PAPER

Analisis Data Outlier

Pada Data Pengeluaran Rumah Tangga

Di Kota Kupang, Nusa Tenggara Timur Tahun 2005

Dengan Metode ROBPCA

oleh

Suryana

NRP. 1306 201 710

PROGRAM STUDI MAGISTER
JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA

2007

1. Pendahuluan

Analisis komponen utama (*principal component analysis*=PCA) merupakan teknik statistik yang telah banyak digunakan. Pembahasan materi ini dapat ditemukan pada buku-buku analisis multivariat baik teori maupun aplikasi. Pemanfaatan PCA utamanya dalam menyusutkan dimensi data yang saling berkorelasi satu sama lain.

Perkembangan PCA dimulai sejak diperkenalkan pertama kali oleh Pearson pada tahun 1901. Sejalan dengan perkembangan teknologi komputer dan kemajuan di bidang matematika, PCA hingga kini masih terus mengalami perkembangan. Perkembangan selanjutnya, diperkenalkan generalisasi dari PCA oleh Loève pada tahun 1963.

Perkembangan PCA selanjutnya dipengaruhi adanya kebutuhan model PCA yang robust terhadap data pencilan (*outlier*). PCA klasik (CPCA) sangat dipengaruhi oleh kehadiran pencilan karena CPCA didasarkan pada matrik kovarian yang juga sangat sensitif terhadap keberadaan data pencilan. Untuk mengatasi masalah ini, matrik kovarian diestimasi estimasi kovarian yang robust dengan M-estimator (Devlin, dkk, 1975), Minimum Covariance Determinant (Rousseeuw, 1984), atau S-estimator (Croux dan Haesbroek, 1999). Ketiga metode ini baik jika digunakan untuk kasus jumlah variabel $p >$ jumlah observasi n. Untuk kasus ini, Li dan Chen (1985) memperkenalkan robust PCA dengan *Projection Pursuit* (PP).

Konsep PP selanjutnya diperbaiki dengan Algoritma C-R, diperkenalkan oleh Croux dan Ruiz-Gazen (1996). Algoritmanya bekerja dengan baik, namun pada dimensi yang tinggi, algoritma C-R lambat dalam waktu komputasi dan tidak stabil secara numerik. Pada tahun 2004, Hubert dkk menggabungkan konsep PP dan estimator kovarian yang robust dengan nama ROBPCA.

Tertarik dengan konsep ROBPCA yang diajukan Hubert dkk, paper ini mengkaji konsep ROBPCA dan membandingkan hasilnya dengan CPCA. Kedua metode PCA tersebut diaplikasikan pada data pengeluaran rumah tangga hasil Survei Sosial Ekonomi Nasional (SUSENAS) tahun 2005 di Kota Kupang, Nusa Tenggara Timur.

2. Metode ROBPCA

Diasumsikan data disimpan dalam sebuah vektor random $\mathbf{X}$ di mana elemen baris menyatakan n observasi dan banyaknya kolom menyatakan p variabel $\boldsymbol{X_1, \ldots, X_p}$. Metode ROBPCA memproses data multivariat dalam tiga tahap. Pertama, data diproses sedemikian rupa sehingga data hasil transformasi terletak dalam subruang yang dimensinya kurang dari $n - 1$. Selanjutnya matrik covarian awal $\mathbf{S_0}$ dibangun dan digunakan untuk menyeleksi jumlah komponen k yang akan diseleksi. Langkah ini menghasilkan subruang berdimensi k yang cocok dengan data.

Langkah selanjutnya, titik-titik data diproyeksikan pada subruang ini di mana lokasi dan *scatter matrix* diestimasi secara robust dari nilai eigen *k* positif $l_1, \ldots, l_k$. Vektor eigen yang bersesuaian nilai eigen yang dihasilkan membentuk komponen utama yang robust.

Tahapan ROBPCA selengkapnya sebagai berikut:

Tahap 1. Sebagaimana diproposed oleh Hubert dkk (2002), kita mulai dengan mereduksi ruang data menjadi subruang **affine** yang direntang oleh *n* observasi. Langkah ini berguna ketika $p \geq n$, tetapi meskipun $p < n$, observasi dapat merentangkan kurang dari ruang dimensi-*p* secara menyeluruh. Cara mudah melakukan hal ini dengan *singular value decompotision* (SVD) dari matrik data terpusat di rata-rata (*the mean centered-data matrix*) menghasilkan

$$\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\mu}_0^t = \mathbf{U}_{n,r_0} \mathbf{D}_{r_0,r_0} \mathbf{V}_{r_0,p}^t \tag{0.1}$$

Di mana $\hat{\mu}_0$ vektor mean, $r_0 = \text{rank}(\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\mu}_0^t)$, **D** matrik diagonal r_0 x r_0, dan $\mathbf{U}^{\text{t}}\mathbf{U} = \mathbf{I}_{r_0} = \mathbf{V}^{\text{t}}\mathbf{V}$, di mana $\mathbf{I}_{r_0}$ matrik identitas berdimensi r_0 x r_0. Pada saat $p > n$, penyelesaian (1.1) menggunakan pendekatan kernel berdasarkan penghitungan vektor dan nilai eigen dari $(\mathbf{X} - \mathbf{1}_n \hat{\mu}_0^t)(\mathbf{X} - \mathbf{1}_n \hat{\mu}_0^t)^{\text{t}}$ (Wu, Massart, dan de Jong 1997). Tanpa kehilangan informasi, sekarang data telah berada di subruang yang dibangkitkan oleh r_0 kolom **V**, yaitu $\mathbf{Z}_{n,r_0} = \mathbf{UD}$.

