TESIS

PELABELAN *OUTLIERS* PADA DATA MULTIVARIAT MENGGUNAKAN METODE *MINIMUM VECTOR VARIANCE* DENGAN KRITERIA DEPTH DAN MAHALANOBIS

Disusun dan diajukan oleh

PUJI PUSPA SARI
H062192006

PROGRAM STUDI MAGISTER STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2021

PELABELAN *OUTLIERS* PADA DATA MULTIVARIAT MENGGUNAKAN METODE *MINIMUM VECTOR VARIANCE* DENGAN KRITERIA DEPTH DAN MAHALANOBIS

TESIS

Sebagai Salah Satu Syarat Untuk Mencapai Gelar Magister

PROGRAM STUDI

MAGISTER STATISTIKA

Disusun dan diajukan oleh:

PUJI PUSPA SARI
H062192006

Kepada

PROGRAM STUDI MAGISTER STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2021

HALAMAN PENGESAHAN

PELABELAN *OUTLIERS* PADA DATA MULTIVARIAT MENGGUNAKAN METODE *MINIMUM VECTOR VARIANCE* DENGAN KRITERIA DEPTH DAN MAHALANOBIS

Disusun dan diajukan oleh:

PUJI PUSPA SARI
H062192006

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Program Studi Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin pada tanggal 6 Juli 2021
Dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping
Dr. Erna Tri Herdiani, S.Si, M.Si	**Dr. Nurtiti Sunusi, S.Si, M.Si**
NIP. 19750429 200003 2001	NIP. 19720117 199703 2002
Ketua Program Studi Magister Statistika,	Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam,
Dr. Dr. Georgina Maria Tinungki, M.Si	**Dr. Eng. Amiruddin, S.Si, M.Si**
NIP. 19620926 198702 2 001	NIP. 19720515 199702 1 002

LEMBAR PERNYATAAN KEASLIAN PENELITIAN

Yang bertanda tangan dibawah ini :

Nama	: Puji Puspa Sari
NIM	: H062192006
Program Studi	: Magister Statistika
Jenjang	: S2

Menyatakan dengan ini bahwa karya tulis saya yang berjudul :

PELABELAN *OUTLIERS* PADA DATA MULTIVARIAT MENGGUNAKAN METODE *MINIMUM VECTOR VARIANCE* DENGAN KRITERIA DEPTH DAN MAHALANOBIS

Adalah karya tulis saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain. Bahwa Tesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila dikemudian hari terbukti atau dapat dibuktikan sebagian atau keseluruhan Tesis ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 6 Juli 2021

Yang Menyatakan,

Puji Puspa Sari

KATA PENGANTAR

Puji syukur atas kehadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menuyusun dan menyelesaikan tesis ini. Penulis menyadari sepenuhnya bahwa apa yang dikemukakan dalam tesis ini masih jauh dari kesempurnaan yang merupakan sebagai akibat dari keterbatasan kemampuan serta berbagai kesulitan yang penulis hadapi dalam penyusunan tesis ini.

Keberhasilan penulis dalam menyusun tesis ini tidak lepas dari kemudahan yang di berikan oleh Allah SWT dan bantuan dari berbagai pihakyang bersifat moril maupun materil. Oleh karena itu, penulis memanjatkan doa kepada Tuhan Yang Maha Esa agar memberikan rahmat-Nya kepada pihak yang banyak membantu dalam penyelesaian tesis ini. Penulis juga percaya bahwa tesis ini dapat selesai bukan hanya dengan kekuatan pikiran penulis semata akan tetapi karena bantuan dari berbagai pihak juga, baik selama proses perkuliahan bahkan sampai proses pengerjaan tesis di Program Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Namun demikian, penulis dengan senang hati menerima kritik dan saran yang bersifat membangun dari pembaca karya tulis ini demi sempurnanya tesis ini.

Terima kasih yang tak terhingga kepada kedua orang tua tercinta dan kedua adik-adikku atas doa yang tak pernah putus, dukungan serta segala kebaikan mereka yang sampai kapan pun takkan pernah bisa terbalaskan atas kasih saying yang tiada henti dalam penyelesaian tesis ini. Selanjutnya saya ingin menyampaikan juga rasa hormat dan terima kasih kepada :

1. **Ibu Prof. Dr. Dwia Aries Tina Palubuhu, MA**. selaku Rektor Universitas Hasanuddin
2. **Bapak Dr. Eng. Amiruddin** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam beserta seluruh jajarannya.
3. **Ibu Dr. Nurtiti Sanusi, S.Si., M.Si**. selaku Ketua Departemen Statistika, Pembimbing Pertama dan Penasehat Akademik penulis yang senantiasa

memberikan ilmu, dukungan, dan motivasi serta kemudahan kepada penulis dalam berbagai hal selama menjalani pendidikan di Departemen Statistika.

4. **Ibu Dr**. **Dr. Georgina Maria Tinungki, M.Si** selaku ketua Program Studi Magister Statistika Universitas Hasanuddin yang juga menjadi salah satu tim penguji tesis.
5. **Ibu Dr. Erna Tri Herdiani S.Si,M.Si** selaku Pembimbing Utama yang telah bersabar dan bersedia meluangkan banyak waktunya untuk membimbing penulis dan memberikan masukan dalam penyelesaian tesis ini.
6. **Ibu Dr. Anna Islamiyati S.Si.,M.Si** dan **Bapak Dr. La Podje Talangko, M.S** selaku penguji penulis yang telah bersedia memberikan masukan-masukan dan arahan dalam penyusunan tesis ini.
7. Bapak dan Ibu Dosen di Departemen Statistika Fakultas MIPA Universitas Hasanuddin, yang dengan tulus ikhlas memberikan ilmu pengetahuan dan pengalaman yang dimilikinya selama perkuliahan berlangsung sehingga memberikan banyak manfaat bagi penulis untuk saat ini maupun di masa mendatang.
8. Kepada **Edy Abdurrahman Syahrir, S.PWK., M.URP** terima kasih untuk doa, dukungan, kebersamaan, dan banyak motivasi yang diberikan pada penulis.
9. Teman-teman Mahasiswa Program Magister Statistika Angkatan 1 terima kasih atas nasehat dan dukungan luar biasa kepada penulis.
10. Sahabat-sahabatku di Asrama Ramsis Unhas **Waode Sitti Purnama Sari, Syam Nurmila H, Orin Widiarti, Monika Sroyer, Werlin Tabuni, Christina Yuanita Kafiar, dan Perawati** terima kasih banyak untuk semuanya, semoga silaturahmi kita tetap berlanjut setelah melewati jenjang pendidikan.

