PRISMA 2 (2019): 52-57
PRISMA, PROSIDING SEMINAR NASIONAL MATEMATIKA
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Metode *Robust Principle Component Analysis* (RPCA) dengan Algoritme Proyeksi dan Matriks Ragam Peragam

Ratna Nur Mustika Sanusi[a,*], Dewi Retno Sari Saputro[b]

[a,b]*Program Studi Matematika FMIPA Universitas Sebelas Maret, Jl. Ir. Sutami No. 36A, Surakarta, Indonesia*

**Alamat Surel: mustikaratna68@student.uns.ac.id*

Abstrak

Metode *principle component analysis* (PCA) merupakan analisis *multivariate* yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel tersebut sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya. Selainnya PCA dapat mengatasi multikolinearitas dengan vektor eigen. Namun metode tersebut sangat sensitif terhadap pencilan sehingga akan menghasilkan penduga parameter berbias. Kelemahan pada metode PCA dapat diatasi dengan metode yang merupakan kombinasi dari konsep *projection pursuit* dengan penduga kovarian *minimum covariance determinant* yaitu RPCA. Kehadiran pencilan mampu diatasi oleh RPCA karena vektor ciri pada komponen utama tidak terpengaruh oleh adanya pencilan. Akar ciri komponen utama pertama tidak bertambah seiring dengan penambahan proporsi pencilan sehingga proporsi keragaman kumulatif data yang mampu direpresentasikan oleh komponen utama pertama cenderung stabil. Pada penelitian ini dilakukan kajian tentang metode RPCA yang *robust* terhadap pencilan dengan algoritme proyeksi dan matriks ragam peragam.

Kata kunci:

Principle component analysis, robust principle component analysis, pencilan, algoritme proyeksi, matriks ragam peragam.

1. Pendahuluan

Principal component analysis (PCA) merupakan teknik analisis statistik untuk mentransformasi variabel yang saling berkorelasi satu dengan yang lain menjadi variabel baru yang tidak berkorelasi lagi atau dapat dikatakan mereduksi dimensi data. PCA menyelesaikan hasil dari amatan data apabila data tersebut terdiri atas variabel-variabel yang jumlahnya sangat banyak sehingga diperoleh variabel-variabel baru yang jumlahnya lebih sedikit namun tetap mampu menjelaskan variansi data (Johnson dan Wichern, 2007). Perkembangan PCA dimulai sejak 1901 yang diperkenalkan pertama kali oleh Pearson. PCA didasarkan asumsi bahwa data tidak mengandung pencilan (*outlier)* sehingga perkembangan PCA selanjutnya dipengaruhi adanya kebutuhan suatu model PCA yang *robust* terhadap data pencilan *(outlier). Classical principal component analysis* sangat dipengaruhi oleh kehadiran pencilan *(outlier)* karena CPCA dibentuk berdasarkan pada matriks kovarian yang sensitif terhadap keberadaan data *outlier* (Hubert *et al.*, 2005). Oleh karena itu, diperkenalkan suatu konsep *robust* PCA atau RPCA.

Masih menurut RPCA adalah suatu metode yang kekar (*robust*) untuk PCA terhadap keberadaan *outlier* pada data (Hubert *et al*, 2004). Metode RPCA menggabungkan konsep *Projection Pursuit* (PP) dengan penaksir *robust Minimum covariance determinant* (MCD). Keunggulan RPCA yaitu dapat menyelesaikan masalah multikolinearitas dan pencilan dengan membuat komponen utama yang baru sebagai variabel bebas yang selanjutnya diregresikan dengan variabel terikat yang menghasilkan model yang *robust* terhadap pencilan.