Tahap 2. Pada tahan ini dicari paling sedikit $h < n$ titik-titik data pencilan. Besarnya $h = \max\{[\alpha n], [(n + k_{\max} + 1)/2\}$, di mana $k_{\max}$ menyatakan banyaknya komponen *k* optimal. Parameter α dapat diambil antara 0.5 dan 1.

Untuk menemukan paling sedikit *h* titik-titik data pencilan, dilakukan melalui:

1. Untuk setiap titik data $\mathbf{x}_i$ dihitung *outlyingness* dengan rumusan Stahel-Donoho sebagai berikut:

$$\text{outl}_{\Lambda}(\mathbf{x}_i) = \max_{\mathbf{v} \in B} \frac{\left| \mathbf{x}_i^{\text{t}}\mathbf{v} - \text{med}(\mathbf{x}_j^{\text{t}}\mathbf{v}) \right|}{\text{mad}(\mathbf{x}_j^{\text{t}}\mathbf{v})} \tag{0.2}$$

   Di mana: *B* mengandung semua vektor bukan nol, $\text{med}(\mathbf{x}_j^{\text{t}}\mathbf{v})$ adalah median dari $\left[\mathbf{x}_j^{\text{t}}\mathbf{v},\ j = 1, \ldots, n \right]$ dan $\text{mad}(\mathbf{x}_j^{\text{t}}\mathbf{v}) = \text{med}\left| \mathbf{x}_j^{\text{t}}\mathbf{v} - \text{med}(\mathbf{x}_j^{\text{t}}\mathbf{v}) \right|$. Hasil (0.2) belum ortogonal. Untuk menjadikan ortogonal (0.2) dimodifikasi menjadi:

$$\text{outl}_{\text{O}}(\mathbf{x}_i) = \max_{\mathbf{v} \in B} \frac{\left| \mathbf{x}_i^{\text{t}}\mathbf{v} - t_{MCD}(\mathbf{x}_j^{\text{t}}\mathbf{v}) \right|}{S_{MCD}(\mathbf{x}_j^{\text{t}}\mathbf{v})} \tag{0.3}$$

   Di mana t_{MCD} dan S_{MCD} adalah mean dan standar deviasi yang *robust* dari *minimum covarian determinan* . Jika semua S_{MCD} tidak nol maka dapat dihitung (0.3) untuk semua titik data dan menganggap *h* observasi adalah pencilan yang paling kecil. Indeksnya

kemudian di simpan dalam himpunan H_0. Sebaliknya, jika S_{MCD} ada nilai nol maka akan ditemukan suatu *hyperplane* H_v yang ortogonal dengan **v**. Pada saat ini terjadi, semua titik data diproyeksikan pada H_v. Langkah ini diselesaikan dengan Reflection Algorithm oleh Hubert dkk (2002) sampai didapatkan himpunan H_0 yang beranggotakan h titik data dengan pencilan paling sedikit.

2. Dari H_0 didapatkan mean μ_1^t dan covarian $\mathbf{S}_0$. Dengan cara yang serupa, dicari vektor eigen dan nilai eigen yang bersesuaian dari matrik covarian $\mathbf{S}_0$. $\mathbf{S}_0$ selanjutnya didekomposisi spektral

$$\mathbf{S}_0 = \mathbf{P}_0 \mathbf{L}_0 \mathbf{P}_0^t \tag{0.4}$$

Dengan **L** = diag($\tilde{l}_1,...,\tilde{l}_n$) dan $r \le r_1$.

Matrik covarian $\mathbf{S_0}$ digunakan untuk menentukan berapa banyak komponen utama yang diperlukan. Salah satu metode penentuan jumlah komponen utama adalah dengan *Screeplot*.

3. Sebagai langkah akhir, dilakukan proyeksi titik data pada subruang yang dibangkitkan oleh k_0 komponen vektor eigen dari $\mathbf{S_0}$. Hasilnya dinyatakan sebagai

$$\mathbf{X}^*_{n,k_0} = \left(\mathbf{X}_{n,r_1} - \mathbf{1}_n \hat{\mu}_1^{\mathbf{t}}\right)\mathbf{P}_{r_1,k_0} \tag{0.5}$$

Di mana $\mathbf{P}_{r_1,k_0}$ terdiri dari k_0 kolom pertama dari $\mathbf{P_0}$ persamaan (0.4).

