ANALISIS REGRESI KOMPONEN UTAMA ROBUST UNTUK DATA MENGANDUNG PENCILAN

By Notiragayu Notirgayu

J. Sains MIPA, April 2008, Vol. 14, No. 1, Hal.: 45 - 50
ISSN 1978-1873

ANALISIS REGRESI KOMPONEN UTAMA *ROBUST* UNTUK DATA MENGANDUNG PENCILAN

Notiragayu* dan Khoirin Nisa

Jurusan Matematika, FMIPA Universitas Lampung
**Alamat korespondensi e-mai* : *Notiragayu@unila.ac.id*

Diterima 5 Februari 2008, perbaikan 17 Maret 2008, disetujui untuk diterbitkan 19 March 2008

ABSTRACT

Principal Component Regression (PCR) is one of the widely used statistical techniques for regression analysis with colinearity, and a robust technique on PCR when data contains outlier is an important problem. In this paper we consider the problem of robust PCR based on Minimum Volume Ellipsoid (MVE) estimator and Least Trimmed Square (LTS) regression. We aimed to look at the behavior of the principal component regression coefficient resulted by MVE-LTS and compare them with classical estimator through the bias and the mean square error. The result shows that PCR using MVE-LTS is very robust.

Keywords *: Principal component regression, colinearity, robust*

1. PENDAHULUAN

Salah satu masalah yang sering dihadapi dalam analisis regresi berganda adalah multikolinieritas, yaitu adanya hubungan linier (korelasi) antara dua peubah bebas atau lebih dalam suatu persamaan regresi. Hal ini menyebabkan matriks X^TX menjadi singular sehingga persamaan normalnya tidak lagi mempunyai jawaban yang tunggal. Akibatnya dugaan koefisien regresi kuadrat terkecil [$b = (X^TX)^{-1} X^TY$] memiliki simpangan baku yang besar yang menjadikannya tidak lagi terandalkan[1)].

Regresi komponen utama (RKU) merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah multikolinearitas. Metode ini mengatasi multikolinieritas dengan dua tahapan, tahap pertama analisis komponen utama terhadap peubah-peubah bebas X_i, dan tahap kedua analisis regresi terhadap komponen-komponen utama dengan peubah respon Y. Pada RKU klasik, kedua tahap dilakukan secara tradisional, yaitu komponen utama dibentuk menggunakan vektor eigen dari matriks peragam sampel (S) klasik dan diregresikan terhadap Y dengan metode kuadrat terkecil. Namun analisis seperti ini sangat sensitif terhadap pencilan (*outliers*) dan akan menghasilkan dugaan parameter yang bias akibat terpengaruh oleh data pencilan[2)]. Cara yang paling sederhana untuk mengatasi pencilan pada RKU adalah dengan menggunakan metode *robust* pada kedua tahap.

Analisis komponen utama *robust* dapat dilakukan dengan beberapa cara, diantaranya dengan algoritma proyeksi[3)], dengan matriks peragam *robust*[4)], atau gabungan dari keduanya[5)]. Sedangkan analisis regresi *robust* dapat dilakukan dengan menggunakan metode *alternating regression*[6)], metode kuadrat terpangkas terkecil[7)], metode median kuadrat terkecil[8)], metode nilai mutlak terkecil[9)], dsb. Dalam tulisan ini, pada tahapan pertama kami memilih untuk menggunakan matriks peragam *robust* dengan menggunakan metode Volume Ellipsoid Minimum (VEM) yang diperkenalkan oleh Peter J. Rousseeuw pada tahun 1985[10)], dan pada tahapan kedua menggunakan metode Kuadrat Terpangkas Terkecil (KTT) yang diperkenalkan oleh Reusseeuw dan Leroy pada tahun 1987[11)] dan dikenal sangat tegar serta memiliki sifat-sifat statistik yang baik[12)]. Untuk memperlihatkan ketegaran analisis RKU dengan metode VEM-KTT terhadap data mengandung pencilan, kami melakukan simulasi Monte Carlo dan membandingkannya dengan RKU klasik.

1.1. Regresi Komponen Utama

Regresi komponen utama merupakan regresi dari peubah tak-bebas terhadap komponen-komponen utama yang tidak saling berkorelasi, dimana setiap komponen utama merupakan kombinasi linear dari semua peubah bebas yang telah dispesifikasikan sejak awal. Untuk persamaan regresi dalam bentuk peubah asal *X* dapat ditulis sebagai berikut:

$Y = {}_0 + {}_1 X_1 + {}_2 X_2 + \dots + {}_p X_p +$

dengan *Y* merupakan peubah tak-bebas, X_i peubah bebas ke-i (i = 1, 2, …, p), ${}_i$ adalah parameter-parameter regresi, dan merupakan galat.