To cite this article:
Sanusi, R.N.M. & Saputro, D.R.S. (2020). Metode Robust Principle Component Analysis (RPCA) dengan Algoritme Proyeksi dan Matriks Ragam Peragam. *PRISMA, Prosiding Seminar Nasional Matematika 3*, 52-57

Pada penelitian ini dilakukan kajian tentang cara mengidentifikasi dan mengatasi multikolinearitas, serta mengatasi masalah pencilan menggunakan metode *robust principle component analysis* (RPCA) dengan penduga *minimum covariance determinant* (MCD) dan teknik *projection pursuit* (PP). Estimator MCD merupakan salah satu metode estimator *robust* paling popular untuk mengestimasi multivariat (Eric, 2009). Teknik PP merupakan metode reduksi dimensi yang lebih baik dibandingkan dengan metode yang lain karena PP membentuk komponen utama yang saling bebas dan memiliki korelasi yang tinggi dengan variabel terikat. Pembentukan komponen utama menggunakan matriks ragam peragam merupakan metode yang baik karena dapat menghasilkan varian dengan keragaman terbesar dari variabel terikat. Kecocokan RPCA pun dikaji guna mendapatkan metode yang tepat untuk mengatasi kedua masalah tersebut. Oleh karena itu, penelitian ini dilakukan dengan tujuan untuk menyelesaikan masalah analisis yang mengandung pencilan dengan metode RPCA yang *robust* terhadap pencilan dengan algoritme proyeksi dan matriks ragam peragam.

2. Pembahasan

Robust merupakan metode yang pada awalnya dipublikasikan oleh Andrews (1972) yang selanjutnya dikembangkan oleh Ryan (1997), sebagai metode yang digunakan saat terdapat data *outlier* pada suatu amatan. Hingga saat ini belum ada pembaharuan metode tersebut sehingga *robust* dianggap kompatibel dalam mengatasi masalah kehadiran *outlier* yang dapat menghasilkan model yang *robust* terhadap *outlier*. *Outlier* dapat terjadi pada variabel bebas dan variabel terikat. Jika variabel terikat terdapat *outlier* disebut *outlier* univariat. Sebaliknya, jika variabel bebas terdapat *outlier* disebut *outlier* multivariat. Baik o*utlier* univariat maupun multivariat, keduanya merupakan masalah dalam penelitian.

2.1. Principle Component Analysis (PCA)

PCA merupakan salah satu analisis multivariat yang bermanfaat untuk mereduksi dimensi data. Secara umum metode PCA merupakan metode analisis multivariat yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel tersebut sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya (Garthwaite, 2002). Sering ditemukan data yang berukuran besar atau data yang terdiri atas banyak variabel dan diharapkan hasil analisis terhadap data tersebut menghasilkan banyak informasi yang diperlukan. Akibat ukuran data yang besar, analisis data sulit dilakukan karena masalah dimensi data dan terdeteksinya data *outlier*. Dengan demikian, analisis komponen utama robust memegang dua peranan yaitu untuk mereduksi dimensi data dan menghasilkan RPCA yaitu komponen utama yang tidak terpengaruh terlalu banyak dengan keberadaan *outlier*.

Pada metode PCA, misalkan vektor variabel asal adalah $\boldsymbol{X_1, X_2, \dots, X_p}$ adalah variabel random yang mempunyai distribusi tertentu dengan vektor rataan $\boldsymbol{\mu}$ **dan** matriks varian-kovarian $\boldsymbol{\Sigma}$. *Principle component* (PC) merupakan kombinasi linier terboboti dari variabel-variabel asal yang mampu menjelaskan data secara optimal. PC ke-j dari p variabel dituliskan sebagai

$$Y_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{pj}x_p = \boldsymbol{a_j'X}$$

Dan variansi PC ke-j adalah $Var(Y_j) = \lambda_j, j = 1, 2, \dots, p$ dengan $\lambda_1, \lambda_2, \dots, \lambda_p$ adalah akar ciri yang diperoleh dari persamaan $|\sum -\lambda I| = 0$ dengan $\lambda_1 \geq \lambda_2 \geq \ \dots \geq \ \lambda_p \geq 0$. Vektor ciri **a** sebagai pembobot dari transformasi linear variabel asal yang diperoleh dari $\left(\sum -\lambda_j I\right)a_j = 0$.

