

ANALISIS ANGKA KEMATIAN BAYI (AKB) DI KALIMANTAN BARAT DENGAN *ROBUST PRINCIPAL COMPONENT ANALYSIS (ROBPCA)*

Lina Astuti, Yundari

INTISARI

Principal Component Analysis (PCA) merupakan teknik analisis multivariat yang bertujuan mereduksi dimensi data dimana sejumlah variabel bebas yang masih saling berkorelasi, dengan mempertahankan sebesar mungkin varians data menjadi kumpulan data baru yang tidak berkorelasi lagi. PCA sangat dipengaruhi oleh kehadiran pencilan (*outlier*) karena PCA didasarkan pada matriks kovarians yang sensitif terhadap keberadaan data *outlier*. *Robust Principal Component Analysis (ROBPCA)* merupakan suatu analisis komponen utama yang robust terhadap keberadaan *outlier* dalam data pengamatan. Dalam analisis ini menggabungkan konsep *Projection Pursuit (PP)* dan *Minimum Covariance Determinant (MCD)*. Tujuan penelitian ini adalah mendapatkan komponen utama dengan data yang terindikasi masalah multikolinearitas dan *outlier*, serta mendapatkan model persamaan terbaik yang diterapkan pada data Angka Kematian Bayi (AKB). Variabel-variabel bebas yang memengaruhi AKB yaitu, jumlah ibu hamil, jumlah persalinan yang ditolong tenaga kesehatan, jumlah tenaga medis, jumlah ibu hamil yang mengalami komplikasi kebidanan, dan presentase penduduk miskin. Pada proses awal data pengamatan dilakukan uji multikolinearitas dan selanjutnya menentukan komponen utama dengan PCA. Setelah mendapatkan komponen utama, dilakukan analisis regresi. Pendeteksian *outlier* menggunakan jarak robust Mahalanobis. Ketika terdapat *outlier*, maka proses dilanjutkan menggunakan ROBPCA. Hasil analisis menunjukkan bahwa ROBPCA dapat menghasilkan 2 komponen utama dari 5 variabel asal. Berdasarkan penelitian model regresi PCA dan ROBPCA sama-sama memiliki nilai *R-square* sebesar 0,4206 artinya bahwa penelitian ini mampu menjelaskan 42,06% varians Y. Akan tetapi pada *Residual Standard Error (RSE)* untuk model regresi PCA sebesar 3,526 lebih besar daripada model regresi ROBPCA yaitu sebesar 0,827. Model terbaik yang didapatkan untuk analisis Angka Kematian Bayi di Kalimantan Barat adalah ROBPCA.

Kata Kunci : PCA, robust, ROBPCA, outlier, *R-square*, RSE

PENDAHULUAN

Angka Kematian Bayi (AKB) di Kalimantan Barat sebesar 7 per 1000 menurut Dinas Kesehatan Provinsi Kalimantan Barat tahun 2018. *Millenium Development Goals (MDGs)* memiliki target yaitu menurunkan AKB, suatu komitmen bersama masyarakat internasional untuk mempercepat pembangunan manusia dan mengentaskan kemiskinan. Estimasi AKB menjadi penting mengingat AKB merupakan salah satu indikator pembangunan bidang kesehatan di suatu wilayah. Pentingnya mengestimasi AKB guna menunjang pembangunan ini, maka studi kasus yang dipilih dalam penelitian ini yaitu menganalisis AKB per kabupaten/kota di Provinsi Kalimantan Barat dengan menggunakan analisis multivariat. Analisis multivariat merupakan metode pengolahan variabel dalam jumlah banyak, dimana tujuannya adalah untuk mencari pengaruh variabel-variabel tersebut terhadap suatu objek secara simultan. Metode yang digunakan dalam analisis ini adalah *Robust Principal Component Analysis (ROBPCA)*.

Robust merupakan metode yang pada awalnya dipublikasikan oleh Andrews (1997) yang selanjutnya dikembangkan oleh Ryan (1997), sebagai metode yang digunakan saat terdapat data *outlier* pada suatu pengamatan. Hingga saat ini belum ada pembaharuan metode tersebut sehingga *robust* dianggap kompatibel dalam mengatasi masalah kehadiran *outlier* yang dapat menghasilkan model yang *robust* terhadap *outlier*. ROBPCA adalah suatu metode yang kekar (*robust*) untuk

Principal Component Analysis (PCA) terhadap keberadaan *outlier* pada data pengamatan. Metode ROBPCA menggabungkan konsep *Projection Pursuit* (PP) dengan penaksir *robust Minimum Covariance Determinant* (MCD). Penelitian ini diharapkan mendapatkan komponen utama dengan data yang terindikasi masalah multikolinearitas dan *outlier*, serta mendapatkan model persamaan terbaik yang diterapkan pada data Angka Kematian Bayi (AKB) di Kalimantan Barat pada tahun 2018.