Peubah baru sebagai komponen utama (K) adalah hasil transformasi dari peubah asal (X) yang modelnya dalam bentuk matriks adalah K = ***A X***, dan komponen utama ke-*j* ditulis :

$K_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p = \underline{a_j}'\,\underline{x}$ (1)

dimana vektor pembobot $\underline{a_j}'$ diperoleh dengan memaksimumkan keragaman komponen utama ke-*j*, yaitu $S^2_{K_{j}} = \underline{a_j}'\, S\, \underline{a_j}$ dengan kendala $\underline{a_j}'\,\underline{a_j} = 1$ serta $\underline{a_h}'\,\underline{a_j} = 0$, untuk $h \neq j$. Vektor pembobot $\underline{a_j}'$ diperoleh dari matriks peragam yang diduga dengan matriks ***S***, yaitu $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$. Vektor $\underline{a_j}'$ yang memenuhi kendala di atas adalah vektor eigen dari matriks peragam .

Model regresi komponen utama dapat ditulis sebagai berikut :

$$Y = \;_0 + \;_1 K_1 + \;_2 K_2 + \ldots + \;_m K_m + \;,$$

dengan m p

1.2. Metode Kuadrat Terpangkas Terkecil (KTT)

Metode KTT menduga koefisien regresi dari data yang mengandung pencilan dengan meminimumkan jumlah kuadrat galat terhadap sebaran data yang sudah terpangkas (*trimmed*) atau sebaran terwinsorkan (*winsorized distribution*). Dengan kata lain, metode KTT menduga koefisien regresi dengan menggunakan metode MKT terhadap subhimpunan data terbaik H , yaitu

$$= \arg\min\left(\sum_{i=1}^{h} e_i^{\,2}\right) = \arg\min\left(\sum_{i=1}^{h}(y_i \quad \hat{y}_i)^2\right),$$

$$\frac{(3n + p + 1)}{4} \quad h \quad n\,.$$

dengan h merupakan banyaknya anggota dalam subhimpunan H. Penentuan subhimpunan H terbaik dilakukan dengan menggunakan algoritma resampling[13] dari seluruh kemungkinan subhimpunan yang didapat dibentuk yaitu sebanyak $\binom{n}{h}$. Subhimpunan H yang diperoleh merupakan sebaran data yang sudah terpangkas (*trimmed distribution*)[14].

1.3. Metode Volume Ellipsoid Minimum (VEM)

Metode Volume Ellipsoid Minimum (VEM) merupakan metode penduga *robust* untuk vektor nilai tengah dan matriks peragam. Pada prinsipnya metode ini adalah mencari ellipsoid dengan volume paling minimum yang melingkupi suatu subhimpunan dari minimal h pengamatan. Subhimpunan berukuran h ini disebut *halfset* karena h sering dipilih lebih dari setengah n pengamatan . Penduga nilai tengah adalah pusat ellipsoid secara geometris dan penduga matriks peragam adalah matriks pembentuk ellipsoid.

Mencari penduga VEM secara esensial dilakukan dengan dua proses. Bagian pertama mencari *halfset* terbaik yang memuat h pengamatan. Bagian kedua mencari volume paling minimum dari ellipsoid yang melingkupi *halfset*. Untuk sebuah *halfset* terdiri dari banyak ellipsoid yang dapat melingkupinya. Kedua proses ini dilakukan secara iterative menggunakan algoritma resampling seperti dalam pendugaan matriks peragam *robust* menggunakan metode determinan peragam minimum[15].

Penduga VEM didefinisikan sebagai pasangan ($\overline{\mathbf{X}}, \mathbf{S}$) di mana $\overline{\mathbf{X}}$ adalah vektor-p dan **S** adalah matriks semi-definit positif pxp yang memenuhi :

$$\#\{i \mid (x_i \quad \overline{\mathbf{X}})'\mathbf{S}^{\;1}(x_i \quad \overline{\mathbf{X}}) \quad c^2\} \quad h$$

dengan n adalah jumlah pengamatan, p adalah jumlah peubah, h = [(n+p+1)/2], c suatu konstanta dan x_i adalah data pengamatan[16].

