PENENTUAN INDIKATOR KEMISKINAN PENDUDUK INDONESIA TAHUN 2017 DENGAN PEMODELAN *SPARSE PRINCIPAL COMPONENT ANALYSIS*

Georgina Maria Tinungki
Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin Makassar

email korespondensi: ina_matematika@yahoo.co.id

ABSTRAK

Data terbaru dari Badan Pusat Statistika (BPS) bulan Maret 2018, menyatakan bahwa jumlah penduduk miskin (penduduk dengan pengeluaran per kapita per bulan di bawah Garis Kemiskinan) di Indonesia mencapai 25,95 juta orang (9,82%), berkurang sebesar 633,2 ribu orang dibandingkan dengan kondisi September 2017 yang sebesar 26,58 juta orang (10,12%). Data Indikator Kemiskinan Penduduk Indonesia ini terdiri atas 13 variabel dan 34 observasi. Dalam hal ini, kita berhadapan dengan data yang berukuran besar yaitu, data yang terdiri dari banyak variabel yang diharapkan hasil analisis terhadap data tersebut menghasilkan banyak informasi yang diperlukan dalam hal pengambilan keputusan. Akan tetapi, terkadang peneliti mengalami kesulitan dalam melakukan interpretasi informasi dari hasil Principal Component Analysis (PCA) yang diperoleh. Di sini Sparse Principal Component Analysis menggabungkan kekuatan PCA standar/klasik, reduksi data, dengan pemodelan sparseness, sehingga hasil reduksi dimensi data akan mudah diinterpretasi. Data yang diperoleh adalah data sekunder yaitu data Indikator Kemiskinan Penduduk Indonesia Tahun 2017 yang diperoleh dari Badan Pusat Statistika pada situs http://www.bps.go.id. Hasil yang diperoleh menunjukkan penerapan metode Sparse PCA pada data Indikator Kemiskinan Penduduk Indonesia Tahun 2017 dalam mereduksi 13 (tiga belas) variabel, menghasilkan empat (4) variabel baru Principle Component (PC) yang telah mampu menjelaskan 81% dari total variansi data

Kata kunci : indikator kemiskinan, pemodelan *sparseness*, *principal component analysis*

PENDAHULUAN

Berdasarkan Ensiklopedia Internasional definisi dari kemiskinan: "*Proverty is scarcity, death, or state of one who lacks a certain amount of material possessions or money*". Dengan kata lain kemiskinan adalah tidak memiliki apa-apa, atau orang yang tidak memiliki harta benda atau uang. BPS mengartikan kemiskinan sebagai ketidakmampuan untuk memenuhi standar minimum kebutuhan dasar yang meliputi kebutuhan makanan maupun non-makanan (*Alexandre, 2007*).

Menghadapi masalah indikator kemiskinan, kita dihadapkan dengan data yang berukuran besar yaitu, data yang terdiri dari banyak variabel yang menghasilkan banyak informasi yang diperlukan dalam hal pengambilan keputusan. Namun seringkali peneliti mengalami kesulitan dalam melakukan interpretasi informasi yang diperoleh. Sebagai solusinya suatu metode yang dapat menjembatani masalah tersebut adalah model *Sparse Principal Component Analysis* yang menggabungkan kekuatan PCA standar/klasik, reduksi data, dengan pemodelan *sparseness*, sehingga hasil reduksi dimensi data akan mudah diinterpretasi (*Andreas, 2013).*

Sparse PCA diperkenalkan oleh Zou *et al*. pada tahun 2004. *Sparse* PCA menggabungkan kekuatan PCA klasik, reduksi data, dan pemodelan *sparseness,* yang mengeluarkan variabel yang tidak efektif dari model PCA dengan mengecilkan beban variabel tersebut menjadi nol (Hsu *et al*., 2014). Oleh karena itu, *Sparse* PCA memiliki kelebihan dalam membuat interpretasi PC menjadi lebih mudah. Selanjutnya Li *et al.*

pada tahun 2018 juga menerapkan metode *Sparse* PCA pada data berdimensi tinggi. Li *et al* (2018) membandingkan metode PCA dan *Sparse* PCA, hasil dari penelitian ini menunjukkan metode *Sparse* PCA dapat secara efektif memasukkan informasi-informasi penting ke PC yang terbentuk serta komponen beban (nilai *loading)* yang dihasilkan lebih mudah diinterpretasikan (Li *et al*. 2018).

METODE PENELITIAN

1. Indikator Kemiskinan

Indikator utama kemiskinan menurut BAPPENAS dapat dilihat dari; (1) kurangnya pangan, sandang, dan perumahan yang tidak layak; (2) terbatasnya kepemilikan tanah dan alat-alat produktif; (3) kurangnya kemampuan membaca dan menulis; (4) kurangnya jaminan dan kesejahteraan hidup; (5) kerentanan dan keterpurukan dalam bidang sosial dan ekonomi; (6) ketakberdayaan atau daya tawar yang rendah; dan (7) akses terhadap ilmu pengetahuan yang terbatas.

Menurut Bank Dunia indikator kemiskinan yaitu:

a. Kepemilikan tanah dan modal yang terbatas
b. Terbatasnya sarana dan prasarana yang dibutuhkan, pembangunan yang bias kota
c. Perbedaan kesempatan di antara anggota masyarakat
d. Perbedaan sumber daya manusia dan sektor ekonomi
e. Rendahnya produktivitas
f. Budaya hidup yang jelek
g. Tata pemerintahan yang buruk
h. Pengelolaan sumber daya alam yang berlebihan

Dari sisi makanan, BPS menggunakan indikator yang direkomendasikan oleh Widyakara Pangan dan Gizi tahun 1998 yaitu kebutuhan gizi 2.100 kalori per orang per hari, sedangkan dari sisi kebutuhan non-makanan tidak hanya terbatas pada sandang dan papan melainkan termasuk pendidikan dan kesehatan. Model ini pada intinya membandingkan tingkat konsumsi penduduk dengan suatu garis kemiskinan (GK), yaitu jumlah rupiah untuk konsumsi per orang per bulan. Sedangkan data yang digunakan adalah data makro hasil Survei Sosial dan Ekonomi Nasional (BPS, 2017).

