan. bagian 2, berisikan tentang teori dasar dan penurunan dari kriteria yang diusulkan. Bagian 3, berisikan studi kasus untuk metode yang telah dimodifikasi, baik data real atau pun data simulasi. Akhir dari paper ini adalah bagian 4 yang berisikan kesimpulan.

2. Modifikasi Penaksiran Robust

Metode penaksiran robust untuk lokasi dan dispersi telah lama berkembang pesat. Pada paper ini, metode terbaru yang robust akhir-akhir ini adalah metode minimum vektor variance (MVV). Berdasarkan Djauhari & Umbara (2006), bahwa $|\boldsymbol{M_i}|$ dapat menjadi kriteria baru dalam mengurutkan data multivariat. Oleh karena itu, penulis mencoba untuk memodifikasi MVV melalui perubahan urutan data multivariatnya.

2.1 Metode *Minimum Vector Variance* (MVV)

Kriteria MVV untuk menaksir lokasi dan dispersi, pertama kali diperkenalkan oleh Herwindiati (2006) dengan mempertimbangkan data set $\boldsymbol{X} = \left\{\vec{\boldsymbol{X}}_1, \vec{\boldsymbol{X}}_2, \cdots, \vec{\boldsymbol{X}}_n\right\}$ dari satu observasi dengan variabel p dan $\boldsymbol{H} \subseteq \boldsymbol{X}$. Misalkan T_{MVV} dan C_{MVV} adalah taksiran MVV untuk parameter lokasi dan matriks variansi kovariansi. Taksiran diperoleh berdasarkan himpunan *H*. Jumlah elemen dari *H* adalah $\boldsymbol{h} = \left[\frac{\boldsymbol{n}+\boldsymbol{p}+1}{2}\right]$ data yang akan memberikan matriks variansi kovariansi C_{MVV} dengan nilai $\boldsymbol{tr}\left(\boldsymbol{C}^2_{\boldsymbol{MVV}}\right)$ minimum untuk semua kemungkinan himpunan yang mengandung h data. Oleh karena itu, secara berturut-turut taksiran MVV untuk parameter lokasi dan matriks variansi kovariansi adalah sebagai berikut:

$$\boldsymbol{T_{MVV}} = \frac{\mathbf{1}}{\boldsymbol{h}}\sum_{\boldsymbol{i}\in\boldsymbol{H}}\vec{\boldsymbol{X}}_{\boldsymbol{i}} \text{ dan } \boldsymbol{C_{MVV}} = \frac{\mathbf{1}}{\boldsymbol{h}-\mathbf{1}}\sum_{\boldsymbol{i}\in\boldsymbol{H}}\left(\vec{\boldsymbol{X}}_{\boldsymbol{i}} - \boldsymbol{T_{MVV}}\right)\left(\vec{\boldsymbol{X}}_{\boldsymbol{i}} - \boldsymbol{T_{MVV}}\right)^{\boldsymbol{t}}.$$

Adapun algoritma untuk MVV adalah sebagai berikut:

1. Ambillah himpunan data yang terdiri dari $\boldsymbol{h} = \left[\frac{\boldsymbol{n}+\boldsymbol{p}+\mathbf{1}}{\mathbf{2}}\right]$ data, sebutlah himpunan data ini dengan nama H_{old}.
2. Hitunglah vektor mean $\bar{\vec{\boldsymbol{X}}}_{\boldsymbol{H_{old}}}$ dan matriks kovariansi $\boldsymbol{S_{H_{old}}}$ untuk semua data H_{old}. Selanjutnya untuk i= 1, 2, ..., n hitunglah

$$d_{H_{old}}^2 = d_{H_{old}}^2\left(\vec{X}_i, \bar{\vec{X}}_{H_{old}}\right) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)^t S_{H_{old}}^{-1}\left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right).$$

3. Urutkan hasil perhitungan dari yang terkecil ke yang terbesar. Urutan ini akan memberikan permutasi indeks observasi $\boldsymbol{\pi}$. Misalkan hasil pengurutan data tersebut adalah

   $d_{H_{old}}^2(\pi_1) \le d_{H_{old}}^2(\pi_2) \le \cdots \le d_{H_{old}}^2(\pi_n)$.

