

BAB I

PENDAHULUAN

1.1 Latar Belakang

Analisis multivariat merupakan cabang statistika yang digunakan untuk menganalisis data yang melibatkan lebih dari satu variabel respon secara simultan. Analisis ini banyak diterapkan dalam berbagai bidang, seperti bioinformatika, genetika, biostatistika, ekonomi, keuangan, dan ilmu sosial, yang umumnya menghasilkan data dengan banyak variabel yang saling berkaitan. Dengan mempertimbangkan hubungan antarvariabel secara simultan, analisis multivariat mampu memberikan informasi yang lebih komprehensif dibandingkan analisis univariat sehingga banyak digunakan dalam penelitian yang melibatkan lebih dari satu variabel respon [1][18].

Salah satu metode dalam analisis multivariat adalah *Multivariate Analysis of Variance* (MANOVA), yaitu pengembangan dari *Analysis of Variance* (ANOVA) yang digunakan untuk menguji kesamaan parameter lokasi multivariat pada dua atau lebih kelompok berdasarkan beberapa variabel respon secara simultan. Berbeda dengan ANOVA univariat, MANOVA mempertimbangkan hubungan antarvariabel respon dalam proses pengujian sehingga dapat memberikan informasi mengenai perbedaan antarkelompok secara simultan [1][18].

Penerapan MANOVA klasik memiliki beberapa asumsi, antara lain normalitas multivariat, homogenitas matriks kovarians, independensi pengamatan, serta kondisi jumlah variabel yang lebih kecil daripada jumlah observasi. Namun, pada penelitian modern, khususnya di bidang bioinformatika, genomik, dan genetika, sering dijumpai data dengan jumlah variabel yang jauh lebih besar daripada jumlah observasi (*large p, small n*) [5][9]. Pada kondisi tersebut, matriks kovarians sampel menjadi singular sehingga tidak memiliki invers. Akibatnya, statistik uji MANOVA klasik, seperti *Wilks' Lambda*, *Pillai's Trace*, *Hotelling–*

Lawley Trace, dan *Roy's Largest Root*, tidak dapat diterapkan karena memerlukan invers matriks kovarians sampel [3][9][13][20]. Kondisi tersebut menunjukkan perlunya metode alternatif yang dapat digunakan untuk menguji kesamaan parameter lokasi multivariat pada data berdimensi tinggi.

Salah satu contoh data berdimensi tinggi yang banyak digunakan dalam bidang bioinformatika adalah data *RiboNucleic Acid Sequencing (RNA-Seq)*. Teknologi *RNA-Seq* memungkinkan pengukuran tingkat ekspresi ribuan hingga puluhan ribu gen secara simultan pada setiap sampel biologis sehingga menghasilkan data dengan dimensi yang sangat tinggi. Data *RNA-Seq* dapat digunakan untuk mempelajari karakteristik molekuler berbagai jenis kanker serta membandingkan profil ekspresi gen antarkelompok sampel. Namun, jumlah sampel yang tersedia umumnya lebih sedikit dibandingkan jumlah gen yang diamati sehingga dapat memenuhi karakteristik *large p, small n*. Karakteristik tersebut menjadikan data *RNA-Seq* sesuai digunakan sebagai objek penelitian dalam penerapan metode MANOVA berdimensi tinggi [9][18][25].

Untuk mengatasi keterbatasan MANOVA klasik pada kondisi tersebut, telah dikembangkan berbagai metode MANOVA berdimensi tinggi yang tidak bergantung pada invers matriks kovarians sampel. Salah satu pendekatan yang digunakan adalah median geometrik sebagai estimator parameter lokasi multivariat. Median geometrik diperoleh dengan meminimumkan jumlah jarak Euclidean suatu titik terhadap seluruh pengamatan dan dapat digunakan sebagai estimator lokasi pada data berdimensi tinggi [8][16]. Selanjutnya, prosedur *wild bootstrap* dapat digunakan untuk membentuk distribusi empiris statistik uji sehingga nilai kritis dan *p-value* dapat ditentukan melalui distribusi hasil *resampling* [15][23].