Dalam kehidupan masyarakat yang tergolong penduduk miskin berdasarkan kemampuannya memenuhi kebutuhan hidupnya, menurut Badan Pusat Statistik adalah:

- Penduduk dikatakan sangat miskin apabila kemampuan memenuhi konsumsi makanan hanya mencapai 900/kalori/orang/hari ditambah kebutuhan dasar atau setara dengan Rp. 120.000/orang/bulan.

- Penduduk dikatakan miskin apabila kemampuan memenuhi konsumsi makanan hanya mencapai antara 1900/2100 kalori/orang/hari ditambah kebutuhan dasar atau setara dengan Rp. 120.000-Rp. 150.000/orang/bulan.
- Penduduk dikatakan mendekati miskin apabila kemampuan memenuhi konsumsi makanan hanya mencapai 2100/23000 kalori/orang/hari dan kebutuhan dasar atau setara dengan Rp. 150.000-Rp. 175.000/orang/bulan.

2. Sparse Principal Component Analysis

Salah satu bentuk pengembangan terbaru dari PCA adalah *Sparse* PCA. *Sparse* PCA menggabungkan kelebihan PCA klasik, reduksi data, dengan pemodelan *sparseness*, yang mengeluarkan variabel yang tidak efektif dari model PCA dengan mengecilkan nilai *loading* dari variabel-variabel ini menjadi nol (Hsu *et al*., 2014). Oleh karena itu, *Sparse* PCA memiliki kelebihan dalam membuat interpretasi komponen utama menjadi lebih mudah. Cara sederhana untuk melakukannya adalah dengan mengatur semua pembebanan dengan nilai absolut lebih kecil dari ambang batas tertentu menjadi nol. Metode ini disebut *thresholding* sederhana (Johnstone dan Lu, 2004).

Zou *et al*. (2004) memperkenalkan *Sparse* PCA menggunakan metode *Elastic Net* yang dikembangkan dari metode *Least Absolute Shrinkage and Selection Operator* (LASSO) untuk menghasilkan PC yang dimodifikasi dari nilai-nilai *loading* yang *sparse*. Zou *et al*., (2004) mengemukakan PCA dapat diformulasikan sebagai masalah optimasi pada regresi, sehingga nilai-nilai *loading* dapat diperoleh dengan menerapkan batasan *Elastic Net* pada koefisien regresi β.

Elastic Net merupakan suatu metode seleksi dengan menggabungkan regresi ridge dan LASSO. Dengan kata lain, *Elastic Net* menggabungkan batasan L_1*-norm* dan L_2*-norm* kuadrat pada β. Penggabungan dua batasan tersebut diharapkan dapat menyeimbangkan kelemahan dari masing-masing metode (*ridge* dan LASSO) dengan batasan L_1*-norm* menghasilkan model yang lebih sederhana karena terjadi penyusutan beberapa β yang tepat nol, sedangkan batasan L_2*-norm* kuadrat menghasilkan model yang tidak menyeleksi variabel namun meningkatkan efek pengelompokan dan penyusutan β (Ramadhini, 2014).

Batasan L_1*-norm* dan L_2*-norm* kuadrat menyusutkan penduga β ke arah nol. Meskipun kedua metode adalah metode penyusutan, namun batasan L_1*-norm* dan L_2*-norm* kuadrat memberikan pengaruh yang berbeda. Menggunakan batasan L_2*-norm* kuadrat cenderung menghasilkan β yang kecil tapi tidak nol, sedangkan batasan L_1*-norm* cenderung menghasilkan beberapa koefisien regresi tepat nol dan sebagian koefisien regresi lainnya bernilai kecil. Kombinasi batasan L_1*-norm* dan L_2*-norm* kuadrat

memberikan hasil diantara keduanya yaitu menghasilkan beberapa koefisien regresi tepat nol tetapi tidak sebanyak dengan hanya menggunakan batasan *L_1-norm* (Ramadhini, 2014). Metode *Sparse* PCA dengan menempatkan batasan *L_1-norm* dan *L_2-norm* kuadrat ini dapat menghasilkan nilai-nilai *loading* yang *sparse* dan persentase *varians* yang lebih tinggi daripada metode *Sparse* PCA yang hanya menempatkan batasan *L_1-norm.*

3. Metode Least Absolute Shrinkage and Selection Operator dan Elastic Net

Metode LASSO diperkenalkan pertama kali oleh Tibshirani pada tahun 1996. Penduga koefisien LASSO diperoleh dengan cara pemrograman kuadratik. LASSO merupakan salah satu teknik regresi pereduksian variabel bebas. LASSO menyusutkan koefisien regresi dari variabel yang memiliki korelasi tinggi dengan galat, menjadi tepat nol atau mendekati nol (Tibshirani, 1996). LASSO merupakan sebuah metode *Penalized Least Squares* (PLS) yang mengubah kendala dalam regresi *ridge* menjadi dalam bentuk *L_1-norm* yang disebut juga dengan istilah regularisasi *L_1.*

Misalkan terdapat sekumpulan data terdiri atas n pengamatan dan p variabel bebas**.** Misal $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ adalah variabel respon dan $\mathbf{X} = \left[\mathbf{x_1}, \mathbf{x_2}, \dots \mathbf{x_p}\right]$ adalah model matriks dengan $\mathbf{x_j} = \left[x_{1j}, x_{2j}, \dots x_{nj}\right]^{\mathrm{T}}$, $j = 1, \dots p$ variabel bebas. $\widehat{\beta}_{lasso}$ diperoleh dengan meminimumkan kriteria LASSO melalui persamaan (2.1) (Tibshirani, 1996).