Pada tahapan awal analisis, data pengamatan distandarisasikan karena memiliki satuan data yang memiliki *range* berbeda. Setelah standarisasi satuan data, dilakukan uji multikolinearitas untuk mendeteksi adanya hubungan antar variabel bebas lainnya. Setelah dilakukan uji multikolinearitas, selanjutnya mencari komponen utama dengan *Principal Component Analysis* (PCA). Setelah mendapatkan komponen utama, tahap selanjutnya yaitu melakukan analisis regresi dengan komponen utama yang terpilih. Selanjutnya mendeteksi *outlier* dengan mengukur jarak *robust* Mahalanobis pada data pengamatan. Setelah *outlier* terdeteksi, maka proses dilanjutkan dengan mencari komponen utama yang *robust* dengan *Robust Principal Component Analysis* (ROBPCA). Setelah komponen utama yang *robust* didapatkan, maka proses selanjutnya melakukan analisis regresi pada komponen utama *robust* yang terpilih. Tahapan terakhir melakukan uji multikolinearitas untuk melihat apakah masalah multikolinearitas sudah teratasi dengan menggunakan ROBPCA.

PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis adalah suatu teknik analisis statistik untuk mentransformasi variabel-variabel asli yang masih saling berkorelasi satu dengan yang lain menjadi satu set variabel baru yang tidak berkorelasi lagi. Secara umum tujuan utama dari PCA adalah mereduksi dimensi data. PCA merupakan solusi bagi proses pengumpulan data dimana data tersebut terdiri dari variabel-variabel yang jumlahnya sangat banyak sehingga diperoleh variabel-variabel baru yang jumlahnya lebih sedikit tetapi mampu menjelaskan varians data [1].

Perkembangan PCA dimulai sejak 1901 yang diperkenalkan pertama kali oleh Pearson. Sejalan dengan perkembangan teknologi komputer dan kemajuan di bidang matematika, PCA hingga kini masih terus mengalami perkembangan. Perkembangan selanjutnya, diperkenalkan generalisasi dari PCA pada tahun 1963 oleh Loeve. Perkembangan PCA selanjutnya dipengaruhi adanya kebutuhan suatu model PCA yang *robust* terhadap data pencilan (*outlier*). PCA sangat dipengaruhi oleh kehadiran pencilan (*outlier*) karena PCA didasarkan pada matriks kovarians yang sensitif terhadap keberadaan data *outlier* [2].

Misal diberikan matriks berordo $n \times p$ dengan n banyaknya sampel dan p variabel yang telah dilakukan standarisasi data. Banyaknya variabel yang dinyatakan sebagai berikut:

$$\mathbf{Z}_{n \times p} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \cdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}$$

Dengan: $\mathbf{Z} = [Z_1, Z_2, \dots, Z_p]$

$$\mathbf{Z}_i = [z_{i1}, z_{i2}, \dots, z_{ip}]$$

$$i = 1, 2, \dots, n$$

Langkah-langkah metode PCA adalah sebagai berikut:

1. Menghitung matriks varian kovarians $\hat{\mathbf{S}}$.
2. Dari matriks varian kovarians sampel dapat diperoleh nilai eigen (*eigen value*) yaitu $\lambda_1, \lambda_2, \dots, \lambda_p$ dimana $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Untuk menghitung nilai eigen dengan cara menyelesaikan persamaan karakteristik dari matriks varian kovarians adalah sebagai berikut:

$$|\hat{S} - \lambda I| = 0 \tag{1}$$

3. Setelah didapat nilai eigen akan ditentukan banyaknya komponen utama. Terdapat tiga cara yang digunakan untuk menentukan jumlah komponen utama, yaitu:
 - a. Nilai eigen lebih dari satu.
 - b. Penentuan banyaknya komponen utama yang dipilih juga dilakukan dengan memperhatikan patahan siku dari *screeplot*.
 - c. Total varians yang dijelaskan lebih dari 80%.
4. Untuk menentukan model komponen utama, terlebih dahulu dihitung vektor eigen untuk setiap nilai eigen. Untuk nilai eigen (λ) terdapat vektor eigen yang bersesuaian v_1, v_2, \dots, v_p dengan memenuhi persamaan berikut:

$$(\hat{S} - \lambda I)Z = 0 \tag{2}$$

5. Membentuk komponen utama dengan menggunakan persamaan:

$$\begin{aligned} K_1 &= v_{11}Z_1 + v_{12}Z_2 + \dots + v_{1p}Z_p \\ K_2 &= v_{21}Z_1 + v_{22}Z_2 + \dots + v_{2p}Z_p \\ &\vdots \\ K_p &= v_{p1}Z_1 + v_{p2}Z_2 + \dots + v_{pp}Z_p \end{aligned} \tag{3}$$

dengan : K = komponen utama

v_p = vektor eigen

Z_p = variabel yang sudah di standarisasi

METODE PROJECTION-PURSUIT (PP)