Akhir kata, penulis menyampaikan terima kasih kepada semua pihak yeng telah banyak membantu, semoga Tuhan Yang Maha Esa memberikan balasan yang berlipat ganda. Penulis juga berharap semoga tesis ini memberikan manfaat bagi pengembangan ilmu pengetahuan dan menambah referensi di bidang Statistika serta berguna bagi masyarakat.

ABSTRAK

Pencilan adalah data yang memiliki karakteristik unik yang terlihat sangat jauh berbeda dari observasi yang lain. Pencilan dapat mengakibatkan penyimpangan terhadap hasil analisis dan interpretasi data. Sehingga dibutuhkan metode dalam melabel pencilan menggunakan pendekatan statistik. Salah satu metode dalam pelabelan pencilan adalah algoritma *Minimum Vector Variance* (MVV). MVV merupakan metode *robust* karena memiliki beberapa keunggulan yaitu nilai *breakdown point* yang tinggi, bersifat *affine equavariance*, dan efisiensi komputasi dengan tingkat kompleksifitas yang rendah. Dalam penelitian ini dilakukan modifikasi pada algoritma MVV menggunakan Depth dan Mahalanobis pada pengurutan data multivariat. Dari modifikasi tersebut diperoleh 3 metode yakni Mahalanobis MVV (MMVV), Depth Mahalanobis MVV (DMMVV), dan *Robust* Depth Mahalanobis MVV (RDMMVV). Tujuan dari metode ini adalah melabel pencilan menggunakan ketiga metode tersebut. Pelabelan pencilan pada ketiga metode diterapkan pada data simulasi yang berdistribusi normal multivariat. Hasil penelitian menunjukkan ketiga metode mendeteksi observasi yang sama sebagai pencilan yaitu pengamatan ke 8, 20, 39, 42, 51, 66, 72, 85, 99, dan 117. Dalam penelitian ini metode RDMMVV lebih efisien jika didasarkan pada waktu perhitungan mengeksekusi program dibandingkan dengan MMVV dan DMMVV.

Kata Kunci : Depth, Mahalanobis, MVV, Pencilan, *Robust*

ABSTRACT

Outliers are data that have unique characteristics that look very much different from other observations. Outliers may result in deviations from the results of data analysis and interpretation. So, we need a method in labeling outliers using a statistical approach. One of the outlier labeling methods is the Minimum Vector Variance (MVV) algorithm. MVV is a robust method because it has several advantages, namely a high breakdown point value, affin equivariance, and computational efficiency with a low level of complexity. In this study, the MVV algorithm was modified using Depth and Mahalanobis in multivariate data sorting. From these modifications, three methods were obtained, namely Mahalanobis MVV (MMVV), Depth Mahalanobis MVV (DMMVV), and Robust Depth Mahalanobis MVV (RDMMVV). The purpose of this method is to label outliers using all three methods. Outlier labeling in the three methods was applied to simulation data with multivariate normal distribution. The results showed that the three methods of detecting the same observations with outliers were observations 8, 20, 39, 42, 51, 66, 72, 85, 99, and 117. In this study, the RDMMVV method is more efficient if it is based on the calculation of program execution time than the MMVV and DMMVV methods.

Keywords: Depth, Mahalanobis, MVV, Outliers, Robust

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I PENDAHULUAN

1. 1 Latar Belakang

Deteksi pencilan atau *outlier* pada data multivariat menjadi salah satu bahasan yang menarik dan menantang bagi para peneliti. Hal ini karena deteksi pencilan untuk data multivariat membutuhkan penyelesaian yang lebih kompleks dibanding data univariat. Pelabelan *outlier* pada pengamatan univariat telah dilakukan oleh Hawkins (1980) yang mendeteksi suatu *outlier* tunggal pada sampel dengan melihat pengamatan yang menyimpang dari pengamatan lain. *Outlier* pada univariat juga dapat ditemukan pada saat melihat distribusi nilai dalam ruang dimensi tunggal (Endang, 2019). Selain itu, Manoj dan Kannan (2013) mengatakan bahwa pada kasus univariat cukup mudah untuk menemukan *outlier* yakni dengan cara menvisualisasikan data menggunakan *scatterplot* atau *boxplot*. Menurut Kelly M *et al* (2019) mengidentifikasi pencilan pada data multivariat tidak cukup baik jika hanya menggunakan metode pendekatan univariat, karena *outlier* pada multivariat berada dalam ruang *n*-dimensi. Pencilan atau *outlier* didefinisikan sebagai suatu data yang menyimpang dari sekumpulan data yang lain (Ferguson, 1961). Sedangkan, menurut Barnett (1984) pencilan adalah pengamatan yang tidak mengikuti sebagian besar pola data dan terletak jauh dari pusat data. Peneliti lain mendefinisikan pencilan sebagai data yang memiliki karakteristik unik yang terlihat sangat jauh berbeda dari observasi yang lain (Hair *et al*,1995). Johnson dan Wichern (1996) mendefinisikan pencilan adalah data dalam satu karakteristik yang tidak biasa, yakni relatif paling kecil atau paling besar dibandingkan dengan yang lain. Konsep dasar dalam pendeteksian *outlier* pada data multivariat adalah dengan mengukur jarak setiap titik ke pusat datanya (Erna, 2017).

