p-ISSN: 1979-5661
e-ISSN: 2622-321X

Metode ROBPCA (Robust Principal Component Analysis) dan Clara (Clustering Large Area) pada Data dengan Outlier

(Studi Kasus Data Laporan Indeks Kebahagiaan Dunia Tahun 2018)

Bekti Endar Susilowati [1)], Pardomuan Robinson Sihombing [2)]
[1]*Badan Pusat Statistik Kabupaten Sleman, Jalan Purbaya, Mlati, Kabupaten Sleman, Yogyakarta*
[1]bekti@bps.go.id

[2]*Badan Pusat Statistik, Jalan Dr. Sutomo No 6-8, Jakarta Pusat*
[2]robinson@bps.go.id

Abstract

PCA is one of multivariate analysis used for deputizing variables using less number of Principal Components without losing much information. In other words, it is used for explaining the underlying variance-covariance structure of the large data set of variables through a few linear combinations of these variables. PCA is significantly influenced by the outliers, since the covariant matrix are sensitive to outliers. Thus, the analysis for this study was conducted by using a PCA that is robust to outliers, namely ROBPCA or Hubert PCA. Then, the principal components formed were used as inputs in cluster analysis using the Clara method. Clara is one of the k-medoids methods that is robust to outliers and is appropriate for large data analysis. In the case study of the compiling variables of happiness index based on The World Happiness Report (WHR)2018 using the Clara method with Manhattan distance, the best average value of Overall Average Silhouette Width in the 5 clusters were obtained.

Keywords: robust, outlier, ROBPCA, Clara

Abstrak

PCA merupakan salah satu analisis multivariat yang digunakan untuk mengganti variable dengan *Principal Component* yang sedikit jumlahnya namun tidak terlalu banyak informasi yang hilang. Atau dengan kata lain, *it used to explain the underlying variance-covariance structure of the large data set of variables through a few linear combination of these variables.* PCA sangat dipengaruhi oleh kehadiran *outlier* karena didasarkan pada matriks kovarian yang sensitive terhadap *outlier*. Oleh karena itu, pada analisis ini akan digunakan PCA yang robust terhadap outlier yaitu ROBPCA atau PCA Hubert. Selanjutnya, dari *Principal Component* yang terbentuk digunakan sebagai input (masukan) untuk *cluster analysis* dengan metode Clara. *Clara* merupakan salah satu metode k-medoids yang robust terhadap *outlier* dan baik digunakan pada data dalam jumlah besar. Dalam studi kasus terhadap variabel penyusun indeks kebahagiaan berdasarkan *The World Happiness Report* 2018 dengan metode Clara yang menggunakan jarak manhattan didapatkan nilai rata-rata *Overall Average Silhouette Width* yang terbaik pada 5 cluster.

Kata kunci: robust, outlier, ROBPCA, Clara

1. PENDAHULUAN

The World Happiness Report (WHR) diterbirtkan setiap tahun oleh Perserikatan Bangsa-Bangsa (PBB) sejak tahun 2011 melalui inisiatif global *Sustainable Development Solution Network (SDSN)*. Laporan tersebut merupakan upaya pengukuran tingkat kebahagiaan penduduk negara-negara anggota sebagai pedoman dalam menetapkan kebijakan publik terkait kesejahteraan penduduk. Berdasarkan WHR tahun 2018, Finlandia menempati peringkat teratas disusul oleh Norwegia dan Denmark. Sedangkan untuk peringkat terbawah ditempati oleh Burundi. Sementara itu, untuk Negara ASEAN Singapura menempati peringkat teratas (peringkat 34 dunia), disusul Malaysia(peringkat 35 dunia), sementara Indonesia menempati peringkat ke 96 dunia.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Data yang dikaji dalam menyusun WHR antara lain *kekuatan ekonomi (GDP per capita), social support, Healthy life expectancy at birth, Freedom to make life choices, Generosity, Perceptions of corruption, Positive Affect, Negative Affect, Confidence in National Government, GINI index (World Bank estimate) average 2000-15, and gini of household income reported in Gallup, by wp5-year.* Berdasarkan hasil penelitian sebelumnya (Sobiroh, 2015), ROBPCA yang menggabungkan konsep *Projection Pursuit (PP)* dan *Minimum Covariance Determinant (FAST*-MCD dengan teorema C-step) memberikan kesimpulan lebih baik daripada *Classic Principal Component Analysis (CPCA)* karena mampu menghasilkan jumlah komponen utama lebih sedikit namun telah mampu mejelaskan sebesar 84,79% dari total variasi sampel.

Pada penelitian kali ini akan dilakukan analisis komponen utama (*PCA)* dari variabel-variabel tersebut kemudian dilanjutkan dengan analisis cluster terhadap negara-negara anggota. Untuk melakukan analisis dari kedua meotode tersebut, dipilih metode yang *robust* terhadap *outlier* sehingga diharapkan didapatkan hasil analisis yang lebih akurat. Pada metode PCA dipilih salah satu metode yang robust terhadap outlier yaitu ROBPCA(Robust PCA) sedangkan untuk analysis Cluster dipilih metode Clara yang dalam penelitian sebelumnya disebutkan bahwa metode tersebut *robust* terhadap *outlier* dan efektif digunakan dalam data yang cukup besar.

