PENERAPAN MODEL INFERENSI BAYESIAN DENGAN VARIATIONAL BAYESIAN PRINCIPAL COMPONENT ANALYSIS (VBPCA) DALAM MENGATASI MISSING DATA ANALISIS KOMPONEN UTAMA

APPLICATION OF BAYESIAN INFERENCE MODEL VARIATIONAL BAYESIAN PRINCIPAL COMPONENT ANALYSIS (VBPCA) FOR HANDLING MISSING DATA IN PRINCIPAL COMPONENT ANALYSIS

Ricky Yordani
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 16-05-2016, diterima untuk diterbitkan tanggal: 29 Agustus 2016

Abstrak

Dalam Analisis komponen utama (AKU) yang standar muncul salah satu masalah yaitu AKU tidak jelas dalam mengatasi adanya gugusan data yang tidak lengkap. Prosedur standarnya pada masalah tersebut adalah dengan menghilangkan observasinya (prosedur *listwise deletion*) atau mengisinya dengan rata-rata variabel, hal ini dapat mengakibatkan hilangnya informasi dari observasi tersebut. Metode lain yang digunakan adalah dengan mengintegrasikan *Expectation Maximization* (EM) ke dalam metode *Probabilistic Principal Component Analysis* (PPCA). Tetapi metode PPCA dapat pula menghasilkan prediksi respon yang *overfitting*. Dalam penelitian ini dibahas tentang *Variational Bayesian Principal Component Analysis* (VBPCA) sebagai metode pengembangan dari metode PPCA dengan memasukkan informasi prior dari distribusi parameter model komponen utama. Dari studi simulasi dengan konsep *missing at random* (MAR), ukuran kecocokan antara nilai respon dengan prediksinya dengan melalui ukuran NRMSEP menghasilkan metode VBPCA lebih baik dibandingkan PPCA.

Kata Kunci: Variational Bayesian PCA, Analisis Komponen Utama, Nilai Hilang, Data Tidak Lengkap

Abstract

In standard Principal Component Analysis (PCA) comes one problem in addressing the set of incomplete data. The standard PCA procedure on incomplete data is to eliminate (listwise deletion procedure) or using the mean of the variable, this procedure may result in loss information from these observations. Another method used is to integrate Expectation Maximization (EM) to the method of Probabilistic Principal Component Analysis (PPCA). But PPCA can produce overfitting response prediction. In this study discussed the Variational Bayesian Principal Component Analysis (VBPCA) which is a method of development of PPCA method by incorporating prior information from the distribution of the principal components of the model parameters. From the simulation studies by eliminating the data through the concept of missing at random (MAR), obtained results that the value of the correlation scores principal components complete data with the principal component score predicted results PPCA method is superior when compared with VBPCA, as well as to the value of the correlation scores for the various percentages are generally incomplete data. However, judging from the size of a match between the response to predictions by the size normalized root mean square error of prediction (NRMSEP) VBPCA method produces better than PPCA.

Keywords: *Variational Bayesian PCA, Principal Component Analysis, Missing Value, Incomplete Data*

PENDAHULUAN

Para ahli metodologi telah mempelajari penanganan ketidaklengkapan data atau lebih popular dengan istilah *missing data* dalam kurun waktu ke belakang, dan telah banyak artikel yang membahas tentang topik ini. Para peneliti telah ditantang dengan adanya ketidaklengkapan data (*missing data*) semenjak awal penelitian dilaksanakan (Graham, 2009). Menurut seorang ahli metodologi William Shadish seperti dikutip oleh Baraldi dan Enders (2010) menyatakan bahwa missing data dalam penelitian termasuk perihal statistik yang sangat penting dan termasuk dalam masalah desain. Cara terbaik yang berhubungan dengan *missing data* dalam penelitian memerlukan dua informasi pokok yang harus dilaporkan dalam setiap studi penelitian: (a) tingkat dan sifat dari data yang tidak tersedia (b) prosedur yang digunakan dalam mengatasi data yang tidak tersedia termasuk dasar pemikiran untuk menggunakan metode yang digunakan untuk mengatasinya (Schlomer dkk, 2010).

Analisis Komponen Utama (AKU)– *Principal Component Analysis* (PCA)- merupakan metode analisis multivariat yang bertujuan memperkecil dimensi variabel asal sehingga diperoleh variabel baru (komponen utama). Komponen utama tersebut tidak saling berkorelasi tetapi menyimpan sebagian besar informasi yang terkandung pada variabel asal.

AKU dihadapkan dengan masalah yang muncul saat terdapat gugusan data yang tidak lengkap, misalnya saat ada beberapa nilai yang hilang. Ketidaklengkapan data dapat mengakibatkan kurang akurat dalam menganalisis dan mengevaluasi hasil penelitian. Walaupun begitu, banyak peneliti tidak menyadari pentingnya melaporkan dan mengelola nilai hilang, dan biasanya editor tidak mendesak peneliti menyediakan informasi penting ini (Schlomer dkk, 2010).

Prosedur AKU klasik dalam mengatasi masalah ini biasanya dengan menghilangkan observasi yang mengandung data yang tidak lengkap. Langkah tersebut merupakan suatu tindakan yang dapat membuang informasi ketika observasi yang terdapat nilai hilang memiliki proporsi yang tinggi tetapi hanya satu atau dua variabel yang nilainya hilang.

Beberapa teknik menganalisis AKU pada saat nilai hilang mempunyai mekanisme MAR diantaranya dipaparkan oleh Ilin dan Raiko (2010), yang dibedakan menjadi dua teknik, yaitu : (1) Teknik *Least Squares*, dan (2) Model Probabiliti untuk AKU. Teknik *Least Square* mencakup berbagai teknik, diantaranya *The Cost Function*, *Algoritma W-X* dan *Gradient Descent Algorithm*. Berdasarkan Ilin dan Raiko (2010), teknik-teknik tersebut cukup baik bila nilai yang hilangnya sedikit, tapi tidak dapat diaplikasikan saat nilai yang hilangnya banyak karena dapat mengakibatkan *overfitting*. Diantara akibat *overfitting* saat nilai hilangnya banyak adalah metode tersebut mampu menganalisis komponen utama tetapi model yang dibentuk tidak cukup bagus karena parameter yang dihasilkan tidak mampu memprediksi nilai yang hilang secara akurat.

Model Probabiliti untuk AKU dengan adanya nilai yang hilang antara lain *Probabilistic PCA* (PPCA) dan *Variational Bayes PCA* (VBPCA). VBPCA diperkenalkan oleh Bishop tahun 1999 untuk memilih banyaknya komponen dalam AKU, yang kemudian tahun 2003 oleh Oba dkk dilakukan penerapan metode VBPCA untuk

Metode AKU data tidak lengkap yang memberikan estimasi akurasi rata-rata bagus adalah VBPCA dan PPCA (Stacklies dkk, 2007). Berdasarkan Ilin dan Raiko (2010), PPCA merupakan metode yang cukup bagus digunakan pada data yang tidak lengkap, tetapi dapat terjadi *overfitting* saat data yang lengkapnya sedikit. *Overfitting* tersebut misalnya komponen utama yang terbentuk hanya

dipengaruhi variabel observasi tertentu saja, sehingga diperlukan informasi yang dapat mengatasi kelemahan tersebut. Informasi tersebut dapat berupa informasi tambahan melalui fungsi kepadatan prior dalam *Bayesian Framework* yang mengasumsikan semua parameter akan dianggap sebagai variabel acak. Sehingga dengan mengkombinasikan informasi tambahan *prior* sebelum melakukan AKU diharapkan mampu mengatasi kelemahan *overfitting y*ang terdapat pada PPCA.

METODOLOGI

Asumsi Mekanisme Ketidaklengkapan Data *(Missing Data)*

Asumsi mekanisme diperlukan karena metode analisis data akan berbeda berdasarkan asumsinya dalam menangani data tidak lengkap (missing data). Mekanisme ini menjelaskan hubungan antara variabel yang diukur dengan kemungkinan dari tidak tersedianya data. Mekanisme masing-masing asumsi mempunyai syarat-syarat probabilistik dan matematis yang secara pokok dapat memberikan perbedaan alasan penyebab data tidak tersedia (Baraldi dan Enders, 2010). Mekanisme *missing data* ini oleh Little dan Rubin (1987) diklasifikasikan dalam tiga kategori, yaitu:

1. *Missing completely at random* (MCAR); adalah kasus saat pola kemungkinan data tidak tersedia pada suatu variabel Y tidak berkaitan dengan variabel lain atau terhadap variabel Y sendiri.
   Dalam prakteknya, sulit untuk menentukan apakah data yang tidak tersedia mengikuti pola MCAR. Teknik menentukan MCAR telah diberikan oleh Schlomer dkk (2010).
2. *Missing at random* (MAR); adalah mekasnisme kasus saat terjadi kaitan antara variabel data yang memiliki nilai data yang tidak tersedia (*incomplete*) dengan variabel data yang tersedia nilai datanya. Pada kasus ini, nilai dari sebuah variabel yang tidak tersedia dipengaruhi oleh nilai-nilai dari variabel lainnya tetapi tidak dari variabel data.
3. *Nonignorable*; adalah mekanisme missing data dengan kemungkinan respon yang tidak tersedia jelas tergantung dengan variabel yang tidak lengkap tersebut. Dikenal juga dengan NMAR "*Not Missing at Random*" (Schlomer dkk, 2010) atau ada yang menyebutnya dengan MNAR "*Missing Not at Random*" (Baraldi dan Enders, 2010).