2. METODE PENELITIAN

Dalam penelitian ini kami menggunakan data yang dibangkitkan dengan menggunakan perangkat lunak SAS/IML versi 8. Data terdiri dari 5 peubah bebas X_i dan sebuah peubah tak bebas Y. Untuk menciptakan multikolineritas dalam data ini membangkitkan Xi dari sebaran normal peubah ganda dengan vektor nilai tengah = 0 dan matriks peragam nondiagonal yang dibangkitkan secara acak namun dirancang agar memiliki nilai eigen pertama yang proporsinya melebihi 75%. Sedangkan Y dibangkitkan sebagai kombinasi liner dari X_i ditambah dengan unsur galat. Model regresi komponen utama yang digunakan adalah : $Y = \;K + \;$, dimana K merupakan komponen utama pertama dengan proporsi keragaman di atas 75%. Data pencilan dibangkitkan dari sebaran normal dengan = 10 dan =1.

Simulasi Monte Carlo dilakukan sebanyak 100 replikasi untuk ukuran sampel besar (mengingat metode VEM dan KTT bersifat tak bias asimtotik) yaitu n = 100, 300, dan 500. Pada setiap sampel diberikan kontaminasi pencilan sebesar = 5%, 10%, 15% dan 20%. Nilai bias dan KTG dihitung atas 100 himpunan sampel sebagai berikut :

$$\text{Bias}(\;) = \left|\left(\frac{1}{100}\sum_{s=1}^{100}\hat{\;}^{(s)}\right) \quad \hat{\;}^{(0)}\right|$$

$$\text{MSE}(\;) = \frac{1}{100}\sum_{s=1}^{100}\left(\hat{\;}^{(0)} \quad \hat{\;}^{(s)}\right)^2$$

Dengan $\hat{\;}^{(0)}$ adalah koefisien regresi untuk data tanpa pencilan, dan $\hat{\;}^{(s)}$ adalah koefisien regresi untuk data yang telah diberi kontaminasi pencilan.

3. HASIL DAN PEMBAHASAN

Pengaruh pencilan pada analisis RKU pada setiap tahapan dijelaskan sebagai berikut. Pada tahap pertama, pencilan mempengaruhi matriks peragam dari data X yang secara tak langsung mempengaruhi nilai eigennya. Hal ini mengakibatkan buruknya representasi komponen utama terhadap sebaran data awal karena komponen utama dibentuk berdasarkan vektor eigen matriks peragam seperti pada persamaan (1). Dengan menggunakan metode VEM, nilai eigen yang diperoleh mendekati nilai eigen dari data awal, sehingga skor komponen utama yang dihasilkannya jauh lebih baik untuk menjelaskan sebaran data awal dibandingkan metode klasik. Pada Tabel 1 disajikan nilai eigen untuk matriks peragam dari **X** pada n=100.

Pada Tabel 1 di bawah terlihat bahwa metode *robust* jauh lebih baik dibandingkan dengan metode klasik. Semakin besar prosentase pencilan mengakibatkan penyimpangan yang semakin besar terhadap nilai eigen metode klasik, sementara nilai eigen metode *robust* hanya sedikit terpengaruh. Untuk ukuran sampel dan prosentase pencilan lainnya memberikan hasil yang serupa seperti di atas sehingga tidak ditampilkan di sini.

Tabel 1. Nilai eigen dari matriks peragam dari **X** dengan n=100

Nilai Eigen	data awal	data dengan 5% pencilan		data dengan 10% pencilan	
		Klasik	VEM	Klasik	VEM
1	14,588879	31,143393	14,054304	49,868027	13,877666
2	2,7551894	10,041588	2,6013773	11,675904	2,6451922
3	7,06E-16	18,183967	0,0032795	2,0697765	0,0030162
4	-2,72E-16	0,0937708	0,002237	0,0853384	0,0028041
5	-2,18E-15	0,0476747	0,0021113	0,058925	0,0017368

Tabel 1 (lanjutan)

Data dengan 15% pencilan		Data dengan 20% pencilan	
klasik	VEM	klasik	VEM
70,452.787	13,452839	90,000848	12,033517
11,58036	2,6848096	12,210814	2,7617919
2,0126182	0,0028809	21,895263	0,002751
0,1756584	0,0023371	0,2636281	0,0023845
0,0713062	0,0022188	0,1998619	0,0021649