2. METODE PENELITIAN

Sumber Data dan Variabel Penelitian

Data diperoleh dari *The World Happiness Report 2018.* Data yang akan digunakan sebanyak 141 negara dengan 11 variabel. (Statistical Appendix 1 for Chapter 2 of World Happiness Report 2018). Adapun variabel yang digunakan dalam penelitian ini adalah:

a. *GDP per capita* atau pendapatan per kapita adalah besarnya pendapatan rata-rata penduduk di suatu negara. Pendapatan per kapita didapatkan dari hasil pembagian pendapatan nasional suatu negara dengan jumlah penduduk negara tersebut.

b. *Social support* (dukungan sosial) merupakan rata-rata nasional dari respon dalam bentuk biner (0 atau 1).

c. *Healthy life expectancy at birth;* Variabel *Healthy life expectancy at birth (*harapan hidup sehat) dihitung berdasarkan data dari World Health Organization (WHO), the World Development Indicators (WDI), and jurnal-jurnal statistik.

d. *Freedom to make life choice;* Freedom to make life choices (kebebasan untuk membuat pilihan hidup) rata-rata respons nasional terhadap pertanyaan "Apakah Anda puas atau tidak puas dengan kebebasan Anda untuk memilih apa yang Anda lakukan dengan hidup Anda?"

e. *Generosity;* Variabel Generosity (kemurahan hati) merupakan rata-rata nasional dari respon Sudahkah Anda menyumbangkan uang untuk kegiatan amal dalam sebulan terakhir?" pada PDB per kapita.

f. *Perceptions of corruption; Perceptions of corruption (persepsi tentang korupsi)* merupakan variabel yang mengukur rata-rata nasional dari respons survei terhadap dua pertanyaan: "Apakah korupsi tersebar luas di seluruh pemerintah atau tidak" dan "Apakah korupsi tersebar luas di dalam bisnis atau tidak?".

g. *Positive Affect; Positive Affect* (pengaruh positif) merupakan variabel yang didefinisikan sebagai rata-rata dari tiga ukuran efek positif: kebahagiaan, tawa dan kesenangan.

h. *Negative Affect*; *Negative Affect* (pengaruh negatif) merupakan variabel yang didefinisikan sebagai rata-rata dari tiga ukuran efek negative yaitu kekhawatiran, kesedihan dan kemarahan.

i. *Confidence in National Government* merupakan variabel yang merngukur kepercayaan terhadap pemerintah.

j. *GINI index (World Bank estimate) average 2000-15;* Indeks Gini merupakan indikator yang menunjukkan tingkat ketimpangan pendapatan secara menyeluruh. Nilai Koefisien Gini berkisar antara 0 hingga 1. Koefisien Gini bernilai 0 menunjukkan adanya pemerataan pendapatan yang sempurna, atau setiap orang memiliki pendapatan yang sama.

k. *Gini of household income reported in Gallup, by wp5-year;* Variabel ini merupakan indeks Gini dari pendapatan rumah tangga. Variabel pendapatan dibuat dengan mengonversi mata uang lokal ke Dolar Internasional (ID) menggunakan rasio paritas daya beli.

ROBPCA (ROBust PCA)

ROBPCA (ROBust PCA) atau disebut juga *Hubert PCA* ditemukan oleh Hubert dkk (2005) sebagai perkembangan dari *robust PCA.* Metode tersebut merupakan gabungan konsep *Projection*

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Pursuit(PP) dan estimator kovarian yang robust yaitu *Minimum Covariant Determinant* (MCD) yang dimodifikasi bersama dengan Van Driessen menjadi *Fast Minimum Covariance Determinant* (FAST-MCD) pada tahun 1999. Metode tersebut digunakan untuk mendapatkan komponen utama (*PCA)* yang tidak terpengaruh terlalu banyak dengan kehadiran data *outlier*.

Untuk analisis selanjutnya, robust principal komponen yang telah diperoleh dengan metode ROBPCA digunakan sebagai input (masukan) untuk cluster analysis. Berdasarkan penelitian sebelumnya (Muslim, 2018), *cluster analysis* dengan Clara method menyimpulkan bahwa Clara method dengan jarak manhattan lebih robust dibandingkan dengan K-means dengan jarak Manhattan dan Pam. Oleh karena itu pada penelitian kali akan mengaplikasikan Clara method dengan jarak manhattan untuk studi kasus data dengan variabel-variabel yang bersumber dari *The World Happiness Report* (WHR) 2018.

Matriks kovariansi didefinisikan sebagai berikut:

$$\sum = cov(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad (1)$$

Matriks korelasi didefinisikan sebagai berikut:

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{12}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{1p}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{12}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{2p}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{2p}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix} \quad (2)$$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \quad (3)$$

Mahalanobis Distance

Jarak mahalanobis pada data multivariate digunakan untuk mendeteksi outlier, yang diperoleh dengan menghitung jarak tiap observasi terhadap pusat datanya.

$$d_{MD}^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) > \chi^2_{p;(1-\alpha)} \quad (4)$$

Deteksi outlier dengan jarak mahalanobis kurang maksimal ketika datanya mengandung lebih dari satu outlier, sehingga dikembangkan jarak mahalanobis robust (robust distance/RD) didasarkan pada penaksir robust untuk vector rata-rata dan matriks kovariansi. Pengamatan x_i dikatakan outlier apabila jika:

$$d_{RD}^2 - (x_i - \mu_{MCD})' \Sigma_{MCD}^{-1} (x_i - \mu_{MCD}) > \chi^2_{p;(1-\alpha)} \quad (5)$$

Dengan μ_{MCD} dan Σ_{MCD} merupakan vector rata-rata dan matriks kovariansi dari sebagiam data **X** yang mempunyai determinan matriks kovariansi terkecil.