Analisis Komponen Utama (AKU)

Analisis Komponen Utama adalah metode Analisis Statistik Multivariat yang bertujuan memperkecil dimensi variabel asal sehingga diperoleh variabel baru (komponen utama) yang tidak saling berkorelasi tetapi menyimpan sebagian besar informasi yang terkandung pada variabel asal. Langkah tersebut dicapai dengan mentransformasikan ke gugus variabel baru (komponen utama), yang tidak saling berkorelasi, serta berurutan sehingga urutan beberapa gugus awal komponen baru mempertahankan paling banyak keragaman dalam variabel asal.

Komponen utama diekstrak sedemikian rupa sehingga komponen utama pertama, yang dinyatakan dengan X_1 merupakan kombinasi linear dari variabel pengamatan Y_j, untuk j = 1, 2, …, d, yaitu :

$$X_1 = u_{11}Y_1 + u_{12}Y_2 + \ldots + u_{1d}Y_d \quad (1)$$

dengan bobot u_{11}, u_{12}, …., u_{1d} dipilih untuk memaksimumkan rasio dari varian X_1 dengan variasi total, dengan batasan:

$$\sum_{j=1}^{d} u_{1j}^2 = 1 \quad (2)$$

sehingga komponen utama kedua X_2, sampai dengan komponen utama terakhir adalah kombinasi linear dari variabel Y, yaitu:

$$X_d = u_{d1}Y_1 + u_{d2}Y_2 + \cdots + u_{dd}Y_d \quad (3)$$

dengan batasan

$$\sum_{j=1}^{d} u_{dj}^2 = 1 \quad (4)$$

sehingga sifat varian komponen utamanya

$$\text{Var}(X_1) \geq \text{Var}(X_2) \geq \ldots \geq \text{Var}(X_d) \quad (5)$$

***Probabilistic Principal Component Analysis* (PPCA)**

PPCA merupakan metode AKU pada data tidak lengkap dengan menggunakan mekanisme algoritma *expectation maximization* (EM) dalam mengestimasi nilai data yang tidak tersedia. Algoritma EM tidak memerlukan matriks kovarian sampel dan dapat digunakan pada data yang high dimensional.

AKU sebagai kasus terbatas dari model linier Gaussian dengan kovarian dari *noise* **v** menjadi sangat kecil dan sama pada semua kovarian. Secara matematik, AKU didapatkan dengan mengambil limit $R = \lim_{\epsilon \to 0} \epsilon \mathbf{I}$. Hal tersebut mempunyai efek *likelihood* (kemungkinan) dari titik $\boldsymbol{y}$ yang didominasi tunggal oleh jarak kuadrat antara $\boldsymbol{y}$ dengan pembangkitan kembali $\mathbf{W}\boldsymbol{x}$. Arah dari kolom pada $\mathbf{W}$ yang meminimumkan error ini dikenal dengan sumbu utama. Inferensi sekarang menjadi proyeksi sederhana *least square*., yaitu:

$$p(x \mid y) = N\left(\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}\right)^{-1}\mathbf{W}^{\mathsf{T}} y, 0\right) = \delta\left(x - \left(\mathbf{W}\mathbf{W}^{\mathsf{T}}\right)^{-1}\mathbf{W}^{\mathsf{T}} y\right) \quad (6)$$

ketika noise menjadi terlalu kecil, posteriornya melampaui titik menjadi titik tunggal dan kovariannya menjadi nol.

Persamaan (7) dapat digunakan sebagai **e-step** untuk mengestimasi titik yang tidak lengkap atau tidak diketahui, kemudian menggunakan Persamaan (8) untuk mendapatkan **m-step** untuk mendapatkan **W**. Algoritmanya dinyatakan sebagai :

$$\textbf{e-step } \mathbf{X} = (\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\mathbf{Y} \quad (7)$$

$$\textbf{m-step} \qquad \mathbf{W^{new}} = \mathbf{Y}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1} \quad (8)$$

dengan **Y** adalah matrik berukuran $n \times d$ data yang diobservasi dan **X** adalah matriks berukuran $n \times k$ dari titik yang tidak diketahui. Kolom dari matriks **W** akan mencakup bagian dari k pertama sumbu-sumbu utama. Untuk menghitung nilai eigen dan vektor eigen, data dapat diproyeksikan ke dalam *subspace* berdimensi k dan sumbu orthogonal yang terurut untuk kovarian dalam subspace dapat dihasilkan.

Maksud algoritma tersebut adalah memperkirakan orientasi untuk *subspace* utama. Memperbaiki hasil dari perkiraan *subspace* dan memproyeksikan data $\boldsymbol{y}$ ke dalamnya untuk mendapatkan nilai dari titik $\boldsymbol{x}$ yang tersembunyi. Kemudian memperbaiki nilai-nilai dari titik yang tersembunyi dan memilih orientasi subspace yang meminimumkan error kuadrat yang dihasilkan dari data tersebut.

Mekanisme VBPCA

Seperti halnya PPCA, VBPCA menggunakan mekanisme algoritma EM dengan mengkombinasikan metode penaksiran bayesian dalam mengestimasi nilai data yang ditaksirnya. VBPCA dikembangkan khususnya untuk mengestimasi nilai yang tidak lengkap dan didasarkan atas kerangka kerja *Variational Bayes* (VBF) dengan *automatic relevance determination* (ARD).

Pada VBPCA, ARD menjadikan perbedaan skala terhadap komponen utama, skor dan nilai eigen ketika dibandingkan dengan AKU klasik ataupun PPCA. Metode ini yang mendasari antar komponen utama tidak perlu saling ortogonal.

Estimator VBPCA

Misalkan keseluruhan data dinyatakan dalam matriks Y yang berukuran (n×d), dengan n menandakan banyaknya sampel dan d menyatakan banyaknya variabel. Dengan yij menandakan baris ke-i sampel dan ke-j variabel dari matriks Y.

Dijelaskan oleh Oba dkk (2003), proses AKU pada data yang tidak lengkap dengan VBPCA terdiri atas tiga proses dasar, yaitu :

1. Analisis Faktor dengan Komponen Utama ;
2. Estimasi Bayesian;
3. Algoritma expectation-maximization (EM)

(15)

Analisis Faktor dengan Komponen Utama

Data yang memiliki data tidak lengkap dalam analisis faktor komponen utama bagian yang tidak lengkap dari y* diestimasi dari variabel observasi yang lengkap y^{obs} dengan menggunakan hasil dari AKU. Dimisalkan $\mathbf{w}_l^{obs}$ dan $\mathbf{w}_l^*$ sebagai bagian dari sumbu utama $\mathbf{w}_l$, yang masing-masing menyatakan data yang lengkap dan yang tidak lengkap dalam **y**. Kemudian, misalkan $W=(W^{obs},W^*)$ dengan masing-masing W^{obs} atau W* menyatakan matriks yang kolomnya berisi vektor $w_l^{obs}, \ldots, w_k^{obs}$ atau $w_1^*, \ldots, w_k^*$. Nilai skor faktor x=(x1,⋯,xk) dari vektor y didapatkan dengan meminimumkan kesalahan sisa:

$$err=||yobs-Wobs\ x||^2 \quad (9)$$

solusi dengan least square adalah:

$$x=(W^{obs\top} W^{obs})^{-1} W^{obs\top} y^{obs} \quad (10)$$

dengan menggunakan x, bagian yang tidak lengkap diestimasi dengan :

$$y^*=W^*\ x \quad (11)$$

Estimasi Bayesian

Dalam PPCA, model probabilistik dibentuk berdasarkan asumsi residual error $\boldsymbol{\varepsilon}$ dan skor faktor $x_l(1\leq l\leq k)$ berdistribusi normal dalam loading ke-k persamaan tersebut menjadi :

$$x\sim N\ (x;0,Ik\ ) \quad (12)$$

$$\varepsilon\sim Nd\ (\varepsilon;0,(1/\tau)\ Id\ ) \quad (13)$$

dengan τ menyatakan skalar dari invers varian ε. Sehingga dalam model PPCA, fungsi log-likelihood adalah :

$$\ln p(y,x|\theta) \equiv \ln p(y,x|W,\mu,\tau) \quad (14)$$

$$= -\frac{\tau}{2}\|\mathbf{y}-\mathbf{W}\mathbf{x}-\boldsymbol{\mu}\|^2 - \frac{1}{2}\|\mathbf{x}\|^2 + \frac{d}{2}\ln\tau - \frac{k+d}{2}\ln 2\pi$$

dengan θ ≡ {W,μ,τ} sebagai gugusan parameter.

Penambahan estimasi Bayesian untuk PPCA yang menjadikan perbedaan dengan VBPCA adalah mendapatkan distribusi posterior θ dan X menggunakan teorema Bayes.

p(θ) dinamakan distribusi *prior*, yang menyatakan pilihan awal dari parameter θ. Distribusi prior merupakan bagian dari model dan harus didefinisikan sebelum melakukan estimasi.