Tahap 3. Pada tahapan ini dilakukan pengukuran Score Distance, SDi dan Orthogonal Distance, Oi untuk dapat menentukan jenis observasi. Untuk setiap observasi ke i, dapat ditampilkan pada sumbu-x jarak SDi dengan rumusan

$$SD_i = \sqrt{t_i^t L^{-1} t_i} \tag{0.6}$$

Dan pada sumbu-*y* jarak OD_i dengan rumusan

$$OD_i = \left\| \mathbf{x}_i - \hat{\mu}_x - \mathbf{P}_{p,k}\mathbf{t}_i \right\| \tag{0.7}$$

Kombinasi kedua jarak ini menghasilkan jenis observasi seperti berikut:

Jarak	SD kecil	SD besar
OD besar	Outlier ortogonal	Titik leverage – PCA buruk
OD kecil	Observasi regular	Titik leverage – PCA baik

3. Aplikasi ROBPCA

Pada sesi ini dilakukan penerapan metode ROBPCA pada data pengeluaran rumah tangga hasil Survei Sosial Ekonomi Nasional tahun 2005 di Kota Kupang, Nusa Tenggara Timur. Analisis didahului dengan Analisis Komponen Utama Klasik (*Classic Principal Component Analysis, CPCA*). Jumlah responden rumah tangga sebanyak 607. Variabel yang dianalisis adalah :

X1 = Pengeluaran rumah tangga untuk konsumsi padi-padian.
X2 = Pengeluaran rumah tangga untuk konsumsi umbi-umbian
X3 = Pengeluaran rumah tangga untuk konsumsi ikan
X4 = Pengeluaran rumah tangga untuk konsumsi daging
X5 = Pengeluaran rumah tangga untuk konsumsi telur dan susu
X6 = Pengeluaran rumah tangga untuk konsumsi sayur-sayuran
X7 = Pengeluaran rumah tangga untuk konsumsi kacang-kacangan
X8 = Pengeluaran rumah tangga untuk konsumsi buah-buahan
X9 = Pengeluaran rumah tangga untuk konsumsi minyak dan lemak
X10 = Pengeluaran rumah tangga untuk konsumsi bahan minuman
X11 = Pengeluaran rumah tangga untuk konsumsi bumbu-bumbuan
X12 = Pengeluaran rumah tangga untuk konsumsi konsumsi lainnya
X13 = Pengeluaran rumah tangga untuk konsumsi makanan dan minuman jadi
X14 = Pengeluaran rumah tangga untuk konsumsi tembakau dan sirih
X15 = Pengeluaran rumah tangga untuk konsumsi perumahan dan fasilitas rumahtangga
X16 = Pengeluaran rumah tangga untuk konsumsi aneka barang dan jasa
X17 = Pengeluaran rumah tangga untuk konsumsi pakaian alas kaki dan tutup kepala
X18= Pengeluaran rumah tangga untuk konsumsi barang tahan lama
X19 = Pengeluaran rumah tangga untuk konsumsi pajak, Pungutan dan Asuransi
X20 = Pengeluaran rumah tangga untuk konsumsi keperluan pesta dan upacara

3.1 Analisis Komponen Utama Klasik

Hasil pengukuran variabel sangat penting dieksplorasi untuk menentukan keputusan apakah analisis komponen utama didasarkan pada data mentah atau perlu dilakukan modifikasi tertentu untuk keperluan analisis.

Variable	Total	Mean	SE Mean	StDev	Minimum	Maximum	Range
X1	607	46252	2253	55499	0.000000000	467143	467143
X2	607	2331	193	4748	0.000000000	43929	43929
X3	607	20934	1010	24884	0.000000000	235714	235714
X4	607	12408	910	22411	0.000000000	214286	214286
X5	607	15289	878	21625	0.000000000	214286	214286
X6	607	19187	613	15093	0.000000000	98571	98571
X7	607	7093	398	9816	0.000000000	120000	120000
X8	607	7493	577	14222	0.000000000	96429	96429
X9	607	7946	249	6134	0.000000000	38571	38571
X10	607	11571	504	12406	0.000000000	150000	150000
X11	607	5051	214	5279	0.000000000	52500	52500
X12	607	5299	356	8770	0.000000000	60000	60000
X13	607	24989	1705	42004	0.000000000	330000	330000
X14	607	17172	1362	33548	0.000000000	360000	360000
X15	607	91062	4739	116754	6333	1751667	1745333
X16	607	56937	2656	65442	1750	754167	752417
X17	607	9757	597	14699	0.000000000	208333	208333
X18	607	8560	871	21470	0.000000000	204167	204167
X19	607	4053	524	12918	0.000000000	237500	237500
X20	607	8113	1783	43938	0.000000000	833333	833333

Pengukuran variabel pada Tabel 4.1 menghasilkan range yang nyata berbeda. Selain itu pengukuran 14 variabel pertama dalam satuan minggu sedang 6 variabel sisanya dalam bulan. Menurut Johnson (2002), dengan sifat data seperti itu analisis komponen utama sebaiknya didasarkan pada data yang distandardisasi. Deskriptif statistik dari data hasil standardisasi tampak sebagai berikut.