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{lasso}} = \underset{\beta}{\arg\min} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x_j} \boldsymbol{\beta_j} \right\|^2 + \ell \sum_{\mathrm{j}=1}^{\mathrm{p}} \left| \boldsymbol{\beta_j} \right| \right\} \qquad (2.1)$$

dengan ℓ adalah parameter *tuning* yang menentukan penyusutan koefisien LASSO, $\ell \geq 0$.

Metode LASSO secara bersamaan menghasilkan model yang *sparse* dan akurat, sehingga menjadikan LASSO sebagai metode pemilihan variabel yang menguntungkan. Namun, berdasarkan penelitian yang dilakukan oleh Zou dan Hastie (2003), LASSO memiliki beberapa keterbatasan. Salah satunya adalah jumlah variabel yang dipilih oleh LASSO dibatasi oleh jumlah pengamatan. Misalnya, jika diterapkan pada data *microarray* dimana ada ribuan (gen) $(p > 1000)$ dengan kurang dari 100 sampel (n <100), LASSO hanya dapat memilih paling banyak n variabel (jenis gen).

Oleh karena itu, Zou dan Hastie (2003) melakukan pengembangan dari metode LASSO untuk mengatasi keterbatasannya yang kemudian dikenal sebagai metode

Elastic Net. Untuk setiap ℓ_1 dan ℓ_2 non-negatif, penduga *Elastic Net* $\hat{\beta}_{en}$ diberikan melalui persamaan (2.24).

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{en}} = (1 + \ell_2) \underset{\beta}{\arg\min} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x_j} \boldsymbol{\beta_j} \right\|^2 + \ell_2 \sum_{\mathrm{j}=1}^{\mathrm{p}} |\boldsymbol{\beta_j}|^2 + \ell_1 \sum_{\mathrm{j}=1}^{\mathrm{p}} |\boldsymbol{\beta_j}| \right\} \tag{2.2}$$

Ketika $p > n$ dan $\ell_2 > 0$, *Elastic Net* berpotensi mencakup semua variabel pada *fitted model*, sehingga keterbatasan LASSO dapat dihilangkan. Keuntungan lain yang ditawarkan oleh metode *Elastic Net* adalah dari segi pengelompokkan variabel. *Elastic Net* cenderung memilih sekelompok variabel yang berkolerasi tinggi. Sebaliknya, jika dalam satu kelompok variabel memiliki korelasi berpasangan yang tinggi, LASSO cenderung memilih hanya satu variabel dari kelompok dan tidak memperhatikan variabel lainnya.

4. Pendekatan *Sparse*

Perhatikan bahwa setiap PC merupakan kombinasi linier dari p variabel, sehingga pembebanannya (nilai-nilai *loading*) dapat diperoleh dengan cara meregresikan PC pada p variabel.

Teorema 1 (Zou *et al*., 2006)

Untuk setiap i, $\boldsymbol{W_i}$ adalah PC ke-*i*. Misalkan $\mathbf{X}_{n\times p}$ $(n > p)$ adalah matriks yang memiliki rank penuh (*full-rank*), ℓ suatu bilangan non-negatif dan penduga *ridge* $\hat{\beta}_{ridge}$ diberikan oleh persamaan (2.3) berikut,

$$\widehat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\arg\min} \{\|\boldsymbol{W_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \ell\|\boldsymbol{\beta}\|_2^2\} \tag{2.3}$$

$\hat{\boldsymbol{v}}$ diperoleh dengan menormalisasikan $\widehat{\boldsymbol{\beta}}_{ridge}$, $\hat{\boldsymbol{v}} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}}{\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}\|}$, dengan $\hat{\boldsymbol{v}} = \mathbf{v_j}$

Teorema 1 menunjukkan hubungan antara PCA dan metode regresi. Peregresian PC pada variabel-variabel dibahas oleh Cadima dan Jolliffe pada tahun 1995. Cadima dan Jolliffe (1995) berfokus pada pendekatan PC dengan subset dari variabel k. Zou *et al*., (2004) melakukan pendekatan KU menggunakan regresi *ridge*. Setelah normalisasi, koefisien saling bebas terhadap ℓ. Oleh karena itu penalti regresi *ridge* dalam bentuk *L_2-norm* kuadrat tidak digunakan untuk melakukan penalisasi koefisien regresi, tetapi untuk memastikan rekonstruksi PC.

Selanjutnya, tambahkan *L_1-norm* ke persamaan (2.3) sehingga diperoleh persamaan (2.4) berikut (Zou, et al., 2006).

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\arg\min} \{\|\boldsymbol{W_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \ell\|\boldsymbol{\beta}\|_2^2 + \ell_1\|\boldsymbol{\beta}\|_1\} \tag{2.4}$$

dan $\hat{\mathbf{v}}_\mathbf{i} = \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}$, merupakan penduga dari $\mathbf{v_i}$ dan $\mathbf{X}\hat{\mathbf{v}}_\mathbf{i}$ adalah penduga KU ke-i. Persamaan (2.4) disebut sebagai persamaan *naive elastic net* (Zou dan Hastie, 2003)

dimana persamaan (2.4) merupakan *elastic net* tanpa faktor penskalaan $(1+\ell)$. Faktor penskalaan tidak mempengaruhi $\hat{\mathbf{v}}_\mathbf{i}$ karena akan digunakan *fitted coefficients* yang dinormalisasi, sehingga pada persamaan (2.1), $(1+\ell)$ dapat dihilangkan dan diperoleh hasil yang sama dengan persamaan (2.2). Atau dengan kata lain, persamaan (2.4) adalah regularisasi *naive elastic net* yang merupakan kombinasi dari *L_1-norm* dan *L_2-norm* kuadrat.