Dimisalkan fungsi prior dalam VBPCA adalah:

$$p(\boldsymbol{\mu}|\tau) = N\left(\boldsymbol{\mu}\middle|\bar{\boldsymbol{\mu}}_0,(\gamma_{\mu_0}\tau)^{-1}\mathbf{I}_m\right) \quad (16)$$

$$p(\mathbf{w}_j|\tau,\boldsymbol{\alpha}_j) = N\left(\mathbf{w}_j\middle|\mathbf{0},(\boldsymbol{\alpha}_j\tau)^{-1}\mathbf{I}_m\right) \quad (17)$$

$$p(\tau) = \mathcal{G}\left(\tau\middle|\bar{\tau}_0,\gamma_{\tau_0}\right) \quad (18)$$

dengan :

$\mathcal{G}(\tau|\bar{\tau},\gamma_\tau)$ merupakan distribusi gamma dengan hyperparameter $\bar{\tau}$ dan γ_τ:

$$p(\boldsymbol{\theta},\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$$

$$\mathcal{G}(\tau|\bar{\tau},\gamma_\tau) \equiv \frac{(\gamma_\tau\bar{\tau}^{-1})^{(\gamma_\tau)}}{\Gamma(\gamma_\tau)}\exp[-\gamma_\tau\bar{\tau}^{-1}\tau + (\gamma_\tau-1)\ln\tau] \quad (19)$$

dengan Γ(.) merupakan fungsi Gamma, sehingga kemudian prior μ, τ dan W adalah p(W|τ,α) dinyatakan dengan hyperparameter α∈R^k.

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \equiv p(\boldsymbol{\mu},\mathbf{W},\tau|\boldsymbol{\alpha}) = p(\boldsymbol{\mu}|\tau)p(\tau)\prod_{j=1}^{k}p\left(\mathbf{w}_j\middle|\tau,\alpha_j\right) \quad (20)$$

Variabel yang digunakan dalam prior tersebut γ_{μ_0}, $\bar{\boldsymbol{\mu}}_0, \gamma_{\tau_0}$ dan $\bar{\tau}_0$ merupakan hyperparameter yang mendefinisikan prior. Nilai aktualnya harus sudah ditentukan sebelum melakukan estimasi. $\gamma_{\mu_0} = \gamma_{\tau_0} = 10^{-10}$, $\bar{\boldsymbol{\mu}}_0 = 0$ dan $\bar{\tau}_0 = 1$ yang bersesuaian pada prior non-informatif.

Distribusi posterior dari parameter dinyatakan dengan :

$$q(\theta) = p(\theta|Y,\alpha_{\mathrm{ML-II}}) \quad (21)$$

Prior hirarki $p(\mathbf{W}|\tau,\boldsymbol{\alpha})$ dikenal sebagai ARD (*automatic relevance determination*). Sumbu utama ke-*j* dari $\mathbf{w}_j$ mempunyai prior Gaussian dengan varian $1/(\boldsymbol{\alpha}_j\tau)$ tergantung oleh hyperparameter $\boldsymbol{\alpha}_j$ yang ditentukan oleh estimasi ML tipe-II dari data.

Ukuran Performasi Metode

Menentukan jumlah komponen utama optimal yang mencakup informasi

relevan dengan mengurangi adanya noise salah satunya dapat dilakukan dengan melakukan *Cross Validation*. Q^2 merupakan ukuran yang dapat digunakan untuk melakukan internal *cross validation*. Ukuran tersebut dapat mengestimasi struktur level dari gugusan data dan dapat optimal dalam memilih banyaknya komponen.

Langkah I $SSY = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij}\right)^2$

Langkah II

$$PRESS = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij-}\hat{y}_{ij}\right)^2$$

Langkah III `

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij-}\hat{y}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{d}\left(y_{ij}\right)^2} \quad (22)$$

Nilai maksimum dari Q^2 adalah 1, yang berarti nilai pengamatan aktual sama persis dengan prediktornya; yaitu $Y = \hat{Y}$.

Ukuran lain yang digunakan untuk dapat menentukan banyaknya komponen yang optimal yaitu NRMSEP (*normalized root mean square error of prediction*). NRMSEP menormalkan perbedaan kuadrat untuk variabel tertentu antara nilai estimasi dan nilai sebenarnya dengan varian variabel tersebut. Dasar pemikiran dari ukuran ini adalah dapat memperlihatkan bahwa kesalahan dari prediksi dapat otomatis membesar bila variannya membesar.

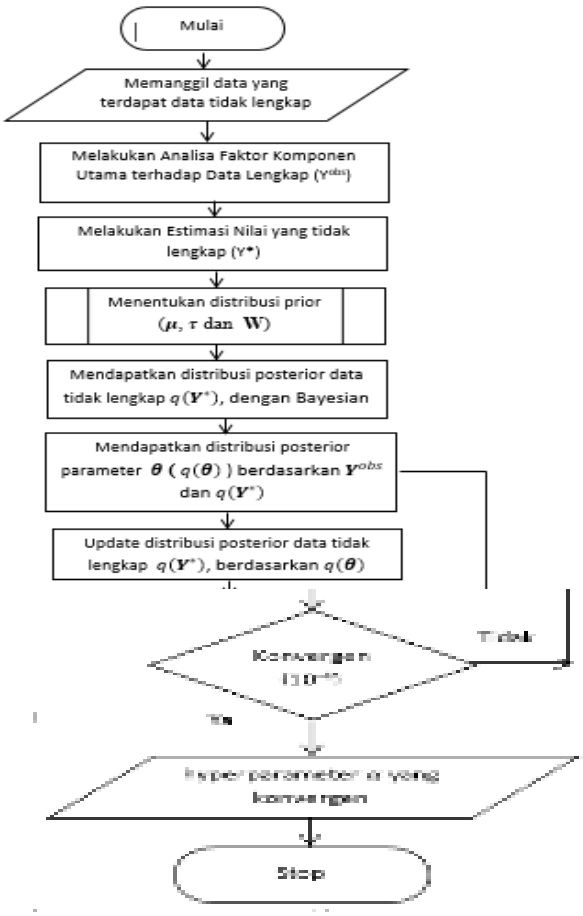
$$\mathbf{NRMSEP}_k = \sqrt{\frac{1}{g}\sum_{j\epsilon G}^{P}\frac{\sum_{i\epsilon o_j}^{U}\left(y_{ij-}\hat{y}_{ijk}\right)^2}{o_j S_{y_j}^2}} \quad (23)$$

dengan G menyatakan gugusan data yang memiliki data tidak lengkap (dimisalkan sebanyak P), sedangkan g merupakan jumlah variabel yang tidak lengkap. O_j merupakan gugusan dari observasi yang memiliki nilai yang tidak lengkap dalam variabel j (dimisalkan sebanyak U) dan o_j merupakan jumlah observasi yang memilki nilai yang tidak lengkap dalam variabel j. $\hat{y}_{ijk}$ menyatakan estimasi nilai ke-i dari variabel ke-j saat menggunakan loading ke-k. $S_{y_j}^2 = \sum_{i=1}^{n}\left(y_{ij} - \bar{y}_j\right)^2 / (n-1)$ merupakan varian dari variabel tersebut. Sehingga NRMSEP akan menjadi kecil bila varian internalnya besar.

Algoritma VBPCA

Algoritma VBPCA terdiri atas beberapa tahapan langkah yang mencakup tiga proses dasar dalam mengestimasi AKU pada data tidak lengkap adalah sebagai berikut :

1. Menginput data berpasangan berdasarkan obyek dan variabel yang akan dianalisis komponen utamanya yang memiliki nilai data tidak lengkap
2. Melakukan analisis faktor komponen utama, dengan data yang tidak lengkap $\mathbf{Y}^*$ diestimasi berdasarkan $\mathbf{Y}^{obs}$ dengan menggunakan hasil dari AKU. $\mathbf{Y}^* = \mathbf{W}^*\boldsymbol{x}$
3. Tentukan distribusi *prior* ($\boldsymbol{\mu}$, τ dan $\mathbf{W}$)
4. Memperoleh distribusi posterior data tidak lengkap $q(\boldsymbol{Y}^*)$
5. Mendapatkan distribusi posterior parameter $\boldsymbol{\theta}$ $(q(\boldsymbol{\theta}))$ berdasarkan data observasi $\mathbf{Y}^{obs}$ dan berdasarkan distribusi posterior yang baru dari data yang tidak lengkap $q(\mathbf{Y}^*)$
6. Distribusi posterior data tidak lengkap $q(\mathbf{Y}^*)$ kemudian diestimasi berdasarkan distribusi posterior $q(\boldsymbol{\theta})$ yang baru didapatkan sebelumnya pada tahap (5).
7. Hyperparameter $\boldsymbol{\alpha}$ kemudian diupdate berdasarkan $q(\boldsymbol{\theta})$ dan $q(\mathbf{Y}^*)$ yang baru (kedua tahapan hasil di atas (5) dan (6)).
8. Ulangi proses hingga langkah 5 sampai 7 konvergen.