Cheng, Lin, dan Peng (2024) mengembangkan metode MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*. Dalam metode tersebut, statistik uji dibangun berdasarkan selisih estimator median geometrik antarkelompok dan distandardisasi menggunakan *Spatial Sign covariance matrix* (SSCM) serta estimator momen negatif pertama. Statistik uji yang terbentuk selanjutnya digunakan dalam prosedur pengujian bertipe maksimum

(*maximum-type statistic*), sedangkan *wild bootstrap* digunakan untuk memperoleh distribusi empiris statistik uji sebagai dasar penentuan nilai kritis dan *p-value* [19].

Penerapan metode tersebut memerlukan beberapa tahapan komputasi yang saling berkaitan, mulai dari estimasi median geometrik menggunakan algoritma Weiszfeld, pembentukan *Spatial Sign covariance matrix* (SSCM), perhitungan estimator momen negatif pertama dan parameter standardisasi, pembentukan statistik uji pasangan (*pairwise test statistic*) dan statistik uji global, hingga pelaksanaan *wild bootstrap*. Oleh karena itu, diperlukan implementasi komputasi yang sistematis agar setiap tahapan metode dapat diterapkan secara konsisten pada data berdimensi tinggi [19].

Berdasarkan kebutuhan tersebut, penelitian ini mengimplementasikan metode MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* menggunakan bahasa pemrograman Python pada data ekspresi gen *RNA-Seq*. Data yang digunakan terdiri atas lima kelompok kanker, yaitu *Breast Invasive Carcinoma* (BRCA), *Colon Adenocarcinoma* (COAD), *Kidney Renal Clear Cell Carcinoma* (KIRC), *Lung Adenocarcinoma* (LUAD), dan *Prostate Adenocarcinoma* (PRAD). Tahapan penelitian meliputi prapemrosesan data, estimasi median geometrik, pembentukan parameter standardisasi, perhitungan statistik uji pasangan dan statistik uji global, prosedur *wild bootstrap*, serta pengambilan keputusan berdasarkan nilai kritis *bootstrap* dan *p-value* empiris [19][22][23][25].

Berdasarkan uraian tersebut, penelitian mengenai *Multivariate Analysis of Variance* (MANOVA) Berdimensi Tinggi dengan Pendekatan Median Geometrik dan *Bootstrapping* penting dilakukan untuk mengkaji penerapan metode tersebut pada data ekspresi gen *RNA-Seq*. Penelitian ini mencakup formulasi matematis dan implementasi komputasi metode yang dikembangkan oleh Cheng, Lin, dan Peng (2024), sehingga dapat memberikan gambaran mengenai penerapan MANOVA berdimensi tinggi untuk menguji kesamaan parameter lokasi multivariat pada data dengan jumlah variabel yang lebih besar daripada jumlah sampel.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana formulasi matematis MANOVA berdimensi tinggi berbasis median geometrik?
2. Bagaimana prosedur pengujian MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*?
3. Bagaimana hasil penerapan MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* pada data *RNA-Seq* berdimensi tinggi?

1.3 Batasan Masalah

Agar ruang lingkup penelitian lebih terarah dan tidak menyimpang dari tujuan penelitian, batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah data ekspresi gen *RiboNucleic Acid Sequencing (RNA-Seq)* yang memiliki karakteristik data multivariat berdimensi tinggi (*high-dimensional data*), yaitu jumlah variabel (p) lebih besar daripada jumlah sampel (n).
2. Penelitian difokuskan pada pengujian kesamaan parameter lokasi multivariat pada lima kelompok data ekspresi gen kanker *RNA-Seq* menggunakan metode MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*.
3. Estimasi parameter lokasi multivariat dilakukan menggunakan median geometrik sesuai dengan metode yang dikembangkan oleh Cheng, Lin, dan Peng (2024).
4. Statistik uji yang digunakan merupakan statistik uji bertipe maksimum (*maximum-type statistic*) yang dibentuk berdasarkan statistik uji pasangan

(*pairwise test statistic*) sebagaimana dirumuskan dalam metode Cheng, Lin, dan Peng (2024).

