Karakteristik Estimator Analisis Komponen Utama untuk Mengestimasi Model Variabel Laten Menggunakan Metode *High-Dimensional AIC*

Lukman*

Departemen Pendidikan Matematika
Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam
Universitas Pendidikan Indonesia
*Surel: lukman12@upi.edu

ABSTRAK. Makalah ini bertujuan untuk mengetahui sifat estimator Analisis Komponen Utama (AKU) untuk mengestimasi model variabel laten. Metode yang digunakan adalah metode *High-Dimensional AIC* (*HAIC*) dengan simulasi data berdistribusi Bernoulli. Tahapannya adalah: (1) menentukan matriks AKU; (2) membuat model estimator AKU untuk mengestimasi variabel laten dengan menggunakan HAIC; (3) mensimulasikan data distribusi Bernoulli dengan pengulangan 1.000.748 kali. Hasil simulasi menunjukkan model estimator AKU bekerja dengan baik.

Kata Kunci: *AKU, HAIC,* Model Variabel Laten.

The Characteristic of Principle Component Analysis Estimator to Estimate Latent Variable Model Method Using High-Dimensional AIC

ABSTRACT. *This paper aims to determine the properties of Principle Component Analysis (PCA) estimator to estimate latent variable models. The method used is the High-Dimensional AIC (HAIC) method with simulation of Bernoulli distribution data. Stages are: (1) determine the matrix PCA; (2) create a model of the PCA estimator to estimate the latent variables by using HAIC; (3) simulated the Bernoulli distribution data with repetition 1,000,748 times. The simulation results show that the PCA estimator models work well.*

Keywords *: PCA, HAIC, Latent Variable Models.*

1. PENDAHULUAN

Analisis Komponen Utama (AKU) adalah salah satu metode teknik reduksi dimensi, yaitu metode penyajian data kategorik baris dan kolom dalam tabel kontingensi dua arah. Menurut Greenacre[1], ada dua hal penting dalam teknik reduksi AKU, yaitu: pertama tentang pemilihan data matriks subruang optimal yang disebut stabilitas internal dan teknik pengambilan sampel populasi disebut stabilitas eksternal. Greenacre menggunakan metode *Jackknifing* untuk menyelidiki stabilitas internal dan metode *boot strapping* untuk menyelidiki stabilitas eksternal. Lynn dan McCulloch[2] menggunakan estimator AKU dalam menyelidiki model variabel laten.

Wegelin, Packer, dan Richardson[3] menyelidiki keberadaan variabel laten dalam matriks kovarians antara dua blok variabel menggunakan matriks kovarian silang. Ada tiga model laten yang dimunculkan, yaitu: model peringkat regresi, model korelasi-laten berpasangan, dan model diagonal-laten. Kemudian, Ogura dan Fujikoshi[4] mengusulkan metode *AIC* (*Akaike Information Criterion*) dan *HAIC* (*High-Dimensional AIC*) untuk mengestimasi dimensionalitas.

Dalam tulisan ini, akan diteliti bagaimana sifat estimator AKU dalam mengestimasi model variabel laten dengan menggunakan metode HAIC[5]. Simulasi numerik menggunakan data distribusi Bernoulli dengan pengulangan 1.000.748 kali.

2. METODOLOGI

Penelitian ini termasuk penelitian dasar atau penelitian murni karena bertujuan untuk mengembangkan teori yang sudah ada atau mencari teori baru dan atau melengkapi teori yang sudah ada di bidang statistika. Metode yang digunakan adalah metode non eksperimental, yang dibuat dengan mengkaji teori-teori yang berkaitan dengan pembobotan[6] dan aplikasinya dalam Metode AKU, dan Teori AKU

Matriks AKU

Misalkan y_{ij} adalah bilangan real positif untuk setiap i =1, 2, ...,I dan j=1, 2, ...J. Matriks AKU adalah

$$\mathrm{Y} = \left(y_{ij} - \bar{y}_{.j}\right);\ \mathrm{Y}'\mathrm{Y}\mathbf{v} = \lambda\mathbf{v} \qquad \ldots (1)$$

dimana **v** dan λ masing-masing adalah vektor singular dan nilai singular.