Terdapat tiga macam jenis outlier sebagai berikut (Sobiroh, 2015):

a. Good Leverage, merupakan pengamatan yang berada di ruang distribusi tetapi sudah tidak berada di daerah mayoritas data
b. Bad Leverage, merupakan pengamatan yang tidak berada baik dalam ruang distribusi maupun daerah mayoritas data
c. Orthogonal Leverage, merupakan pengamatan yang memiliki jarak pengamatan sangat besar dari daerah mayoritas data sehingga pengamatan tersebut sudah tidak dapat dilihat dalam ruang distribusinya.

Minimum Covariance Determinant (MCD) dan Projection Pursuit

Misal $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]'$ adalah himpunan data sejumlah n pengamatan yang terdiri dari p variabel dimana n ≥ p+1.

$$t = \frac{1}{h} \sum_{i=1}^{h} x_i \quad (6)$$

p-ISSN: 1979-5661
e-ISSN: 2622-321X

$$C = \frac{1}{h}\sum_{i=1}^{h}(x_i - t_1)(x_i - t_1)' \quad (7)$$

t dan **C** merupakan matriks definit positif simetri berdimensi pxp dari suatu sub sampel berukuran h pengamatan dimana $\frac{n+p+1}{2} \leq h \leq n$ yang meminimumkan det (C). Metode MCD mencari himpunan bagian dari **X** sejumlah h elemen dimana h adalah integer terkecil dari (n+p+1)/2. Dimisalkan bahwa himpunan bagian itu adalah $\boldsymbol{X}_h$. Untuk mendapatkan penaksir MCD perlu dicari C_h^n kombinasi. Jika n kecil maka penaksir MCD cukup mudah ditemukan. Akan tetapi masalh muncul ketika n cukup besar karena terdapat banyak kombinasi sub sampel yang harus ditemukan untuk memperoleh penaksir MCD. *Projection Pursuit (PP)* bertujuan untuk mendapatkan struktur pada data peubah ganda dengan memproyeksikannya pada subruang berdimensi lebih rendah (Hubert, 1985). Seperti CPCA, metode tersebut mencari suatu arah dengan penyebaran maksimal data diproyeksikan di dalamnya.

PCA yang diperkenalkan pertama kali oleh Pearson pada tahun 1901, merupakan suatu analisis multivariate yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel tersebut. Hal ini bertujuan agar dihasilkan dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya. Dalam perkembangannya, dipengaruhi adanya kebutuhan suatu model PCA yang robust terhadap data *outlier.* PCA yang juga dikenal dengan *Clasisical Principal Component Analysis* (CPCA) sangat dipengaruhi oleh kehadiran *outlier* karena didasarkan pada matriks kovarian yang peka terhadap *outlier.* Untuk mengatasi hal tersebut beberapa ahli menggantikan matriks kovarian klasik dengan estimator kovarian robust.

Sebagai perkembangan dari *robust PCA,* Hubert dkk (2005) menemukan ROBPCA atau disebut juga *Hubert PCA.* Metode tersebut merupakan gabungan konsep *Projection Pursuit(PP)* dan estimator kovarian yang robust yaitu *Minimum Covariant Determinant* (MCD) yang dimodifikasi bersama dengan Van Driessen menjadi *Fast Minimum Covariance Determinant* (FAST-MCD) pada tahun 1999. Metode tersebut digunakan untuk mendapatkan komponen utama (*PCA)* yang tidak terpengaruh terlalu banyak dengan kehadiran data *outlier*.

Algoritma ROBPCA:

1. Memilih $\frac{1}{2} < \alpha < 1$ untuk mendapatkan
$$\mathrm{h} = \max\{\alpha_n, [(h+p+1)/2]\} \quad (8)$$
2. Menghitung outlyingness setiap data x_i dengan rumus Stanhel-Donoho
$$O(x_i) = max_{v \in B}\frac{|x_i^t v - \hat{\mu}_{MCD} x_j^t v|}{S_{MCD}(x_j^t v)} \quad (9)$$
3. Matriks covarian S_0 dikomposisi sehingga diperoleh komponen utamanya.
4. Pada n, k dari algoritma ke-3 dihitung kembali penduga nilai tengah $(\hat{\mu}_2)$ dan matriks kovarian MCD (S_i) menggunakan FAST-MCD yang diadaptasi. Komponen utama akhir adalah vector eigen dari matriks kovarian tersebut (S_i)

Berdasarkan studi-studi sebelumnya, ROBPCA merupakan suatu pendekatan *PCA* yang lebih efektif dalam untuk data yang mengandung *outlier.*

Jarak Euclidean dan Manhattan

Jarak Euclidean merupakan jarak terpendek antar 2 titik, digunakan untuk menghitung jarak Euclidean antara suatu objek dengan pusat klaster.

$$d_{euc}(x_{ij}, c_{kj}) = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - c_{kj})^2} \quad (10)$$