Gambar 1 Diagram Alur Algoritma VBPCA

Data dan Variabel Aplikasi Analisis VBPCA

Untuk memberikan gambaran mengenai aplikasi metode VBPCA, maka dilakukan AKU dengan berbagai persentase keadaan data tidak lengkap. Pada penerapan aplikasi dilakukan pada data lengkap olahan Survei Sosial Ekonomi Nasional (SUSENAS) tahun 2007 tentang indikator-indikator tujuan pembangunan millennium pada kabupaten/kota di Kawasan Indonesia Timur (10 provinsi dengan 112 kabupaten/kota) dengan dilakukan simulasi terdapat beberapa nilai yang tidak tersedia sehingga menjadikan gugusan data tersebut tidak lengkap. Ukuran observasi kabupaten/kota sebanyak 112, dengan variabel yang dianalisis sebanyak 30 variabel.

Variabel indikator yang digunakan antara lain sebanyak 30 variabel variabel (PDDK, IND_1, IND_2, … , IND_29) sebagai berikut :

PDDK : Jumlah penduduk Kabupaten/Kota
IND_1 : Penduduk berumur 0-6 tahun yang pernah pra sekolah
IND_2 : Angka Partisipasi Sekolah penduduk berumur 7-12 tahun
IND_3 : Angka Partisipasi Sekolah penduduk berumur 13-15 tahun
IND_4 : Angka Partisipasi Sekolah penduduk berumur 16-18 tahun
IND_5 : Angka Partisipasi Sekolah penduduk berumur 19-24 tahun
IND_6 : Angka Partisipasi Murni SD/MI
IND_7 : Angka Partisipasi Murni SMP/Kejuruan/MTs
IND_8 : Angka Partisipasi Murni SMA/Kejuruan/MA
IND_9 : Angka Partisipasi Murni Perguruan Tinggi
IND_10 : Angka Partisipasi Kotor SD/MI
IND_11 : Angka Partisipasi Kotor SMP/Kejuruan/MTs
IND_12 : Angka Partisipasi Kotor SMA/Kejuruan/MA
IND_13 : Angka Partisipasi Kotor Perguruan Tinggi
IND_14 : Penduduk berumur 15-24 tahun yang buta huruf
IND_15 : Penduduk berumur 15 tahun ke atas yang buta huruf
IND_16 : Balita yang ditolong kelahirannya oleh tenaga medis
IND_17 : Anak berumur 12-23 bulan yang diimunisasi Campak
IND_18 : Anak berumur 1-4 tahun yang diimunisasi lengkap
IND_19 : Anak berumur 6 bulan ke atas yang diberi ASI saja selama 6 bulan atau lebih
IND_20 : Anak berumur 0-4 tahun yang memiliki akte kelahiran
IND_21 : WUS yang sedang ber-KB

IND_22 : WUS yang sedang ber-KB hormonal
IND_23 : WUS yang sedang ber-KB mantap
IND_24 : WUS yang sedang ber-KB dibanding laki-laki yang sedang ber-KB
IND_25 : Remaja berumur 10-18 yang pernah kawin
IND_26 : Angka ketergantungan penduduk
IND_27 : Angka ketergantungan penduduk muda
IND_28 : Angka ketergantungan penduduk tua
IND_29 : Rasio jenis kelamin

Aplikasi penelitian data tidak lengkap pada variabel indikator tujuan pembangunan millenium tersebut menggunakan persentase data tidak lengkap mulai dari 5%, 15%, 40% dan 60% dengan mekanisme data tidak lengkapnya mengikuti MAR. Penelitian ini akan diuji coba pada level data tidak lengkap 60% apakah hasil PPCA masih dapat memberikan akurasi yang tepat dan selanjutnya akan dibandingkan dengan VBPCA sebagai metode yang lebih maju, sehingga akan didapat perbandingan performasi metode dari VBPCA dengan metode sebelumnya. Presentase simulasi data tidak lengkap dimaksudkan untuk membandingkan antara metode PPCA dengan metode VBPCA dan juga metode AKU tradisional dalam melakukan analisis saat terdapat data tidak lengkap pada beberapa tingkat persentase data tidak lengkap.

Simulasi data dengan mekanisme MAR dilakukan dengan mengurutkan keseluruhan variabel berdasarkan variabel tertentu, yang kemudian menghilangkan nilai yang terendah dari beberapa variabel berdasarkan presentase data tidak lengkapnya. Simulasi ini menggunakan bantuan paket software *open source* statistik R versi 3.2.3, adapun langkah-langkah simulasi MAR tersebut adalah sebagai berikut:

1. Menginisiasi dan memanggil data lengkap dan menyimpannya sebagai objek dengan nama IndoTim07_L.
```
> IndoTim07_L <-read.table("
. . .")# titik diisiirektori
```
2. Membuat objek bayangan sebagai penyimpan sementara objek data lengkapnya.
```
> IndoTim07_D<-IndoTim07_L
> IndoTim07_D<-
as.data.frame(IndoTim07_D)
> attach(IndoTim07_D)
```
3. Membuat objek untuk menyimpan hasil simulasi data tidak lengkap, sebagai contoh untuk data tidak lengkap 40% adalah Indv_TL.40.
```
> Indv_TL.40<-IndoTim07_L
```
4. Mencari nilai baris sebagai batas tidak lengkap.
```
> a.40<-0.4 # presentase MAR
40%
> n.40<-
  length(IndoTim07_D[,3])*a.
  40 #      kolom 1 dan ke 2
  dari data merupakan kode
  provinsi dan kab, maka
  tidak perlu diikutkan,
  sehingga proses komputasi
  dimulai pada kolom ke-3
  dan seterusnya
> Tr.40<-round(n.40)        #
  mencari letak baris ke
  beberapa sebagai batas
  treshoold presentase data
  tidak lengkap
```
5. Mengurutkan variabel tertentu, dengan mekanisme MAR (variabel nilai data yang tidak tersedia berkorelasi terhadap variabel lain yang lengkap) sehingga akan diurutkan beberapa variabel sebagai patokan variabel lain untuk dihilangkan datanya. Variabel yang dijadikan patokan urutan antara lain variabel PDDK, IND_2, IND_6, IND_14, IND_26.
```
> # Mengurutkan berdasarkan
order ketiga yaitu varbl
IND_6 kolom 9
> IndoTim07_D_3<-
IndoTim07_D[order(IND_6),]
> Indv_D_3_40<-
IndoTim07_D_3[Tr.40,9]
> j.3<-
   (1:length(IndoTim07_D[,9])
   )
   [(IndoTim07_D[,9])<=Indv_D
   _3_40]
```

```
# mengidentifikasi urutan
  ke berapa dari var IND_6
  yang lebih kecil dari
  batas threshold
  presentase data tidak
  lengkap
```

6. Membuat variabel dummy, dengan 1 merupakan yang disimulasikan datanya tidak tersedia sedangkan 0 datanya tersedia.

```
> k=1
> for (k in
  length(IndoTim07_D[,9]))
  IndoTim07_D[1:length
  (IndoTim07_D[,9]),9] <-0
  #OK, dummy kode 0 dan 1
> IndoTim07_D[j.3,9]<-1 #
kode 1 tuk yang
disimulasikan hilang
> co.3<-
  c(11,14,19,28,31,32)#
  kolom variabel yang nilai
  tidak tersedia akibat
  dipengaruhi order 3
```

7. Mengidentifikasi yang berkode 1 menjadi 99999

```
> Indv_TL.40[,co.3]<-
  Indv_TL.40[,co.3]*(1-
  IndoTim07_D[,9])+
  IndoTim07_D[,9]*99999
```

8. Mengganti yang bernilai 99999 dengan kode NA *(not available)*

```
>
Indv_TL.40[,4:32][Indv_TL.40
[,4:32]==99999]=NA
```

9. Menyimpan data hasil simulasi MAR 40% ke file

```
>
  write.table(Indv_TL.40,"F:
  /Tesis/1 11
  Des/Indv_TL.40.Final.xls",
  sep="\t", row.names=FALSE)
```

Setelah dilakukan simulasi data tidak lengkap dari variabel indikator tersebut, dilakukan uji *Little's MCAR* agar didapatkan data tidak lengkap hasil simulasi telah sesuai dengan yang diharapkan bahwa nilai variabel yang tidak tersedia dipengaruhi variabel lain. Kemudian dilakukan AKU dengan menggunakan metode tradisional listwise, metode imputasi rata-rata, PPCA dan VBPCA yang dilanjutkan dengan melakukan perbandingan hasil metode-metode tersebut.

Adapun langkah-langkah lengkap komputasi dalam penelitian ini dengan aplikasi R adalah sebagai berikut:

1. Mengaktifkan library pcaMethods dengan perintah

```
> library(pcaMethods)
```

2. Mengimput nilai-nilai dari variabel indikator yang akan dianalisis sebanyak 30 (*d*) variabel **Y**, dan sebanyak 112 kabupaten/kota (*n*).
3. Mengaplikasikan mekanisme data tidak lengkap MAR dengan beberapa tingkatan persentase 5%, 15%, 40% dan 60% data tidak lengkapnya terhadap keseluruhan data saat lengkap.
4. Melakukan uji mekanisme ketidaklengkapan data.
5. Melakukan proses standarisasi data menjadi variabel baku karena datanya memiliki satuan dan besaran yang tidak sama.
6. Melakukan estimasi AKU dengan data awal yang lengkap.
7. Melakukan estimasi dengan metode tradisional saat AKU memiliki data tidak lengkap, yang kemudian dilakukan estimasi PPCA dan VBPCA untuk mendapatkan banyaknya komponen utama dengan menggunakan fungsi pca() terhadap beberapa tingkat persentase mekanisme data tidak lengkap di atas.
8. Menghitung Q^2 dan $NRMSEP_k$ beberapa metode dan beberapa tingkat persentase mekanisme data tidak lengkap.
9. Membandingkan hasil dari Q^2 dan $NRMSEP_k$ dari beberapa metode dan beberapa tingkat persentase mekanisme data tidak lengkap.