4. Bentuklah suatu himpunan baru yang terdiri dari h observasi dengan indeks $\boldsymbol{\pi}(\mathbf{1}), \boldsymbol{\pi}(\mathbf{2}), \cdots, \boldsymbol{\pi}(\boldsymbol{h})$, dan berilah nama H_{new}.
5. Hitunglah $\bar{\vec{\boldsymbol{X}}}_{\boldsymbol{H_{new}}}$, $\boldsymbol{S_{H_{new}}}$ dan $d_{H_{new}}^2\left(\vec{X}_i, \bar{\vec{X}}_{H_{new}}\right)$ seperti pada tahap 2.
6. Jika $\boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{new}}}^{\mathbf{2}}\right) = \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{old}}}^{\mathbf{2}}\right)$ maka proses dikerjakan. Jika $\boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{new}}}^{\mathbf{2}}\right) < \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{old}}}^{\mathbf{2}}\right)$ maka proses dilanjutkan sampai iterasi ke-k mencapai $\boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{new}}}^{\mathbf{2}}\right) = \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{old}}}^{\mathbf{2}}\right)$.
7. Jika $\boldsymbol{S_{H_i}}$ adalah matriks kovariansi dari iterasi ke-*k*. Pada akhir iterasi ke-*k* akan dimiliki $\boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_1}}^{\mathbf{2}}\right) \ge \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_2}}^{\mathbf{2}}\right) \ge \cdots \ge \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_{k-1}}}^{\mathbf{2}}\right) = \boldsymbol{Tr}\left(\boldsymbol{S}_{\boldsymbol{H_k}}^{\mathbf{2}}\right)$.

2.2 Metode Modifikasi *Minimum Vector Variance*

Sebenarnya ide dasar dari modifikasi metode penaksiran robust ini adalah karena penulisan dari jarak mahalanobis yang mengandung invers dari matriks variansi kovariansi sampel, dimana menurut Ye & Chen (2001) dapat membuat lambatnya waktu perhitungan pada saat variabel yang dilibatkan cukup besar misalnya ratusan. Sedangkan Djauhari (2003) telah membuat suatu penulisan lain dari jarak mahalanobis yang tidak melibatkan invers matriks variansi kovariansi. Oleh karena itu, penulis tertarik untuk merubah kriteria MVV pada tahap 2 dari bentuk persamaan (1) dengan persamaan (2).

Hasil yang diusulkan dapat juga digunakan untuk menaksir lokasi dan dispersi dengan algoritma sama dengan MVV tetapi kriteria urutan data multivariatnya saja yang berbeda. Selanjutnya metode ini penulis namakan Modifikasi Minimum Vektor Variance (MVVM). Persamaan (2) menyatakan

$$d_i^2 = 1 - \frac{|M_i|}{|S|},\; M_i = \begin{pmatrix} 1 & \left(\vec{X}_i - \bar{\bar{X}}\right)^t \\ \left(\vec{X}_i - \bar{\bar{X}}\right) & S \end{pmatrix},$$

sehingga akan diperoleh suatu hubungan, jika $d_i^2 \le d_j^2; i \ne j$ maka $|\boldsymbol{M_i}| \ge |\boldsymbol{M_j}|$. Hal ini pun telah diungkapkan oleh Djauhari & Umbara (2006). Dengan demikian, pada paper ini pun akan diselidiki bagaimana algoritma MVV apabila urutan data multivariat dengan menggunakan persamaan (3), yang selanjutnya metode ini disebut dengan MVVML.