Tabel 2. Nilai dugaan koefisien RKU dengan jumlah pencilan 5%

n	Koefisien RKU		
	^(0)	^ klasik	^ VEM KTT
100	0,86566	2,10824	1,06507
300	0,87215	2,15147	0,87497
500	0,90962	2,18290	0,89653

Tabel 3. Nilai bias dan KTG koefisien RKU

N	Prosentase pencilan	Bias		KTG	
		Klasik	Robust	Klasik	Robust
100	0%	0,000405	0,087775	0,000009	0,007850
	5%	1,212908	0,033743	1,471467	0,001265
	10%	1,327582	0,079398	1,762496	0,006457
	15%	1,346974	0,085516	1,814347	0,001663
	20%	1,357721	0,109189	1,843410	0,012100
300	0%	0,000132	0,161447	0,000003	0,029577
	5%	1,282254	0,039634	1,644248	0,002745
	10%	1,358500	0,058209	1,805530	0,003438
	15%	1,346712	0,123704	1,813635	0,015334
	20%	1,342722	0,132010	1,842902	0,001529
500	0%	0,000058	0,128654	0,000002	0,022731
	5%	1,256842	0,063360	1,579686	0,004417
	10%	1,319958	0,048786	1,712291	0,002529
	15%	1,309704	0,059196	1,745324	0,004278
	20%	1,327861	0,078258	1,763215	0,006942

Pada tahap kedua, pencilan pada peubah respon Y mempengaruhi dugaan koefisien regresi. Sebagai contoh pada Tabel 2 di atas disajikan dugaan koefsien RKU pada setiap ukuran sampel dengan jumlah pencilan sebesar 5%.

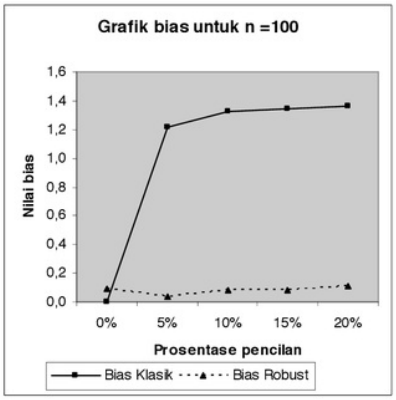
Pada Tabel 2 terlihat bahwa dugaan metode *robust* lebih mendekati dugaan data awal dbadngkan dengan metode klasik. Penyimpangan terhadap dugaan koefisien regresi akan menghasilkan model yang salah sehingga berdampak buruk pada peramalan.

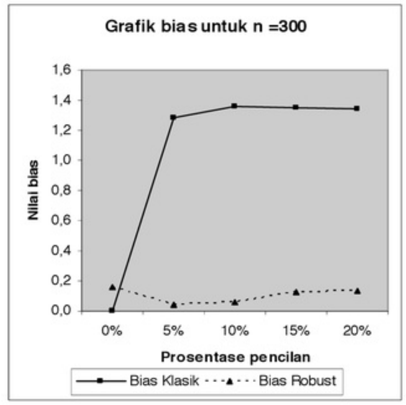
Untuk membandingkan nilai dugaan koefisien RKU secara keseluruhan kami melakukan replikasi sebanyak 100 kali untuk setiap ukuran sampel dan prosentase pencilan yang telah ditentukan. Berdasarkan nilai-nilai dugaan dari 100 replikasi diperoleh nilai bias dan MSE seperti pada Tabel 3.

Untuk data tanpa pencilan (0%), nilai bias dan KTG yang dihasilkan oleh metode *robust* VEM-KTT lebih besar dari metode klasik yang berarti bahwa metode klasik lebih baik dari metode *robust*. Namun nilai bias dan KTG metode VEM-KTT pada pencilan 0% masih relatif kecil yaitu bias < 0,2 dan KTG < 0,03 sehingga masih cukup baik sebagai penduga RKU.

Untuk data mengandung pencilan 5% - 20%, nilai bias dan KTG metode klasik lebih besar dari bias dan KTG metode *robust*. Dan perhatikan bahwa penambahan pencilan dalam data meningkatkan nilai bias dan KTG metode klasik yang berarti bahwa nilai dugaanya semakin buruk. Sedangkan nilai bias dan KTG metode *robust* VEM-KTT dan nilainya stabil < 0,2 untuk bias dan < 0,03 untuk KTG. Semakin kecil nilai bias suatu metode pendugaan menunjukkan bahwa nilai dugaan yang dihasilkan mendekati nilai parameter sebenarnya. Dan semakin kecil nilai KTG suatu metode pendugaan menunjukkan bahwa nilai dugaan yang dihasilkan semakin stabil. Untuk mempermudah pembandingan, data bias dan KTG pada Tabel 3 disajikan dalam bentuk grafik pada Gambar 1- 6 di bawah.