HASIL DAN PEMBAHASAN

Simulasi Kejadian Data Tidak Lengkap

Pembahasan pada bagian ini mengandung beberapa hal, diantaranya : Gambaran Umum Data, Uji MAR Data Tidak Lengkap, Ukuran Performasi Metode (Cross Validation (Q^2),

NRMSEP), skor observasi dan skor loading penimbang variabel terhadap komponen utama dari beberapa metode.

Gambaran Umum Data

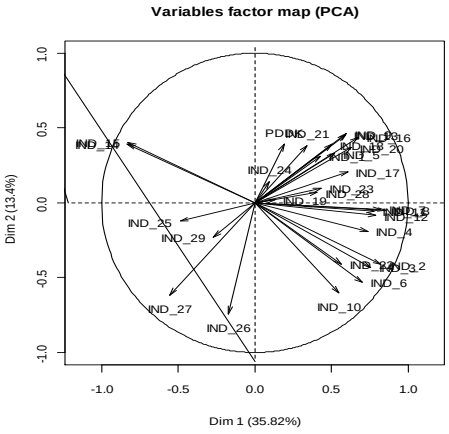
Saat data tersebut lengkap statistik rata-rata dan standar deviasi awal ditampilkan di Tabel 1, dengan total observasi saat lengkap akan terdapat sebanyak 112 nilai dari masing-masing variabel.

Dengan dilakukan simulasi data tidak lengkap sebanyak 5% MAR, pola data tidak lengkap ditampilkan di Tabel 2 Pada Tabel 2, setiap baris menandakan pola respon observasi kabupaten/kota terhadap nilai variabelnya Saat terdapat nilai dari variabel maka berkode 1, sedang saat tidak terdapat nilainya (NA-*not available*) berkode 0. Untuk baris pertama menandakan bahwa observasi kabupaten/kota ini mempunyai nilai variabel yang lengkap dengan banyaknya kasus atas pola ini sebanyak 89 observasi kabupaten/kota (kolom di kiri), atau dari keseluruhan 112 observasi kabupaten/kota terdapat hanya 89 observasi kabupaten/kota yang memiliki data yang lengkap.

Tabel 1 Statistik Rata-Rata dan Standar Deviasi Data Lengkap

VARIABLE	MEAN	STDEV	VALID N
"PDDK"	187375.54	157172.02	112
"IND_1"	4.89	3.26	112
"IND_2"	94.33	7.26	112
"IND_3"	82.99	9.70	112
"IND_4"	53.85	14.24	112
"IND_5"	10.62	7.96	112
"IND_6"	90.67	6.57	112
"IND_7"	59.02	14.10	112
"IND_8"	42.19	14.13	112
"IND_9"	6.62	7.67	112
"IND_10"	110.48	9.27	112
"IND_11"	75.57	16.66	112
"IND_12"	56.86	18.67	112
"IND_13"	10.94	10.51	112
"IND_14"	4.57	8.30	112
"IND_15"	10.57	12.42	112
"IND_16"	54.45	21.59	112
"IND_17"	78.13	15.98	112
"IND_18"	56.92	22.83	112
"IND_19"	27.77	13.29	112
"IND_20"	25.92	15.41	112

Uji MAR Data Tidak Lengkap



Gambar 2. Plot Variabel Hasil AKU pada Dua Komponen Utama Pertama Data Lengkap

Kolom di kanan menandakan pola jumlah nilai variabel yang NA (berkode 0) dalam baris observasi. Baris terakhir tabel yang terdapat di bawah menandakan banyaknya baris dari suatu nilai variabel yang NA (kode 0) dalam setiap variabel Indikator.

Sehingga dapat dilihat dari tabel, untuk variabel ke-17 terdapat nilai yang NA sebanyak 12 baris observasi. Untuk pola dari baris observasi kabupaten/kota, terdapat 5 observasi kabupaten/kota yang memiliki nilai NA sebanyak 12 dalam variabelnya (Tabel 4.2 baris ke-7 dan ke-8), dengan dua pola yang berbeda dalam adanya nilai data yang NA.

Simulasi juga dilakukan dengan presentase MAR 15%, 40 % dan 60%. Simulasi data yang dihasilkan perlu dilakukan uji mekanisme MAR data tersebut sesuai dengan asumsi dalam melakukan metode AKU dengan PPCA dan VBPCA.

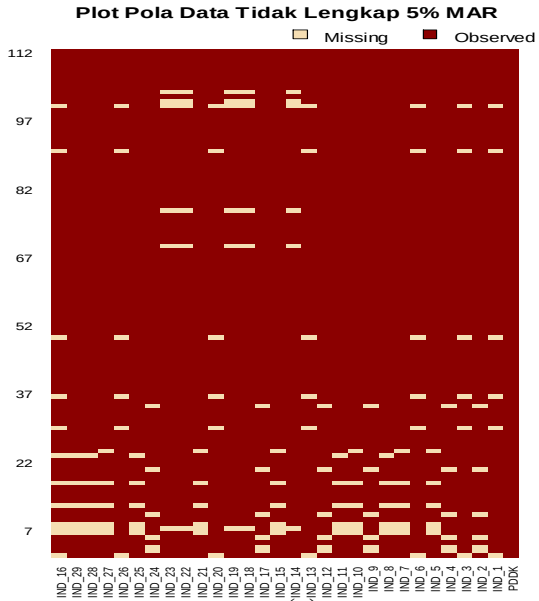
Dengan bantuan aplikasi yang telah memiliki mekanisme uji data tidak lengkap, analisis uji Little's MCAR dapat dilakukan pada data yang telah disimulasikan tidak lengkap seperti terdapat dalam Lampiran 1.

Dengan nilai data yang tidak lengkap telah diketahui sebelumnya (karena dalam kasus ini, penelitian menggunakan hasil dari simulasi) dapat diverifikasi bahwa data tidak lengkap mengikuti mekanisme MAR. Sedangkan pada saat tidak diketahui

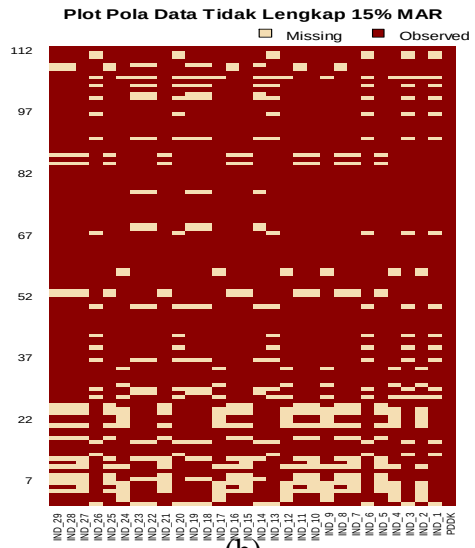
nilai data sebenarnya dari variabel nilai data yang tidak lengkap, mekanisme antara MAR dan NMAR tidak dapat dilakukan verifikasi (Enders, 2010).

Pada Gambar 3 ditampilkan beberapa visualisasi dari pola data tidak lengkap dalam penelitian ini. Untuk visualisasi persentase data tidak lengkap lainnya dapat dilihat dalam lampiran 3.

Tabel 3 menjelaskan bahwa metode imputasi dengan rata-rata masih tidak lebih baik dari metode dengan PPCA dan VBPCA. Dapat dilihat juga bahwa pada berbagai kondisi simulasi data tidak lengkap korelasi yang terbesar didominasi oleh metode dengan PPCA.



(a)



(b)

Gambar 3 Plot Pola Data Hasil Simulasi (a) Lengkap Data 5% Tidak Lengkap, (b) Data 15% Tidak

Tabel 3 diperlihatkan korelasi skor observasi dan skor loading AKU yang dihasilkan dari data awal lengkap dengan data lengkap hasil estimasi dengan VBPCA.

Pengembang metode estimasi VBPCA menyatakan bahwa metode VBPCA menghasilkan estimasi nilai taksiran nilai data tidak lengkap lebih baik dari pada metode AKU yang lain.

Pada Tabel 4 nilai korelasi warna biru menandakan angka korelasi yang dihasilkan tersebut lebih tinggi dibanding korelasi VBPCA pada Tabel 3, sedangkan nilai korelasi warna merah menandakan angka korelasi tersebut tidak signifikan pada tingkat signifikansi α=5%. Terlihat angka korelasi yang dihasilkan secara dominan lebih baik dari pada hasil VBPCA pada Tabel 4.

Pada Tabel 5 diberikan hasil metode dalam melakukan AKU pada data tidak lengkap. Pada kolom (7) $VBPCA_b$, kolom tersebut menerapkan tehnik AKU dengan VBPCA, tetapi setelah didapatkan perkiraan keseluruhan nilai data lengkap kemudian dilakukan AKU terhadap data lengkap tersebut.