Jarak Manhattan (*city block distance)* diibaratkan sebagai jarak blok antara 2 titik suatu kota.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

$$d_{man}(x_{ij}, c_{kj}) = \sum_{i=1}^{p}\sum_{i=1}^{n} |x_{ij} - c_{kj}| \quad (11)$$

$d_{euc}(x_{ij}, c_{kj})$: jarak Euclidean antara pengamatan ke-i variabel ke-j ke pusat klaster ke-k pada variabel ke-j

$d_{man}(x_{ij}, c_{kj})$: jarak Manhattan antara pengamatan ke-i variabel ke-j ke pusat klaster ke-k pada variabel ke-j

x_{ij}: objek pada pengamatan ke-i variabel ke-j

c_{kj}: pusat klaster ke-k pada variabel ke-j
p :banyak variabel
n : banyak pengamatan

CLARA

Clara (Kaufman dan Rousseew, 1990) merupakan salah satu macam pengelompokan data dengan medoid sebagai pusat klasternya. Medoid merupakan objek yang letaknya terpusat pada suatu klaster, atau dengan kata lain merupakan suatu objek yang merepresentasikan anggota pada suatu data dan memiliki rata-rata perbedaan (dissimailarity) yang paling kecil dengan anggota-anggota lain.

Berbeda dengan metode medoid lainnya, yaitu Pam, metode Clara memiliki sifat robust terhadap outlier dan dapat digunakan untuk data dalam jumlah besar. Metode ini lebih efisien dalam hal waktu komputasi dan dalam penyimpanan data set yang besar.

Clara menggunakan pendekatan sampling, kemudian menerapkan algoritma Pam untuk mendapatkan medoid yang optimal untuk sampel. Kualitas medoid yang dihasilkan diukur dengan rata-rata perbedaan jarak antara setiap objek di data set dan medoid pada sampel. Dengan mengambil sampel cesara acak, medoid dari sampel diharapkan akan mendekati nilai medoid dari data set.

Algoritma Clara (Muslim, 2018):

1. Menentukan banyaknya klaster (k),
2. Membagi data set secara acak dalam beberapa sub set dengan ukuran tetap, dimana ukuran sampel setiap sub set minimal 40+2*k,
3. Menentukan medoid awal ,
4. Menghitung jarak non-medoid dengan medoid setiap klaster,
5. Menempatkan objek berdasarkan jarak terdekat dengan medoid,
6. Menghitung total jarak yang diperoleh,
7. Memilih secara acak objek non-medoid pada masing-masing klaster sebagai kandidat medoid baru,
8. Menghitung jarak setiap objek non medoid dengan kandidat medoid baru dan menempatkan objek berdasarkan jarak terdekat dengan medoid baru tersebut.
9. Menghitung selisih total jarak kandidat medoid baru dengan total jarak pada medoid lama. Jika total jarak setiap objek dengan kandidat medoid baru kurang dari total jarak setiap objek dengan medoid lama, maka kandidat medoid baru menjadi medoid baru,
10. Mengulangi kembali langkah 7-9,
11. Menghitung jarak antara semua *non medoid* dengan objek yang menjadi *medoid*, hingga diperoleh sub set dengan jumlah terkecil adalah yang dipilih.

3. HASIL DAN PEMBAHASAN

Analisis Deskriptif

Dari *summary* data pada Tabel 1, terlihat bahwa terdapat beberapa *missing values (NA)*. Untuk melanjutkan analisis berikutnya, perlu dilakukan penanganan terhadap *missing values*, yaitu dengan mengestimasi nilai-nilai tersebut dengan menggunakan EM-Algorithm. Metode EM-Alghorithm merupakan suatu metode dengan Expectation-Step dilanjutkan Maximation-Step dengan iteratif maximum likelihood. Metode ini mengasumsikan sebuah distribusi dari data hilang secara parsial dan berdasarkan fungsi *likelihood* dari distribusi tersebut. Dari pengujian dapat disimpulkan bahwa mekanisme *missing values* yaitu dengan MCAR dan data tersebut berdistribusi normal, sehingga

p-ISSN: 1979-5661
e-ISSN: 2622-321X

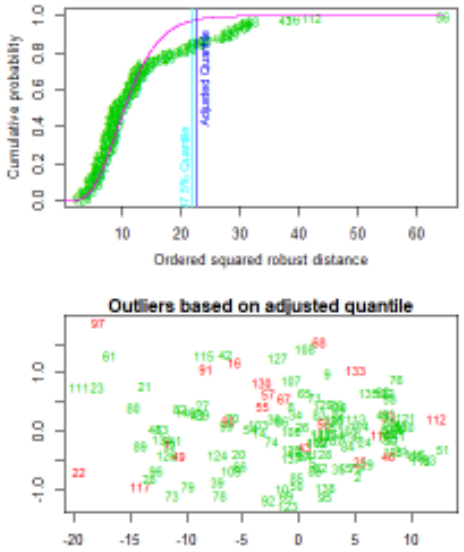
untuk menaksir *missing values* dapat dilakukan dengan EM Alghoritma. Selanjutnya dilakukan uji normalitas multivariat, didapatkan asumsi normalitas multivariat terpenuhi.