Selanjutnya, Zou *et al*. (2006) menyusun suatu algoritma untuk meminimukan kriteria *Sparse* PCA berdasarkan persamaan (2.26). Misal $\widehat{\mathbf{B}} = [\widehat{\boldsymbol{\beta}}_\mathbf{1}, \widehat{\boldsymbol{\beta}}_\mathbf{2}, \dots \widehat{\boldsymbol{\beta}}_\mathbf{k}]$ merupakan matriks berukuran $p \times k$, p jumlah variabel bebas dan k jumlah KU yang terpilih, dan $\widehat{\boldsymbol{\beta}}_\mathbf{j}$ suatu vektor penduga *naive elastic net*, subtitusi $\boldsymbol{W}_\mathbf{j} = \mathbf{X}\boldsymbol{\alpha}_\mathbf{j}$ untuk setiap $j = 1, \dots k$, ke persamaan (2.4) sehingga diperoleh persamaan (2.5) (Zou *et al*., 2006).

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ \|\boldsymbol{W}_\mathbf{j} - \mathbf{X}\boldsymbol{\beta}_\mathbf{j}\|^2 + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\}$$

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ \|\mathbf{X}\boldsymbol{\alpha}_\mathbf{j} - \mathbf{X}\boldsymbol{\beta}_\mathbf{j}\|^2 + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\}$$

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ \|\mathbf{X}(\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j})\|^2 + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\}$$

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ (\|\mathbf{X}(\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j})\|)(\|\mathbf{X}(\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j})\|) + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\}$$

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ (\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j})^\mathrm{T} (\|\mathbf{X}\|\|\mathbf{X}\|)(\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j}) + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\}$$

dengan mensubstitusikan $\|\mathbf{X}\| = \sqrt{\mathbf{X}^\mathbf{T}\mathbf{X}}$, maka diperoleh persamaan (2.5).

$$\widehat{\boldsymbol{\beta}}_\mathbf{j} = \underset{\beta_j}{\arg\min} \left\{ (\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j})^\mathrm{T} \mathbf{X}^\mathbf{T}\mathbf{X} (\boldsymbol{\alpha}_\mathbf{j} - \boldsymbol{\beta}_\mathbf{j}) + \ell \|\boldsymbol{\beta}_\mathbf{j}\|_2^2 + \ell_{1,j} \|\boldsymbol{\beta}_\mathbf{j}\|_1 \right\} \quad (2.5)$$

dengan $\|\boldsymbol{\beta}_\mathbf{j}\|_2^2 = \sum_{i=1}^{p} \beta_{ij}^2$ dan $\|\boldsymbol{\beta}_\mathbf{j}\|_1 = \sum_{i=1}^{p} |\beta_{ij}|$.

Catatan:

ℓ digunakan untuk semua k PC, sedangkan nilai $\ell_{1,j}$ dapat berbeda-beda untuk setiap $j = 1, \dots k$ PC (Zou et al., 2006).

5. Memilih Parameter Tuning

Pemilihan $\ell_{1,j}$ pada persamaan (2.3) dapat dilakukan dengan menggunakan *Cross Validation* (CV). CV yang sebaiknya digunakan adalah 5-*fold* atau 10-*fold* karena memberikan dugaan sisaan prediksi yang mempunyai bias tinggi namun memberikan MSE kecil dan juga variansi yang lebih kecil (Ramadhini, 2014).

Zou dan Hastie (2005) memilih parameter *tuning* $\ell_{1,j}$ menggunakan 10-*fold* CV. Adapun untuk pemilihan ℓ, ditentukan oleh peneliti (dapat mempertimbangkan prinsip *parsimony*). *ℓ yang dipilih adalah ℓ yang menghasilkan model matriks loading yang sparse dan proporsi keragaman kumulatif yang dihasilkan ≥ 80%.*

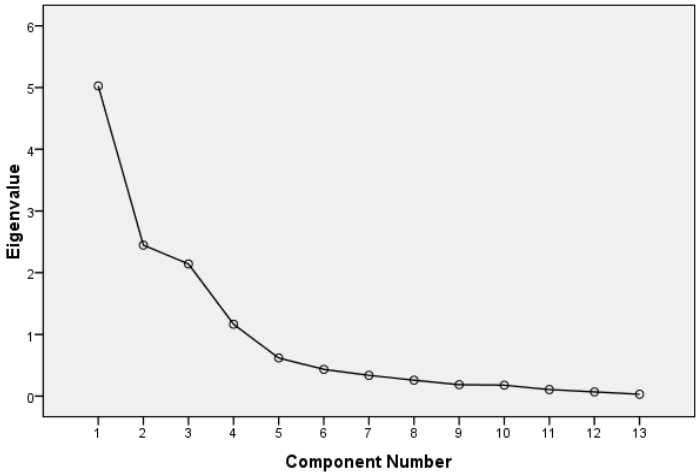
HASIL DAN PEMBAHASAN

1. Penentuan Jumlah Komponen Utama yang Terbentuk menggunakan *Principal Component Analysis*

Untuk menentukan jumlah PC yang terbentuk, terdapat 3 kriteria yang bisa digunakan yaitu dengan melihat *scree plot*, nilai *eigen*, dan keragaman kumulatif yang dapat dijelaskan oleh PC.

a. Scree Plot

Scree plot adalah *plot* antara nilai *eigen* dengan banyaknya KU yang terbentuk. Jumlah KU yang terbentuk dapat diketahui dengan memperhatikan patahan siku dari *scree plot. Scree plot* data dengan metode PCA dapat dilihat pada Gambar 4.1.



Gambar 1
Scree Plot data dengan metode PCA

Pada Gambar 1, terlihat bahwa kurva mulai meluruh setelah PC ke-4. Sehingga dapat disimpulkan bahwa dengan mengambil empat KU, sudah mencukupi untuk mewakili ketiga belas variabel lama.

b. Nilai eigen

PC yang terbentuk dapat ditentukan dengan memilih PC yang memiliki nilai *eigen* lebih besar dari 1. Nilai *eigen* dengan metode PCA disajikan dalam Tabel 1.