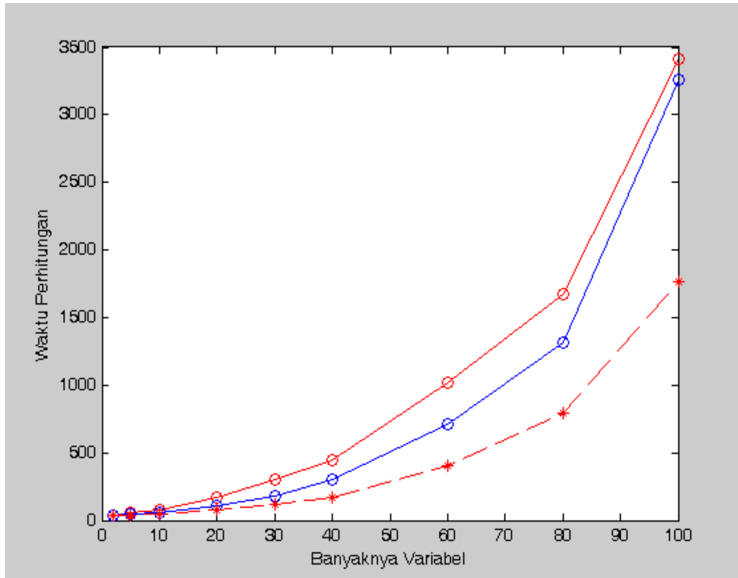
3. Studi Kasus

Penulis akan menunjukkkan keuntungan dari metode MVVM dalam pelabelan outlier multivariat. Pertama-tama akan diaplikasikan untuk data simulasi. Banyaknya data yang diambil 1075 data, mereka dibangkitkan dari distribusi multivariat gabungan, dimana 1000 data adalah berdistribusi normal standar $N_p\left(\vec{0}, I\right)$, selanjutnya disebut data clean dan sisanya 75 data berdistribusi $N_p\left(\vec{3}, I\right)$ yang selanjutnya disebut data kontaminan. Proses pembangkitan dilakukan sebanyak 500 kali. Hasilnya adalah sebagai berikut.

Tabel 1. Waktu Perhitungan dalam Detik.

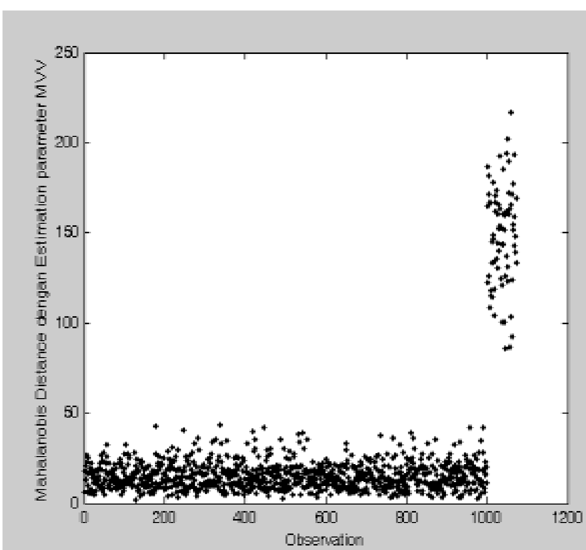
No.	Variabel	Metode (Detik)		
		MVV	**MVVM**	**MVVML**
1	2	40,8430	37,6410	35,907
2	5	55,2500	45,1250	39,313
3	10	81,2030	61,3430	47,875
4	20	167,4690	110,3280	76,422
5	30	297,0470	182,4530	118,141
6	40	446,2500	303,0310	172,391
7	60	1018,8	712,9380	406,234
8	80	1669,5	1314,3	794,360
9	100	3409,3	3257,5	1766,2

Jika Tabel 1 dinyatakan dalam bentuk grafik, hasilnya dapat dilihat dalam gambar berikut.

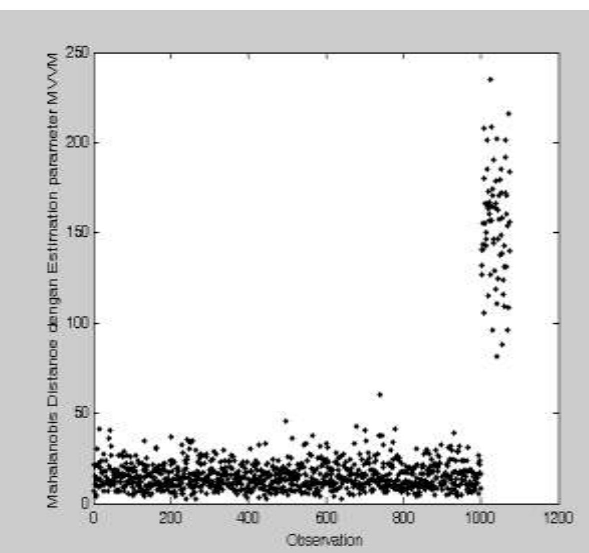


Gambar 1. Waktu perhitungan dalam detik.