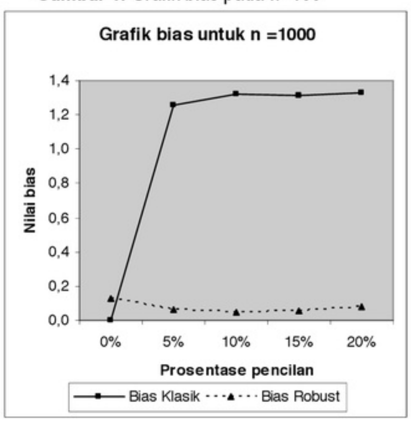
Pada Gambar 1- Gambar 3 dapat dilihat grafik nilai bias dari metode klasik sangat baik untuk pencilan 0% namun meningkat pesat pada saat data diberi pencilan 5% dan seterusnya. Sedangkan bias VEM-KTT pada pencilan 0% lebih besar dari nilai bias pada pencilan 5%-20%. Demikian pula nilai KTG pada Gambar 4 – Gambar 6, nilai KTG metode klasik memiliki pola yang sama dengan nilai biasnya, sedangkan nilai KTG metode *robust* VEM-KTT cenderung stabil < 0,03.

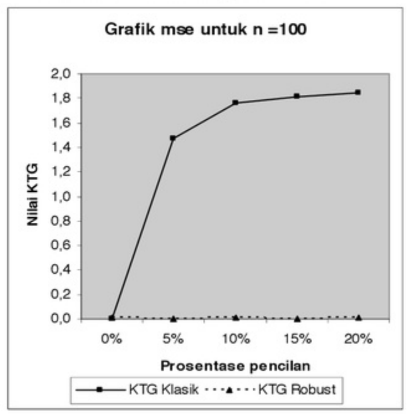




Gambar 1. Grafik bias pada n=100

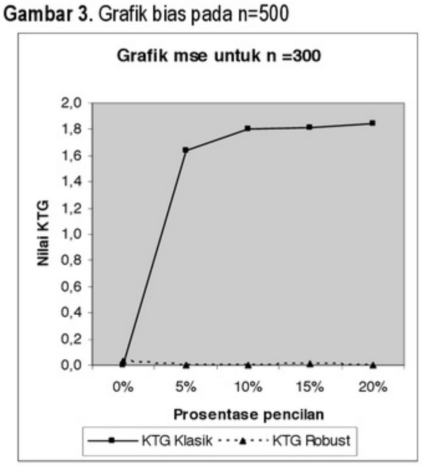
Gambar 2. Grafik bias pada n=300

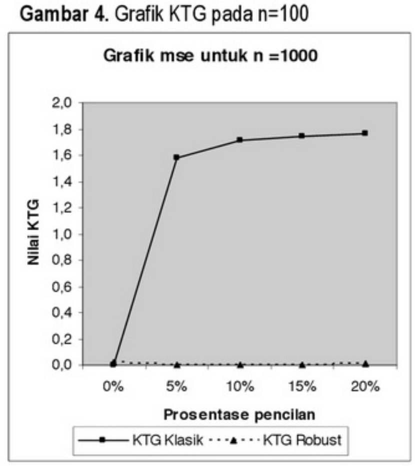




Gambar 3. Grafik bias pada n=500

Gambar 4. Grafik KTG pada n=100





Gambar 5. Grafik KTG pada n=300

Gambar 6. Grafik KTG pada n5100

4. KESIMPULAN

Analisis Regresi Komponen Utama terhadap data mengandung pencilan memerlukan metode *robust* agar menghasilkan dugaan yang tak bias dan stabil. Berdasarkan hasil simulasi data diperoleh bahwa metode VEM-KTT menghasilkan dugaan RKU yang sangat baik dan stabil untuk data dengan pencilan 5% - 20 %.

