ANALISIS KOMPONEN UTAMA DENGAN MENGGUNAKAN MATRIK VARIAN KOVARIAN YANG *ROBUST*

Irwan Sujatmiko, Susanti Linuwih, dan Dwi Atmono A.W.
Jurusan Statistika ITS
Kampus ITS Sukolilo Surabaya 60111

Abstract. The present of outlier data causes the estimator of variance-covariance be overestimated. As a consequent, in the principle component analysis, the variability of the data in the main component becomes biger than as expected. To cope this condition, one can use robust estimator, i.e. MVE and MCD. Using simulation of Monte Carlo Experiments, the Principal Component Analysis using estimator MCD has the better performance than the estimator MVE and also classic estimator.

Keywords: Outlier, Principal Component Analysis, Minimum Volume Ellipsoid, Minimum Covariance Determinant, Monte Carlo Experiments.

1. PENDAHULUAN

Analisis Komponen Utama (AKU) merupakan suatu metode untuk menjelaskan struktur matrik varian-kovarian data melalui sejumlah kecil komponen yang tidak saling berkorelasi, dimana komponen tersebut merupakan kombinasi linier dari variable-variabel asal sedemikian hingga mempunyai variansi yang maksimal [5]. Seringkali AKU digunakan sebagai langkah pertama dalam analisis data, seperti dalam regresi komponen utama, analisis *discriminant*, analisis *cluster* dan berbagai metode analisis *multivariate* lainnya.

Namun, matrik varian-kovarian ternyata sangat sensitif terhadap pengamatan *outlier*. Pada kenyataannya, data lapangan seringkali mengandung beberapa pengamatan *outlier* dan biasanya tidak mudah untuk dipisahkan dari data. Oleh karena itu, penyusutan dimensi data berdasarkan AKU yang klasik (AKU-kla) menjadi tidak dapat diandalkan lagi apabila pengamatan *outlier* muncul dalam data.

Tujuan dari AKU yang *robust* (AKU-rob) adalah mendapatkan komponen utama yang tidak banyak dipengaruhi oleh pengamatan *outlier*. Pendekatan pertama dilakukan dengan menggantikan matrik varian-kovarian yang klasik dengan suatu estimator yang *robust*. Berbagai metode untuk mendapatkan estimator yang *robust* telah banyak ditawarkan. Namun, *breakdown point* (tingkat *robust*) yang dimiliki tidak lebih dari $1/(p+1)$, dimana p merupakan dimensi data [3]. Hal ini berarti bahwa untuk dimensi yang besar, suatu kelompok pencemar yang sangat kecil dapat menghasilkan estimasi yang masih dipengaruhi oleh pengamatan *outlier*. *Minimum Volume Ellipsoid* (MVE) dan *Minimum Covariance Determinant* (MCD) dikenalkan oleh [4] sebagai suatu estimator yang *affine equivariant* dengan *breakdown point* yang lebih tinggi. Pendekatan lainnya untuk mendapatkan komponen utama yang *robust* adalah dengan menggunakan *Projection Pursuit* [1]. Selanjutnya, [2] menawarkan pendekatan dua tahap dengan menggabungkan kedua pendekatan tersebut. Penelitian ini mencoba membandingkan performansi antara AKU yang menggunakan estimator MVE dan MCD dengan AKU yang klasik.

2. ANALISIS KOMPONEN UTAMA YANG *ROBUST*

Untuk mendapatkan AKU yang *robust* dapat dilakukan dengan menggantikan matrik varian kovarian yang klasik dengan suatu estimator yang *robust*. Estimator *robust* yang diinginkan adalah memiliki *breakdown point* (tingkat *robust*) yang

tinggi. Secara sederhana, *breakdown point* merupakan proporsi terkecil dari pencemar yang mampu mempengaruhi estimator, yaitu nilainya mengalami pergeseran yang cukup jauh dari $t(\mathbf{X})$. Sifat lainnya yang diinginkan adalah *affine equivariant*. Suatu estimator bagi ukuran pemusatan, yaitu $t(\mathbf{X})$ dan ukuran simpangan, yaitu $\mathbf{C}(\mathbf{X})$ adalah *affine equivariant* jika dan hanya jika untuk setiap vektor baris $\mathbf{b} \in \Re^p$ (konstanta dalam ruang berdimensi p) dan setiap matrik non-singular $\mathbf{A}_{p\times p}$ berlaku,

$$t(AX+b) = At(X)+b, \qquad (2.1)$$

$$C(AX+b) = AC(X)A^t. \qquad (2.2)$$

Estimator *Robust Minimum Volume Ellipsoid* (MVE) dan *Minimum Covariance Determinant* (MCD) telah dibahas [4].

