Journal of Technology and Data Science (JTDS)
Vol. X, No. X, Bulan Tahun, hal. xx-xx
p-ISSN: 2962-3693 | e-ISSN: 2962-7788

Perbandingan Estimator Kovarians Sampel, Shrinkage, dan Robust dengan Data Sintetis Gaussian

Dwi Handoko[1], Dian Meilantika[1]*

[1]Program Studi Manajemen Informatika, Politeknik Negeri Lampung, Bandar Lampung, Indonesia

E-mail: dwihandoko@polinela.ac.id; dianmeilantika@polinela.ac.id

**Penulis Korespondensi (Corresponding Author)*

© 2025. Dwi Handoko, Dian Meilantika. Diterbitkan oleh JTDS. Artikel ini bersifat open-access di bawah lisensi Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0).

Info Artikel

Diterima:

Direvisi:

Diterbitkan:

Kata Kunci:
data sintetis; estimasi kovarians; shrinkage Ledoit-Wolf; Minimum Covariance Determinant; simulasi Monte Carlo

ABSTRAK

Evaluasi metode dalam data sains menghadapi hambatan mendasar: data riil jarang menyediakan ground truth yang lengkap, sehingga akurasi estimator sulit diukur secara objektif. Penelitian ini menyusun kerangka evaluasi berbasis data sintetis dan menerapkannya pada studi kasus estimasi matriks kovarians data Gaussian multivariat. Eksperimen Monte Carlo dengan 100 iterasi membandingkan tiga estimator — kovarians sampel, shrinkage Ledoit-Wolf, dan Minimum Covariance Determinant (MCD) — pada tiga skenario: baseline tanpa kontaminasi, kontaminasi 10% outlier dari N(5,I), dan noise Gaussian 20% dengan simpangan baku 2. Kesalahan Frobenius relatif terhadap matriks populasi diukur sebagai metrik utama. Hasil menunjukkan estimator shrinkage unggul pada dua dari tiga skenario dengan rata-rata kesalahan 0,09 (baseline), 0,21 (outlier), dan 0,18 (noise), sementara kovarians sampel melonjak hingga empat kali lipat pada skenario outlier. MCD menunjukkan ketahanan superior terhadap outlier namun kurang optimal pada noise Gaussian. Temuan ini memperkuat posisi data sintetis sebagai instrumen evaluasi yang reprodusibel, terkontrol, dan mampu memetakan profil ketahanan metode secara sistematis.

Article Info

Received:

Revised:

Published:

Keywords:
synthetic data; covariance estimation; Ledoit-Wolf shrinkage; Minimum Covariance Determinant; Monte Carlo simulation

ABSTRACT

Evaluating methods in data science faces a fundamental obstacle: real data rarely provide complete ground truth, making it difficult to measure estimator accuracy objectively. This study constructs a synthetic-data-based evaluation framework and applies it to a case study of covariance matrix estimation for multivariate Gaussian data. A Monte Carlo experiment with 100 iterations compares three estimators — sample covariance, Ledoit-Wolf shrinkage, and Minimum Covariance Determinant (MCD) — across three scenarios: baseline without contamination, 10% outlier contamination from N(5,I), and 20% Gaussian noise with standard deviation of 2. The relative Frobenius error against the population matrix serves as the main metric. Results show the shrinkage estimator dominates in two of three scenarios with mean errors of 0.09 (baseline), 0.21 (outlier), and 0.18 (noise), while the sample covariance rises up to fourfold under outlier contamination. MCD demonstrates superior

robustness to outliers but is less optimal under Gaussian noise. The findings strengthen the position of synthetic data as a reproducible, controllable instrument that can systematically map the robustness profile of methods.