Variable	Tabel	Mean	SE Mean	StDev	Minimum	Maximum	Range
Z1	607	-5.73489E-16	0.0406	1.0000	-0.8334	7.5838	8.4172
Z2	607	6.28837E-16	0.0406	1.0000	-0.4910	8.7616	9.2526
Z3	607	1.46584E-16	0.0406	1.0000	-0.8413	8.6312	9.4725
Z4	607	-2.01445E-16	0.0406	1.0000	-0.5536	9.0078	9.5615
Z5	607	-2.72135E-17	0.0406	1.0000	-0.7070	9.2023	9.9093
Z6	607	5.25120E-16	0.0406	1.0000	-1.2713	5.2595	6.5308
Z7	607	-3.89229E-16	0.0406	1.0000	-0.7226	11.5022	12.2248
Z8	607	1.43332E-16	0.0406	1.0000	-0.5268	6.2533	6.7801
Z9	607	-4.51462E-16	0.0406	1.0000	-1.2955	4.9930	6.2885
Z10	607	-6.49112E-16	0.0406	1.0000	-0.9326	11.1578	12.0905
Z11	607	-7.45172E-16	0.0406	1.0000	-0.9568	8.9885	9.9453
Z12	607	-3.45779E-16	0.0406	1.0000	-0.6043	6.2375	6.8418
Z13	607	7.06697E-16	0.0406	1.0000	-0.5949	7.2615	7.8565
Z14	607	9.29161E-17	0.0406	1.0000	-0.5119	10.2191	10.7309
Z15	607	3.86626E-16	0.0406	1.0000	-0.7257	14.2231	14.9488
Z16	607	-1.69136E-16	0.0406	1.0000	-0.8433	10.6542	11.4975
Z17	607	-6.96058E-16	0.0406	1.0000	-0.6638	13.5095	14.1732
Z18	607	1.38669E-16	0.0406	1.0000	-0.3987	9.1106	9.5093
Z19	607	-1.62522E-16	0.0406	1.0000	-0.3138	18.0710	18.3848
Z20	607	6.23416E-18	0.0406	1.0000	-0.1846	18.7813	18.9660

Selanjutnya dihitung matrik kovarian sampel $\mathbf{S}_z$. Hasilnya ditampilkan pada output berikut. Ternyata, matrik kovarian sampel dari data terstandardisasi sama dengan matrik korelasi sampel **R**. Dengan demikian total sampel matrik kovarian sampel $\mathbf{S_z}$ sama dengan jumlah variabel yaitu 20.

Covariances: Z1, Z2, Z3, Z4, Z5, Z6, Z7, Z8, ...

	Z1	Z2	Z3	Z4	Z5	Z6	Z7
Z1	1.00000						
Z2	0.40966	1.00000					
Z3	0.09222	0.11082	1.00000				
Z4	0.36856	0.37736	0.34463	1.00000			
Z5	0.52576	0.36575	0.24618	0.51113	1.00000		
Z6	0.06773	0.17231	0.42277	0.31190	0.20187	1.00000	
Z7	0.47737	0.36093	0.26579	0.44340	0.52201	0.27906	1.00000
Z8	0.70090	0.48863	0.18786	0.47979	0.55300	0.15416	0.59662
Z9	0.43498	0.38989	0.37772	0.41085	0.40682	0.43768	0.48827
Z10	0.46518	0.35461	0.21581	0.38385	0.34672	0.29669	0.31387
Z11	0.09991	0.29581	0.34392	0.34671	0.19798	0.53751	0.20907
Z12	0.54377	0.37645	0.16542	0.37986	0.45480	0.15794	0.52156
Z13	0.48492	0.34590	0.18915	0.43533	0.46721	0.37679	0.46993
Z14	0.14517	0.02514	0.20647	0.19928	0.12683	0.10675	0.06084
Z15	0.23644	0.31849	0.31001	0.31589	0.37842	0.28142	0.33608
Z16	0.23399	0.23001	0.30071	0.34373	0.26559	0.38516	0.25288
Z17	0.45324	0.30189	0.18179	0.37821	0.44978	0.25794	0.33070
Z18	0.36690	0.22119	0.19490	0.30718	0.30261	0.24266	0.21073
Z19	0.11577	0.07255	0.19937	0.14778	0.12188	0.11534	0.05765
Z20	-0.01957	0.04045	0.12905	0.04899	0.34979	0.08684	0.11140

	Z8	Z9	Z10	Z11	Z12	Z13	Z14
Z8	1.00000						
Z9	0.44544	1.00000					
Z10	0.42574	0.45860	1.00000				
Z11	0.14587	0.46965	0.40178	1.00000			
Z12	0.53357	0.46022	0.38914	0.13249	1.00000		
Z13	0.50201	0.43228	0.48807	0.26524	0.48604	1.00000	
Z14	0.01429	0.08191	0.28778	0.18826	0.18496	0.22479	1.00000
Z15	0.35810	0.27632	0.27146	0.22300	0.25445	0.40111	0.07844
Z16	0.29142	0.39981	0.41035	0.29757	0.28080	0.34676	0.01416
Z17	0.46782	0.35975	0.32778	0.19432	0.42200	0.46895	0.11093
Z18	0.36924	0.26638	0.24834	0.12932	0.26296	0.31540	0.00166
Z19	0.23230	0.20021	0.10106	0.14296	0.09180	0.17215	-0.03965
Z20	0.04127	0.03141	-0.01016	0.02019	-0.01173	0.01405	0.02969

	Z15	Z16	Z17	Z18	Z19	Z20
Z15	1.00000					
Z16	0.30168	1.00000				
Z17	0.33098	0.40304	1.00000			
Z18	0.22408	0.33324	0.38920	1.00000		
Z19	0.34857	0.15784	0.19921	0.19155	1.00000	
Z20	0.22572	0.04920	0.08769	0.06810	0.08210	1.00000

Tampak bahwa seluruh variabel saling berkorelasi satu sama lain. Untuk mengeksplorasi data ini, AKU sesuai dengan persyaratan karena data bersifat multivariate. Langkah berikutnya adalah menghitung eigenvalue sampel dari matrik kovarian sampel $\mathbf{S_z}$.