Elemen ke-(i, j) dari matriks Y′Y adalah[2]

$$\sum_i y_{ij}y_{ik} - \frac{1}{n}\sum_i y_{ij}\sum_i y_{ik}$$

$$\frac{1}{n}\mathrm{Y'Y} \to \sum$$

$$\frac{1}{n}\sum_i y_{ij}y_{ik} \to \frac{1}{n}\sum_i E\left(y_{ij}y_{ik}\right) \text{ dan } \frac{1}{n}\sum_i y_{ij} \to \frac{1}{n}\sum_i E\left(y_{ij}\right) \quad \ldots(2)$$

Untuk data berdistribusi Bernoulli[4], element ke-*j* dari persamaan (2) adalah

$$\lambda v_j = \frac{1}{n}\sum_i v_j e_j\left(1+e_{ij}\right)^{-2} + \frac{1}{n}\sum_i\left[\left(1+e_{ij}\right)^{-1}\sum_k v_k\left(1+e_{ij}\right)^{-1}\right] - \frac{1}{n}\sum_i\left(1+e_{ij}\right)^{-1}\frac{1}{n}\sum_i\left[\sum_k v_k\left(1+e_{ij}\right)^{-1}\right]$$

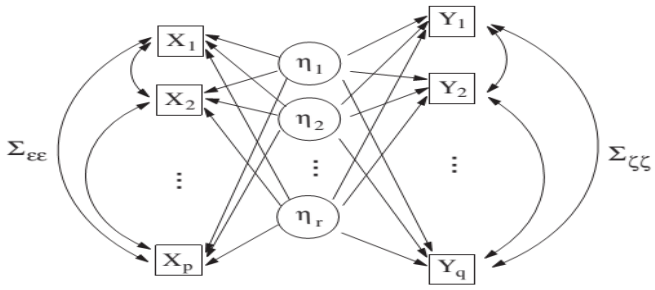
…(3)

Model-model Variabel Laten

Model variabel laten yang dikaji dalam makalah ini adalah model korelasi laten simetris dengan satu pasang variabel laten, yaitu model regresi peringkat tereduksi, model korelasi laten berpasangan, dan model laten diagonal. Dalam pembahasan ini, model yang digunakan adalah model diagonal laten dengan r-variat. Model laten diagonal simetris *r*-variat adalah himpunan distribusi pada vektor-*r* laten η, vektor kesalahan *p* yang diamati ε, vektor-*q* yang diamati Y, dan vektor-*q* kesalahan ζ, ditentukan sebagai berikut[3] :

$$\left.\begin{array}{l}\mathbf{x} = \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\epsilon} \text{ and } \mathbf{y} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta},\\ \text{dimana:}\\ \mathbf{Var}\,(\boldsymbol{\eta}) = \mathbf{I}_r,\ \mathbf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\varepsilon\varepsilon} \in I\!R^{(p\times p)},\ \mathbf{Var}(\boldsymbol{\zeta}) = \boldsymbol{\Sigma}_{\zeta\zeta} \in I\!R^{(q\times q)},\\ \varepsilon,\ \eta,\ \text{dan } \zeta \text{ bebas linier,}\\ \mathbf{A} \in \mathbf{R}^{(p\ \mathrm{x}\ r)} \text{ and } \mathbf{B} \in \mathbf{R}^{(q\ \mathrm{x}\ r)}\end{array}\right\} \quad \ldots (3)$$

Parameter-parameternya adalah matriks A, B, $\sum_{\varepsilon\varepsilon}$, dan $\sum_{\zeta\zeta}$, dengan kendala $\sum_{\varepsilon\varepsilon}$ dan $\sum_{\zeta\zeta}$ keduanya semidefinite positif. Model laten diagonal simetris *r*-variat adalah kasus khusus dari model korelasi laten berpasangan r-variate dimana ξ ≅ ω, Gambar 1.



Gambar 1. Model Diagonal Laten dengan r-variat

Metode HAIC

Metode HAIC merupakan salah satu metode estimasi reduksi dimensi terbaru yang dikemukakan oleh Ogura dan Fujikoshi [4]. Metode tersebut merupakan pengembangan dari metode AIC yang dikemukakan oleh Fujiksoshi [5]. Misalkan $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_{m-1}^2$ adalah korelasi laten diagonal laten dari matriks populasi pada dimensi dasar-*m*. Korelasi ini diestimasi oleh $rho_1^2 \geq rho_2^2 \geq \cdots \geq rho_{m-1}^2$ dari sampel acak distribusi Bernoulli. Misalkan M_k adalah model berdimensi *k*, yaitu,