Metode PCA didasarkan pada matriks kovarian sebagai indeks proyeksi yang unsur-unsurnya berupa variansi dan kovarian dari sekumpulan variabel. Berikut merupakan struktur matriks kovarian.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1j} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{pj} \end{bmatrix}$$

Struktur matriks kovarian berdasarkan sebagian kecil komponen yang tidak saling berkorelasi, sedemikian sehingga mempunyai variansi yang maksimal. Oleh karena itu, matriks kovarian sensitif terhadap kehadiran pencilan sehingga menghasilkan penduga parameter bias. Komponen yang terdapat pada matriks kovarian dapat diperoleh dengan rumus berikut

$$\sigma_{pj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)(x_{ij} - \bar{x}_j)$$

Metode reduksi dimensi yang digunakan dalam penelitian ini adalah metode projection pursuit (PP) yang telah diperkenalkan oleh Li dan Chen (1985) karena metode ini menggunakan penduga kovarian yang kekar.

Kehadiran pencilan dapat diatasi dengan menambahkan estimasi kovarian yang robust terhadap pencilan. Matriks kovarian yang robust dapat diestimasi dengan M-estimator (Devlin, dkk, 1975), minimum covariance determinant (Rousseeuw, 1984), atau S-estimator (Croux dan Haesbroek, 1999). Ketiga metode ini baik jika digunakan untuk kasus jumlah variabel p > jumlah observasi n. Pada penelitian ini digunakan estimator MCD untuk mengatasi kehadiran pencilan. Kelemahan pada metode PCA dapat diatasi dengan metode yang merupakan kombinasi dari konsep projection pursuit dengan penduga kovarian minimum covariance determinant adalah RPCA (Hubert *et al*, 2004).

2.2. Projection Pursuit (PP)

PP adalah adalah metode pereduksian dimensi yang bersifat *underdispersived* berdasarkan pencarian suatu proyeksi informasi utama data berdimensi besar. Metode reduksi dimensi dengan konsep PP dikembangkan oleh Li dan Chen pada tahun 1985. Metode PP bertujuan untuk mendapatkan struktur pada data peubah ganda dengan memproyeksikannya pada subruang berdimensi lebih rendah. *Projection-pursuit* tepat digunakan untuk menganalisis data dengan jumlah peubah yang besar. Subruang berdimensi rendah dipilih dengan memaksimumkan indeks proyeksi tertentu (Hubert, 1985). Metode ini mencari suatu arah dengan penyebaran maksimal data diproyeksikan di dalamnya. PP ini merupakan pengembangan dari metode PCA yang menggunakan penduga ragam sebagai indeks proyeksi. Metode PCA dikembangkan menjadi metode PP berdasarkan kebutuhan untuk menghasilkan komponen utama yang kekar.

Pengembangan metode tersebut terdapat pada perubahan penduga ragam sebagai indeks proyeksi yang diganti dengan penduga kovarian yang *robust* V. Rousseeuw dan Croux (1993) mengasumsikan terdapat n pengamatan pada p variabel $x_1, \dots, x_n \in \mathbb{R}^p$ yang direpresentasikan dalam matriks $\boldsymbol{X}$, vektor ciri ke-*k* dituliskan sebagai

$$a_k = argmax\ V\ (a^t x_1, \dots, a^t x_n)$$

dengan $\|a\| = 1$ dan $\boldsymbol{V}$ sebagai penduga matriks varians-kovarian yang kekar atau indeks proyeksi. Akar ciri ke-*k* didefinisikan sebagai

$$\lambda_k = V_n^2(\ (a_k^t x_1, \dots, a_k^t x_n)$$

skor komponen utama selanjutnya sebagai proyeksi dari kovarian kekar dapat diambil dari dekomposisi *spectral.* Penduga matriks varian-kovariansi *robust* **V** ditulis sebagai