Pada umumnya, *outlier* dapat terjadi karena kesalahan manusia, kesalahan instrumen, perilaku curang, perubahan perilaku sistem atau kesalahan sistem, dan penyimpangan alami di dalam populasi. Keberadaan pencilan pada data dapat mengakibatkan penyimpangan terhadap hasil analisis data seperti penyimpangan terhadap hasil uji statistik berdasarkan parameter rata-rata dan kovarians. Adanya *outlier* berpengaruh pada estimasi nilai parameter yang bersifat bias, sehingga

menyebabkan interpretasi hasil yang diperoleh menjadi tidak akurat. Akan tetapi pada saat tertentu, *outlier* tidak dapat dihapus karena mengandung informasi penting yang tidak dapat diberikan oleh observasi lain (Hadi, 1992). Oleh karena itu, perlu dilakukan identifikasi terhadap keberadaannya.

Ada beberapa metode yang digunakan untuk mendeteksi atau melabeli pencilan diantaranya adalah Jarak Mahalanobis yang umum digunakan pada data multivariat dengan menggunakan vektor rata-rata dan matriks varians kovarians (Rousseeuw & Van Zomeren, 1990), metode *Minimum Volume Ellipsoid* (MVE) yang didasarkan pada estimator volume terkecil *ellipsoid* yang mencakup h dari n pengamatan. Penaksir lokasi MVE merupakan pusat *ellipsoid* dengan penaksir scatter berkoresponden menjadi matriks bentukan (*shape matrix*) (Aelst dan Rousseeuw, 2009), metode *Minimum Covariance Determinan* (MCD) digunakan untuk mendeteksi pencilan berdasarkan nilai determinan matriks varians-covarians. Prinsip metode MCD adalah mendapatkan subhimpunan dari keseluruhan pengamatan yang matriks varians-kovariansnya memiliki determinan terkecil diantara semua kombinasi kemungkinan data. Pendeteksian outlier dengan metode MCD dilakukan berdasarkan jarak robust dan nilai *cut-offnya* (Butler *et al*, 1993)), algoritma *Fast Minimum Covariance Determinant* (FMCD) merupakan pengembangan dari metode MCD, ide dasarnya menggunakan teknik MCD dengan menggunakan iterasi yang selektif (*selective iteration*). Untuk data yang kecil MCD dapat ditentukan dengan eksak oleh FMCD. Sedangkan, untuk data yang besar FMCD akan memberikan kesimpulan yang sangat akurat pada data yang memuat h atau lebih observasi (Rousseeuw dan Van Driessen, 1999), dan metode *Minimum Vector Variance* (MVV) merupakan modifikasi algoritma FMCD dengan menggunakan ukuran *Vector Variance* (VV) yang minimum (Herwindiati, 2006).

MVV merupakan metode *robust* terhadap pencilan dibandingkan dengan metode lainnya yang telah disebutkan sebelumnya, karena memiliki beberapa keunggulan yaitu nilai *breakdown point* yang tinggi, bersifat *affine equavariance*, dan efisiensi komputasi dengan tingkat kompleksifitas yang rendah. Kriteria pengurutan metode MVV menggunakan jarak Mahalanobis menggunakan vektor mean sampel dan

matriks variansi kovariansi sampel (Herwindiati, 2006). Salah satu keunggulan MVV dibanding FMCD adalah metode MVV dapat digunakan pada data dengan determinan sama dengan nol karena tidak terpengaruh oleh nilai varians/kovarians yang sama dengan nol (Faruq, 2008). Beberapa peneliti yang menggunakan MVV diantaranya adalah Ali *et al* (2014) Menggunakan MVV dengan pembobot *reweighted*, menghasilkan skema *reweighted* yang mampu mempertahankan *breakdown point* 0,5 dan mencapai efisiensi yang lebih tinggi pada distribusi normal, Herwindiati (2009) menerapkan *Robust Principal Componen Analisys* untuk mereduksi data yang telah dianalisis menggunakan metode MVV, Yahaya *et al* (2011) menggunakan T^2 Hotteling pada MVV dan menghasilkan penelitian yang menunjukkan bahwa peta kendali MVV memiliki kinerja kompetitif relatif terhadap diagram kontrol, Erna *et al* (2019) menggunakan metode MVV dengan mengkontaminasi data dengan pencilan untuk melihat keakuratan metode MVV.

Kriteria jarak Mahalanobis merupakan salah satu metode untuk mengukur atau menghitung jarak tiap observasi terhadap pusat datanya (Mahalanobis, 1936). Sedangkan, kriteria *depth* adalah gagasan tentang pengukuran jarak berdasarkan kedalaman data yang dikemukakan oleh Tukey (1975) sebagai alat grafis untuk memvisualisasikan kumpulan data bivariat dan sejak itu diperluas ke kasus multivariat (Donoho dan Gasko, 1992). Kedalaman titik relatif terhadap kumpulan data mengukur seberapa dalam titik itu berada di dalam sebaran data. Konsep *Depth* dalam data merupakan kriteria baru dalam mengurutkan data multivariat. Djauhari (2003) dalam tulisannya telah menyatakan bahwa kuadrat jarak mahalanobis dapat dinyatakan bukan dalam bentuk invers matriks varians kovarians, tetapi dalam bentuk kriteria baru menggunakan *Depth*. Menurut Ye & Chen (2001) perhitungan komputasi menggunakan jarak Mahalanobis akan mengalami kendala jika jumlah variabel mencapai ratusan. Penelitian lainnya mengatakan bahwa jarak Mahalanobis dapat dimodifikasi menggunakan *statistical depth function* menjadi *robust depth* Mahalanobis yang dapat memberikan hasil *affine invarian* lebih baik daripada jarak Mahalanobis (Djauhari & Umbara, 2006). Selain itu, penelitian terdahulu yang menerapkan fungsi *depth* pada jarak Mahalanobis diantaranya adalah Zuo dan Serfling

(2000) menghasilkan sifat-sifat yang lebih baik dibanding Mahalanobis, sedangkan Suwanda (2019) menerapkan fungsi *depth* pada bagan kendali dengan menggunakan kriteria *Vector Variansi*.