Dengan membandingkan hasil yang diperoleh dengan data lengkap saat tujuh komponen utama pertama yang mampu menjelaskan 79,18% keragaman data, pilihan dalam menggunakan teknik klasik *listwise deletion* dan imputasi rata-rata nilai data yang tidak lengkap bukanlah merupakan suatu pilihan yang patut dipertimbangkan karena hasil variasi yang dapat dijelaskan dari estimasinya masih jauh dibawah metode PPCA dan VBPCA.

Terlihat bahwa antara dua metode PPCA dan VBPCA merupakan pilihan metode yang cukup bagus dalam menangani AKU pada data tidak lengkap. Perbandingan metode keduanya dapat dilihat dalam perbandingan nilai Cross Validation dan NRMSEP.

Gambar 5 ditampilkan berbagai hasil dari NRMSEP dari berbagai kemungkinanan data tidak lengkap. Gambar 5 memperlihatkan bahwa prediksi akar rata-rata kesalahan kuadrat error yang dinormalkan antara metode PPCA dan VBPCA, yang menghasilkan kesimpulan bahwa menggunakan metode VBPCA lebih baik dibandingkan PPCA dalam berbagai presentase data tidak lengkap. Untuk Grafik data tidak lengkap 40% dapat dilihat di Lampiran 3.

Pada Gambar 6 tentang grafik Q^2, dapat diperlihatkan bahwa nilai Q^2 hasil PPCA berada di bawah nilai Q^2 saat menggunakan metode VBPCA. Gambar tersebut juga menyiratkan bahwa pada berbagai presentase data tidak lengkap,
hasil dari VBPCA lebih baik dibandingkan dengan PPCA. Q^2 manandakan nilai dari pengamatan aktual sama persis dengan prediktornya sebesar nilai tersebut.

KESIMPULAN

Beberapa kesimpulan yang dapat dirangkum dari hasil dan pembahasan sebelumnya antara lain :

1. VBPCA dan PPCA merupakan metode yang lebih baik dibandingkan *listwise deletion* dan imputasi rata-rata.
2. Metode-metode yang diuji coba dalam penelitian ini tidak mampu menghasilkan estimasi yang akurat saat data 40% tidak lengkap atau lebih dari itu.
3. Hasil NRMSEP yang dihasilkan antara PPCA dan VBPCA tidak begitu berbeda saat KU di bagian awal, tapi akan langsung berbeda saat KU berikutnya, dengan VBPCA memberikan hasil yang lebih baik.
4. Hasil Q^2 Cross Validation PPCA dan VBPCA berhimpit pada KU pertama, tapi kemudian berbeda saat Q^2 selanjutnya dengan VBPCA memberikan hasil yang lebih baik.

DAFTAR PUSTAKA

Bakshi, B. R., Nounou, M. N., Goel, P. K., dan Shen, X. (tanpa tahun). Bayesian Principal Component Analysis. Melalui < http://classifion.com/References/BayesianPCA.pdf > [15/9/15]

Banda, J.P. 2003. Nonsampling Errors in Surveys. *Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Survey*. Melalui <http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_7.pdf >[12/01/16]

Baraldi, A. N. dan Enders, C. K. 2010. An introduction to Modern Missing Data Analyses. *Journal of School Psychology* 48, 5–37.

Bentley, J. P. 2009. Missing Data: An Introduction (with a focus on multiple imputation). Workshop offered by the Mississippi Center for Supercomputing Research and the UM Office of Information Technology. Melalui <http://www.mcsr.olemiss.edu /mathematica/Missing%20Data%20-%20An%20introduction.pdf.copy > [6/8/15]

Bernard, C., Michel dan Jegou, H. 2008. Chris Bishop's Pattern Recognition and Machine Learning, Ch. XII. Contiuous Latent Variables. Melalui <http://lear.inrialpes.fr/~jegou/bishopreadinggroup/chap12.pdf> [18/9/15]

Bishop, C. M. 1999. Variational Principal Component. *Ninth International Conference on Artificial Neural Networks*, ICANN, IEE Vol I, 509-514.

Bishop, C. M dan Tipping, M. E. 1999. Probabilistic Principal Component Analysis. *Journal of The Royal Statistical Society*, Series B, 61, Part3, 611-622.

Bolstad, W. M. 2004. *Introduction to Bayesian Statistics*. New Jersey : John Wiley & Sons, Inc.

Borman, S. 2009. The Expectation Maximization Algorithm A Short Tutorial. Melalui <http://www.seanborman.com/publications/EM_algorithm.pdf> [11/01/16]

BPS. 2011. *Perkembangan Beberapa Indikator Utama Sosial-Ekonomi Indonesia (Mei)*. Jakarta-Indonesia : BPS.

Chen, H. 2001. Principal Component Analysis With Missing Data and Outliers. Melalui <http://www.nec-labs.com/~haifeng/mypubs/tutorialrpca.pdfhttp://www.nec-labs.com/~haifeng/mypubs/tutorialrpca.pdf> [10/8/15]

Enders, C. K. 2010. *Applied Missing Data Analysis*. New York: Guilford Press.

Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.

Gold, M. S., Bentler, P. M. dan Kim, K. H. 2002. A Comparison of Maximum-Likelihood and Asymptotically Distribution-Free Methods of Treating Incomplete Non-Normal Data. Melalui <http://statistics.ucla.edu/system/resources/BAhbBlsHOgZmSSIBkjIwMTIvMDUvMjEvMTVfNDdfNTFfNjYyX0FfQ29tcGFyaXNvbl9vZl9NYXhpbXVtX0xpa2VsaWhvb2RfYW5kX0FzeW1wdG90aWNhbGx5X0Rpc3RyaWJ1dGlvbl9GcmVlX01ldGhvZHNfb2ZfVHJlYXRpbmdfSW5jb21wbGV0ZV9Ob25fTm9ybWFsX0RhdGEucGRmBjoGRVQ/A%20Comparison%20of%20Maximum-Likelihood%20and%20Asymptotically%20Distribution-Free%20Methods%20of%20Treating%20Incomplete%20Non-Normal%20Data.pdf. > [24/6/15]

Hogg, R. V., McKean, J. W., dan Craig, A. T. 2005. *Introduction to Mathematical Statistics*. New Jersey : Prentice Hall. Sixth Edition.

Ilin, A. dan Raiko, T. 2010. Pratical Approach to Principal Component

Analysis in the Presence of Missing Values. *Journal of Machine Learning Research*, 11, 1957-2000.

Izenman, A. J. 2008. *Modern Multivariate Statistical Technique Regression Classification and Manifold Learning*. New York : Springer Text in Statistics.

Jaya, I G. N. M. 2010. *Modul Komputasi Statistik dengan Software R.* Jurusan Statistika Fakultas MIPA Unpad. Edisi Kedua.

Jolliffe, I. T. 2002. *Principal Component Analysis.* New York : Springer-Verlag, 2nd Edition.

Little, R. J. A. 1998. A Test of Missing Completely at Random for Multivariate Data With Missing Value. *Journal of The American Statistical Association*, Vol. 83, No.404, 1198-1202. American Statistical Association.

Little, R. J. A., dan Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. Hoboken, NJ: Jhon Wiley & Sons.

Luttinen, J. dan Illin, A. 2009. Transformation for variational factor analysis to speed up learning. *Neurocomputing.*

Rubin, D. B. 1976. Inference and Missing Data. *Bioinformatics*, Vol. 63, No.3, 581-592. Great Britain : Biometrika Trust.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. dan Ishii, S. 2003. A Bayesian Missing Value Estimation Method for Gene Expression Profile Data. *Bioinformatics*, Vol. 19, No.16, 2088-2096. Oxford University Press.

Oba, S., Sato, M., dan Ishii, S. 2003. Prior Hyperparameters in Bayesian PCA. *Joint International Conference ICANN/ICONIP*, LNCS 2714, 271-279. Berlin : Springer-Verlag.

Scheffer, J. 2002. Dealing with Missing Data. *Res. Lett. Inf. Math. Sci*, Vol. 3, 153-160.

Schlomer, G. L., Bauman, S. dan Card, N. A. 2010. Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology* Vol. 57, No. 1, 1–10.

Sharma, S. 1996. *Applied Multivariate Techniques*. New York : Jhon Wiley & Sons Inc.

Stacklies, W. dan Redestig, H. 2016. The pcaMethods Package. Melalui <https://www.bioconductor.org/packages/3.3/bioc/manuals/pcaMethods/man/pcaMethods.pdf > [13/5/16]

Stacklies, W., Redestig, H., Scholz, M., Walther, D. dan Selbig, J. 2007. pcaMethods -a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, Vol. 23 No. 9, 1164-1167.

Takane, Y. dan Takane, Y. O. 2003. Relationships between Two Methos for Dealing with Missing Data in Principal Component Analysis. *Behaviormetrika*, 30, 145-154.

Widiastuti, S. dkk. 2003. Analisis Komponen Utama. *Makalah Metode Penelitian dan Telaah Pustaka*. IPB.