Eigenanalysis of the Covariance Matrix

Eigenvalue	**6.7668**	**1.8338**	**1.3739**	**1.1785**	**1.0130**	0.9221	0.8204	0.7280
Proportion	0.338	0.092	0.069	0.059	0.051	0.046	0.041	0.036
Cumulative	0.338	0.430	0.499	0.558	**0.608**	0.654	0.695	0.732

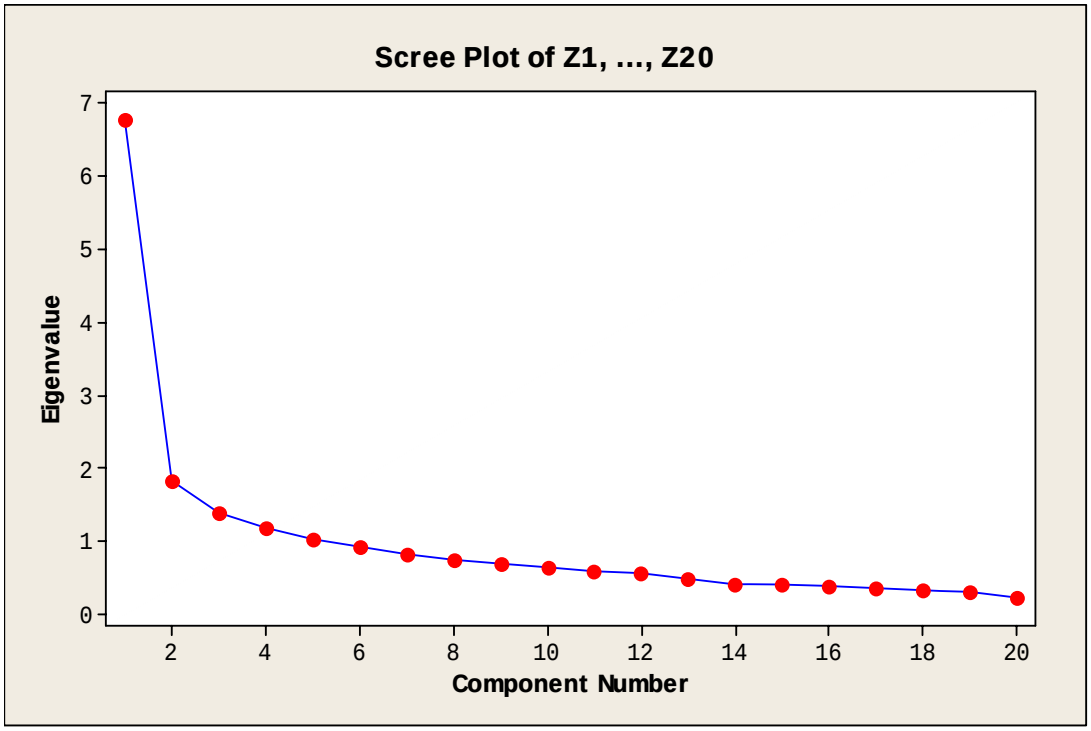
Eigenvalue	0.6774	0.6445	0.5930	0.5690	0.4873	0.4131	0.4012	0.3719
Proportion	0.034	0.032	0.030	0.028	0.024	0.021	0.020	0.019
Cumulative	0.766	**0.798**	**0.828**	0.856	0.880	0.901	0.921	0.940

Eigenvalue	0.3483	0.3341	0.2932	0.2305
Proportion	0.017	0.017	0.015	0.012
Cumulative	0.957	0.974	0.988	1.000

Dalam menentukan jumlah komponen utama yang ideal dapat digunakan beberapa kriteria. Pertama, kriteria total variansi sampel yang dapat dijelaskan lebih dari 80 persen (Johnson, 2002). Komponen utama pertama menjelaskan 33,8 persen dari total varian sampel. Komponen utama kedua secara kumulatif menjelaskan 43 persen dari total varian sampel. Sampai dengan komponen utama ke-10 secara kolektif menjelaskan 79,8 persen dari total varian sampel. Pada komponen utama ke-11 varian yang dapat dijelaskan sudah lebih dari 80 persen dari total varian sampel. Dengan demikian, menurut kriteria ini dibutuhkan 11 komponen utama untuk mereduksi 20 variabel tanpa banyak kehilangan informasi.