$$C_V = \sum_{k=1}^{p} \lambda_k\, \boldsymbol{a}_k \boldsymbol{a}_k^t$$

Croux C *et al.* (2007) menjelaskan bahwa algoritma grid bertujuan untuk mencari dimensi yang memiliki kemungkinan arah maksimal dari vektor ciri dengan lebih teliti. Ide dasar dalam mencari arah maksimal dari vektor ciri adalah iterasi berulang untuk optimasi di dalam ruang dua dimensi. Matriks $\boldsymbol{X}$ dalam keadaan telah terurut sehingga $S_j = S(\boldsymbol{x_{1j}}, \dots, \boldsymbol{x_{nj}})$ untuk $j = 1, \dots, p$, sehingga $S(\boldsymbol{e_1}) \geq S(\boldsymbol{e_2}) \dots \geq S(\boldsymbol{e_p})$ dengan $\boldsymbol{e_1}, \dots, \boldsymbol{e_p}$ merupakan vektor basis kanonik.

- Dimulai dengan $\widehat{\boldsymbol{a}} = \boldsymbol{e_1}$
- Untuk $i = 1, \dots, N_g$
  Selanjutnya untuk $j = 1, \dots, p$
  a. Maksimumkan fungsi tujuan pada dimensi yang mencakup $\boldsymbol{e_j}$ dan $\widehat{\boldsymbol{a}}$ menggunakan pencarian grid dari fungsi $\theta \to S(\cos\ \theta\ \widehat{\boldsymbol{a}} + \sin\theta\ \boldsymbol{e_j})$ dengan sudut θ berada dalam interval $\left[\frac{-\pi}{(2^i)}, \frac{\pi}{(2^i)}\right]$. Sudut θ_0 merupakan sudut maksimum yang dicapai untuk seluruh titik-titik grid.

b. Perbarui $\widehat{\boldsymbol{a}} \rightarrow \cos\ \theta_0\ \widehat{\boldsymbol{a}} + \sin\theta_0\ \boldsymbol{e_j}$)

Arah pertama $\widehat{\boldsymbol{a}}$ merupakan salah satu variabel yang memiliki disperse terbesar. Selama satu iterasi, dilakukan urutan pencarian grid lebih dari $p-1$ dimensi, koordinat j akan memperbarui koordinat j sebelumya dari $\widehat{\boldsymbol{a}}$. Ketika iterasi kedua dimulai, titik $\widehat{\boldsymbol{a}}$ sudah berada pada arah yang benar, tapi masih perlu penyempurnaan dalam solusi lokal. Oleh karena itu, pencarian grid tidak akan mencari di seluruh dimensi, tetapi hanya selama dimensi berada pada interval $\left[\frac{-\pi}{4}, \frac{\pi}{4}\right]$. Setelah setiap iterasi dilakukan, pencarian terbatas pada interval yang lebih sempit dari sudutnya, namun nilai grid untuk setiap interval N_g akan tetap konstan pada setiap iterasi. Sehingga beberapa iterasi pertama memungkinkan untuk menemukan wilayah dengan dimensi yang terdapat nilai maksimum dan siklus berikutnya bertujuan untuk meningkatkan presisi dan solusi yang dihasilkan.

2.3. Robust Principle Component Analysis (RPCA)

RPCA diperkenalkan oleh Li dan Chen, yang dapat digunakan untuk banyak variabel maupun amatan. Selanjutnya pada tahun 1996, Croux dan Ruiz–Gazen memperkenalkan PP. Algoritme RPCA bekerja dengan baik, namun pada dimensi yang tinggi, algoritme berjalan lambat dalam waktu komputasi dan tidak stabil secara numerik (Croux and Ruiz, 2005). Pada tahun 2004, Hubert *et al.* (2005) menggabungkan konsep PP dan estimator kovarian yang robust dengan nama RPCA. Metode RPCA dapat diterapkan pada data amatan yang mempunyai jumlah variabel (p) lebih kecil dari jumlah observasi (n) atau suatu data dengan jumlah variabel (p) lebih besar dari jumlah observasi (n). Ketika jumlah variabel (p) lebih kecil dari jumlah observasi (n), maka dilakukan dengan Dekomposisi nilai singular.