Tabel 3 Korelasi AKU Data Lengkap dengan Simulasi Kondisi Data Tidak Lengkap dengan Berbagai Metode Estimasi pada Hasil Skor Observasi dan Loading Variabel

Perbandingan Korelasi	Kondisi Simulasi Data Tidak Lengkap	Komponen Utama																											
		Ke-1				Ke-2				Ke-3				Ke-4				Ke-5				Ke-6				Ke-7			
		Metode				Metode				Metode				Metode				Metode				Metode				Metode			
		I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Skor Observasi	Data Tidak Lengkap 5 %		0.935	-0.988	-0.983		0.855	0.982	-0.913		0.808	0.953	0.856		-0.701	-0.763	0.754		-0.674	0.388	-0.706		-0.818	0.483	0.824		0.792	-0.637	0.763
	Data Tidak Lengkap 15 %		0.877	0.953	-0.936		0.761	-0.979	-0.846		0.621	0.921	-0.684		0.671	-0.820	-0.716		-0.485	-0.537	0.530		-0.586	0.549	-0.655		0.692	0.580	-0.687
	Data Tidak Lengkap 40 %		0.836	0.871	-0.910		-0.689	0.841	0.800		0.526	0.747	0.390		0.246	-0.209	0.295		-0.018	0.012	-0.183		-0.040	0.249	0.240		0.123	0.305	0.187
	Data Tidak Lengkap 60 %		0.462	0.054	-0.362		0.314	0.109	0.171		-0.216	0.385	-0.063		0.312	-0.610	-0.292		0.004	-0.099	-0.153		-0.015	0.368	0.101		-0.068	-0.136	
Loading Variabel	Data Tidak Lengkap 5 %	0.918	0.953	-0.968	-0.970	-0.824	0.851	0.957	0.893	0.741	0.739	0.940	0.838	-0.751	-0.687	-0.754	0.752	-0.524	-0.689	0.383	-0.705	-0.729	-0.886	0.548	0.899	0.073	0.763	-0.618	0.745
	Data Tidak Lengkap 15 %	0.914	0.929	0.954	-0.952	-0.854	0.812	-0.945	-0.845	-0.740	0.514	0.888	-0.611	-0.427	0.643	-0.800	-0.684	-0.311	-0.656	-0.753	0.669	-0.722	-0.702	0.684	-0.782	-0.016	0.659	0.556	-0.640
	Data Tidak Lengkap 40 %	-0.170	0.819	0.658	-0.821	-0.151	-0.698	0.672	0.817	-0.370	0.324	0.607	0.168	0.204	0.159	-0.223	0.281	0.416	-0.247	-0.127	-0.335	-0.154	-0.043	0.282	0.262	0.279	-0.019	0.144	0.369
	Data Tidak Lengkap 60 %	0.389	0.593	0.364	-0.515	0.098	0.218	-0.075	0.087		-0.199	0.476	0.011		0.335	-0.683	-0.381		-0.107	0.005	-0.191		-0.008	0.553	0.157		-0.024	-0.308	

catt:
Metode:

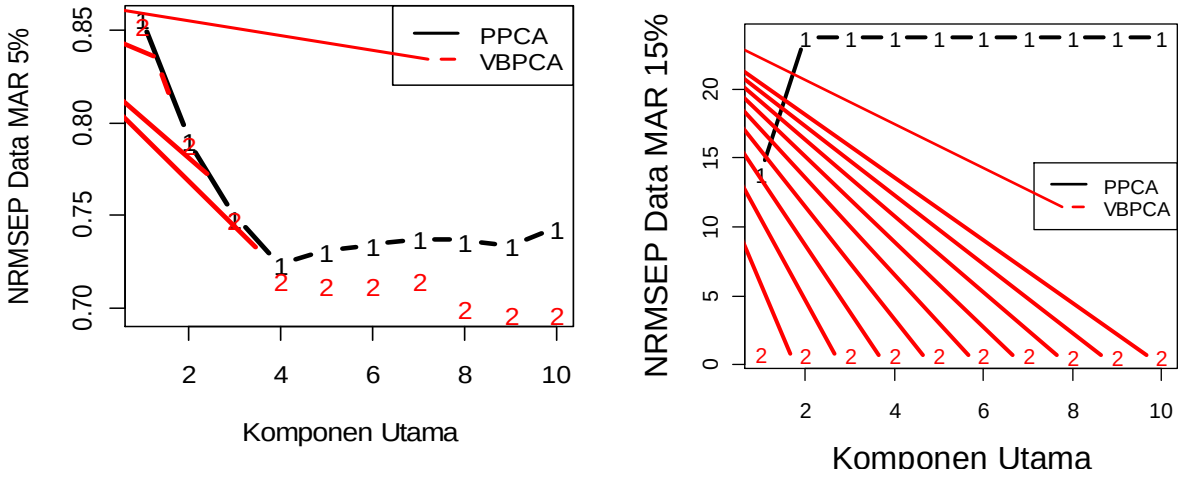
I	Listwise Deletion
II	Imputasi Rata-Rata
III	PPCA
III	VBPCA

Tabel 4 Korelasi AKU Data Lengkap dengan AKU Data Lengkap Hasil Metode Estimasi VBPCA pada Hasil Skor Observasi dan Loading Variabel

Perbandingan Korelasi	Kondisi Simulasi Data Tidak Lengkap	Komponen Utama						
		Ke-1	Ke-2	Ke-3	Ke-4	Ke-5	Ke-6	Ke-7
		$VBPCA_b$	$VBPCA_b$	$VBPCA_b$	$VBPCA_b$	$VBPCA_b$	$VBPCA_b$	$VBPCA_b$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Skor Observasi	Data Tidak Lengkap 5 %	0.9887	-0.9604	0.9218	-0.9595	-0.5679	-0.5947	0.7039
	Data Tidak Lengkap 15 %	0.9541	-0.9440	0.9035	-0.8952	-0.5989	-0.6861	0.6638
	Data Tidak Lengkap 40 %	0.9114	-0.8977	0.5769	0.2155	0.0309	0.1402	-0.1848
	Data Tidak Lengkap 60 %	0.1608	0.0818	-0.4092	0.5418	-0.0575	-0.1081	0.1178
Loading Variabel	Data Tidak Lengkap 5 %	0.9733	-0.9352	0.9135	-0.963	-0.562	-0.6626	0.7069
	Data Tidak Lengkap 15 %	0.9529	-0.9184	0.8758	-0.8683	-0.7652	-0.8254	0.6370
	Data Tidak Lengkap 40 %	0.8024	-0.8213	0.4387	0.1784	0.1099	0.1127	-0.0283
	Data Tidak Lengkap 60 %	0.356	0.0433	-0.5928	0.5755	-0.0802	-0.1721	0.3083

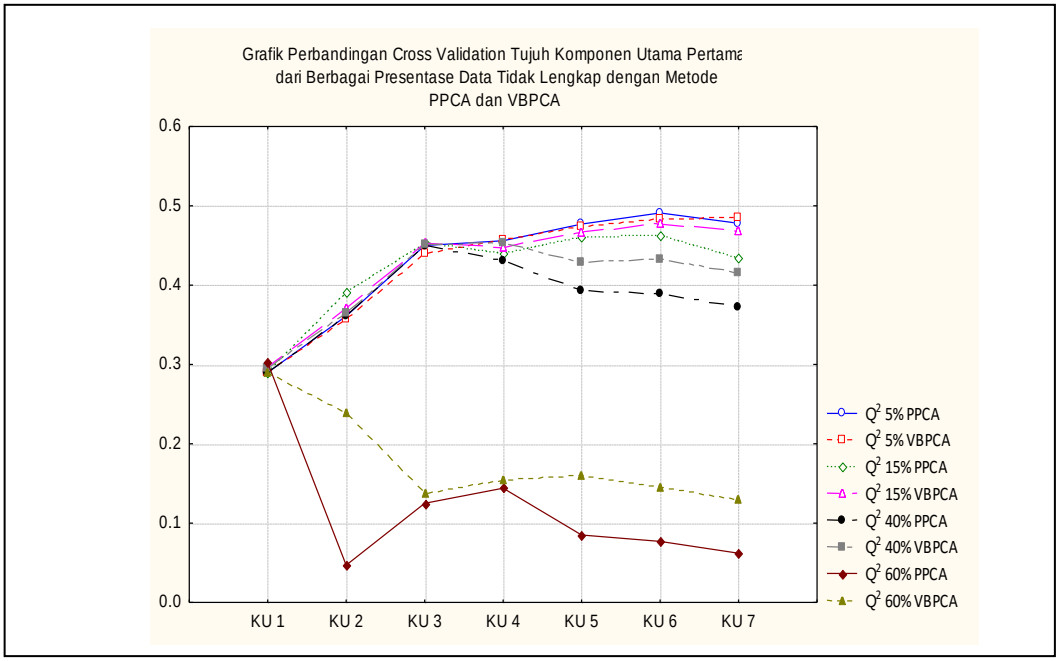
Tabel 5 Perbandingan Hasil Metode Analsis Komponen Utama Saat Disimulasikan Beberapa Presentase Kondisi Data Tidak Lengkap