Berdasarkan beberapa penelitian tersebut, peneliti tertarik untuk menggunakan kriteria *Depth* dan Mahalanobis pada metode *Minimum Vector Variance* untuk pelabelan pencilan pada data multivariat.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian yang telah dipaparkan, maka rumusan masalah pada penelitian ini adalah bagaimana pelabelan pencilan menggunakan Jarak Mahalanobis, *Depth* Mahalanobis, dan *Robust Depth* Mahalanobis pada metode *Minimum Vector Variance* dalam data multivariat?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, maka tujuan penulisan penelitian ini adalah memperoleh hasil pelabelan pencilan menggunakan Jarak Mahalanobis, *Depth* Mahalanobis, dan *Robust Depth* Mahalanobis pada metode *Minimum Vector Variance* dalam data multivariat.

1.4 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberi pemahaman tentang penerapan modifikasi kriteria pengurutan data pada algoritma *Minimum Vector Variance* (MVV) menggunakan Jarak Mahalanobis, *Depth* Mahalanobis, dan *Robust Depth* Mahalanobis untuk melabeli pencilan pada data multivariat.

1.5 Batasan Masalah

Agar tidak menimbulkan penafsiran yang lebih luas, beberapa asumsi diberikan sebagai batasan masalah dari penelitian ini yaitu data yang digunakan merupakan data simulasi multivariat dan berdistribusi normal.

BAB II TINJAUAN PUSTAKA

Pada bagian ini dijelaskan teori pendukung yang akan digunakan pada bagian hasil dan pembahasan. Sebagian besar teori yang digunakan pada penelitian ini dikutip dari buku karangan Johnson, R.A. & Wichern, D. W tahun 2007 tentang *Applied Multivariate Statistical Analysis 6th edition* serta jurnal-jurnal pendukung yang menyangkut penelitian ini.

2.1 Pencilan (*Outlier*)

Pencilan atau *Outlier* menurut (Hawkins, 1980) didefinisikan sebagai pengamatan yang sangat menyimpang dari pengamatan lainnya. Pencilan pada data dapat menyebabkan bias pada penduga rata-rata dan kovarian data multivariat (Wang dkk, 2015). Johnson dan Wichern (2007) mendefinisikan pencilan adalah data dalam satu karakteristik yang tidak biasanya, dia relatif paling kecil atau paling besar dibandingkan dengan yang lain. Pencilan juga memberi efek *masking* (mengaburkan data) dan *swamping* (kesalahan mengidentifikasi data non *outlier* sebagai *outlier*). Kehadiran *outlier* pada data dapat disebabkan karena kesalahan prosedural seperti kesalahan dalam entri atau kesalahan dalam pengkodean. Oleh karena itu, perlu dilakukan identifikasi terhadap keberadaannya (Noeryanti, 2001).

2.2 Data Multivariat

Data multivariat adalah data yang dikumpulkan dari dua atau lebih observasi dengan mengukur observasi tersebut dengan beberapa karakteristik. Data multivariat dapat dituliskan dalam bentuk matriks, sehingga disebut matriks data multivariat. Dalam analisis multivariat sering kali dihadapkan pada masalah pengamatan yang dilakukan pada suatu periode waktu untuk $p > 1$ variabel atau karakter. Akan digunakan notasi x_{ij} yang mendefinisikan objek ke-i pada variabel ke-j. Menurut Johnson dan Wichern (2007: 5), secara umum sampel data multivariat dapat disajikan dalam bentuk sebagai berikut:

	Var-1	Var-2	⋯	Var- j	⋯	Var- p
Objek-1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1p}
Objek-2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2p}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
Objek- i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{ip}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
Objek- n	x_{n1}	x_{n2}	⋯	x_{nj}	⋯	x_{np}

Atau dapat ditulis dalam bentuk matriks $\boldsymbol{X}$ sebagai berikut

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Keterangan:

x_{ij} = adalah objek ke- i pada variabel ke- j

n = adalah banyaknya item atau objek

p = adalah banyaknya variabel

Dapat juga dinotasikan dengan $X = x_{ij}; i1,2, \dots, n \; dan \; j = 1,2, \dots, p$

2.3 Distribusi Normal Multivariat

Distribusi normal multivariat merupakan perluasan dari fungsi distribusi normal univariat untuk $p \geq 2$. Vektor random yang terdiri atas p kompenen $X = (X_1, X_2, \dots, X_p)^t$ dikatakan berdistribusi normal dengan vektor mean μ dan matriks variansi-kovariansi Σ, jika fungsi kepadatan peluang bersama $X_1, X_2, \dots, X_p$ adalah Johnson dan Wichern (2007):

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp^{\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\}} \quad (1)$$

Selanjutnya vektor random X yang berdistribusi normal p -variat tersebut diberi lambang $X \sim N_p(\mu, \Sigma)$.