Metode *Projection-Pursuit* (PP) termasuk kedalam kelompok metode pereduksian dimensi yang berdasarkan pencarian suatu proyeksi informasi utama dari data berdimensi besar [4]. Ide dasar metode *Projection-Pursuit* (PP) adalah untuk memperoleh informasi penting dalam data berdimensi besar (banyak variabel bebas) melalui proyeksi ke data berdimensi lebih kecil dengan cara memaksimalkan fungsi objektif yang disebut indeks proyeksi (*Projection Index*).

Metode *Projection-Pursuit* (PP) berbeda dengan metode *Principal Component Analysis* (PCA) dalam hal prosedur reduksi dimensi. Pereduksian dalam PCA berdasarkan varians terbesar dalam data asal. Komponen-komponen dari variabel-variabel asal dengan varians terbesar. Sedangkan untuk metode *Projection-Pursuit* (PP) berdasarkan pemaksimalan indeks proyeksi, sehingga informasi yang ada dalam data asal seperti keadaan data non linear akan tercermin dalam data hasil proyeksi. Hasil reduksi dimensi PP kemudian digunakan untuk melakukan pemodelan *Projection-Pursuit Regression* (PPR), sedangkan hasil reduksi dimensi PCA digunakan untuk melakukan pemodelan *Principal Component Regression* (PCR).

MINIMUM COVARIANCE DETERMINANT (MCD)

Metode MCD merupakan penaksir *robust* untuk rata-rata dan matriks kovarians dengan mencari sebagian data yang mempunyai kovarians minimum yang digunakan untuk mengidentifikasi *outlier*, menentukan jarak dan residu *robust* yang akan digunakan dalam pembobotan data dan penentuan parameter regresi. Metode ini bertujuan untuk mendapatkan suatu sub sampel berukuran h dari keseluruhan pengamatan n , yang matriks varian kovariansnya memiliki determinan terkecil diantara semua kombinasi kemungkinan data, dengan:

$$h = \frac{n+k+1}{2} \tag{4}$$

dengan k menyatakan banyak variabel. Jika nilai h merupakan pecahan maka nilai dari h dibulatkan kebawah [5]. Misalkan terdapat sampel acak, yaitu $\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_p$ diambil dari distribusi yang mempunyai vektor rata-rata $\bar{\mathbf{M}}$ dan matriks varian kovarians $\hat{\mathbf{S}}$, dengan matriks score komponen utama yang di notasikan dengan $\mathbf{M} = (\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_p)$,

$$\mathbf{K}_1 = \begin{pmatrix} k_{11} \\ k_{21} \\ \vdots \\ k_{n2} \end{pmatrix}, \mathbf{K}_2 = \begin{pmatrix} k_{12} \\ k_{22} \\ \vdots \\ k_{n2} \end{pmatrix} \text{ sampai } \mathbf{K}_m = \begin{pmatrix} k_{1p} \\ k_{2p} \\ \vdots \\ k_{np} \end{pmatrix}. \text{ Penduga MCD untuk } \bar{\mathbf{M}} \text{ dan } \hat{\mathbf{S}} \text{ masing-masing}$$

adalah $\bar{\mathbf{M}}_{MCD}$ dan $\hat{\mathbf{S}}_{MCD}$ dengan:

$$\bar{\mathbf{M}}_{MCD} = \frac{1}{h} \sum_{i=1}^h \mathbf{K}_i \quad (5)$$

$$\hat{\mathbf{S}}_{MCD} = \frac{1}{h} \sum_{i=1}^h (\mathbf{K}_i - \bar{\mathbf{M}}_{MCD})(\mathbf{K}_i - \bar{\mathbf{M}}_{MCD})^T \quad (6)$$

dan determinan matriks varians-kovarians \mathbf{S}_{MCD} minimum diantara semua kemungkinan h . Jika n kecil ($n \leq 600$) maka pendugaan MCD mudah dan relatif lebih cepat untuk ditemukan. Sedangkan jika n besar ($n > 600$) maka banyak sekali kombinasi subhimpunan yang harus ditemukan untuk mendapatkan pendugaan MCD. Keterbatasan ini kemudian diatasi dengan algoritma yang dikenal dengan istilah FAST-MCD [6]. Berdasarkan Rousseeuw dan Van Driessen tahun 1998, algoritmanya adalah sebagai berikut:

1. Ambil himpunan bagian dari matriks \mathbf{M} yang terdiri atas $h = \frac{n+k+1}{2}$ buah data dan disimbolkan dengan H_1 .
2. Hitung vektor rata-rata $\bar{\mathbf{M}}_{i1}$ dan matriks varians-kovarians $\hat{\mathbf{S}}_1$.
3. Kemudian hitung jarak mahalnobis $MD_1 = \sqrt{(\mathbf{L}_i - \bar{\mathbf{M}})^T \hat{\mathbf{S}}^{-1} (\mathbf{L}_i - \bar{\mathbf{M}})}$ dengan \mathbf{L}_i merupakan vektor komponen utama pada pengamatan ke- i .
4. Urutkan \mathbf{L}_i berdasarkan nilai MD_1 dari yang terkecil ke nilai yang terbesar.
5. Definisikan himpunan bagian baru yang dinotasikan dengan $\mathbf{H}_2 = \{\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_h\}$.
6. Hitung $\bar{\mathbf{M}}_{i2}$, $\hat{\mathbf{S}}_2$, dan MD_2 .
7. Ulangi langkah 1 sampai langkah 6 hingga ditemukan $\det(\hat{\mathbf{S}}_2) \leq \det(\hat{\mathbf{S}}_1)$.

Karena $\bar{\mathbf{M}}_{MCD}$ dan $\hat{\mathbf{S}}_{MCD}$ merupakan penduga untuk MCD maka $\bar{\mathbf{M}}_{MCD} = \bar{\mathbf{M}}_2$ dan $\hat{\mathbf{S}}_{MCD} = \hat{\mathbf{S}}_2$. Metode mempunyai kemampuan mengukur jarak *robust* yang dapat digunakan untuk mendeteksi *outlier leverage*.

Jarak *robust* merupakan suatu pendekatan untuk mendeteksi *outlier* pada data multivariat, yaitu dengan menggunakan penduga dari $\bar{\mathbf{M}}_{MCD}$ dan $\hat{\mathbf{S}}_{MCD}$ pada metode *robust*. Sehingga metode ini mampu meminimumkan pengaruh dari adanya efek *masking* dan *swamping* dalam pendeteksian *outlier*. Terdapat beberapa penyebab munculnya *outlier*, salah satunya *outlier* yang disebabkan oleh variabel bebas, dinamakan *outlier leverage*. *Outlier leverage* dideteksi dengan menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke- i . Jarak *robust* didefinisikan pada persamaan berikut:

$$RD_i = \sqrt{(\mathbf{L}_i - \bar{\mathbf{M}}_{MCD})^T \hat{\mathbf{S}}^{-1}_{MCD} (\mathbf{L}_i - \bar{\mathbf{M}}_{MCD})}, \quad i = 1, 2, \dots, n \quad (7)$$

dengan RD_i merupakan jarak *robust* untuk setiap pengamatan ke- i . \mathbf{L}_i merupakan vektor komponen utmama pada pengamatan ke- i , $\bar{\mathbf{M}}_{MCD}$ merupakan vektor rata-rata dari \mathbf{K}_j dengan metode MCD, $\hat{\mathbf{S}}_{MCD}$ merupakan matriks varian kovarians sampel dengan MCD, $\hat{\mathbf{S}}^{-1}_{MCD}$ merupakan invers dari matriks $\hat{\mathbf{S}}$ dengan, $\mathbf{L}_i = (k_{i1}, k_{i2}, \dots, k_{im})$, $\bar{\mathbf{M}}_{MCD} = (\bar{K}_1, \bar{K}_2, \dots, \bar{K}_m)$, dan $\hat{\mathbf{S}}_{MCD} =$

$$\begin{pmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & s_{mm} \end{pmatrix}$$

pendeteksian *outlier leverage* menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke- i dapat dituliskan sebagai berikut (Chen, 2002):

$$leverage = \begin{cases} \text{jika } RD_i \leq C, \text{ maka pengamatan bukan } outlier \text{ (diberi kode 0)} \\ \text{jika } RD_i > C, \text{ maka pengamatan merupakan } outlier \text{ (diberi kode 1)} \end{cases}$$

dengan $C = \sqrt{\chi^2_{p;\alpha}}$, C dinyatakan sebagai nilai *cut-off*. Dalam hal ini, nilai *cut-off* merupakan suatu nilai yang digunakan untuk menentukan apakah suatu pengamatan dideteksi sebagai *outlier* atau bukan. Notasi $\chi^2_{p;\alpha}$ merupakan nilai χ^2 yang membuat luas di ujung kanan distribusinya sama dengan α dan RD_i merupakan jarak *robust* untuk setiap pengamatan ke- i .