Tabel 1. Analis Deskriptif

Variable	**x1**	**x2**	**x3**	**x4**
Median	9.544	0.829	65.13	0.812
Max.	11.465	0.967	76.54	0.985
NA's	7	1	0	1
Variable	**x5**	**x6**	**x7**	**x8**
Median	-0.035	0.781	0.712	0.281
Mean	-0.011	0.735	0.707	0.292
NA's	8	12	1	1
Variable	**x9**	**x10**	**x11**	
Median	0.476	0.368	0.439	
Mean	0.496	0.384	0.459	
NA's :	13	16	0	

Deteksi Outlier

Fungsi aq.plot pada R (packages mvoutlier) menggambarkan jarak Mahalanobis kuadrat robust dari pengamatan terhadap fungsi distribusi empiris dari jarak Mahalanobis. Perhitungan jarak didasarkan pada Estimator MCD. Dari plot terlihat cukup banyak observasi yang merupakan outlier.



Gambar 1 Plot Uji Outlier

Hasil ROBPCA

Dari ROBPCA didapatkan Principal Component 1 (PC1) menjelaskan 51,01%, kemudian ditambahkan PC2 sudah mampu menjelaskan sebesar 77,20 % dan ditambahkan PC3 menjadi sebesar 92,6 %. Nilai eigen >1 yang diperoleh yang diperoleh dari matriks korelasi dibandingkan dengan nilai 1 dengan alasan karena ketika komponen utama diperoleh dari matriks korelasi (standardized data) variansi dari masing-masing variabelnya sama dengan 1. Jika suatu komponen utama tidak dapat menerangkan variansi melebihi dirinya sendiri, maka komponen utama tersebut tidak signifikan atau dengan kata lain, komponen utama yang memiliki nilai eigen <1 dapat diabaikan.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Tabel 2. Proporsi Kumulatif dan Nilai Eigen ROBPCA

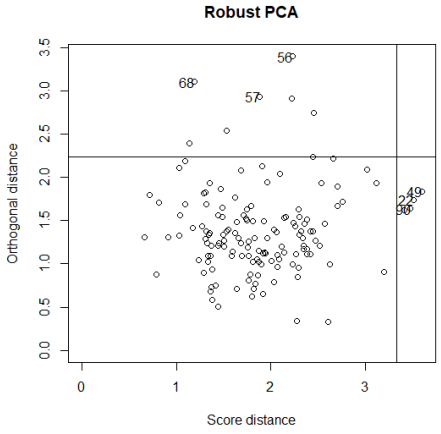
ROBPCA	PC1	PC2	PC3
Cumulative Proportion	0.5101	0.7720	0.926
λ_1	λ_2	λ_3	λ_4
4.7325703	2.4298481	1.4285390	0.6862558

Dari output tersebut didapatkan $\lambda_1, \lambda_2, \lambda_3$ yang memiliki nilai eigen >1 sehingga untuk selanjutnya didapatkan 3 principal komponen. Hasil dari ROBPCA lebih baik, jika dibandingkan dengan hasil yang didapatkan dari Classic Principal Component Analysis sebagai berikut:

Tabel 3. Proporsi Kumulatif PCA Classic

CPCA	PC1	PC2	PC3	PC4	PC5
Cumulative Proportion	0.3935	0.6030	0.7309	0.7991	0.85879

Dari PCA didapatkan Principal Component yang lebih banyak, Principal Componen ke 3 hanya dapat dapat menjelaskan sebesar 73% sedangkan 85,89% dijelaskan oleh Principal Componen ke 5.



Gambar 2. Output RobPCA

Dari plot di atas terlihat bahwa nomor observasi 56, 57 dan 68 dan lain-lain merupakan observasi yang outlier orthogonal karena mempunyai OD (*Orthogonal Distance*) besar *(>cut-off point*), dan SD (Score Distance) kecil (< cut-off point). Nomor 22, 49, dan 96 merupakan bad leverage point karena OD bernilai kecil namun SD besar. Nilai Loading ROBPCA merupakan bentuk standardisasi, maka persamaan Principal Component sebagai berikut:

$$Y_1 = e_1'Z = -0.42752694Z_1 - 0.41890063\ Z_2 - 0.41660572Z_3 - 0.27802083Z_4 - 0.10265343Z_5 + 0.21175268Z_6 - 0.25418623Z_7 + 0.38125181Z_8 + 0.01212936Z_9 + 0.13381094Z_{10} + 0.30267638\ Z_{11}$$

$$Y_2 = e_2'Z = 0.13711767\ Z_1 + 0.05796348Z_2 + 0.18385556Z_3 - 0.46310641Z_4 - 0.32435379Z_5 + 0.23825719Z_6 - 0.40141868\ Z_7 - 0.00568559Z_8 - 0.47805209Z_9 - 0.27843609Z_{10} - 0.31680164Z_{11}$$

$$Y_3 = e_3'Z = 0.12414565\ Z_1 + 0.17582854Z_2 + 0.08799225Z_3 + 0.13112034Z_4 - 0.42608314Z_5 + 0.27371537Z_6 + 0.33210065Z_7 + 0.02486175Z_8 - 0.38743172Z_9 + 0.60164699Z_{10} + 0.22198799Z_{11}$$

Nilai Y_1 dapat diterangkan paling baik oleh variabel Z_1 (*log GDP per capita*), Z_2(*social support*), Z_3(*Healthy life expectancy at birth*) dengan korelasi antara Y_1 dengan ketiga variabel tersebut adalah negatif. Hal ini berarti apabila ketiga variabel tersebut besar maka Y_1 bernilai kecil. Demikian pula

p-ISSN: 1979-5661
e-ISSN: 2622-321X

dengan variabel-variabel lain, jika bertanda negatif berarti korelasi antara Y_1 dengan variabel tersebut negatif, yang berarti jika variabel tersebut bernilai besar maka maka Y_1 akan bernilai kecil meskipun dengan rentang yang tidak begitu besar.