2.4 Determinan Matriks

Determinan suatu matriks adalah suatu fungsi skalar dengan domain matriks bujur sangkar. Dengan kata lain, determinan merupakan pemetaan dengan domain berupa matriks bujur sangkar, sementara kodomain berupa suatu nilai skalar. Tiap matriks bujur sangkar $p \times p$, $A = [a_{ij}]$ ditetapkan memiliki skalar khusus yang disebut determinan dari A yang dinotasikan

$$\det(A) \text{ atau } |A| \text{ atau } \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{vmatrix}$$

Susunan skalar $p \times p$ yang dibatasi oleh dua garis lurus yang disebut determinan berorde p (Lip Schutz dan Lipson, 2002). Determinan matriks A berukuran $p \times p$ dapat dihitung dengan mengalikan entri pada suatu baris atau kolom dengan masing-masing kofaktor dan menjumlahkan hasil perkalian tersebut.

2.5 Vektor Rata-Rata

Vektor rata-rata untuk data populasi dapat ditulis dalam bentuk matriks sebagai berikut (Johnson& Winchen, 2007):

$$\text{Mean} = E(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_P) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_P \end{bmatrix} = \boldsymbol{\mu} \qquad (2)$$

Vektor rata-rata sampel sebagai berikut:

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{1i} \\ \frac{1}{n}\sum_{i=1}^{n} x_{2i} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{pi} \end{pmatrix} \qquad (3)$$

Sehingga, mean sampel adalah $\bar{X} = [\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_p]$

2.6 Matriks Variansi Kovariansi

Matriks variansi kovariansi merupakan suatu matriks simetris yang berisi variansi pada diagonal utamanya dan kovariansi pada elemen lainnya. Koefisien variansi menggambarkan sebuah indeks dari hubungan linear antara dua variabel. Matriks varians kovarians data populasi dapat dituliskan sebagai berikut (Johnson& Winchen, 2007):

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t$$

$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_P - \mu_P \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, .., X_P - \mu_P]\right)$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \dots & (X_1 - \mu_1)(X_P - \mu_P) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \dots & (X_2 - \mu_2)(X_P - \mu_P) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \dots & (X_P - \mu_P)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \dots & E(X_1 - \mu_1)(X_P - \mu_P) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \dots & E(X_2 - \mu_2)(X_P - \mu_P) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p - \mu_p)(X_1 - \mu_1) & E(X_p - \mu_p)(X_2 - \mu_2) & \dots & E(X_P - \mu_P)^2 \end{bmatrix}$$

Sehingga, matriks untuk varians kovarians pada data populasi didefinisikan sebagai berikut:

$$\boldsymbol{\Sigma}_{\text{p}\times\text{p}} = \text{cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1\text{p}} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2\text{p}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\text{p}1} & \sigma_{\text{p}2} & \dots & \sigma_{\text{pp}} \end{bmatrix} \quad (4)$$

dengan matriks varians kovarians untuk data sampel sebagai berikut:

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t$$

$$= \frac{1}{n-1}((x_i - \bar{x})(x_1 - \bar{x})^t + (x_2 - \bar{x})(x_2 - \bar{x})^t + \cdots + (x_n - \bar{x})(x_n - \bar{x})^t)$$

$$= \frac{1}{n-1}\left[\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{p1} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]\left[\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{p1} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]^t + \left[\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{p2} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]\left[\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{p2} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]^t + \ldots + \left[\begin{pmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{pn} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]\left[\begin{pmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{pn} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right]^t$$

$$= \frac{1}{n-1}\begin{pmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \\ \vdots \\ x_{p1} - \bar{x}_p \end{pmatrix}\left(x_{11} - \bar{x}_1 x_{21} - \bar{x}_2 \ \ldots \ x_{p1} - \bar{x}_p\right) + \begin{pmatrix} x_{12} - \bar{x}_1 \\ x_{22} - \bar{x}_2 \\ \vdots \\ x_{p2} - \bar{x}_p \end{pmatrix}\left(x_{12} - \bar{x}_1 x_{22} - \bar{x}_2 \ \ldots \ x_{p2} - \bar{x}_p\right) + \ldots + \begin{pmatrix} x_{1n} - \bar{x}_1 \\ x_{2n} - \bar{x}_2 \\ \vdots \\ x_{pn} - \bar{x}_p \end{pmatrix}\left(x_{1n} - \bar{x}_1 x_{2n} - \bar{x}_2 \ \ldots \ x_{pn} - \bar{x}_p\right)$$

Jadi, matriks untuk varians kovarians data sampel adalah sebagai berikut:

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} \tag{5}$$

2.7 Vektor Variansi

Menurut Herwindiati (2007) ada dua ukuran penyebaran pada multivariat, yaitu Total Variansi (TV) dan General Variansi (GV). Misalkan $\boldsymbol{X}$ adalah sebuah vektor random dengan matriks variansi kovariansi (Σ), maka TV dan GV dari Σ dapat didefinisikan:

$$TV = Tr(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \tag{6}$$

$$GV = \det(\Sigma) = |\Sigma| = \sigma_{11}.\sigma_{22}.\ldots.\sigma_{pp} \tag{7}$$

Berdasarkan definisi tersebut TV hanya melibatkan variansi tanpa melibatkan struktur kovariansinya. TV merupakan *trace* dari diagonal utama yang dijumlahkan. Sedangkan, GV adalah ukuran yang melibatkan variansi dan kovariansinya sehingga GV dapat ditemukan dalam aplikasi yang lebih luas. GV merupakan determinan yang

berasal dari diagonal utama yang dikalikan. Meskipun GV ataupun CD (*Covariance Determinant*) mempunyai aplikasi yang lebih luas dibandingkan dengan TV, namun bukan berarti GV tidak mempunyai kelemahan. Kelemahan dari GV adalah ketika nilai CD = 0 atau matriks variansi kovariansi singular, maka GV atau CD tidak dapat digunakan untuk pemecahan multivariat. Oleh karena itu, masalah ini dapat diselesaikan dengan Total Variansi.