Kriteria kedua adalah ukuran relatif dari eigenvalue lebih besar dari satu. Berdasarkan kriteria ini, dengan hanya lima komponen informasi dari data sudah dapat disarikan meskipun hanya mampu menjelaskan 60,8 persen variansi data. Kriteria ketiga adalah dengan mengamati Scree Plot dari eigenvalue dan jumlah komponen. Untuk menentukan jumlah komponen dengan memperhatikan patahan siku dari Scree Plot. Tampak setelah komponen kedua, perubahan nilai

eigenvalue antar komponen cukup kecil. Oleh karena itu, menurut kriteria ketiga ini cukup dibutuhkan dua komponen utama.



Gambar 1. Screeplot CPCA dgn standardisasi.

Berikut ditampilkan sampel komponen utama kesatu sampai dengan kesebelas berikut korelasi antara variabel z_i dengan masing-masing komponen utamanya. Korelasi yang diberi warna merah menunjukkan asumsi bahwa korelasi antara komponen utama dengan variabel z tidak ditolak pada alpha 5 persen. Tampak bahwa hanya dengan komponen utama pertama semua variabel *z* mempunyai korelasi yang sigifikan pada alpha 5 persen.

Variabel	PC1	$R_{y1,Zk}$	PC2	$R_{y2,Zk}$	PC3	$R_{y3,Zk}$	PC4	$R_{y4,Zk}$	PC5	$R_{y5,Zk}$	PC6	$R_{y6,Zk}$
Z1	-0.26	-0.69	0.36	0.48	0.10	0.12	0.03	0.03	-0.09	-0.09	0.02	0.02
Z2	-0.22	-0.58	0.12	0.17	0.06	0.08	0.05	0.06	0.32	0.32	0.21	0.20
Z3	-0.17	-0.44	-0.39	-0.53	-0.09	-0.11	-0.14	-0.15	-0.01	-0.01	0.05	0.05
Z4	-0.26	-0.68	-0.05	-0.06	0.04	0.05	-0.10	-0.11	0.03	0.03	0.00	0.00
Z5	-0.27	-0.70	0.15	0.21	-0.20	-0.24	-0.31	-0.34	0.09	0.09	-0.12	-0.12
Z6	-0.19	-0.49	-0.46	-0.62	0.05	0.06	0.07	0.07	0.16	0.16	-0.17	-0.17
Z7	-0.26	-0.68	0.14	0.19	0.00	0.00	-0.16	-0.17	0.34	0.34	0.06	0.05
Z8	-0.29	-0.75	0.29	0.40	-0.06	-0.07	0.08	0.09	0.08	0.08	0.12	0.12
Z9	-0.27	-0.71	-0.13	-0.17	0.12	0.15	0.10	0.11	0.25	0.25	0.08	0.08
Z10	-0.25	-0.65	-0.05	-0.07	0.30	0.35	0.00	0.00	-0.17	-0.17	0.02	0.02
Z11	-0.18	-0.47	-0.44	-0.59	0.21	0.25	0.02	0.02	0.19	0.19	0.13	0.12
Z12	-0.26	-0.66	0.24	0.32	0.15	0.17	-0.05	-0.05	-0.02	-0.02	0.03	0.03
Z13	-0.28	-0.72	0.05	0.07	0.10	0.11	-0.01	-0.01	-0.17	-0.18	0.01	0.01
Z14	-0.08	-0.22	-0.12	-0.17	0.35	0.41	-0.52	-0.57	-0.56	-0.56	0.08	0.08
Z15	-0.21	-0.55	-0.11	-0.15	-0.36	-0.42	-0.05	-0.06	-0.10	-0.10	0.31	0.29

Z16	-0.21	-0.56	-0.20	-0.27	-0.03	-0.03	0.29	0.32	-0.06	-0.06	-0.36	-0.34
Z17	-0.25	-0.64	0.09	0.12	-0.10	-0.12	0.13	0.14	-0.25	-0.26	-0.26	-0.25
Z18	-0.19	-0.50	0.03	0.05	-0.18	-0.21	0.31	0.33	-0.25	-0.25	-0.44	-0.42
Z19	-0.11	-0.28	-0.12	-0.17	-0.41	-0.48	0.30	0.32	-0.34	-0.34	0.56	0.54
Z20	-0.05	-0.14	-0.08	-0.11	-0.54	-0.63	-0.51	-0.55	0.10	0.10	-0.24	-0.23