Kondisi	Perbandingan Nilai	Listwise Deletion	Imputasi Rata-Rata	PPCA	VBPCA	$VBPCA_b$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Data Tidak Lengkap 5%	Variasi yang dijelaskan dari 7 Komponen Utama Pertama	73.96	70.03	79.95	76.35	79.75
	Cross Validation (Q^2) 7 Komponen Utama Pertama			0.478	0.486	
	NRMSEP 7 Komponen Utama Pertama			0.738	0.715	
Data Tidak Lengkap 15%	Variasi yang dijelaskan dari 7 Komponen Utama Pertama	74.72	70.03	80.18	76.98	80.67
	Cross Validation (Q^2) 7 Komponen Utama Pertama			0.434	0.458	
	NRMSEP 7 Komponen Utama Pertama			23.776	0.725	
Data Tidak Lengkap 40%	Variasi yang dijelaskan dari 7 Komponen Utama Pertama	93.56	62.28	83.74	69.72	82.79
	Cross Validation (Q^2) 7 Komponen Utama Pertama			0.372	0.416	
	NRMSEP 7 Komponen Utama Pertama			0.898	0.831	
Data Tidak Lengkap 60%	Variasi yang dijelaskan dari 7 Komponen Utama Pertama		61.07	84.06	60.09	81.05
	Cross Validation (Q^2) 7 Komponen Utama Pertama			0.062	0.130	
	NRMSEP 7 Komponen Utama Pertama			1.024	0.885	



(a) (b)

Gambar 5 Grafik Output NRMSEP pada (a) Data Tidak Lengkap 5% dan (b) Data Tidak Lengkap 15%.



Gambar 6. Grafik Perbandingan Nilai Cross Validation (Q^2) pada Berbagai Data Tidak Lengkap dengan Metode PPCA dan VBPCA

Lampiran 1

Uji Little's MCAR

a. Data 5 % tidak lengkap

Hipotesis :

H_0: Pola data tidak lengkap 5% mengikuti mekanisme MCAR

H_1: Pola data tidak lengkap 5% tidak mengikuti mekanisme MCAR.

Hipotesis awal menyatakan pola data tidak lengkap 5% mengikuti mekanisme MCAR, sedangkan hipotesis alternatif berarti pola data tidak lengkap 5% tersebut tidak mengikuti pola MCAR.

Tingkat signifikansi :

α = 5 %

Statistik Uji :

$$\chi^2_{\text{hitung}} = \sum_{j=1}^{J} m_j\left(\hat{\mu}_j - \hat{\mu}_j^{(\text{ML})}\right)^{\text{T}} \hat{\Sigma}_{\text{j}}^{-1}\left(\hat{\mu}_j - \hat{\mu}_j^{(\text{ML})}\right)$$

dengan :

m_j = banyaknya baris dalam pola data tidak lengkap ke - j

$\hat{\mu}_j$ = rata-rata variabel dalam pola data tidak lengkap ke - j

$\hat{\mu}_j^{(\text{ML})}$ = rata-rata variabel hasil estimasi kemungkinan maksimum

$\hat{\Sigma}_j$ = matriks kovarian hasil estimasi kemungkinan maksimum

J = banyaknya pola data tidak lengkap, dengan j = 1, . . J

Daerah Penolakan :

$\chi^2_{\text{hitung}} > \chi^2_{(0.05;df)}$ atau p-value < α

dengan :

$df = \sum d_j - d$

d_j = banyak variabel yang datanya lengkap pada pola ke – j

Hipotesis awal (H_0) akan ditolak jika $\chi^2_{\text{hitung}} > \chi^2_{(0.05;df)}$ dengan $\chi^2_{(0.05;df)} = \chi^2_{(0.05;169)}$= 200,333909, atau nilai bila p-value kurang dari α. Sebaliknya, hipotesis awal akan tidak ditolak apabila p-value melebihi α.

Interprestasi Hasil tes Little's MCAR

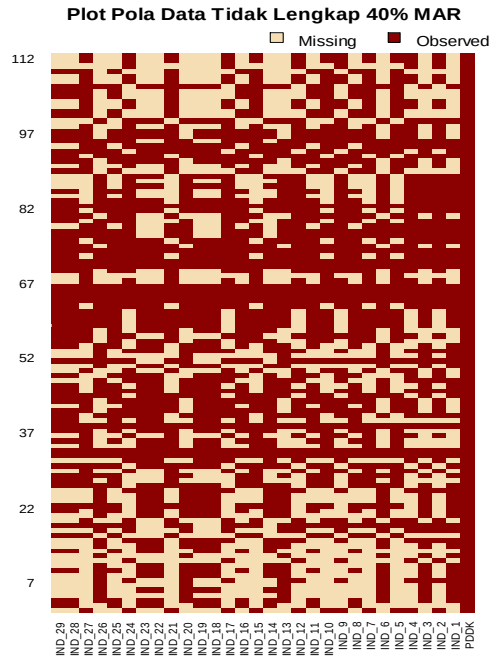
Pada gambar di bawah ini menunjukkan p-value hasil uji Little's MCAR Test bernilai 0.000 < α atau hasil χ^2_{hitung}= 355,737 > 200,333909, maka dilakukan tolak hipotesis awal dan dinyatakan bahwa pola data tidak lengkap 5% tidak mengikuti MCAR.

IND_28	144815.277	4.91170	8.72906	-.17250	.83124	-6.86969
IND_29	-373781.224	-10.76730	.33239	18.72325	-1.17107	-5.91119

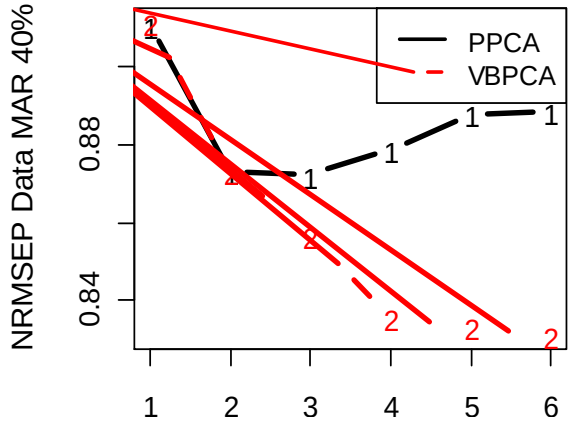
a. Little's MCAR test: Chi-Square = 355.737, DF = 169, Sig. = .000

Lampiran 2

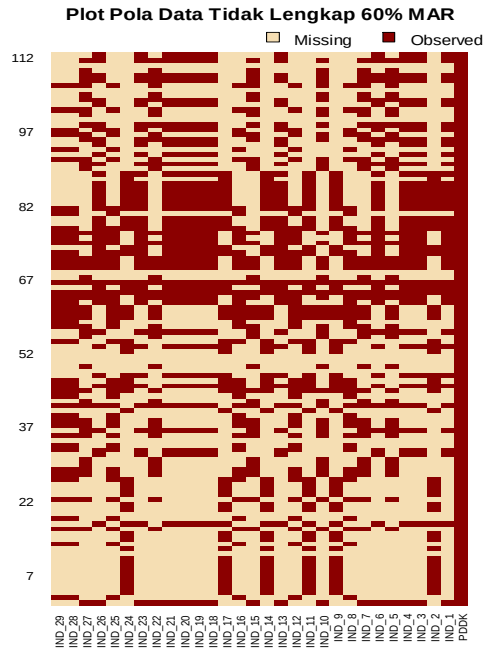
Plot Pola Data untuk Data 40% dan 60% Tidak Lengkap



Lampiran 3



Grafik hasil Output NRMSEP dengan metode PPCA dan VBPCA pada data tidak lengkap 40%



Lampiran 4

Uji signifikansi korelasi skor observasi hasil AKU data lengkap dengan skor observasi pada data tidak lengkap 5% MAR

Hipotesis :

H_0 : $\rho = 0$; tidak terdapat hubungan signifikan antara skor observasi data lengkap dengan skor observasi saat disimulasikan data 5% tidak lengkap.

H_1 : $\rho \neq 0$; terdapat hubungan signifikan antara skor observasi data lengkap dengan skor observasi saat disimulasikan data 5% tidak lengkap.

Dilakukan uji dua pihak karena hanya ingin diketahui apakah ada hubungan signifikan antara skor observasi data lengkap dengan skor observasi saat disimulasikan data 5% tidak lengkap.

Tingkat signifikansi :

α = 5 %

Statistik Uji :

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

dengan :

r = korelasi yang terjadi antara kedua variabel yang akan diuji.

Daerah Penolakan :

$t_{\text{hitung}} > t_{(0.05;df)}$ atau p-value < α

dengan :

df = $n-2$

Hipotesis awal (H_0) akan ditolak bila p-value kurang dari α. Sebaliknya, hipotesis awal akan tidak ditolak apabila p-value melebihi α.

Interprestasi Hasil Uji Signifikansi Korelasi

Pada hasil olah yang ditampilkan dalam gambar 4.5 menunjukkan p-value hasil uji bernilai 0.000 < α untuk semua metode yang digunakan saat data tidak lengkap 5%, maka dilakukan tolak hipotesis awal dan dinyatakan bahwa korelasi antara skor observasi data lengkap dengan saat simulasi data tidak lengkap 5% terjadi korelasi signifikan. Kecuali metode listwise tidak dilakukan hipotesis, karena skor observasi yang dihasilkan tidak mencakup keseluruhan observasi yang ada.