Tabel 1
Nilai *Eigen* dengan Metode PCA

Nilai eigen dengan metode PCA			
PC ke-	Nilai Eigen	PC ke-	Nilai Eigen
1	**5.025562**	8	0.259007
2	**2.44679**	9	0.18576
3	**2.141998**	10	0.177708
4	**1.164892**	11	0.106607
5	0.619113	12	0.068969
6	0.434757	13	0.031239
7	0.337597		

Sumber: Data Diolah, 2018

Tabel 1 menunjukkan bahwa terdapat 4 PC yang memiliki nilai *eigen* lebih besar dari 1. PC_1 memiliki nilai *eigen* sebesar 5.025562. Selanjutnya, PC_2 memiliki nilai *eigen* sebesar 2.44679, PC_3 memiliki nilai *eigen* sebesar 2.141998 dan PC_4 memiliki nilai *eigen* sebesar 1.164892. Adapun KU_5 hingga PC_{13} memiliki nilai *eigen* kurang dari 1. Sehingga dapat disimpulkan bahwa jumlah PC yang terbentuk menurut kriteria nilai *eigen* adalah 4 PC.

Persentase Proporsi Keragaman Kumulatif

KU yang terbentuk dapat ditentukan dengan memilih PC yang dapat menjelaskan proporsi keragaman sampel secara kumulatif minimal 80%. Tabel 2 menunjukkan bahwa PC pertama menjelaskan 38,66% dari total variansi sampel. PC kedua menjelaskan 18,82% dari total variansi sampel. Adapun PC ketiga dan keempat masing-masing mampu menjelaskan 16,48% dan 8,96% dari total variansi sampel. Secara kumulatif, keempat PC pertama ini telah mampu menjelaskan 82.92% dari total variansi sampel. Sehingga, berdasarkan kriteria persentase proporsi keragaman kumulatif, dibutuhkan 4 PC untuk mereduksi 13 variabel tanpa kehilangan informasi penting yang terkandung di dalamnya.

Tabel 2
Proporsi Keragaman dan Keragaman Kumulatif dengan Metode PCA

PC ke-	Proporsi keragaman	Proporsi keragaman kumulatif
1	0.3866	**0.3866**
2	0.1882	**0.5748**
3	0.1648	**0.7396**
4	0.0896	**0.8292**
5	0.0476	0.8768
6	0.0334	0.9102
7	0.0260	0.9362
8	0.0199	0.9561
9	0.0143	0.9704
10	0.0137	0.9841
11	0.0080	0.9923
12	0.0053	0.9976
13	0.0024	1.0000

Sumber: Data Diolah, 2018

Jumlah PC yang dipilih adalah sebanyak empat (4) berdasarkan ketiga kriteria pemilihan jumlah KU yaitu *scree plot*, nilai *eigen* dan persentase keragaman kumulatif.

1. Matriks Loading menggunakan Metode Principal Component Analysis

Berdasarkan ketiga kriteria pemilihan jumlah PC yaitu *scree plot*, nilai eigen dan persentase keragaman kumulatif, diperoleh empat PC sebagai variabel baru. Keempat variabel ini memiliki nilai eigen yang telah terurut. Untuk menentukan variabel lama yang

paling banyak menjelaskan masing-masing keempat variabel baru ini, kita dapat melihat nilai *loading* tiap komponen. Nilai *Loading* tiap KU menggunakan metode PCA disajikan pada Tabel 3. Nilai *Loading* pada Tabel 3 menjelaskan hubungan (korelasi) antara variabel lama dengan variabel baru yang dibentuk dengan metode PCA.

Tabel 3
Matriks *Loading* Menggunakan Metode PCA

Variabel	Komponen Utama			
	PC_1	PC_2	PC_3	PC_4
X_1	-0.251	0.013	-0.437	0.286
X_2	0.188	0.322	-0.350	-0.427
X_3	0.409	0.016	-0.127	-0.037
X_4	-0.304	0.048	-0.300	0.205
X_5	0.373	-0.225	0.119	-0.149
X_6	-0.243	0.409	-0.217	-0.288
X_7	0.350	-0.079	-0.120	0.232
X_8	0.107	-0.368	-0.335	-0.387
X_9	0.231	-0.300	-0.308	0.010
X_{10}	0.323	0.082	-0.014	0.505
X_{11}	0.143	0.299	0.462	-0.234
X_{12}	0.200	0.472	-0.021	0.270
X_{13}	0.303	0.353	-0.289	-0.034

Sumber: Data diolah, 2018

Berdasarkan Tabel 4.3, nilai *loading* beberapa variabel tidak memiliki perbedaan yang signifikan sehingga sulit untuk menentukan variabel mana yang membentuk PC. Oleh karena itu, tahapan analisis yang digunakan untuk mengatasi masalah tersebut adalah dengan menerapkan metode *sparse* PCA yang memodifikasi matriks *loading* dari Tabel 3.

2. Pemilihan Parameter Tuning ℓ dan $\ell_{1,j}$

Pemilihan model terbaik penduga *Elastic Net* pada metode *Sparse* PCA dapat menggunakan *Cross Validation* (CV) dengan memilih $\ell_{1,j}$ yang memiliki nilai CV terkecil. Pada penelitian ini, diuji beberapa nilai ℓ *untuk mendapatkan model yang sederhana namun tetap* mampu menjelaskan keragaman data asli minimal 80%. Hasil dari pengujian nilai ℓ ini dapat dilihat pada Tabel 4.