Variabel	**PC7**	**$R_{y7,Zk}$**	**PC8**	**$R_{y8,Zk}$**	**PC9**	**$R_{y9,Zk}$**	**PC10**	**$R_{y10,Zk}$**	**PC11**	**$R_{y11,Zk}$**
Z1	0.02	0.02	-0.12	-0.11	-0.23	-0.19	0.03	0.02	-0.12	-0.09
Z2	-0.46	-0.42	-0.30	-0.26	0.27	0.22	-0.11	-0.09	-0.19	-0.14
Z3	0.52	0.47	-0.20	-0.17	0.04	0.03	-0.32	-0.26	-0.23	-0.18
Z4	0.13	0.12	-0.35	-0.29	0.42	0.35	-0.11	-0.09	0.49	0.38
Z5	-0.01	-0.01	-0.08	-0.07	-0.04	-0.03	0.04	0.03	0.24	0.19
Z6	0.05	0.04	0.20	0.17	0.08	0.06	0.41	0.33	-0.10	-0.08
Z7	0.32	0.29	0.22	0.19	0.04	0.03	0.08	0.06	-0.08	-0.06
Z8	0.08	0.08	-0.11	-0.09	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03
Z9	0.12	0.11	0.01	0.01	-0.42	-0.35	0.02	0.02	0.00	0.00
Z10	-0.34	-0.31	0.02	0.01	-0.29	-0.24	-0.25	-0.20	-0.15	-0.11
Z11	-0.26	-0.23	-0.19	-0.16	-0.11	-0.09	0.23	0.19	0.18	0.14
Z12	0.20	0.18	0.22	0.19	-0.12	-0.10	-0.06	-0.05	-0.05	-0.04
Z13	-0.07	-0.07	0.38	0.33	0.17	0.14	0.33	0.27	-0.04	-0.03
Z14	0.00	0.00	-0.10	-0.08	0.03	0.03	0.01	0.01	-0.07	-0.06
Z15	-0.18	-0.17	0.29	0.25	0.41	0.34	-0.12	-0.10	-0.33	-0.25
Z16	-0.10	-0.09	0.25	0.22	-0.03	-0.02	-0.60	-0.48	0.11	0.08
Z17	-0.08	-0.07	0.15	0.13	0.09	0.07	0.11	0.09	0.42	0.32
Z18	0.07	0.06	-0.45	-0.39	0.06	0.05	0.27	0.21	-0.41	-0.32
Z19	0.06	0.06	-0.08	-0.07	-0.28	-0.23	0.10	0.08	0.22	0.17
Z20	-0.28	-0.25	-0.02	-0.02	-0.34	-0.28	0.01	0.01	-0.04	-0.03

3.2 Analisis Komponen Utama Robust dengan metode ROBPCA

Dibandingkan dengan CPCA, metode ROBPCA menghasilkan jumlah komponen yang lebih sedikit daripada CPCA untuk mereduksi data. Dengan varian yang dapat dijelaskan 80 persen, ROBPCA membutuhkan tiga komponen utama sedang CPCA membutuhkan sebelas komponen. Nilai eigen, kumulatif proporsi eigen terhadap total nilai eigen, dan koefisien komponen tampak pada tabel berikut.

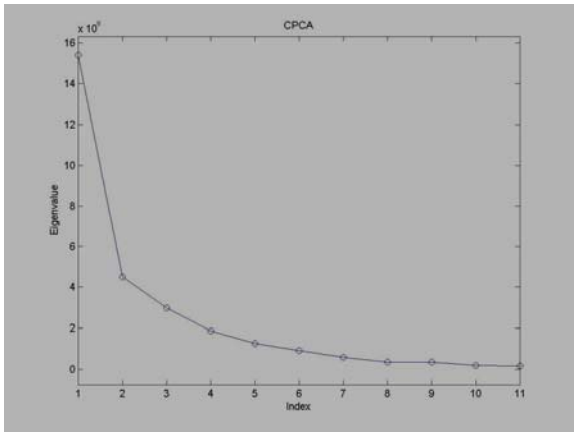
Selain jumlah komponen ROBPCA lebih sedikit daripada CPCA, jumlah orthogonal outlier yang dideteksi CPCA lebih banyak daripada ROBPCA. Demikian juga dengan bad leverage, titik data yang dapat dideteksi CPCA juga lebih banyak. Pada gambar 6, tampak batas cutt off ellipse dari ROBPCA lebih kecil dibandingkan dengan CPCA. Hal ini menunjukkan bahwa ROBPCA lebih robust daripada CPCA. Ellipse yang melebar dipengaruhi oleh mean dari data yang tidak robust terhadap outlier.