5. Distribusi empiris statistik uji diperoleh menggunakan prosedur *wild bootstrap* dengan pembobot acak Rademacher.
6. Implementasi metode dilakukan menggunakan bahasa pemrograman Python.
7. Penelitian ini berfokus pada penerapan MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*. MANOVA klasik digunakan sebagai landasan teoritis untuk menjelaskan keterbatasannya pada data berdimensi tinggi, tetapi tidak diterapkan maupun dibandingkan secara empiris dalam penelitian ini.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dikemukakan, tujuan penelitian ini adalah sebagai berikut.

1. Menjelaskan formulasi matematis MANOVA berdimensi tinggi berbasis median geometrik.
2. Menjelaskan prosedur pengujian MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*.
3. Menganalisis hasil penerapan MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* pada data ekspresi gen *RNA-Seq* berdimensi tinggi.

Adapun manfaat penelitian ini adalah sebagai berikut.

1. Memberikan pemahaman mengenai konsep, formulasi matematis, dan prosedur pengujian MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*.
2. Memberikan gambaran mengenai penerapan MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* pada data *RNA-Seq* berdimensi tinggi.

3. Menambah referensi akademik mengenai implementasi MANOVA berdimensi tinggi pada analisis data multivariat.
4. Menjadi referensi bagi penelitian selanjutnya yang berkaitan dengan pengembangan maupun penerapan metode MANOVA berdimensi tinggi.

1.5 Metode Penelitian

Penelitian ini merupakan penelitian metodologis (*methodological research*) yang dilakukan melalui studi literatur dan implementasi komputasi pada data ekspresi gen *RiboNucleic Acid Sequencing (RNA-Seq)*. Penelitian difokuskan pada pembahasan formulasi matematis dan penerapan metode *Multivariate Analysis of Variance* (MANOVA) berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* untuk menguji kesamaan parameter lokasi multivariat antarkelompok. Implementasi metode dilakukan menggunakan bahasa pemrograman Python berdasarkan metode yang dikembangkan oleh Cheng, Lin, dan Peng (2024).

Tahapan penelitian yang dilakukan adalah sebagai berikut.

1. Studi Literatur

Tahap awal dilakukan melalui studi literatur untuk memperoleh landasan teori yang berkaitan dengan analisis multivariat, MANOVA klasik, data berdimensi tinggi (*high-dimensional data*), MANOVA berdimensi tinggi, median geometrik, *Elliptically Symmetric Distribution* (ESD), *Spatial Sign covariance matrix* (SSCM), aproksimasi Gaussian, *bootstrap*, dan *wild bootstrap*. Referensi diperoleh dari buku, artikel ilmiah, dan jurnal yang relevan dengan topik penelitian.

2. Pengumpulan dan Praproses Data

Data penelitian berupa data ekspresi gen *RNA-Seq* diperoleh dari Kaggle. Data yang digunakan terdiri atas lima kelompok kanker, yaitu BRCA, COAD, KIRC, LUAD, dan PRAD. Praproses data dilakukan dengan menghapus variabel gen yang memiliki nilai ekspresi nol pada seluruh sampel. Selanjutnya, jumlah

sampel pada setiap kelompok disesuaikan menjadi 78 sampel sehingga diperoleh data akhir dengan 390 sampel dan 9.758 gen.