$$M_k: \rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_{m-1}^2 = 0$$

Penggunaan AIC yang dimodifikasi oleh Fujikoshi [5], yaitu HAIC,

$$min_{M_k} n\left\|D_r^{-1/2} P D_c^{-1/2} - \Delta_r^{-1/2} \mathrm{P} \Delta_r^{-1/2}\right\|^2$$

$$= n(rho_{k+1}^2 + \cdots + rho_{m-1}^2)$$

$$\approx -nlog(1 - rho_{k+1}^2) \ldots (1 - rho_{m-1}^2)$$

$$D_k = DIC_k - DIC_{m-1} = -n\log(1 - rho_{k+1}^2) \ldots (1 - rho_{m-1}^2) + 2(r - k - 1)(c - k - 1) \quad \ldots (4)$$

3. TEMUAN DAN PEMBAHASAN

Simulasi Numerik

Simulasi akan dibahas dalam dua kasus, yaitu: (1) misalkan model M_k dengan k = 1, dan (2) model M_k dengan k = 2, atau disebut Model *true* M_k dengan k = 1, dan model *true* Mk dengan k = 2. Lebih tepatnya, elemen struktur probabilitas sejati (kovariansi) didefinisikan sebagai berikut:

$$M_k: \rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_{m-1}^2 = 0$$

Kemudian mengkonstruksi matriks A dari data berdistribusi Multinomial dengan probabilitas P untuk n = 150, 200, dan 250. Dengan cara yang sama juga dilakukan simulasi dengan menggunakan estimator AKU. Simulasi menggunakan program Matlab.

Kasus 1

Misalkan probabilitas populasi (kovarians)

P = 0.1840 0.0340 0.0340 0.0340 0.0340
0.0340 0.0340 0.0340 0.0340 0.0340
0.0340 0.0340 0.0340 0.0340 0.0340
0.0340 0.0340 0.0340 0.0340 0.0340
0.0340 0.0340 0.0340 0.0340 0.0340

Dari pengaturan ini, maka akar laten populasi AKU adalah

$$\rho_1^2 = 0{,}3750^2, \rho_2^2 = 0{,}00^2, \rho_3^2 = 0{,}00^2, \rho_4^2 = 0{,}00^2$$

Dari hasil generalisasi N, yaitu: n = 150, 200, dan 250 dari data berdistribusi multinomial dengan nilai peluang P dan 1.000.748 pengulangan diperoleh

$rho_1^2 = 0{,}0144^2, rho_2^2 = 0{,}0003^2, rho_3^2 = 0{,}00^2, rho_4^2 = 0{,}00^2.$

Hasil simulasi model AKU dapat dilihat pada Tabel 1. Dari hasil simulasi didapatkan rata-rata 64%, sehingga dapat dikatakan model berjalan dengan baik.

Tabel 1. Persentase kasus 1 model (1.000.748 kali)

n/k	0	1	2	3	4
150	0,00%	64,30%	28,56%	7,14%	0,00%
200	0,00%	64,31%	28,55%	7,14%	0,00%
250	0,00%	64,31%	28,55%	7,14%	0,00%

Kasus 2

Misalkan probabilitas populasi (kovarians)

P = 0.1840 0.0340 0.0340 0.0340 0.0340
0.0340 0.1390 0.0150 0.0150 0.0150
0.0340 0.0150 0.0350 0.0350 0.0350

0.0340 0.0150 0.0350 0.0350 0.0350
0.0340 0.0150 0.0350 0.0350 0.0350

Dengan cara yang sama diperoleh akar laten populasi dan akar laten hasil simulasi, yaitu:

$$\rho_1^2 = 0{,}5371^2, \rho_2^2 = 0{,}3574^2, \rho_3^2 = 0{,}00^2, \rho_4^2 = 0{,}00^2$$
$$rho_1^2 = 0{,}0879^2, rho_2^2 = 0{,}0022^2, rho_3^2 = 0{,}0012^2, rho_4^2 = 0{,}00^2.$$

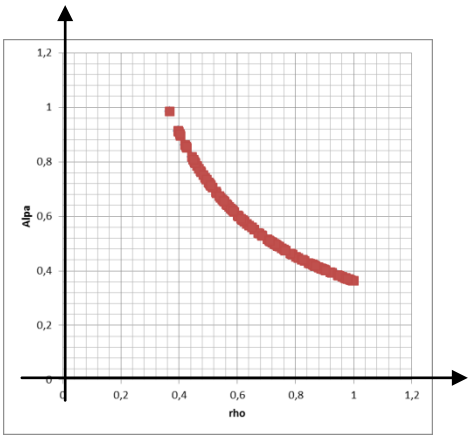
Hasil simulasi model AKU untuk kasus 2 dapat dilihat pada Tabel 2. Dari hasil simulasi didapatkan rata-rata 66%, maka model berjalan dengan baik.