Tabel. Nilai Eigen Dan Koefisien Komponen

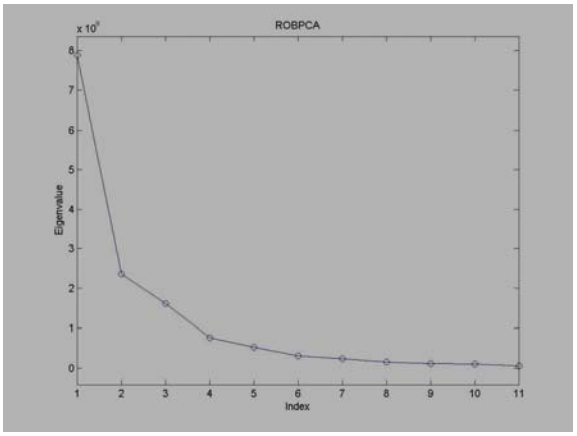
Dengan Metode ROBPCA

Nilai Eigen	Proporsi	Kumulatif Proporsi	PC1	PC2	PC3
7.90	0.53	0.53	-0.20	0.23	-0.76
2.35	0.16	0.68	-0.01	0.00	-0.02
1.95	**0.13**	**0.81**	**-0.07**	**0.00**	**0.02**
0.87	0.06	0.87	-0.09	0.04	-0.15
0.56	0.04	0.91	-0.10	0.07	-0.19
0.37	0.02	0.93	-0.06	0.07	-0.01
0.24	0.02	0.95	-0.03	0.02	-0.03
0.16	0.01	0.96	-0.04	0.02	-0.05
0.13	0.01	0.97	-0.02	0.03	-0.03
0.11	0.01	0.97	-0.03	0.02	-0.01
0.07	0.00	0.98	-0.02	0.02	-0.02
0.06	0.00	0.98	-0.03	0.04	-0.04
0.05	0.00	0.99	-0.22	0.22	-0.35
0.05	0.00	0.99	-0.03	-0.02	-0.07
0.04	0.00	0.99	-0.85	-0.49	0.14
0.04	0.00	1.00	-0.38	0.80	0.44
0.02	0.00	1.00	-0.06	0.05	-0.05
0.02	0.00	1.00	-0.06	0.06	-0.13
0.01	0.00	1.00	-0.03	-0.01	-0.01
0.01	0.00	1.00	-0.03	0.00	-0.03

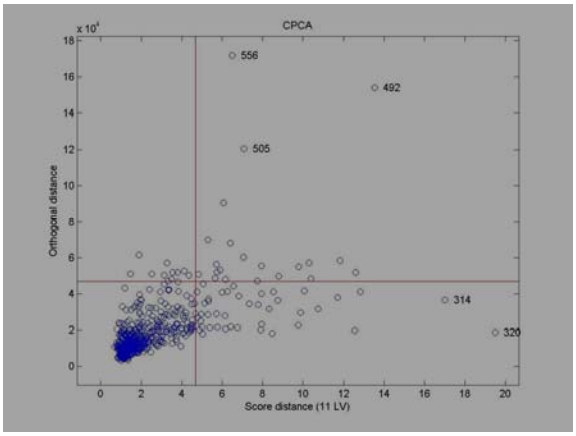
Sumber: Hasil Olah dengan Minitab 6.5.



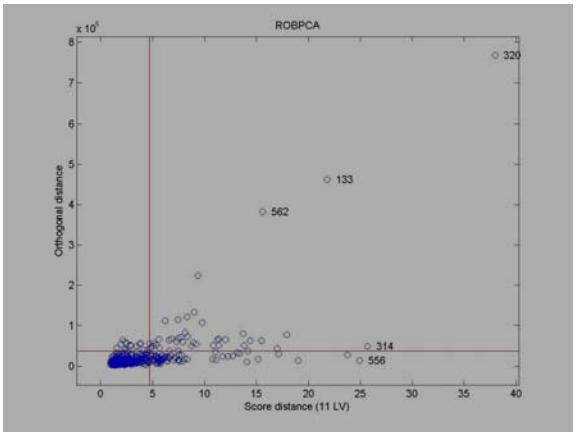
Gambar 2. Screeplot CPCA



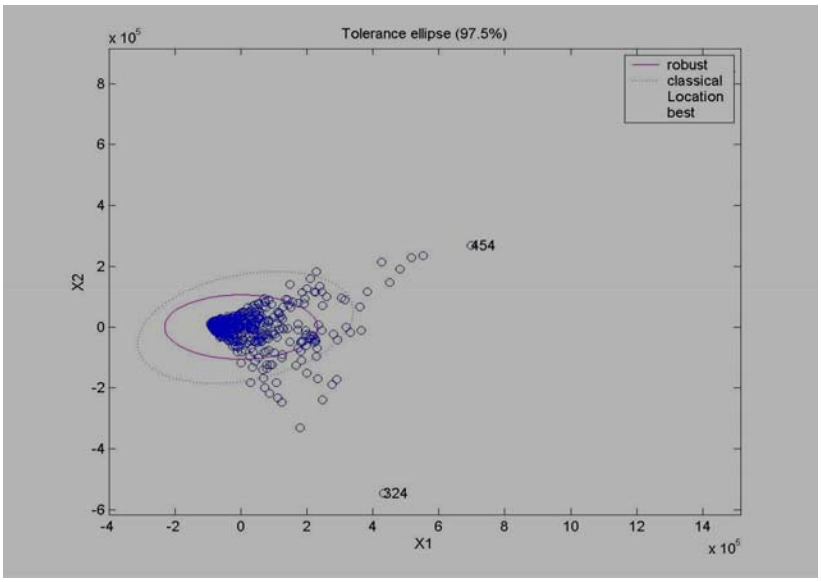
Gambar 3. Screeplot ROBPCA



Gambar 4. Orthogonal Distance dan Score Distance CPCA



Gambar 5. Orthogonal Distance dan Score Distance ROBPCA



Gambar 6. Tolerance ellipse dari CPCA dan ROBPCA

Referensi

1. Hubert, M., Rousseeuw, Peter J., dan Branden, Karlien V. (2004). ROBCA: A New Approach to Robust Principal Component Analysis. *Technometrics*. Feb 2005, 47, No. 1. 64-79.
2. Verboven, S. dan Hubert, M. (2004). LIBRA: a MATLAB Library for Robust Analysis. http://www.wis.kuleuven.ac.be/stat/robust.html.
3. 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.
4. Johnson, R.A and Wichern, D.W., (2002). *Applied Multivariate Statistical Analysis*. 5th Ed. New Jersey: Prentice Hall.