Estimator MVE merupakan pasangan $t(\mathbf{X})$ dan $\mathbf{C}(\mathbf{X})$, dimana $t(\mathbf{X})$ merupakan vektor rata-rata dan $\mathbf{C}(\mathbf{X})$ merupakan matrik $p\times p$ simetris definit-positif, dari suatu sub-sampel berukuran h pengamatan dimana volume ellipsoid dari sub-sampel tersebut adalah yang minimal. Dengan $h_0 \le h \le n$ dan h_0 merupakan nilai integer terkecil dari $((n+p+1)/2)$.

$$MVE \approx \min\left\{m_j^{2p}\det\left(\mathbf{C}(\mathbf{X})_j\right)\right\}^{1/2},$$

$$j = 1,\ldots,\binom{n}{h}. \qquad (2.3)$$

Estimator MCD hampir mirip dengan estimator MVE, perbedaan hanya terletak pada kriteria pemilihannya. Estimator MCD merupakan pasangan $t(\mathbf{X})$ dan $\mathbf{C}(\mathbf{X})$, dimana $t(\mathbf{X})$ merupakan vektor rata-rata dan $\mathbf{C}(\mathbf{X})$ merupakan matrik $p\times p$ simetris definit-positif, dari suatu subsampel berukuran h pengamatan dimana volume ellipsoid dari sub-sampel tersebut adalah yang minimal. Dengan $h_0 \le h \le n$ dan h_0 merupakan nilai integer terkecil dari $((n+p+1)/2)$.

$$MCD \approx \min\left\{\det\left(\mathbf{C}(\mathbf{X})_j\right)\right\},\ j = 1,\ldots,\binom{n}{h}. \qquad (2.4)$$

Menggunakan persamaan eigen matriks varian kovarian yang *robust* tersebut, diturunkan nilai eigen dan vektor eigen yang bersesuaian. Nilai eigen dan vektor eigen yang diperoleh merupakan nilai eigen dan vektor eigen yang *robust*. Komponen utama yang diperoleh merupakan komponen utama yang *robust*.

3. SIMULASI

Untuk membandingkan performansi AKU yang klasik dengan AKU yang *robust*, yaitu AKU-MVE dan AKU-MCD, dilakukan simulasi *Monte Carlo Experiments*. Dilakukan simulasi dengan membangkitkan 100 ($n=100$) pengamatan berdimensi 4 ($p=4$) yang dilakukan secara berulang sebanyak 250 ($r=250$) kali.

Data simulasi yang digunakan dalam *Monte Carlo Experiment* dibangkitkan berdasarkan model,

$$(1-\varepsilon)N_p(\mu,\boldsymbol{\Sigma}) + \varepsilon N_p(\tilde{\mu},\tilde{\boldsymbol{\Sigma}}). \qquad (3.1)$$

untuk berbagai nilai n, p, ε, μ, $\boldsymbol{\Sigma}$, $\tilde{\mu}$ dan $\tilde{\boldsymbol{\Sigma}}$. Yaitu, $n(1-\varepsilon)$ pengamatan dibangkitkan berdasarkan distribusi normal multivariate $N_p(\mu,\boldsymbol{\Sigma})$ dan $n\varepsilon$ pengamatan dibangkitkan berdasarkan distribusi normal multivariate $N_p(\tilde{\mu},\tilde{\boldsymbol{\Sigma}})$.

Contoh dalam simulasi ini digunakan $\mu = [0 \quad 0 \quad 0 \quad 0], \boldsymbol{\Sigma} = Diag[8 \quad 4 \quad 2 \quad 1]$. Dari matriks varian-kovarian tersebut diperoleh nilai eigen $\lambda_1 = 8$, $\lambda_2 = 4$, $\lambda_3 = 2$, dan $\lambda_4 = 1$. Nilai eigen tersebut merupakan nilai yang akan diestimasi dengan menggunakan data hasil bangkitan untuk dibandingkan hasilnya. Sehingga apabila ditentukan tiga ($k=3$) komponen utama, maka proporsi variabilitas data yang dapat diterangkan oleh tiga komponen utama pertama adalah 93,33%. Simulasi dilakukan untuk berbagai situasi berikut.