Tabel 4. Korelasi antara variabel and principal component

	Y_1	Y_2	Y_3
Z_1	-0.930	0.214	0.148
Z_2	-0.911	0.090	0.210
Z_3	-0.906	0.287	0.105
Z_4	-0.605	-0.722	0.157
Z_5	-0.223	-0.506	-0.509
Z_6	0.461	0.371	0.327
Z_7	-0.553	-0.626	0.397
Z_8	0.829	-0.009	0.030
Z_9	0.026	-0.745	-0.463
Z_{10}	0.400	-0.434	0.719
Z_{11}	0.658	-0.494	0.265

Nilai Y_2 dapat diterangkan paling baik oleh variabel Z_4 (*Freedom to make life choices*), Z_7(*Positive affect*), Z_5(*Confidence in national government*), koefisien ketiganya bernilai negative yang berarti korelasi antara Y_2dengan ketiga variabel tersebut adalah negative. Hal ini berarti apabila Y_2 kecil maka ketiga variabel tersebut besar. Demikian pula dengan variabel-variabel lain, jika bertanda positif berarti korelasi antara Y_2dengan variabel tersebut positif, yang berarti jika Y_2 besar berarti maka variabel tersebut bernilai besar meskipun dengan rentang yang tidak begitu besar.

Nilai Y_3 dapat diterangkan paling baik oleh variabel Z_{10} (*gini of household income reported in Gallup*), yang bernilai positif yang berarti korelasi dengan Y_3 adalah positif, jika variabel gini of household income reported in Gallup besar maka Y_3 juga besar. Variabel X_6(*Generosity*) dan X_5(*Confidence in national government*) memiliki koefisien bernilai negative yang berarti korelasi antara Y_3dengan ketdua variabel tersebut adalah negative. Hal ini berarti apabila Y_3 kecil maka kedua variabel tersebut besar. Demikian pula dengan variabel-variabel lain, jika bertanda positif berarti korelasi antara Y_3dengan variabel tersebut positif, yang berarti jika Y_3 besar berarti maka variabel tersebut bernilai besar meskipun dengan rentang yang tidak begitu besar.

Skor komponen utama diperoleh dari mensubtitusikan setiap observasi yang telah distandardisasi ke Y_1, Y_2, Y_3. Dari skor komponen utama tersebut, selanjutnya akan dilakukan *cluster analysis* dengan Clara Method.

CLARA

Dari analisis cluster dengan Clara Method dengan k dari 1 sampai dengan 10 dengan jarak Manhattan diperoleh nilai rata-rata Overall Average Silhouette Width paling tinggi untuk cluster 5 yaitu sebesar 0,40082. Hal ini menunjukkan bahwa terdapat ikatan yang cukup baik antara objek dan klaster yang terbentuk.

Selanjutnya dilakukan profiliasasi pada hasil kuster. Profilisasi dilakukan pada metode Clara dengan jarak Manhattan dengan jumlah klaster sebanyak 5. Pada tahap profilisasi akan dilihat karakteristik dari tiap klaster yang terbentuk, sehingga dapat dilihat kecenderungan tiap klaster.

Pada metode Clara, karakteristik dari klaster yang terbentuk, direpresentasikan dengan medoid tiap klaster. Selanjutnya, untuk menentukan karakteristik tiap klaster dilakukan perbandingan medoid antar klaster dengan memberikan skor setiap klaster untuk masing-masing variabel.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Tabel 5. Output Metode Clara dengan Jarak Manhattan

	Medoid 1	Medoid 2	Medoid3	Medoid4	Medoid5
PC1	0.6597170	-0.2559214	-3.4573324	-1.2281044	1.3392786
PC2	2.3116726	-0.3103989	-1.2828750	1.5429096	-1.1803006
PC3	-0.4269951	2.0638235	-1.0097316	0.3678118	-0.4829762

Skor disesuaikan dengan korelasi (positif atau negatif) antara variabel asal dengan variabel *Principal Component* dan berdasarkan literatur dari makna setiap variabel, didapatkan:

Tabel 6.Skor dari Medoid

Score	Medoid 1	Medoid 2	Medoid 3	Medoid 4	Medoid 5
PC1	2	3	5	4	1
PC2	1	3	5	2	4
PC3	3	1	5	2	4
Total	6	7	15	8	9

Intrepetasi Setiap Medoid

Medoid 1 merupakan cluster dengan skor 6

Medoid 1 merupakan cluster dengan karakteristik yang tersusun sebagai berikut:

- ✓ PC1 diterangkan dengan korelasi yang tinggi untuk variabel X_1(*log GDP per capita*), X_2 (*social support*), X_3 (*healthy life expectancy at birth)* yang berarti variabel dalam medoid tersebut memiliki nilai rendah dan X_8*(negative affect)* yang memiliki korelasi positif dengan nilai yang tinggi.
- ✓ PC2 diterangkan dengan korelasi negative yang tinggi oleh variabel X_4 (*Freedom to make life choices*), X_7(*Positive affect*), X_9(*Confidence in national government*) yang berarti variabel dalam medoid tersebut memiliki nilai kecil.
- ✓ PC3 diterangkan dengan korelasi positif yang tinggi oleh variabel X_{10} (*gini of household income reported in Gallup*) yang berarti dalam medoid tersebut memiliki nilai yang rendah. Koefisien gini yang rendah berarti ketimpangan kecil, dalam hal ini diberikan skor yang tinggi terkait kontribusinya terhadapat indeks kebahagiaan. Jumlah anggota pada cluster ini sebanyak 26 negara, yaitu Afghanistan, Albania, Algeria, Armenia, Belarus, Bosnia Herzegovina, Kroasia, Agypt, Gabon, Georgia, Greece, Iraq, Jordan, Lebanon, Macedonia, Mauritania, Moldova, Montenegro, Palestina, Terrotories, Serbia, Korea Selatan, Tunisia, Turki, Ukraina, Vietnam, Yaman

Medoid 2 merupakan klaster dengan skor 7

Medoid 2 merupakan cluster dengan karakteristik yang tersusun sebagai berikut:

- ✓ PC1 diterangkan dengan korelasi yang tinggi untuk variabel X_1(*log GDP per capita*), X_2 (*social support*), X_3 (*healthy life expectancy at birth)* yang berarti variabel dalam medoid tersebut memiliki nilai tinggi dan X_8*(negative affect)* yang memiliki korelasi positif dengan nilai yang tinggi.
- ✓ PC2 diterangkan dengan korelasi negatif yang tinggi oleh variabel X_4 (*Freedom to make life choices*), X_7(*Positive affect*), X_9(*Confidence in national government*) yang berarti variabel dalam medoid tersebut memiliki nilai besar (skor tinggi).
- ✓ PC3 diterangkan dengan korelasi positif yang tinggi oleh variabel X_{10} (*gini of household income reported in Gallup*) yang berarti dalam medoid tersebut memiliki nilai yang besar. Koefisien gini yang besar berarti ketimpangan tinggi, dalam hal ini diberikan skor yang rendah terkait kontribusinya terhadapat indeks kebahagiaan. Jumlah anggota pada cluster ini sebanyak 24 negara, yaitu Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominica,

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Ekuador, El Savador, Guatemala, Honduras, Jamaica, Libya, Mexico, Namibia, Nicaragua, Panama, Peru, Sri Lanka, Trinidad dan Tobago, Afrika Selatan, Amerika Serikat, Uruguay.

Medoid 3 merupakan klaster dengan skor 15

Medoid 3 merupakan cluster dengan karakteristik yang tersusun sebagai berikut:

- ✓ PC1 diterangkan dengan korelasi yang tinggi untuk variabel X_1(*log GDP per capita*), X_2 (*social support*), X_3 (*healthy life expectancy at birth*) yang berarti variabel dalam medoid tersebut memiliki nilai tinggi dan X_8*(negative affect)* yang memiliki korelasi positif dengan nilai yang rendah.
- ✓ PC2 diterangkan dengan korelasi negative yang tinggi oleh variabel X_4 (*Freedom to make life choices*), X_7(*Positive affect*), X_9*(Confidence in national government*) yang berarti variabel dalam medoid tersebut memiliki nilai besar.
- ✓ PC3 diterangkan dengan korelasi positif yang tinggi oleh variabel X_{10} (*gini of household income reported in Gallup*) yang berarti dalam medoid tersebut memiliki nilai yang rendah. Koefisien gini yang rendah berarti ketimpangan kecil, dalam hal ini diberikan skor yang tinggi terkait kontribusinya terhadapat indeks kebahagiaan. Jumlah anggota pada cluster ini sebanyak 22 negara, yaitu Australia, Austria, Belgium, Denmark, Finlandia, Jerman, Hong Kong, Islandia, Irlandia, Kyrgystan, Luxemburg, Malta, Mauritius, Belanda, Selandia Baru, Norway, Singapura, Arab Saudi, Switzerland, Swedia, United Kingdom, dan Uzbekistan.

Medoid 4 merupakan klaster dengan skor 8

Medoid 4 merupakan cluster dengan karakteristik yang tersusun sebagai berikut:

- ✓ PC1 diterangkan dengan korelasi yang tinggi untuk variabel X_1(*log GDP per capita*), X_2 (*social support*), X_3 (*healthy life expectancy at birth)* yang berarti variabel dalam medoid tersebut memiliki nilai tinggi dan X_8*(negative affect)* yang memiliki korelasi positif dengan nilai yang tinggi.
- ✓ PC2 diterangkan dengan korelasi negatif yang tinggi oleh variabel X_4 (*Freedom to make life choices*), X_7(*Positive affect*), X_9(*Confidence in national government*) yang berarti variabel dalam medoid tersebut memiliki nilai rendah.
- ✓ PC3 diterangkan dengan korelasi positif yang tinggi oleh variabel X_{10} (*gini of household income reported in Gallup)* yang berarti dalam medoid tersebut memiliki nilai yang tinggi. Koefisien gini yang tinggi berarti ketimpangan besar, dalam hal ini diberikan skor yang rendah terkait kontribusinya terhadapat indeks kebahagiaan. Anggota dari Klaster 4 antara lain, Azerbaijan, Bahrain, Bulgaria, Cyprus, Republik Ceko, Estonia, Perancis, Hongaria, Israel, Italia, Jepang, Kazakhstan, Kosovo, Kuwait, Latvia, Lituania, Mongolia, Polandia, Turkmenistan, Romania, Russia, Arab Saudi, Slovakia, Slovenia, Spanyol, Portugal, Thailand, Taiwan.