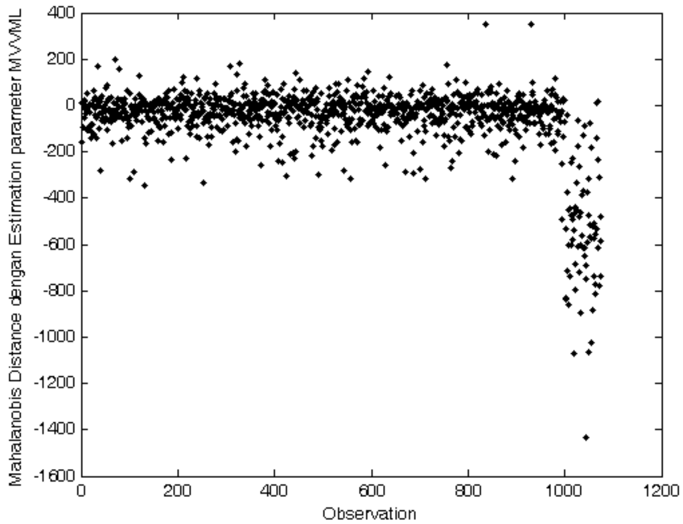
Dari Tabel 1 dan Gambar 1, MVVML merupakan metode tercepat dibandingkan dengan MVV dan MVVM. Ketiga metode ini pun mampu untuk memisahkan data secara jelas antara data kontaminan dengan data clean, walaupun untuk metode MVVML terjadi terbalik dimana data clean ada di atas sedangkan data kontaminant terletak di bawah. Hal tersebut dapat terlihat dalam gambar di bawah ini.



Gambar 2. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVV pada Data Simulasi.

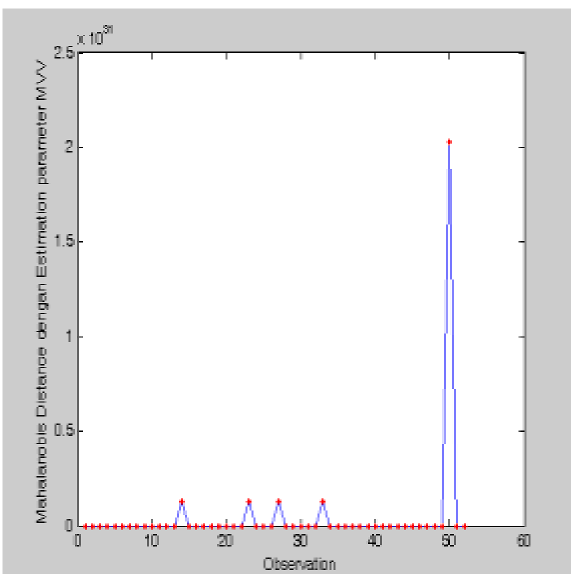


Gambar 3. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVVM pada Data Simulasi

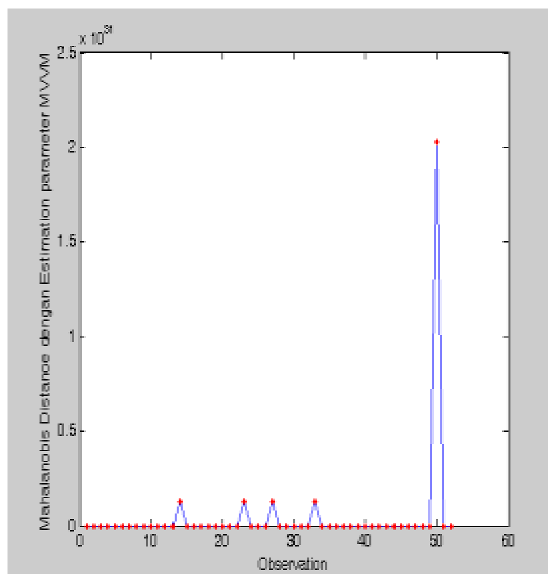


Gambar 4. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVVML.