ROBUST PRINCIPAL COMPONENT ANALYSIS (ROBPCA)

ROBPCA adalah suatu metode yang kekar (*robust*) untuk PCA terhadap keberadaan *outlier* pada data [3]. Metode ROBPCA menggabungkan konsep *Projection Pursuit* (PP) dengan penaksir *robust Minimum Covariance Determinant* (MCD). Keunggulan ROBPCA yaitu dapat menyelesaikan masalah multikolinearitas dan *outlier* dengan membuat komponen utama yang baru sebagai variabel bebas yang selanjutnya diregresikan dengan variabel terikat yang menghasilkan model yang *robust* terhadap *outlier*.

Pada metode ROBPCA terdapat tahapan untuk memproses suatu data multivariat. Langkah-langkah metode ROBPCA adalah sebagai berikut:

1. Mereduksi ruang data, terutama ketika $p \geq n$, dimana n merupakan jumlah observasi dan p adalah jumlah variabel bebas. Langkah ini dilakukan dengan metode *Singular Value Decomposition* (SVD) terhadap matriks data yang telah dipusatkan, dengan rumus berikut:

$$\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\boldsymbol{\mu}}'_0 = \mathbf{U}_{n,r0} \mathbf{D}_{r0,r0} \mathbf{V}'_{r0,p} \quad (8)$$

dimana $\hat{\boldsymbol{\mu}}_0$ merupakan vektor rata-rata klasik, $r0 = rank(\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\boldsymbol{\mu}}'_0)$, \mathbf{D} adalah matriks diagonal berukuran $r0 \times r0$, dan $\mathbf{U}'\mathbf{U} = \mathbf{I}_{r0} = \mathbf{V}'\mathbf{V}$, dimana \mathbf{I}_{r0} adalah matriks identitas berukuran $r0 \times r0$.

2. Menemukan h keterpencilan terkecil, tahap ini dilakukan dengan memilih $\frac{1}{2} < \alpha < 1$ untuk mendapatkan nilai $h = \frac{n+k+1}{2}$, dimana k merupakan jumlah komponen yang akan dihitung.

Selanjutnya keterpencilan dihitung dengan rumus Stahel-Donoho sebagai berikut:

$$Outl_o(\mathbf{X}_i) = \max_{v \in B} \frac{|x'_i v - \hat{\mu}_{MCD}(x'_j v)|}{\hat{S}_{MCD}(x'_j v)} \quad (9)$$

dimana $\hat{\mu}_{MCD}$ dan \hat{S}_{MCD} merupakan penduga nilai tengah dan simpangan baku MCD univariat. Sebanyak h pengamatan dengan nilai keterpencilan terkecil dihitung vektor nilai tengah ($\hat{\mu}_1$) dan matriks varian kovariansnya (\hat{S}_1). Kemudian matriks varian kovarians didekomposisi sehingga diperoleh komponen utamanya. Sebanyak k komponen utama pertama dipilih dan semua data diproyeksikan pada subruang berdimensi- k yang direntang oleh k vektor ciri diperoleh $\mathbf{X}_{n,k}$.

3. Pada $\mathbf{X}_{n,k}$ dari langkah 2, dihitung kembali penduga nilai tengah ($\hat{\mu}_2$) dan matriks varian kovarians MCD (\hat{S}_2).

Perhitungan pada langkah terakhir algoritma ROBPCA memerlukan sejumlah h data dengan matriks varian kovarians yang minimum. Oleh karenanya, digunakan penduga *Minimum Covariance Determinant* (MCD), yang dihitung dengan algoritma FAST-MCD yang diadaptasi.

HASIL DAN PEMBAHASAN

Data pada penelitian ini menggunakan data Angka Kematian Bayi di Kalimantan Barat pada tahun 2018 dengan variabel-variabel yang mempengaruhinya yaitu, jumlah ibu hamil (X_1), jumlah persalinan yang ditolong tenaga kesehatan (X_2), jumlah tenaga medis (X_3), jumlah ibu hamil yang mengalami komplikasi kebidanan (X_4), dan presentase penduduk miskin (X_5). Data angka kematian bayi terdiri dari variabel-variabel yang mempunyai satuan data yang berbeda, maka dilakukan standarisasi satuan data. Setelah data di standarisasi, langkah selanjutnya adalah melakukan uji multikolinearitas pada data pengamatan sebagai berikut:

Tabel 1. Variance Inflation Factors (VIF)