1. $\varepsilon = 0$, (tidak ada pengamatan outlier dalam data).

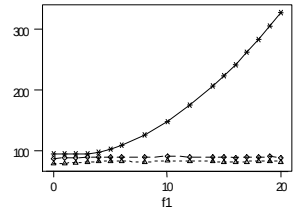
2. $\varepsilon = 10\%$ atau $\varepsilon = 20\%$,
   $\tilde{\mu} = [0 \quad 0 \quad 0 \quad f_1]$, dan $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} / f_2$,
   dengan $f_1 = 0,1,2,\ldots,18,19,20$ dan $f_2 = 1$ atau $f_2 = 15$.

Setiap situasi simulasi, dibandingkan untuk setiap metode (AKU-kla, AKU-MVE, dan AKU-MCD) sebagai berikut.

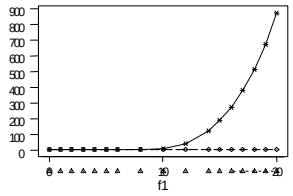
1. Menghitung proporsi variabilitas yang dapat diterangkan berdasarkan estimasi nilai eigen. Dilakukan dengan membandingkan jumlah tiga nilai eigen terbesar hasil estimasi dengan jumlah semua nilai eigen dari populasi.
2. Hasilnya diberikan dalam bentuk rata-rata dari 250 pengulangan.
$$\frac{1}{250}\sum_{r=1}^{250}\frac{\hat{\lambda}_1^{(r)} + \hat{\lambda}_2^{(r)} + \hat{\lambda}_3^{(r)}}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}, \qquad (3.2)$$
   dengan $\hat{\lambda}_i^{(r)}$ merupakan nilai eigen hasil estimasi pada pengulangan ke-r.
3. Untuk 3 nilai eigen terbesar, dihitung juga *mean square error* (MSE), yang dinyatakan sebagai
$$MSE\left(\hat{\lambda}_i\right) = \frac{1}{250}\sum_{r=1}^{250}\left(\hat{\lambda}_i^{(r)} - \lambda_i\right), \qquad i = 1,2,3 \qquad (3.3)$$

Gambar 1 memperlihatkan hasil simulasi untuk situasi $\varepsilon = 10\%$ dengan f_1=0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20 dan $f_2 = 1$. Terlihat bahwa semakin jauh pusat pengamatan outlier bergeser dari pusat distribusi dasar, AKU-kla mengalami lonjakan pada rata-rata variabilitas yang dapat diterangkan oleh tiga komponen utama. Nilai idel untuk proporsi variabilitas yang dapat diterangkan oleh tiga komponen utama pertama adalah 93,33%. Sementara AKU-MVE dan AKU-MCD memberikan nilai yang lebih stabil.

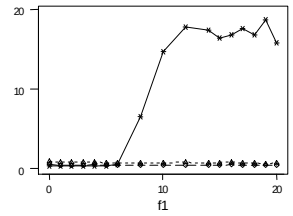
(a)



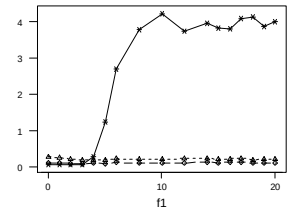
(b)



(c)



(d)



Gambar 1. Pengaruh *shift outlier* pada (a) Proporsi Variabilitas, (b) $MSE\left(\hat{\lambda}_1\right)$,(c) $MSE\left(\hat{\lambda}_2\right)$, (d) $MSE\left(\hat{\lambda}_3\right)$

Keterangan: (*) klasik, (♦) MVE, (Δ) MCD.