Tabel 4
Jumlah Nilai *Loading* Tidak-Nol dan Proporsi Keragaman Kumulatif PC berdasarkan Nilai Parameter Tuning ℓ dan $\ell_{1,j}$

No	Nilai ℓ	$\ell_{1,1}$	$\ell_{1,2}$	$\ell_{1,3}$	$\ell_{1,4}$	Jumlah Nilai Loading $\neq 0$				Proporsi Kumulatif Keragaman
						PC_1	PC_2	PC_3	PC_4	
1	0	1	1	1	1	11	7	8	9	76.4
2	1	0.6464646	0.6969697	0.6767677	1	12	11	8	8	78.8
3	2	0.4848485	0.6565657	0.7676768	1	13	11	8	8	79.2
4	3	0.4040404	0.5252525	0.8282828	1	13	12	8	8	80
5	4	0.363636	0.525253	0.8484848	1	13	12	8	8	80.1
6	5	0.3333333	0.4444444	0.7171717	1	13	12	11	7	80.3
7	10	0.2525253	0.3737374	0.6868687	1	13	13	12	7	80.8
8	50	0.2020202	0.1919192	0.5353535	0	13	13	12	13	83

Sumber: Data diolah, 2018

Pada Tabel 4, model yang paling sederhana akan diperoleh jika nilai $\ell = 0$*. Hal ini ditandai dengan jumlah nilai loading* yang tidak bernilai nol tiap PC nya paling sedikit ditemukan pada model ini. Akan tetapi, proporsi keragaman data yang dapat dijelaskan keempat KU secara kumulatif belum mampu memenuhi kriteria minimal yang diinginkan. Pada nilai $\ell = 3$*, model telah mampu memenuhi* proporsi keragaman data yang dapat dijelaskan keempat PC secara kumulatif yaitu 80%. Begitupula jika nilai $\ell = 4$, $\ell = 5$, $\ell = 10$*, dan* $\ell = 50$ *juga telah memenuhi kriteria* minimal proporsi keragaman *data yang dapat dijelaskan* keempat PC secara kumulatif *yaitu 80.1%, 80.3%, 80.8%, dan 83%. Berdasarkan kelima nilai* ℓ *ini, model yang paling sederhana dihasilkan jika* $\ell = 3$ *dan* $\ell = 4$. Akan tetapi, proporsi keragaman data yang dapat dijelaskan keempat KU secara kumulatif oleh model jika nilai $\ell = 4$ *lebih besar daripada* $\ell = 3$. Sehingga, pada penelitian ini nilai *yang* ℓ *terpilih adalah* $\ell = 4$. Sedangkan untuk ℓ_1 pada model PC_1 yang terpilih 0.363636, ℓ_1 pada model PC_2 yang terpilih adalah 0.525253, ℓ_1 pada model PC_3 yang terpilih adalah 0.8485, dan ℓ_1 pada model PC_4 yang terpilih adalah 1.

3. Matriks Loading menggunakan Algoritma Sparse Principal Component Analysis

Sparse PCA menggabungkan kekuatan PCA klasik, reduksi data, dengan pemodelan yang *sparseness*. *Sparse* PCA mengeluarkan variabel yang tidak efektif dari model PCA dengan mengecilkan nilai *loading* menjadi nol. Tabel 5 merupakan hasil dari matriks *loading* menggunakan algoritma *Sparse* PCA.

Tabel 5
Matriks *Loading* menggunakan Metode *Sparse* PCA

Variabel	*Principle Component*			
	PC_1	PC_2	PC_3	PC_4
X_1	0.111	-0.103	0.555	0
X_2	-0.238	-0.470	0	0.376
X_3	-0.423	-0.067	0	0
X_4	0.215	-0.091	0.424	0
X_5	-0.372	0.195	-0.207	0
X_6	0.220	-0.468	0	0.278
X_7	-0.362	0	0.081	-0.211
X_8	-0.302	0.184	0.153	0.525
X_9	-0.370	0.147	0.256	0.080
X_{10}	-0.249	-0.055	0.080	-0.555
X_{11}	0.034	-0.134	-0.606	-0.009
X_{12}	-0.063	-0.454	0	-0.383
X_{13}	-0.304	-0.464	0	0

Sumber: Data diolah, 2018

Berdasarkankan Tabel 5, terdapat 11 nilai *loading* yang bernilai nol, sehingga terlihat bahwa matriks *loading* yang dihasilkan oleh metode *Sparse* PCA menjadi lebih sederhana. Peneliti lebih memilih model yang lebih sederhana karena dapat memberikan kemudahan dalam melihat hubungan antara PC yang terbentuk dengan variabel lama (variabel bebas). Hal ini menyebabkan variabel-variabel bebas yang tidak

berpengaruh dapat dikeluarkan. Prinsip *parsimony* adalah suatu hal yang penting ketika jumlah variabel bebas besar.

Berdasarkan Tabel 6, diperoleh pengelompokan variabel-variabel asli terhadap PC yang terbentuk. Pengelompokan ini didasarkan pada variabel-variabel lama yang paling banyak menjelaskan tiap KU atau variabel-variabel lama yang memiliki korelasi (nilai *loading*) terbesar pada tiap-tiap KU.

Tabel 4.6
Pengelompokan Variabel Lama, Nilai *Loading*, dan Variansi yang dijelaskan Tiap PC Menggunakan *Sparse* PCA

rinciple Component	Variabel Lama	Nilai Loading	Variansi yang dijelaskan
PC_1	IPM (X_3)	-0.423	35.7%
	Listrik (X_5)	-0.373	
	AH (X_7)	-0.362	
	Imunisasi (X_9)	-0.370	
PC_2	Kepadatan (X_2)	-0.470	18.5%
	Luas Hunian (X_6)	-0.468	
	Pengeluaran Makanan (X_{12})	-0.454	
	Pengeluaran Non Makanan (X_{13})	-0.464	
PC_3	TPAK (X_1)	0.555	16.0%
	Buta Huruf (X_4)	0.424	
	TPT (X_{11})	-0.606	
KU_4	Keluhan Kesehatan (X_8)	0.525	9.9%
	TBAB Sendiri (X_{10})	-0.555	

Sumber: Data diolah, 2018

Pengelompokkan variabel pada Tabel 6 harus memenuhi kriteria minimal nilai *loading* yang digunakan yaitu lebih besar dari 0.3 (tanda positif atau negatif dari nilai *loading* hanya menunjukkan hubungan antara PC dengan variabel asli). Akan tetapi, untuk X_8 dan X_{13} walaupun mempunyai nilai *loading* lebih besar dari 0.3 (lihat Tabel 5) terhadap PC_1, pada Tabel 6 kedua variabel ini tidak dikelompokkan ke dalam KU_1 karena dengan pertimbangan masing-masing variabel X_8 dan X_{13} lebih banyak menjelaskan PC_4 dan PC_2 daripada PC_1. Proporsi keragaman dari keempat PC secara kumulatif menggunakan metode *Sparse* PCA yaitu 80.1%. Hasil ini telah memenuhi standar kriteria persentase proporsi keragaman yang dapat dijelaskan oleh komponen-komponen utama secara kumulatif yakni minimal 80%.