Variabel	VIF
Z_1	1669,3151
Z_2	61,3626
Z_3	2,7136
Z_4	1675,7623
Z_5	1,4587

Dengan Z_1 , Z_2 , Z_3 , Z_4 , dan Z_5 merupakan variabel yang telah di standarisasi. Tabel 1 menunjukkan bahwa terdapat tiga variabel yang mempunyai nilai $VIF > 10$ yaitu Z_1 , Z_2 , Z_4 , sehingga dapat disimpulkan bahwa pada ketiga variabel tersebut terjadi masalah multikolinearitas. Selanjutnya menentukan komponen utama dengan menggunakan *Principal Component Analysis* (PCA). Penentuan jumlah komponen utama yang terbentuk dapat dilakukan menggunakan tiga kriteria. Langkah-langkahnya adalah sebagai berikut:

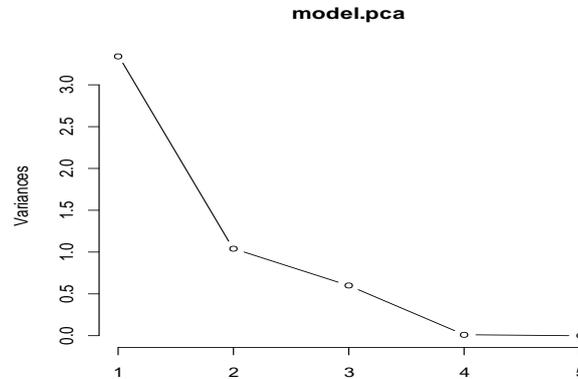
1. Penentuan nilai eigen

Tabel 2. Penentuan Nilai Eigen

Nilai Eigen PCA	
Komponen Utama ke-	Nilai Eigen
1	3,3425
2	1,0434
3	0,6024
4	0,0112
5	0,0003

Berdasarkan Tabel 2 menunjukkan bahwa terdapat dua komponen utama yang memiliki nilai eigen lebih besar satu, sehingga berdasarkan kriteria nilai eigen terdapat dua komponen utama yang terbentuk.

2. Scree Plot



Gambar 1. Scree Plot Penentuan Komponen Utama

Pada Gambar 1 menunjukkan bahwa setelah komponen utama ke dua, plot akan meluruh sehingga berdasarkan kriteria ini dapat disimpulkan bahwa komponen utama yang terbentuk adalah dua komponen utama.

3. Presentase proporsi varians kumulatif

Tabel 3. Proporsi Varians dan Proporsi Varians Kumulatif PCA

	K_1	K_2	K_3	K_4	K_5
Z_1	-0,5289	0,2326	-0,1039	0,3957	0,7061
Z_2	-0,5392	0,1312	-0,059	-0,8296	0,009
Z_3	-0,3305	-0,4815	0,8075	0,081	-0,0154
Z_4	-0,5259	0,2501	-0,118	0,382	-0,7077
Z_5	0,2087	0,7963	0,5654	-0,0498	0,0051
Proporsi Varians	0,6685	0,2087	0,1205	0,00226	0,00006
Proporsi Varians Kumulatif	0,6685	0,8772	0,9977	0,99994	1,00000

Pada Tabel 3 menunjukkan bahwa pada komponen utama pertama menjelaskan 66,85% dari total varians sampel, pada komponen utama ke dua menjelaskan 87,72% dari varians sampel. Berdasarkan kriteria ini, dapat disimpulkan bahwa dua komponen utama telah mampu menjelaskan 87,72% dari total varians sampel. Dari ketiga kriteria tersebut dapat disimpulkan bahwa jumlah komponen utama yang terbentuk dengan metode PCA adalah dua komponen utama. Setelah mendapatkan komponen utama, tahap selanjutnya yaitu melakukan analisis regresi dengan komponen utama yang terpilih sebagai berikut:

Tabel 4. Model Regresi PCA

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	8,1555	0,9424	8,654	0,000003
K_1	1,4643	0,5349	2,737	0,0193
K_2	0,6707	0,9574	2,701	0,0498

Berdasarkan hasil *output* diperoleh F_{hitung} sebesar 3,992 lebih besar dari $F_{tabel(0,05;4;9)}$ sebesar 3,633 atau P -value sebesar 0,04972 lebih kecil dari $\alpha(0,05)$ sehingga dapat disimpulkan bahwa secara

simultan berpengaruh terhadap Y . Berdasarkan nilai R -square sebesar 0,4206 artinya bahwa penelitian ini hanya mampu menjelaskan 42,06% varians Y ditentukan oleh banyaknya komponen utama yang digunakan, sedangkan 57,94% sisanya dipengaruhi oleh varians diluar model dan *Residual Standard Error* (RSE) untuk model regresi PCA sebesar 3,526. Model persamaan regresi untuk Y adalah sebagai berikut:

$$Y = 8,1555 - 1,4643K_1 - 0,6707K_2$$

Langkah selanjutnya adalah mendeteksi adanya *outlier* pada data pengamatan dengan menghitung jarak Mahalanobis sebagai berikut:

Tabel 5. Jarak Mahalanobis dan Jarak Robust setiap pengamatan

Pengamatan per Kab./Kota	Md_i	RD_i
Kab. Sambas	3,0134	1,0522
Kab. Bengkayang	0,94	2,244
Kab. Landak	7,7259	16,4389
Kab. Mempawah	1,3978	0,9835
Kab. Sanggau	3,1488	1,3323
Kab. Ketapang	8,1797	2,8746
Kab. Sintang	2,4632	2,9012
Kab. Kapuas Hulu	2,5363	1,8144
Kab. Sekadau	2,3432	2,9088
Kab. Melawi	1,8481	2,2102
Kab. Kayong Utara	3,2539	1,4630
Kab. Kubu Raya	4,6692	2,347
Kota Pontianak	11,4092	118,2675
Kota Singkawang	12,0705	107674

Nilai *cut-off* pada jarak *robust* adalah $\chi^2_{(p;\alpha)} = \chi^2_{(5;0,95)} = 11,0705$. Nilai jarak *robust* yang lebih besar dari nilai *cut-off* yang dideteksi sebagai *outlier*. Sedangkan nilai jarak *robust* yang lebih kecil dari nilai *cut-off* bukan dideteksi sebagai *outlier*. Setelah menghitung jarak *robust* Mahalanobis untuk setiap kabupaten dan kota di Kalimantan Barat, maka dapat diketahui kabupaten/kota yang terindikasi *outlier*. Pada hasil perhitungan jarak *robust* Mahalanobis diperoleh Kabupaten Landak, Kota Pontianak dan Kota Singkawang terindikasi sebagai data *outlier*. Karena pada data pengamatan terdapat *outlier*, maka proses dilanjutkan menggunakan *Robust Principal Component Analysis* (ROBPCA). Langkah pertama adalah menentukan komponen utama yang *robust* yaitu sebagai berikut:

Tabel 6. Komponen Utama yang robust

	RK_1	RK_2
Z_1	0,546477	-0,00234
Z_2	0,500005	-0,05041
Z_3	0,372117	0,382213
Z_4	0,545463	-0,00256
Z_5	-0,12393	0,922692

Berdasarkan Tabel 6 didapatkan dua komponen utama yang *robust*. Setelah mendapatkan dua komponen utama dengan ROBPCA, dilakukan regresi ROBPCA dari kedua komponen utama tersebut. Model regresi dari ROBPCA adalah sebagai berikut:

Tabel 7. Model Regresi ROBPCA

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	0,00000000003	0,2212	0,000	1,0000
RK_1	-0,3241	0,1277	-2,538	0,0276
RK_2	0,2206	0,2671	2,826	0,0426

Berdasarkan hasil *output* diperoleh F_{hitung} sebesar 3,992 lebih besar dari $F_{tabel(0,05;4;9)}$ sebesar 3,633 atau P -value sebesar 0,0497 lebih kecil dari $\alpha(0,05)$ sehingga dapat disimpulkan bahwa kedua variabel bebas signifikan secara bersama-sama dalam mempengaruhi variabel terikat. Berdasarkan nilai R -square sebesar 0,4206 artinya bahwa penelitian ini mampu menjelaskan 42,06% varians Y ditentukan oleh banyaknya komponen utama *robust* yang terbentuk, sedangkan 57,94% sisanya dipengaruhi oleh varians diluar model dan *Residual Standard Error* (RSE) untuk model regresi ROBPCA sebesar 0,8275. Model persamaan regresi untuk Y adalah sebagai berikut:

$$Y = 0,00000000003 - 0,3241RK_1 + 0,2206RK_2$$

Dengan kombinasi linear variabel-variabel asalnya sebagai berikut:

$$RK_1 = 0,5464Z_1 + 0,500005Z_2 + 0,3721Z_3 + 0,5454Z_4 - 0,1239Z_5$$

$$RK_2 = -0,0023Z_1 - 0,0504Z_2 + 0,3822Z_3 - 0,0025Z_4 + 0,9226Z_5$$

Pada model regresi ROBPCA terlihat bahwa nilai penduga yang dihasilkan belum menunjukkan nilai penduga yang sebenarnya. Untuk memperoleh model yang sesuai maka perlu dilakukan transformasi data sebagai berikut:

1. Mensubstitusikan nilai sehingga diperoleh:

$$\begin{aligned} Y &= 0,00000000003 - 0,3241(0,5464Z_1 + 0,500005Z_2 + 0,3721Z_3 + 0,5454Z_4 - 0,1239Z_5) \\ &\quad + 0,2206(-0,023Z_1 - 0,054Z_2 + 0,3822Z_3 - 0,0025Z_4 + 0,9226Z_5) \\ &= 0,00000000003 - 0,177Z_1 - 0,162Z_2 - 0,1205Z_3 - 0,1767Z_4 + 0,0401Z_5 \\ &\quad - 0,005Z_1 - 0,0111Z_2 + 0,0843Z_3 - 0,0005Z_4 + 0,2035Z_5 \\ &= 0,00000000003 - 0,182Z_1 - 0,1731Z_2 - 0,0362Z_3 - 0,1772Z_4 + 0,2436Z_5 \end{aligned}$$

2. Transformasi dilakukan dengan rumus:

$$x_i = Z_i\sigma + \bar{x}$$

sehingga diperoleh:

$$Y = 0,00000000003 + 7294,98X_1 + 5851,79X_2 + 719,48X_3 + 1454,84X_4 + 8,75X_5$$

Nilai R -square pada model tersebut sebesar 0,4206 artinya bahwa penelitian ini mampu menjelaskan 42,06% varians Y ditentukan oleh banyaknya komponen utama *robust* yang terbentuk, sedangkan 57,94% sisanya dipengaruhi oleh varians diluar model dan *Residual Standard Error* (RSE) untuk model regresi ROBPCA sebesar 0,8275.

Selanjutnya mengatasi masalah multikolinearitas dengan ROBPCA, yaitu sebagai berikut:

Tabel 8. Variance Inflation Factors (VIF) dengan ROBPCA

Variabel	VIF
RK_1	1,0255
RK_2	1,0255

Pada Tabel 8 menunjukkan bahwa masalah multikolinearitas sudah teratasi, karena nilai VIF untuk kedua komponen utama yang *robust* tidak lebih besar dari 10.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan sebelumnya, analisis Angka Kematian Bayi (AKB) dengan ROBPCA dapat diambil kesimpulan yaitu:

1. Dari hasil *Robust Principal Component Analysis* (ROBPCA), bahwa ROBPCA mampu menghasilkan jumlah komponen utama yang lebih sedikit daripada variabel asalnya. Data AKB yang memiliki 5 variabel bebas telah direduksi menjadi 2 variabel komponen utama dimana nilai *R-square* sebesar 0,4206 artinya bahwa penelitian ini mampu menjelaskan 42,06% varians Y ditentukan oleh banyaknya RK_1 dan RK_2 sedangkan 57,94% sisanya dipengaruhi oleh varians diluar model dan *Residual Standard Error* (RSE) untuk model regresi ROBPCA sebesar 0,8275.
2. Penerapan regresi ROBPCA dengan 1 variabel terikat dan 5 variabel bebas pada data AKB di Kalimantan Barat pada tahun 2018 diperoleh model yaitu:

$$Y = 0,00000000003 + 7294,98X_1 + 5851,79X_2 + 719,48X_3 + 1454,84X_4 + 8,75X_5$$

Berdasarkan model, faktor yang berpengaruh positif terhadap AKB yaitu jumlah ibu hamil (X_1), jumlah persalinan ditolong tenaga kesehatan (X_2) jumlah tenaga medis (X_3), jumlah bumil mengalami komplikasi kebidanan (X_4) dan presentase penduduk miskin (X_5).

DAFTAR PUSTAKA

- [1] Johnson, R.A. and Wichern, D.W., (2007), *Applied Multivariate Statistical Analysis*, Ed ke-6, Pearson Prentice Hall, United States of America.
- [2] Hubert, M., Engelen, S., Branden, K.V., (2005), *A Comparison of Three Procedures for Robust PCA in High Dimension*, *Journal of Statistics*, No.2:117-126.
- [3] Hubert, M., Rousseeuw, P.J., and Branden, K.V., (2004), *ROBPCA: A New Approach to Robust Principal Component Analysis*, *Technometrics*, No.47:64-79.
- [4] Friedman, J.K. and Tukey, J.W., (1997), *A Projection Pursuit Algorithm for Exploratory Data Analysis*, *IEEE Transactionson Computers*,82, No.23:881-89.
- [5] Rencher A.C., (2002) *Methods of Multivariate Analysis*. Ed ke-2. Canada: John Wiley Sons.
- [6] Rousseeuw P.J., Driessen K.V.A., (1998) *A Fast Alogarithm for The Minimum Covariance Determinant Estimator*. *Technometrics*, No.41:212-223.

Lina Astuti : Jurusan Matematika FMIPA UNTAN, Pontianak
linaastuti@student.untan.ac.id

Yundari : Jurusan Matematika FMIPA UNTAN, Pontianak
yundari@math.untan.ac.id