Pada metode RPCA terdapat tahapan untuk memproses suatu data multivariat. Diasumsikan data direpresentasikan dengan sebuah vektor random **X** yang terdiri atas elemen baris menyatakan n observasi dan banyaknya kolom menyatakan q variabel $X_1, X_2, \ldots, X_q$.

Proses pertama adalah mereduksi dimensi sedemikian sehingga menjadi kurang dari *n*-1. Proses selanjutnya, menentukan seberapa banyak pencilan pada setiap amatan dengan rumus yang dituliskan sebagai

$$Otl(x_i) = max\frac{|\boldsymbol{a}^t x_i - \hat{\mu}(a^t \boldsymbol{X})|}{\mathbf{S}(a^t \boldsymbol{X})}$$

dengan $\hat{\mu}$ dan $\widehat{\sum}$ adalah penduga rata-rata dan simpangan baku *robust* MCD. Kemudian dibentuk matriks kovarian awal (S_0). Matriks kovarian digunakan untuk menyeleksi jumlah komponen *k*, dari persamaan (Σ-λ_jI)$a_j = 0$ dapat diperoleh nilai λ_j sehingga jumlah komponen *k* akan diperoleh. Kontribusi dari masing-masing komponen utama ke-*j* terhadap total varian dapat diperoleh dari

$\frac{\lambda_j}{tr(\Sigma)} = \frac{\lambda_j}{\lambda_1+\lambda_2+\cdots+\lambda_p}$ dengan j = 1,2,…,p

Proses terakhir adalah memproyeksikan titik-titik data pada subruang yang lokasi dan *scatter matrix*-nya diestimasi secara *robust* dari nilai eigen k positif $l_1, l_2, \ldots, l_k$. dengan $k < m$ dan $m = p + q$ sedemikian sehingga diperoleh penduga pusat *robust* $\hat{\mu}_z$ dari $z_{n,m} = (X_{n,p,} Y_{n,q})$ dan ragamnya S_z.

2.4. Penduga Minimum Covariance Determinant (MCD)

Pendugaan menggunakan MCD menghasilkan akar ciri dan vektor ciri yang *robust*. Menurut Rousseeuw dan Van Driessen (1999), langkah-langkah untuk menghitung penduga median dan ragam peragam MCD sehingga menghasilkan determinan matriks ragam peragam terkecil yang diuraikan sebagai berikut.

- Diambil sejumlah h pengamatan yang berbeda secara acak. Dari n pengamatan akan dihasilkan C_h^n subcontoh dengan banyak elemen

$$h = \frac{n+p+1}{2}$$

dengan $\frac{n+p+1}{2} \leq h \leq n$ atau $h = 0{,}75n$,

- Diambil sejumlah himpunan bagian dari data secara acak. Himpunan bagian tersebut berukuran $(p+1)$ dan diperbesar hingga mencapai h menggunakan C-*step.*
- Dalam setiap h bagian yang terambil, dilakukan dua C-*step.* Tahapan C-*step* adalah melakukan perhitungan penduga median $\hat{\mu}_0$ dan penduga matriks ragam peragam $\hat{\Sigma}_0$ dari h pengamatan, kemudian menghitung jarak dari tiap titik sebagai berikut:

$$d_{(\hat{\mu}_0,\hat{\Sigma}_0)}(x_i) = \sqrt{(x_i - \hat{\mu}_0)'\hat{\Sigma}_0{}^{-1}(x_i - \hat{\mu}_0)}$$

himpunan h baru dibentuk dari h pengamatan dengan jarak pengamatan terkecil.