3. Penyusunan Formulasi Matematis

Tahap ini meliputi penyusunan model dan hipotesis pengujian, estimasi median geometrik menggunakan algoritma Weiszfeld, pembentukan *Spatial Sign covariance matrix* (SSCM), perhitungan estimator momen negatif pertama, pembentukan parameter standardisasi, serta penyusunan statistik uji pasangan (*pairwise test statistic*) dan statistik uji global (*maximum-type statistic*).

4. Landasan Teoretis dan Prosedur *Wild bootstrap*

Pada tahap ini dibahas aproksimasi Gaussian sebagai landasan teoretis distribusi asimtotik statistik uji. Selanjutnya diterapkan prosedur *wild bootstrap* menggunakan pembobot acak Rademacher untuk membentuk distribusi empiris statistik uji. Distribusi empiris tersebut digunakan untuk memperoleh nilai kritis *bootstrap* dan *p-value* empiris.

5. Implementasi Metode dan Analisis Data

Metode yang telah dirumuskan kemudian diimplementasikan menggunakan bahasa pemrograman Python pada data *RNA-Seq*. Tahapan implementasi meliputi estimasi median geometrik setiap kelompok, perhitungan parameter standardisasi, pembentukan statistik uji pasangan dan statistik uji global, pelaksanaan *wild bootstrap*, serta penentuan nilai kritis *bootstrap* dan *p-value* empiris.

6. Interpretasi Hasil

Tahap terakhir dilakukan interpretasi terhadap hasil pengujian berdasarkan statistik uji global, nilai kritis *bootstrap*, dan *p-value* empiris. Selanjutnya ditarik kesimpulan mengenai kesamaan parameter lokasi multivariat pada kelompok kanker yang dianalisis.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun ke dalam lima bab sebagai berikut.

BAB I PENDAHULUAN

Bab ini memuat latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian, serta sistematika penulisan. Pada bab ini dijelaskan latar belakang, permasalahan, tujuan, dan ruang lingkup penelitian mengenai MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*.

BAB II LANDASAN TEORI

Bab ini memuat teori-teori yang mendukung penelitian, meliputi analisis multivariat, MANOVA klasik, data berdimensi tinggi (*high-dimensional data*), MANOVA berdimensi tinggi, *Elliptically Symmetric Distribution* (ESD), median geometrik, aproksimasi Gaussian, serta metode *bootstrap* dan *wild bootstrap*.

BAB III *MULTIVARIATE ANALYSIS OF VARIANCE* (MANOVA) PADA DATA BERDIMENSI TINGGI MENGGUNAKAN PENDEKATAN MEDIAN GEOMETRIK DAN *BOOTSTRAPPING*

Bab ini membahas formulasi matematis dan prosedur pengujian MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*. Pembahasan meliputi formulasi matematis, estimasi median geometrik menggunakan algoritma Weiszfeld, pembentukan *Spatial Sign* covariance matrix, estimasi parameter standarisasi menggunakan estimator momen negatif pertama, pembentukan statistik uji, aproksimasi Gaussian, prosedur *wild bootstrap*, serta tahapan pengujian hipotesis.

BAB IV PENERAPAN *MULTIVARIATE ANALYSIS OF VARIANCE* (MANOVA) PADA DATA BERDIMENSI TINGGI MENGGUNAKAN PENDEKATAN MEDIAN GEOMETRIK DAN *BOOTSTRAPPING*

Bab ini membahas implementasi metode MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap* pada data *RNA-Seq* menggunakan Python. Pembahasan meliputi estimasi median geometrik, perhitungan parameter standardisasi, pembentukan statistik uji pasangan dan statistik uji global, pelaksanaan *wild bootstrap*, pengujian hipotesis, serta interpretasi hasil analisis.

BAB V PENUTUP

Bab ini berisi kesimpulan yang diperoleh berdasarkan hasil penelitian serta saran untuk pengembangan penelitian selanjutnya yang berkaitan dengan penerapan maupun pengembangan metode MANOVA berdimensi tinggi berbasis median geometrik dengan pendekatan *wild bootstrap*.