Tabel 2. Persentase model AKU untuk kasus 2 (1.000.748 kali)

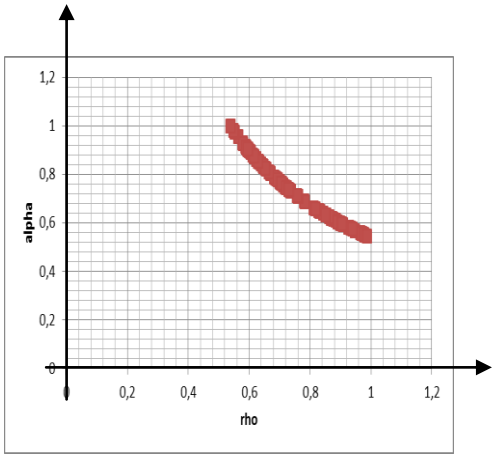
n/k	0	1	2	3	4
150	0,00%	0,03%	65,62%	27,25%	7,10%
200	0,00%	0,00%	65,98%	26,89%	7,09%
250	0,00%	0,00%	66,31%	26,56%	7,08%

Skema ρ dan α pada model laten diagonal berbanding terbalik, dimana α_{maks}=1 dan α_{min} = 0,363 untuk kasus 1, α_{maks}=1 dan α_{min} = 0,5383 untuk kasus 2. Simulasi dilakukan dengan 1.000.748 pengulangan, Kemudian dengan menggunakan metode rijek [6] diperoleh grafik yang tersaji pada Gambar 2 dan Gambar 3.

Hasil estimator AKU hampis sama dengan estimator AK (Analisis Korespondensi) pada data berdistribusi Bernoulli [7]. Artinya untuk data multivariate dengan teknik reduksi akan lebih baik ukuran sampelnya tidak kurang dari 200.



Gambar 2. Hubungan ρ dan α untuk kasus 1



Gambar 3. Hubungan ρ dan α untuk kasus 2

4. KESIMPULAN

Pada analisis estimator AKU untuk $n \to \infty$ (persamaan 3) kurvanya tidak linier[2], artinya sifat estimator bukan merupakan estimator yang baik untuk data tak hingga, tetapi untuk n = 150, 200, 250 dengan 1.000.747 pengulangan menggunakan data berdistribusi Bernoulli estimator AKU berjalan dengan baik. Berdasarkan grafik kurva mulus hasil simulasi numerik diperoleh korelasi laten diagonal (ρ) dengan parameter laten diagonal (α) berbanding terbalik secara eksponensial.

Dari hasil penelitian ini dapat juga disimpulkan bahwa untuk data multivariate yang menggunakan teknik reduksi AKU akan lebih baik ukuran sampelnya tidak kurang dari 200.

5. DAFTAR PUSTAKA

[1] M.J. Greenacre, (1994), *Theory and Application of Correspondence Analysis*, London: Academic Press, h. 233 – 245.

[2] H.S. Lynn and McCulloch, (2000), American Statistical Assosiation Journal of the American Statistical Assosiation Vol.95, No.450.
[3] Wegelin, Thomas S. R., and Asa Packer, (2006), Journal of Multivariate Analysis USA, h. 79 – 102.
[4] T. Ogura, and Y. Fujikoshi, (2013), Proceeding 59th ISI World Statistic Congress Hong Kong, h. 5361 – 5365.
[5] Y. Fujikoshi, R. Enomoto, and T. Sakurai, (2010), AMS Japan, 62H30, h. 1 – 23.
[6] Perros H., Computer Science Department NC State University Raleigh NC (2009).
[7] A.P, Bambang, Lukman, E. Sumiyati, (2016), *The characteristic of correspondence analysis estimator to estimate latent variable model method using high-dimensional AIC*, AIP Conference Proceedings **1708**, 060005 (2016); doi: 10.1063/1.4941168.