- Untuk 10 himpunan bagian h dengan determinan matriks ragam peragam terkecil, dilakukan C-*step* hingga konvergen dan himpunan bagian terakhir disimpan dalam H_1.
- Penduga median $\hat{\mu}_{MCD}$ dan matriks ragam peragam $\hat{\Sigma}_{MCD}$ diperoleh dari H_1 dengan determinan matriks ragam peragam terkecil, selanjutnya, dilakukan tahap pembobotan. Jarak *robust* merupakan suatu pendekatan untuk mendeteksi *outlier* pada data multivariate. *Outlier* yang berapa pada variabel independen dinamakan *outlier laverage. Outlier* ini dapat terdeteksi dengan menggunakan jarak *robust* dengan rumus

$$RD_i = \sqrt{(x_i - \hat{\mu}_{MCD})'\hat{\Sigma}_{MCD}{}^{-1}(x_i - MCD)}$$

dengan RD_i *outlier laverage* dapat diperoleh sehingga

- Komponen utama didefinisikan sebagai *k* vektor ciri dan S_1 yang bersesuaian dengan *k* akar ciri terbesar dari S_1.

3. Simpulan

Berdasarkan pembahasan disimpulkan bahwa adanya multikolinearitas akan menghambat proses analisis dan menghilangkan suatu *outlier* bukanlah suatu keputusan yang tepat. Metode RPCA dapat digunakan untuk mengatasi *outlier* multivariat. Algoritme yang digunakan memenuhi syarat untuk mengatasi masalah tersebut. Algoritme RPCA bekerja dengan baik, namun pada dimensi yang tinggi, algoritme berjalan lambat dalam waktu komputasi dan tidak stabil secara oleh karenanya, perlu digabungkan konsep PP dan estimator kovarian yang robust. Metode RPCA dapat diterapkan pada data amatan yang mempunyai jumlah variabel (p) lebih kecil dari jumlah observasi (n) atau suatu data dengan jumlah variabel (p) lebih besar dari jumlah observasi (n). Pada saat jumlah variabel (p) lebih kecil dari jumlah observasi (n), dapat dilakukan dengan dekomposisi nilai singular.

Daftar Pustaka

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, *28*, 125-136.

Croux, C., & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, *71*(2), 161-190.

Croux, C., Ruiz-Gazen, A. (2005). High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited, *Journal of Multivariate Analysis*. Leuven, Belgium: Department of Applied Economics.

Croux, C., Filzmoser, P., & Oliveira, M. R. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *87*(2), 218-225.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3), 531-545.

Eric, A. Cator., & Hendrik, P. Lopuhaa. (2009). Asymptotic expansion of the minimum covariance determinant estimators. *Journal of Multivariate Analysisi*, *101*(2010), 2372-2388.

G. Li, Z, Chen. (1985). *Journal of the American Statistical Association. 80*, 759-766.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193-218.

Hubert. M., Rousseeuw, Peter J., & Branden, Karlien V. (2004). ROBCA: Anew Approach to Robust Principal Component Analysis. *Technometrics.* Feb 2005, *47*, No. 1. 64-67.

Hubert, M., Rousseeuw, P. J., & Branden, V. K. (2005). *A Comparison of Three Procedures for Robust PCA in High Dimensions.* Leuven, Belgium: Katholieke Universiteit.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivarite Statistical Analysis, Sixth Edition.* New Jersey: United States of America.

Garthwaite, P. H., Jolliffe, I. T., Jolliffe, I. T., & Jones, B. (2002). *Statistical inference*. Oxford University Press on Demand.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, *88*(424), 1273-1283.

Rousseeuw, P.J. (1984). Least Median Aquares regression. *Journal of the American Statistical Association*. *79*, 388.

Rousseuw PJ, Van Driessen K. (1999). A fast algorithm for the minimum covariance determinanr estimator. Technometrics. *41*:212-213.