Medoid 5 merupakan klaster dengan skor 9

Medoid 5 merupakan cluster dengan karakteristik yang tersusun sebagai berikut:

- ✓ PC1 diterangkan dengan korelasi yang tinggi untuk variabel X_1(*log GDP per capita*), X_2 (*social support*), X_3 (*healthy life expectancy at birth*) yang berarti variabel dalam medoid tersebut memiliki nilai rendah dan X_8 (*negative affect*) yang memiliki korelasi positif dengan nilai yang besar.
- ✓ PC2 diterangkan dengan korelasi negative yang tinggi oleh variabel X_4 (*Freedom to make life choices*), X_7(*Positive affect*), X_9(*Confidence in national government*) yang berarti variabel dalam medoid tersebut memiliki nilai tinggi.
- ✓ PC3 diterangkan dengan korelasi positif yang tinggi oleh variabel X_{10} (*gini of household income reported in Gallup*) yang berarti dalam medoid tersebut memiliki nilai yang rendah. Koefisien gini yang rendah berarti ketimpangan rendah, dalam hal ini diberikan skor yang tinggi terkait kontribusinya terhadapat indeks kebahagiaan. Jumlah anggota pada cluster ini sebanyak 41 negara, yaitu Bangladesh, Benin, Botswana, Burkina Faso, Cambodia, Cameroon, Zimbabwe, Chad, Congo (Brazzaville), Congo (Kinshasa), Ethiopia, Ghana,

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Guinea, Haiti, India, Indonesia, Iran, Pantai Gading, Kenya, Laos, Liberia, Madagaskar, Malawi, Mali, Morocco, Mozambik, Myanmar, Nepal, Niger, Nigeria, Pakistan, Filipina, Senegal, Sierra Leone, Sudan Selatan, Tajikistan, Tanzania, Togo, Uganda, Zambia, Republik Afrika Tengah.

4. KESIMPULAN DAN SARAN

Kesimpulan

Metode ROBPCA merupakan suatu model PCA yang *robust* terhadap outlier, dan lebih efisien (mampu menghasilkan lebih sedikit jumlah Principal Component) daripada Classic PCA. Dengan metode ROBPCA yang diterapkan untuk studi kasus variable-variabel penyusun indeks kebahagiaan dari data The World Happiness Report 2018, didapatkan 3 (tiga) Principal Component yang dapat menjelaskan sebesar 92,6 % dari total varians data. Hal ini terbukti ROBPCA lebih efisien daripada metode PCA dengan sebanyak 5 Principal Componen yang dapat menjelaskan sebesar 85,89% dari total varians. Metode Clara merupakan Analysis Cluster dengan pusat cluster medoid yang robust untuk mengelompokkan data dengan outlier dan data dalam jumlah besar. Analysis Cluster dengan Clara method dari Principal Component yang terbentuk menggunakan jarak manhattan didapatkan nilai rata-rata Overall Average Silhouette Width yang terbaik pada 5 cluster. Berdasarkan profiling cluster dengan score antar medoid, didapatkan score paling tinggi pada cluster 3. Berdasrkan literature dari setiap variabel berarti bahwa cluster 3 terdiri dari negara-negara dengan indeks kebahagiaan yang baik. Sedangkan score medoid terendah adalah cluster 1..

Saran

Penerapan metode ROBPCA dan Metode Clara (Clustering Large Area) pada studi kasus yang lain khususnya untuk high dimensional data.

UCAPAN TERIMA KASIH

Terima kasih kepada Bpk Dr. Irlandia Ginanjar,S.Si., M.Si atas masukan dan bimbingannya untuk penelitian ini.

REFERENSI

Hubert, M,,Rousseeuw, P,J,, and Branden, K,,V, 2005, *ROBPCA:* a New Approach to Robust Principal Component Analysis, American Statistical and the American Society for Quality, *Technometric, Vol,47, No,1, Belgium.*

Johnson, R.A. dan Winchern, D.W.2007, *Applied Multivariate Statistical Analysis*. 6th edition, Pearson Education,Inc., USA.

Kassambra, Alboukadel, 2017, Practical Guide to Cluster Analysis in R.STHDA.

Kaufman, L. dan Rousseeuw, P.J., 1990, FindingGroups in Data : *An Introduction to Cluster Analysis*, John Wiley and Sons, Inc, New Jersey.

Muslim,A,B, 2018, *Cluster Analysis using Clara Method for Data with Outlier,* FMIPA UGM, Yogyakarta.

Rencher, A.C., 2002, *Methods of Multivariate Analysis*, Second Edition, John Wiley and Sons, Inc., New York.

Sobiroh,T,R, 2015, *Robust Principal Component Analysis (ROBPCA) for High Dimensional Data with Outlier*, FMIPA UGM, Yogyakarta.

Statistical Appendix 1 for Chapter 2 of World Happiness Report.2018.