Ketepatan estimasi nilai eigen dari sampel terhadap nilai eigen yang dimiliki populasi dapat dinyatakan dengan *Mean Square Error* (MSE). MSE yang diinginkan adalah sekecil mungkin, idealnya adalah 0 (nol). Gambar 1 juga memperlihatkan pergerakan $MSE\left(\hat{\lambda}_1\right)$, $MSE\left(\hat{\lambda}_2\right)$, $MSE\left(\hat{\lambda}_3\right)$ apabila pusat pengamatan outlier bergeser semakin jauh dari pusat distribusi dasar. Bahwa semakin jauh pusat pengamatan outlier bergeser dari pusat distribusi dasar, AKU-kla memiliki kesalahan yang semakin besar untuk $f_1 > 10$. Sementara untuk AKU-MVE dan AKU-MCD memberikan kesalahan estimasi yang kecil dan lebih stabil.

Dari Tabel 1, apabila pengamatan outlier tidak ada dalam data, terlihat bahwa AKU-Kla memberikan rata-rata proporsi variabilitas yang lebih baik dibandingkan AKU yang *robust*, yaitu AKU-MVE dan AKU-MCD. Proporsi variabilitas yang dimiliki AKU-kla untuk $\varepsilon = 0$ adalah 93,64% sangat dekat dengan nilai yang ditaksir, yaitu 93,33%. Namun, apabila pengamatan outlier muncul dalam data, rata-rata proporsi variabilitas AKU-kla mengalami *overestimated*, yaitu nilainya lebih dari 100%. Sementara untuk AKU-MVE dan AKU-MCD memberikan hasil yang lebih *robust*, namun AKU-MVE mengalami *overestimated* apabila proporsi banyaknya pencemar adalah 20% ($\varepsilon = 20\%$) dengan bentuk distribusi data (matrik varian kovarian) yang lebih kecil ($\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}/f_2$ untuk $f_2 = 15$).

Tabel 2. memperlihatkan rata-rata MSE untuk estimasi nilai eigen $\hat{\lambda}_1$, $\hat{\lambda}_2$, dan $\hat{\lambda}_3$. Apabila pengamatan outlier muncul dalam data, rata-rata MSE untuk $\hat{\lambda}_1$ dari MVE memberikan hasil yang paling kecil. Artinya, nilai estimasi memiliki nilai yang cukup dekat dengan nilai sebenarnya, yaitu $\lambda_1 = 8$. Ketika $\varepsilon = 20\%$, $f_1 = 10$ $f_2 = 15$, rata-rata MSE untuk $\hat{\lambda}_1$ metode MVE mengalami peningkatan yang tinggi dibandingkan metode MCD.

Apabila pengamatan outlier muncul dalam data, rata-rata MSE untuk $\hat{\lambda}_2$ memberikan nilai yang kecil untuk metode AKU-MVE dan AKU-MCD. Namun, apabila pengamatan outlier tidak ada dalam data maka ketiga metode memberikan nilai MSE untuk $\hat{\lambda}_2$ yang kecil. Demikian juga MSE untuk $\hat{\lambda}_3$ memberikan hasil yang tidak jauh berbeda, lihat Tabel 2.

4. DATA SUSENAS

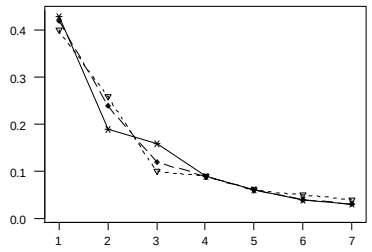
Menggunakan data SUSENAS Kab. Pasuruan tahun 2001, yang terdiri atas 286 pengamatan dengan 7 variabel pengeluaran rumah tangga untuk konsumsi makanan. Sebanyak 286 pengamatan merupakan hasil seleksi dengan tidak menyertakan pengamatan yang mengandung nilai kosong (*missing value*).

Tabel 1. Rata-rata proporsi variabilitas yang dapat diterangkan oleh tiga komponen utama pertama.

Kondisi	AKU-kla	AKU-MVE	AKU-MCD
$\varepsilon = 0$	93,64	87,29	78,39
$\varepsilon = 10\%$ $f_1 = 10$ $f_2 = 1$	147,95	89,23	81,68
$\varepsilon = 10\%$ $f_1 = 10$ $f_2 = 15$	139,86	89,21	81,59
$\varepsilon = 20\%$ $f_1 = 10$ $f_2 = 1$	194,85	91,76	86,44
$\varepsilon = 20\%$ $f_1 = 10$ $f_2 = 15$	178,52	101,05	86,81