4. Interpretasi Principle Component

Berdasarkan Tabel 5 akan dibentuk 4 KU sebagai kombinasi linier variabel-variabel asalnya yaitu sebagai berikut.

$$K_i = \boldsymbol{v}'_i\boldsymbol{X} = v_{1i}X_1 + v_{2i}X_2 + \cdots + v_{pi}X_p \qquad i = 1,2,3$$

$$KU_{\square} = \boldsymbol{v}'_1\boldsymbol{Z} = 0.111Z_1 - 0.238Z_2 - 0.423Z_3 + 0.215Z_4 - 0.372\,Z_5 + 0.220Z_6 - 0.362Z_7 - 0.302Z_8 - 0.370Z_9 - 0.249Z_{10} + 0.034Z_{11} - 0.063Z_{12} - 0.304Z_{13}$$

Nilai dari $\mathrm{KU}_{\mathbf{1}}$ lebih banyak dijelaskan oleh variabel IPM, listrik, angka harapan hidup (AH), dan imunisasi. Hal ini dapat dilihat dari koefisien yang cukup besar dibanding variabel lainnya dan telah memenuhi kriteria nilai minimal *loading.* Koefisien pada keempat variabel tersebut bertanda negatif maka korelasi antara $\mathrm{KU}_{\mathbf{1}}$ dengan variabel IPM, listrik, angka harapan hidup (AH), dan imunisasi adalah negatif. Apabila $\mathrm{KU}_{\mathbf{1}}$ bernilai kecil maka sebaliknya, variabel IPM, listrik, angka harapan hidup (AH), dan imunisasi bernilai besar dan sebaliknya. Sehingga, kesimpulan mengenai variabel IPM, listrik, angka harapan hidup (AH), dan imunisasi juga dapat diambil dengan melihat nilai dari $\mathrm{KU}_{\mathbf{1}}$.

$$\mathrm{PC}_{\square} = \boldsymbol{v'_2 Z} = -0.103Z_1 - 0.470Z_2 - 0.067Z_3 - 0.091Z_4 + 0.195\,Z_5 - 0.468Z_6 + 0.184Z_8 + 0.147Z_9 - 0.055Z_{10} - 0.134Z_{11} - 0.454Z_{12} + 0.464Z_{13}$$

Nilai dari $\mathrm{KU}_{\mathbf{2}}$ lebih banyak dijelaskan oleh variabel kepadatan penduduk, luas hunian, pengeluaran makanan, dan pengeluaran non makanan. Hal ini dapat dilihat dari koefisien ketiga variabel tersebut yang cukup besar dibanding variabel lainnya dan telah memenuhi kriteria nilai minimal *loading.* Koefisien kepadatan penduduk, luas hunian, dan pengeluaran makanan bertanda negatif maka korelasi antara $\mathrm{PC}_{\mathbf{2}}$ dengan variabel kepadatan penduduk, luas hunian, dan pengeluaran makanan adalah negatif. Apabila $\mathrm{PC}_{\mathbf{2}}$ bernilai kecil maka variabel kepadatan penduduk, luas hunian, dan pengeluaran makanan bernilai besar dan sebaliknya. Adapun koefisien pengeluaran non makananan bertanda positif, artinya korelasi antara $\mathrm{PC}_{\mathbf{2}}$ dengan variabel pengeluaran non makananan adalah positif. Apabila $\mathrm{PC}_{\mathbf{2}}$bernilai kecil maka variabel pengeluaran non makananan juga bernilai kecil dan sebaliknya. Sehingga, kesimpulan mengenai variabel kepadatan penduduk, luas hunian, pengeluaran makanan, dan pengeluaran non makanan juga dapat diambil dengan melihat nilai dari $\mathrm{KU}_{\mathbf{2}}$.

$$\mathrm{KU}_3 = \boldsymbol{v'_3 Z} = 0.555Z_1 + 0.424Z_4 - 0.207Z_5 + 0.081Z_7 + 0.153Z_8 + 0.256Z_9 + 0.080Z_{10} - 0.606Z_{11}$$

Nilai dari $\mathrm{PC}_{\mathbf{3}}$ dijelaskan oleh variabel TPAK, buta huruf, dan TPT. Koefisien variabel TPAK dan buta huruf bertanda positif maka korelasi antara $\mathrm{KU}_{\mathbf{3}}$ dengan kedua variabel tersebut adalah positif. Apabila $\mathrm{KU}_{\mathbf{3}}$ bernilai kecil maka variabel TPAK dan buta huruf juga bernilai kecil dan sebaliknya. Koefisien variabel TPT bertanda negatif maka korelasi antara $\mathrm{KU}_{\mathbf{3}}$dengan variabel TPT, adalah negatif. Apabila $\mathrm{PC}_{\mathbf{3}}$bernilai kecil maka variabel TPT bernilai besar dan sebaliknya. Sehingga, kesimpulan mengenai variabel TPAK, buta huruf, dan TPT juga dapat diambil dengan melihat nilai dari $\mathrm{PC}_{\mathbf{3}}$.