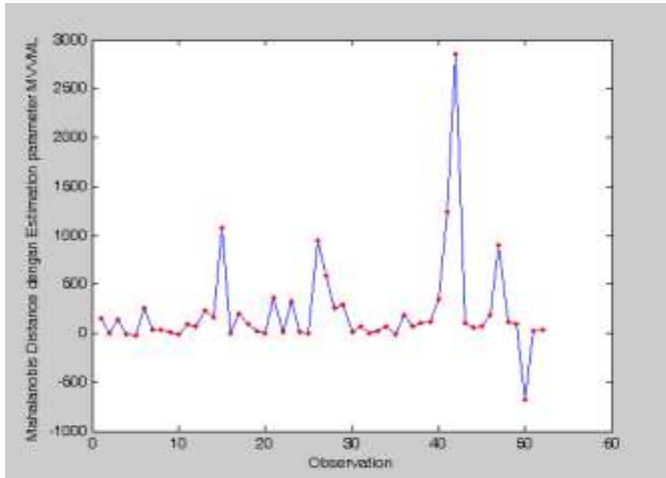
Apabila yang diperlukan dalam suatu kecepatan dalam menghasilkan taksiran lokasi dan dispersi maka sebaiknya menggunakan metode MVVML dalam pengolahan datanya. Kemudian jika penulis aplikasikan metode-metode robust di atas pada data real, hasilnya adalah sebagai berikut. Data yang digunakan adalah data kuesioner yang menilai kinerja perkuliahan di Jurusan Matematika ITB pada tahun 2004, dimana variabel yang terlibat dalam pengolahan data ini terdiri dari 15.



Gambar 5. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVV pada Data Real.



Gambar 6. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVVM pada Data Real.



Gambar 7. Jarak Mahalanobis berdasarkan Penaksiran Parameter MVVML pada Data Real.

Berdasarkan data real, menurut metode MVV dan MVVM, observasi ke-14, 23, 27, 33 dan 50 merupakan yang dicurigai sebagai outlier dalam data real ini. Sedangkan menurut metode MVVML yang dapat dicurigai sebagai outlier dalam data adalah observasi ke-15, 26, 42, 47, dan 50.

4. Kesimpulan

Metode penaksiran robust yang dimodifikasi MVVML akan memberikan kecepatan yang lebih dibandingkan dengan MVV dan MVVM. Selain itu berdasarkan data real, observasi yang memberikan nilai yang paling kecil terlihat jelas dengan ditunjukkan arahnya yang berbeda dengan observasi lainnya yang memberikan nilai lebih bagus.

Ucapan terima kasih

Penulis mengucapkan banyak terima kasih kepada Prof. Dr. Maman A. Djauhari atas dorongannya sehingga memberikan semangat untuk menulis paper ini, tidak lupa juga kepada Direktorat Pendidikan Tinggi yang telah memberikan Dana Penelitiannya melalui BPPS.

Daftar Pustaka

[1] Angiulli, F., Pizzuti, C., 2005, Outlier Mining and Large High Dimensional Data Sets, *IEEE Transaction on Knowledge and Data Engineering*, 17 (2), pp : 203-215.

[2] Djauhari, M.A., 2003, Hubungan norm dan determinan, *Publication intern Data Analysis Research Group*, ITB.

[3] Djauhari, M.A., Umbara, R. F., 2006, On Mahalanobis Depth Function, *paper ini telah dipresentasikan di International Conference on Mathematics and Natural Sciences (ICMNS)*, Bandung, Indonesia, 29-30 November.

[4] Herwindiati, D. E., 2006, A New Criterion in Robust Estimation for Location and Covariance Matrix and Its Application for Outlier Labeling, *Disertasi*, Institut Teknologi Bandung.

[5] Pena, D., Prieto, J.F., 2001, Multivariate Outlier Detection and Robust Covariance Matrix Estimation, *Technometrics*, 43(3), pp : 286-300.

[6] Rousseeuw, P.J., Leroy, A.M., 1987, *Robust Regression and Outlier Detection*, John Wiley & Sons.

[7] Rousseeuw, P.J., Van Zomeren, B.C., 1990, Un masking Multivariate Outliers and Leverage Points, *Journal of The American Statistical Association*, Vol. 85, No. 411, pp : 633-651.

[8] Rousseeuw, P.J., Van Driessen, K., 1999, A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, Vol. 41, No. 3, pp : 212-223.

[9] Werner, M., 2003, Identification of Multivariate Outliers in Large Data Sets, *PhD Thesis*, University of Colorado Denver.

[10] Ye, N., Chen, Q., 2001, An Anomaly Detection Technique Based On A Chi Square Statistic for Detecting Intrusion Into Information Systems, *Quality and Realibility Engineering International, Qual. Realb. Engng. Int.*, 17, 105-112.