Misalkan X dan Y adalah dua vektor random sembarang dengan dimensi berhingga yang mempunyai matriks kovariansi bersama $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, dimana Σ_{11} dan Σ_{22} berturut-turut adalah matriks variansi untuk masing-masing X dan Y. Sedangkan, Σ_{12} dan Σ_{21} adalah matriks kovariansi antara X dan Y. Lazraq dan Clereoux (1989) mendefinisikan ukuran korelasi untuk dua vektor random X dan Y adalah $\rho(X,Y) = \frac{Tr(\Sigma_{12}^2,\Sigma_{21}^2)}{\sqrt{Tr(\Sigma_{11}^2)Tr(\Sigma_{22}^2)}}$.

Berdasarkan definisi tersebut, Herwindiati (2007) menggunakan $Tr(\Sigma_{11}^2)$ dan $Tr(\Sigma_{22}^2)$ berturut–turut sebagai ukuran untuk variansi dari vektor random $\boldsymbol{X}$ dan $\boldsymbol{Y}$, yang selanjutnya disebut Vektor Variansi (VV). Jadi, secara umum jika vektor random $\boldsymbol{X}$ mempunyai matriks variansi kovariansi maka VV dari $\boldsymbol{X}$ adalah $Tr(\Sigma_{11}^2)$ yang merupakan ukuran sebaran multivariat disekitar μ. Djauhari (2005) menyatakan bahwa kelebihan dari vektor variansi adalah:

1. Mampu mengukur Multivariat dispersi walaupun matriks varian kovariannya singular.
2. Proses komputasinya sangat efisien karena hanya menggunakan jumlahan kuadrat dari setiap elemen diagonal utama matriks variansi kovariannya

2.8 Statistik *Robust*

Estimasi klasik sangat bergantung pada asumsi-asumsi dasar seperti normalitas, linearitas dan independensi yang terkadang tidak terpenuhi. *Robust* merupakan sifat kebal (tahan) terhadap adanya data yang tidak seperti biasa/pencilan (Erna, 2019).

Adapun tujuan dari estimator *robust* adalah untuk mengidentifikasi adanya pencilan (Franklin dan Brouduer, 2005).

Robust merupakan alat yang dapat digunakan untuk menganalisa data yang mengandung pencilan dan memberikan hasil yang resisten terhadap adanya pencilan (Ali, 2013). Menurut Johnson & Winchen (2002:1), terdapat 3 kelas masalah yang dapat ditangani dengan menggunakan teknik robust, yaitu:

1) Masalah outlier yang terdapat pada variabel dependen
2) Masalah outlier yang terdapat pada variabel independen
3) Masalah outlier yang terdapat pada keduanya, yaitu variabel dependen dan variabel independen.

Untuk mengatasi masalah pencilan ini, Rousseuw dan Van Driessen memperkenalkan metode FMCD (*Fast Minimum Covariance Determinant*) untuk mendeteksi *outlier* berdasarkan nilai determinan matriks varians-covarians yang minimum. Namun, metode FMCD mempunyai kelemahan ketika nilai determinan matriks varians-covarians sama dengan nol. Pada tahun 2006, Herwindiati memodifikasi algoritma FMCD menjadi lebih efektif dan tingkat kompleksitas yang lebih rendah dengan menggunakan ukuran *vector variance* (VV) yang minimum yang selanjutnya disebut dengan Minimum *Vector Variance* (MVV). Metode MVV terinspirasi oleh algoritma *C-Steps* pada FMCD. Alasan yang mendasar menggunakan metode MVV dalam mendeteksi pencilan adalah karena metode ini *robust* (tegar) terhadap pencilan.

2.9 Jarak Mahalanobis

Jarak Mahalanobis merupakan generalisiasi dari Jarak Eucled yang distandarisasi, selain itu jarak ini dapat mengatasi masalah perbedaan skala dalam data dan mempertimbangakan korelasi antar peubah. Pada metode ini juga didefinisikan sebagai jarak dua titik yang melibatkan kovarians atau korelasi antar peubah (Boni, 2018).

Jarak mahalanobis merupakan metode untuk mendeteksi pencilan pada data multivariat. Jarak mahalanobis diperoleh dengan menghitung jarak tiap observasi

terhadap pusat datanya. Kuadrat jarak mahalanobis dihitung dengan rumusan (Mahalanobis, 1936) yang didefinisikan dengan persamaan (8):

$$d_i^2 = (x_i - \overline{\boldsymbol{x}})\, {}^t\boldsymbol{S^{-1}}(x_i - \overline{\boldsymbol{x}}) > \chi^2_{k(1-\alpha)} \quad (8)$$

dimana

d_i^2 = kuadrat jarak pengamatan ke- i

x_i = nilai pengamatan ke- i

$\bar{x}$ = vektor rata-rata dari pengamatan

S^{-1}= invers matriks variansi kovariansi sampel, dengan:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \overline{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}, dan\ \boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Langkah-langkah mendeteksi pencilan dengan jarak Mahalanobis (Johnson dan Wichern, 2007):

1. Menentukan nilai vektor rata-rata ($\bar{x}$)
2. Menentukan nilai matriks varians kovarians ($\boldsymbol{S}$)
3. Menentukan nilai jarak Mahalanobis pada setiap pengamatan dengan vektor rata-rata: $d_i^2 = (x_i - \overline{\boldsymbol{x}})\,'\boldsymbol{S^{-1}}(x_i - \overline{\boldsymbol{x}})\,, i = 1,2, \dots, p$
4. Mengurutkan nilai d_i^2 dari kecil ke besar $d_1^2 \leq d_2^2 \leq \cdots \leq d_n^2$

Adapun jarak Mahalanobis dievaluasi dengan menggunakan χ^2 pada derajat kebebasan (df=k=p-1) sejumlah variabel yang digunakan dalam penelitian. Identifikasi data pencilan pada pengamatan ke- i didefinisikan sebagai pencilan apabila $d_i^2 > \chi^2_{k(1-\alpha)}$.