Tabel 2. Rata-rata MSE untuk masing-masing nilai eigen dari tiga komponen utama

Kondisi	AKU-kla			AKU-MVE			AKU-MCD		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
$\varepsilon = 0$	1,10	0,32	0,10	1,43	0,46	0,13	2,71	0,93	0,31
$\varepsilon = 10\%$ $f_1 = 10$ $f_2 = 1$	7,02	14,57	4,07	1,40	0,38	0,14	2,31	0,75	0,23
$\varepsilon = 10\%$ $f_1 = 10$ $f_2 = 15$	4,66	11,97	2,61	1,61	0,41	0,12	2,30	0,77	0,24
$\varepsilon = 20\%$ $f_1 = 10$ $f_2 = 1$	87,53	17,64	3,67	1,63	0,37	0,12	1,86	0,53	0,18
$\varepsilon = 20\%$ $f_1 = 10$ $f_2 = 15$	81,42	7,81	1,61	14,36	0,74	0,17	3,72	0,50	0,17



Gambar 2. *Scree plot* untuk data SUSENAS

Keterangan: (*) klasik, (♦) MVE, (Δ) MCD

Identifikasi outlier yang dilakukan, sebanyak 86 pengamatan teridentifikasi sebagai pengamatan outlier berdasarkan metode MVE dan 110 pengamatan teridentifikasi sebagai pengamatan outlier berdasarkan metode MCD.

Berdasarkan *scree plot* pada Gambar 2 terlihat ada penurunan yang tajam dan besar dari komponen pertama ke komponen kedua untuk AKU-kla. Sedangkan dari AKU-MVE dan AKU-MCD terlihat bahwa titik nilai eigen dari komponen pertama ke komponen kedua lebih landai dan lebih kecil. Hal tersebut juga dapat dilihat pada Tabel 3.

5. PENUTUP

Berdasarkan simulasi, AKU-MCD memberikan hasil yang lebih *robust* dibandingkan dua metode lainnya, karena memberikan hasil yang lebih stabil dilihat dari proporsi variabilitas yang dapat diterangkan oleh tiga komponen utama dan berdasarkan MSE nilai eigen. AKU-MVE menjadi tidak *robust* ketika pengamatan outlier dalam data sebesar 20% dengan bentuk distribusi yang lebih kecil dibandingkan bentuk distribusi dari distribusi dasar.

Analisis Komponen Utama yang *Robust* dapat digunakan karena memberikan hasil yang lebih baik. Penelitian dapat dikembangkan untuk berbagai proporsi pengamatan outlier yang lebih besar dari 20% dengan berbagai jenis vektor rata-rata dan bentuk matrik varian kovarian.

5. DAFTAR PUSTAKA

[1] Hubert, M., Rousseeuw, P. J., Verboven, S. (2001), *A Fast Method for Robust Principal Component with Applications to Chemometrics, Revised Version.* www.kuleuven.com, download 15 April 2005.

[2] Hubert, M., Rousseeuw, P. J., Vanden Branden, K. (2005), ROBPCA: *A new Approach to Robust Principal Componen Analysis*, Technometrics, **47**: 64-78.

[3] Maronna, R. A. (1976), *Robust Estimator of Multivariate Location and Scatter*, Ann. Statist., **4**: 51-67.

[4] Rousseeuw, P. J. (1985), *Multivariate Estimation with High Breakdown Point, In: W. Grossmann, G. Pflug, T. Vincze and W. Werz, Eds.*, Mathematical Statistics and Applications, Vol. B. Reidel, Dordrecht.

[5] Timm, N. H. (1975), *Multivariate Analysis with Applications in Education and Psychology*, California: Brooks/ Cole Publishing.

Tabel 3. Eigen value untuk data SUSENAS

EigenValue	Klasik	MVE	MCD
1	39,697,060.00	15,035,340.00	10,776,889.00
2	18,003,091.00	8,639,930.80	7,086,092.90
3	14,544,914.00	4,167,479.10	2,836,911.20
4	8,163,042.10	3,328,196.80	2,365,458.60
5	5,564,782.60	2,101,058.00	1,610,140.20
6	3,804,492.70	1,287,175.60	1,332,402.40
7	2,748,008.00	1,169,023.50	1,070,267.30

Jurnal Matematika Vol. 8, No.2, Agustus 2005: 37-41