$$\mathrm{KU}_{\square} = \boldsymbol{v'_4 Z} = 0.376Z_2 + 0.278Z_6 - 0.211Z_7 + 0.525Z_8 + 0.080Z_9 - 0.555Z_{10} - 0.009Z_{11} - 0.383Z_{12}$$

Nilai dari PC_4 paling banyak dijelaskan oleh variabel keluhan kesehatan dan jumlah penduduk yang memiliki tempat BAB sendiri. Hal ini dapat dilihat dari koefisien variabel keluhan kesehatan dan TBAB sendiri yang lebih besar dari koefisien variabel lainnya. Koefisien keluhan kesehatan bertanda positif maka korelasi antara PC_4 dengan variabel keluhan kesehatan adalah positif. Apabila PC_4 bernilai kecil maka variabel keluhan kesehatan juga bernilai kecil dan sebaliknya. Koefisien TBAB bertanda negatif maka korelasi antara PC_4 dengan variabel TBAB sendiri adalah negatif. Apabila PC_4 bernilai kecil maka variabel TBAB Sendiri bernilai besar dan sebaliknya. Sehingga, kesimpulan mengenai variabel keluhan kesehatan dan jumlah penduduk yang memiliki tempat BAB sendiri juga dapat diambil dengan melihat nilai dari KU_4.

Pada Tabel 6 diketahui bahwa PC pertama (PC_1) memiliki nilai persentase varian sebesar 35.7%. Berdasarkan kriteria nilai *loading*nya, variabel yang membentuk PC_1 yaitu variabel IPM, listrik, angka harapan hidup, dan imunisasi PC_1 memiliki variansi paling besar diantara PC yang terbentuk lainnya, sehingga dapat dikatakan bahwa jumlahan dari variabel IPM, listrik, angka harapan hidup, dan imunisasi. PC_1 menghasilkan variansi yang besar, PC kedua (PC_2) dapat menjelaskan 18.5% dari total varians. Berdasarkan nilai *loading*nya, variabel yang membentuk PC_2 adalah kepadatan, luas hunian, pengeluran makanan, dan pengleuaran non makanan. Selanjutnya PC ketiga (PC_3) mampu menjelaskan variansi sebesar 16,0% dan variabel yang membentuk PC_3 yaitu TPAK, buta huruf, dan TPT. Selanjutnya PC keempat (KU_4) mampu menjelaskan variansi sebesar 11,2% dan variabel yang membentuk PC_4 yaitu keluhan kesehatan dan TBAB sendiri.

KESIMPULAN

Berdasarkan pembahasan, diperoleh kesimpulan bahwa metode *Sparse* PCA dapat diterapkan untuk mereduksi dimensi data Indikator Kemiskinan Penduduk Indonesia Tahun 2017. Penerapan metode *Sparse* PCA pada data Indikator Kemiskinan Penduduk Indonesia Tahun 2017 dalam mereduksi 13 (tiga belas) variabel, menghasilkan empat (4) variabel baru (PC) yang telah mampu menjelaskan 81% dari total variansi data. Penerapan metode *Sparse* PCA pada reduksi dimensi data juga menghasilkan matriks loading yang *sparse* sehingga memudahkan peneliti untuk melakukan interpretasi keempat (4) variabel baru (PC) yang dibentuk.

DAFTAR PUSTAKA

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, Gert R. G. Lanckriet (2007). "A Direct Formulation for Sparse PCA Using Semidefinite Programming" *(PDF). SIAMReview. 49 (3):434448.* arXiv*:*cs/0406021*.* doi*:*10.1137/05064 5506*.*

Andreas M. Tillmann, Marc E. Pfetsch (2013). "The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing". *IEEE Transactions on Information Theory. 60 (2): 1248–1259.* arXiv*:*1205.2081 *.* doi*:*10.1109/TIT.2013.2290112*.*

Badan Pusat Statistik. 2017. *Data Indikiator Kemiskinan Penduduk Indonesia Tahun 2015*.

Cadima, J. F. dan I. T. *Jolliffe*, I. T. *1995*. "Loadings and Correlations in the interpretation of *Principal Components". Journal of Applied Statistics*. 22(2): 203-214.

H. Zou; T. Hastie; R. Tibshirani (2006). "Sparse principal component analysis" *(PDF). Journal of Computational and Graphical Statistics. 15 (2): 262–286.* doi*:*10.1198/106186006x113430

Iain M Johnstone; Arthur Yu Lu (2009). "On Consistency and Sparsity for Principal Components Analysis in High Dimensions". *Journal of the American Statistical Association. 104 (486): 682–693.* doi*:*10.1198/jasa.2009.0121*.* PMC 2898454

Hsu, Y. L., P. Y. Huang, dan D. T. Chen. 2014. Sparse Principal Component Analysis In Cancer Research. *Transl Cancer Res.* 3(3): 182–190*.*

_____ dan _____. 2002. *Applied Multivariate Statistical Analysis Fifth Edition*. New Jersey: Prentice Hall.

_____ dan _____. 1998. *Applied Multivariate Statistical Analysis Second Edition*. New Jersey: Prentice Hall.

Johnstone, I. M. dan A. Y. Lu. 2004. *Sparse Komponen utama Analysis.* California: Standford University and Renaissance Technologies.

Ramadhini, Fitri. 2014. *Penyusutan Koefisien dan Seleksi Variabel Regresi dengan Elastic Net* [skripsi]. Yogyakarta: UGM.

Tantular, B. 2011. Praktikum Analisis Data Multivariat II Menggunakan Software R: Modul 1 Analisis Komponen Utama. *https://berthoveens.files.wordpress.com /2011/07/modul-multi.pdf* dan *bertho@unpad.ac.id*. 20 Desember 2018(09:17).

Zou, H., T. Hastie, dan R. Tibshirani. 2006. "Sparse Principal Component Analysis". *Journal of Computational and Graphical Statistics.* 15(2): 265–286*.*

_____, _____, dan _____. 2004. "*Sparse Principal Component Analysis"*. California: Department of Statistics, Stanford University.