2.10 Depth Mahalanobis

Fungsi Depth merupakan gagasan baru yang dikembangkan secara intensif dalam dekade terakhir di bidang statistik non parametrik, geometri komputasi, aljabar, dan ilmu komputer. Ini terkait erat dengan pengurutan data multivariat, estimasi *robust*, dan deteksi outlier. Salah satu yang paling banyak digunakan dalam statistik dan bidang terkait adalah Mahalanobis *Depth*. Secara teoritis, fungsi depth tersebut dibangun agar

memiliki fungsi yang lebih baik dan efisien secara komputasi. Praktis, peran fungsi kedalaman dalam aplikasi semakin luas dan luas. Dalam penelitian ini, versi sampel kedalaman mahalanobis didefinisikan sebagai berikut (Zuo dan Serfling, 2000):

Misal $X_1, X_2, \dots, X_n$ adalah sampel acak dari distribusi p-variate. Vektor rata-rata sampel dan matriks variansi kovariansi sampelnya adalah:

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i\in n} x_i \tag{9}$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i\in n}(x_i - \overline{\boldsymbol{x}})(\, x_i - \overline{\boldsymbol{x}})^t \tag{10}$$

Versi sampel untuk Mahalanobis *Depth* dari X_i dapat ditulis sebagai berikut (Liu *et al*, 1999):

$$MD_i = \frac{1}{1+(\, x_i - \overline{\boldsymbol{x}})^t \boldsymbol{S}^{-1}(\, x_i - \overline{\boldsymbol{x}})} \tag{11}$$

Persamaan tersebut digunakan untuk mengukur seberapa dalam x terhadap sampel acak $x_1, x_2, \dots, x_n$. Semakin besar nilai MD_i semakin dekat titik x ke pusat $\overline{\boldsymbol{x}}$ dengan $0 > MD_i > 1$.

2.11 Robust Depth Mahalanobis

Robust Depth Mahalanobis merupakan pengembangan dari Depth Mahalanobis. Dalam literatur, misalnya Hadi (1992), Liu et al. (1999), Rousseeuw dan van Driessen (1999), Werner (2003), dan Herwindiati et al. (2006), persamaan jarak dihitung berdasarkan definisinya. Jadi, untuk jarak mahalanobis dan depth mahalanobis dibutuhkan inversi matriks kovarian sampel S untuk menghitungnya. Ini adalah perhitungan yang memakan waktu terutama untuk kumpulan data berdimensi tinggi. Kompleksitas komputasinya dalam hal jumlah operasi dalam algoritmanya termasuk tinggi. Djauhari (2007) mendefinisikan ulang *depth* Mahalanobis dengan memperkenalkan fungsi depth yang baru sebagai berikut:

1. Persamaan ini setara dengan depth Mahalanobis dalam artian bahwa dapat memberikan urutan multivariat yang sama, yaitu urutan titik pusat-*output* yang sama yang dijelaskan oleh poin kedua dan ketiga dalam Definisi. 1 (Liu, 1990) (Liu et al. 1999);

2. Perhitungannya tidak terlalu rumit dibandingkan dengan depth Mahalanobis.

Definisi baru depth Mahalanobis atau *robust* depth Mahalanobis akan dirumuskan menggunakan determinan dari matriks yang dipartisi, sebagai berikut:

Misal $x_1, x_2, \dots, x_n$ adalah sampel acak dari distribusi p-variate, dengan kriteria pengurutan data sebagai berikut:

$$M_i = \begin{bmatrix} 1 & (x_i - \bar{x})^t \\ (x_i - \bar{x}) & S \end{bmatrix} \quad (12)$$

M_i merupakan matriks dengan ukuran $(p+1) \times (\mathrm{p}+1)$ dengan $X_i; i = 1,2, \dots, n$. Keuntungan dari fungsi depth M_i atau robust depth Mahalanobis yang diusulkan terletak pada komputasinya yang tidak terlalu rumit dibandingkan dengan depth Mahalanobis MD_i. Kompleksitas komputasinya yaitu jumlah operasi dalam komputasi M_i lebih kecil dari pada MD_i. Untuk M_i hanya perlu menghitung invers dari S sekali untuk semua item sampel, sedangkan untuk MD_i kita perlu melibatkan S dalam M_i untuk setiap item sampel i.

2.12 Metode *Minimum Vector Variance*

Metode MVV terinspirasi oleh algoritma *C-Steps* pada FMCD. Rousseuw dan Van Driessen memperkenalkan metode FMCD (*Fast Minimum Covariance Determinant*) untuk mendeteksi *outlier* berdasarkan nilai determinan matriks varians-covarians yang minimum. Namun, metode FMCD mempunyai kelemahan ketika nilai determinan matriks varians-covarians sama dengan nol. Herwindiati (2009) memodifikasi algoritma FMCD menjadi lebih efektif dan tingkat kompleksitas yang lebih rendah dengan menggunakan ukuran *vector variance* (VV) yang minimum yang selanjutnya disebut dengan Minimum *Vector Variance* (MVV).