DAFTAR PUSTAKA

1. Myers, R.H. 1990. *Classical and Modern Regression with Applications.* PWS-KENT Publishing Company, Boston.
2. Filzmoser, P. 1999. Robust Principal Component and Factor Analysis in the Geostatistical Treatment of Environmental Data. *Environmetrics*, **10**: 363-375.
3. Croux, C., and Ruiz-Gazen, A. 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, **95**: 206-226.
4. Croux, C., Haesbroeck, G. 2000. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, *Biometrika*, **87** : 603-618
5. Hubert, M., Rousseeuw, P. J., Vanden Branden, K. 2005. ROBPCA: a new approach to robust principal components analysis. *Technometrics*, **47**:64–79.
6. Croux. C , P. Filzmoser, G. Pison & P.J. Rousseeuw. 2004. Fitting Multiplicative Models by Robust Alternating Regressions. *Technical Report* No. 350.
7. Rousseeuw, P.J., Van Aelst. S., Van Driessen, K., Agullio, J. 2004. Robust multivariate regression. *Technometrics* **46**: 293-305.
8. Rousseeuw, P.J. 1984. Least median of squares regression. *Journal of the American Statistical Association.* **79** (388): 871-880.
9. Frome, E. 2003. Least Absolute Values (LAV) Regression. http://www.epm.ornl.gov /~frome/E_L_ Frome LAV Regression.html
10. Hubert, M., Rousseeuw, P.J., Van Aelst, S. 2004. Robustness, *Encyclopedia of Actuarial Sciences*, edited by Sundt, B. and Teugels, J., Wiley, New York, pp. 1515-1529.
11. Rousseeuw, R. J. & A. M. Leroy. 1987. *Robust Regression and Outlier Detection.* New York: Wiley.
12. Filzmoser P. 2005. Identification of Multivariate Outliers: A Performance Study. *Austrian Journal Of Statistics*. **34** (2): 127–138
13. Nisa, K. 2006. Analisis Regresi Robust Menggunakan Metode *Least Trimmed Square* untuk Data Mengandung Pencilan. *Jurnal Ilmiah MIPA.* **IX** (2) : 93-100.
14. Chen, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG procedure. *SUGI Paper* 265-27. SAS Institute: Cary, NC. SAS OnLineDoc. SAS Institute, Cary, NC: IML Robust Regression, http://v8doc.sas.com/ sashtml
15. Rousseeuw, P.J., Van Zomeren, B.C. 1990. Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**:633-639.
16. Nisa, K., Herawati, N., Setiawan, E., Nusyirwan. 2006. Robust Principal Component Analysis Using Minimum Covariance Determinant Estimator. *Proceedings of International Conference on Mathematics and Natural Sciences*. pp 789-792.

ANALISIS REGRESI KOMPONEN UTAMA ROBUST UNTUK DATA MENGANDUNG PENCILAN

ORIGINALITY REPORT

22%

SIMILARITY INDEX

PRIMARY SOURCES

1	ml.scribd.com Internet	147 words — 5%
2	id.scribd.com Internet	38 words — 1%
3	journal.unair.ac.id Internet	32 words — 1%
4	eprints.undip.ac.id Internet	31 words — 1%
5	www.tandfonline.com Internet	29 words — 1%
6	repository.lppm.unila.ac.id Internet	28 words — 1%
7	www.crcsi.com.au Internet	25 words — 1%
8	Rand Wilcox. "Some small-sample properties of some recently proposed multivariate outlier detection techniques", Journal of Statistical Computation and Simulation, 2008 Crossref	24 words — 1%
9	Zuo, Yijun. "Robust Location and Scatter Estimators in Multivariate Analysis", Frontiers in Statistics, 2006. Crossref	22 words — 1%

10	library.binus.ac.id Internet	22 words — 1%
11	scholar.sun.ac.za Internet	20 words — 1%
12	es.scribd.com Internet	19 words — 1%
13	uwaterloo.ca Internet	18 words — 1%
14	ar.scribd.com Internet	17 words — 1%
15	www.sdu.edu.tr Internet	16 words — 1%
16	www.econ.kuleuven.ac.be Internet	16 words — 1%
17	wis.kuleuven.be Internet	15 words — 1%
18	edoc.pub Internet	15 words — 1%
19	loretonet.brinkster.net Internet	13 words — < 1%
20	lestarikanpujii.blogspot.com Internet	12 words — < 1%
21	documents.mx Internet	12 words — < 1%
22	repository.ipb.ac.id Internet	10 words — < 1%
23	www.pjsor.com Internet	9 words — < 1%

24	skripsi-skripsiun.blogspot.com Internet	9 words — < 1%
25	tugasdenny.wordpress.com Internet	8 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		