Kriteria MVV untuk menaksir lokasi dan dispersi, pertama kali diperkenalkan oleh Herwindiati (2006) dengan mempertimbangkan himpunan data $\boldsymbol{X} = \{X_1, X_2, \dots, X_n\}$ dari satu observasi dengan variabel *p* dan $\boldsymbol{H} \subseteq \boldsymbol{X}$**.** Misalkan T_{MVV} dan S_{MVV} adalah taksiran MVV untuk parameter lokasi dan matriks variansi kovariansi. Taksiran diperoleh berdasarkan himpunan H. Jumlah lokasi elemen dari H adalah h= $\frac{(\mathrm{n}+\mathrm{p}+1)}{2}$ data yang akan memberikan matriks variansi kovariansi S_{MVV} dengan nilai

$tr(S^2_{MVV})$ minimum untuk semua kemungkinan himpunan yang mengandung h data. Oleh karena itu secara berturut-turut taksiran MVV untuk parameter lokasi dari matriks adalah sebagai berikut:

$$T_{MVV} = \frac{1}{h}\sum_{i\in H} X_i$$

$$S_{MVV} = \frac{1}{h-1}\sum_{i\in H}(X_i - T_{MVV})(\, X_i - T_{MVV})^t$$

Keterangan:

T_{MVV} = Vektor rata-rata dalam algoritma MVV

S_{MVV} = Matriks Variansi Kovariansi sampel dalam algoritma MVV

h = Dihitung dengan rumus $\frac{n+p+1}{2}$ untuk memperoleh data awal secara acak yang digunakan untuk perhitungan T_{MVV} dan S_{MVV}

n = Jumlah banyaknya pengamatan atau sampel

p = Jumlah banyaknya variable

Algoritma MVV pada dasarnya menghitung nilai objektif dari seluruh kemungkinan subset data yang diperoleh berdasarkan $h = \frac{n+p+1}{2}$ data. Adapun langkah-langkah dalam algoritma MVV adalah sebagai berikut:

1. Mengambil himpunan data yang terdiri dari $h = \frac{(\mathrm{n} + \mathrm{p} + 1)}{2}$ data, sebutlah himpunan data ini dengan H_{old}
2. Menghitung vektor mean $\bar{X}_{H_{old}}$ dan matriks kovariansi $S_{H_{old}}$untuk semua data H_{old}. Selanjutnya untuk $i = 1,2, \dots, n$ hitunglah
$$d^2_{H_{old}} = d^2_{H_{old}}(X_i, \bar{X}_{H_{old}}) = (\, X_i, -\bar{X}_{H_{old}})'\, S^{-1}_{H_{old}}(X_i, -\bar{X}_{H_{old}})$$
3. Mengurutkan hasil perhitungan dari yang terkecil ke yang terbesar. Urutan ini akan memberikan permutasi indeks observasi **$\boldsymbol{\pi}$.** Misalnya hasil pengurutan data tersebut adalah
$$d^2_{H_{old}}(\pi_1) \le d^2_{H_{old}}(\pi_2) \le \dots \le d^2_{H_{old}}(\pi_n)$$
4. Membentuk suatu himpunan baru yang terdiri dari h observasi dengan indeks $\pi(1), \pi(2), \dots, \pi(h)$ dan berilah nama H_{new}

5. Menghitung $\bar{X}_{H_{new}}$, $S_{H_{new}}$ dan $d^2_{H_{new}}(X_i, \bar{X}_{H_{new}})$ seperti pada tahap 2
6. Jika $Tr(S^2_{H_{new}})$=$Tr(S^2_{H_{old}})$ maka proses dikerjakan. Jika $Tr(S^2_{H_{new}})$<$Tr(S^2_{H_{old}})$ maka proses dilanjutkan sampai iterasi ke-k mencapai $Tr(S^2_{H_{new}})$= $Tr(S^2_{H_{old}})$
7. Jika S_{H_i}adalah matriks varians dari iterasi ke-k. Pada akhir iterasi ke-k akan dimiliki $Tr(S^2_{H_1}) \geq Tr(S^2_{H_2}) \geq \ldots \geq Tr(S^2_{H_{k-1}}) = Tr(S^2_{H_k})$
8. Pencilan data ditentukan apabila $Rd_{MVV} \geq \chi^2_{(k;1-\alpha)}$

2.13 Definisi Operasional

Beberapa istilah yang digunakan dalam penelitian ini disajikan dalam tabel sebagai berikut:

Tabel 3.1 Definisi Operasional

Istilah	Uraian
Algoritma	Prosedur atau deretan instruksi yang jelas dalam memecahkan masalah untuk memperoleh keluaran yang diinginkan dari suatu masukkan dalam jumlah waktu yang terbatas.
Data Multivariat	Data yang memiliki variabel ≥ 2
Depth	Metode untuk mengukur jarak berdasarkan kedalaman data serta sebagai alat grafis untuk memvisualisasikan kumpulan data
Determinan	Nilai yang dihitung dengan mengalikan entri pada suatu baris atau kolom dengan masing-masing kofaktor dan menjumlahkan hasil perkalian tersebut.
Estimasi	Suatu pengukuran yang didasarkan pada hasil kuantitatif atau sering disebut sebagai perkiraan
Estimator	Aturan untuk menghitung estimasi kuantitas tertentu berdasarkan data yang diamati dan sering disebut sebagai penaksir

Iterasi	Suatu teknik perulangan yang digunakan pada penulisan program.
Mahalanobis	Metode untuk mengukur atau menghitung jarak tiap observasi terhadap pusat datanya
Matriks	Susunan bilangan, simbol, atau ekspresi yang disusun dalam baris dan kolom sehingga membentuk suatu bangun persegi.
Mean	Nilai rata-rata dari sebuah data
MVV	Metode statistik yang menggunakan *Vector Varians* yang minimum untuk pendetekdian pencilan
Outlier	Titik data yang berbeda secara signifikan dari pengamatan lain dan sering disebut sebagai pencilan
Robust	Metode, alat, atau prosedur yang kebal (tahan) terhadap adanya data yang tidak biasa (pencilan)
Trace	Jumlah dari elemen-elemen diagonal utama dari matriks bujur sangkar
Varians Kovarians	Suatu matriks simetris yang berisi varian pada diagonal utamanya dan kovarian pada elemen lainnya
Vektor	Matriks yang hanya memiliki satu kolom saja atau hanya memiliki satu baris saja dan dinotasikan dengan huruf kecil dan dicetak tebal