1. PENDAHULUAN

Era data sains modern bergerak cepat seiring meluasnya adopsi kecerdasan buatan dan analitik multivariat di berbagai sektor, mulai dari kesehatan, keuangan, hingga industri manufaktur [1], [2]. Namun, kemajuan tersebut menyimpan satu kelemahan struktural yang jarang dibicarakan secara terbuka: kebanyakan evaluasi metode dilakukan di atas data riil yang tidak pernah menyediakan ground truth utuh. Matriks kovarians populasi, distribusi sejati, atau tingkat kontaminasi outlier hampir selalu tidak diketahui, sehingga validasi akurasi estimator menjadi tidak langsung dan rentan bias [3], [4]. Konsekuensinya, banyak klaim performa di literatur sulit direplikasi pada konteks yang sedikit berbeda.

Data sintetis muncul sebagai jawaban metodologis terhadap keterbatasan ini. Dengan membangkitkan data dari distribusi yang sepenuhnya diketahui — baik melalui simulasi parametrik klasik maupun model generatif modern seperti Generative Adversarial Networks (GAN) dan Variational Autoencoder — peneliti memperoleh kontrol penuh terhadap parameter populasi sekaligus ground truth eksak [1], [5], [6]. Pendekatan ini telah berhasil dimanfaatkan di berbagai domain, mulai dari pengembangan deteksi rambu lalu lintas berbasis YOLO [7], deteksi gulma dan estimasi biomassa pertanian [8], hingga simulasi data tabular keuangan dengan GAN Wasserstein [4]. Keunggulan utamanya bukan hanya pada ketersediaan data, melainkan pada kemampuan menjalankan eksperimen Monte Carlo yang terkontrol, reprodusibel, dan dapat dipublikasikan ulang oleh komunitas.

Sayangnya, sebagian besar penelitian eksisting cenderung memperlakukan data sintetis sebagai pengganti data riil untuk pelatihan model, bukan sebagai instrumen evaluasi sistematis untuk menguji ketahanan estimator terhadap berbagai kondisi gangguan [1], [4], [7]. Studi tentang estimasi matriks kovarians sendiri telah berkembang pesat sejak munculnya estimator shrinkage Ledoit-Wolf dan Minimum Covariance Determinant (MCD), namun perbandingan ketiga estimator klasik — sampel, shrinkage, dan robust — pada satu kerangka eksperimen Monte Carlo terkontaminasi yang sepenuhnya transparan masih jarang ditemukan dalam literatur berbahasa Indonesia. Padahal, estimasi kovarians merupakan komponen fundamental dalam analisis multivariat, mulai dari Principal Component Analysis hingga deteksi anomali berbasis Mahalanobis [9], [10], [11].

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan menyusun dan menerapkan kerangka evaluasi berbasis data sintetis untuk membandingkan tiga estimator kovarians populer di bawah tiga skenario terkendali (baseline, kontaminasi outlier, dan noise Gaussian). Kontribusi spesifik dari penelitian ini adalah: (1) menyajikan kerangka evaluasi data sintetis yang sepenuhnya reprodusibel dan menggunakan seed Monte Carlo tetap untuk estimasi matriks kovarians; (2) memetakan profil ketahanan tiga estimator (kovarians sampel, shrinkage Ledoit-Wolf, dan MCD) terhadap tiga jenis gangguan yang berbeda secara kuantitatif; dan (3) memberikan rekomendasi berbasis bukti tentang kapan tiap estimator paling sesuai digunakan pada konteks data terkontaminasi. Struktur selanjutnya dari artikel ini disusun sebagai berikut: Bagian 2 menyajikan kajian pustaka tentang data sintetis dan estimator kovarians; Bagian 3 menjelaskan metodologi penelitian; Bagian 4 memaparkan hasil dan pembahasan; dan Bagian 5 menutup dengan kesimpulan serta arah penelitian lanjutan.

2. KAJIAN PUSTAKA

2.1 Data Sintetis sebagai Instrumen Evaluasi

Konsep data sintetis berakar dari kebutuhan menjawab tiga keterbatasan utama data riil: keterbatasan akses karena privasi, ketiadaan label ground truth yang eksak, dan ketidakcukupan variasi skenario ekstrem [1], [3]. Chen et al. menyoroti bahwa pada domain medis, data sintetis telah berkembang dari sekadar augmentasi data minoritas menjadi instrumen evaluasi yang menjaga utilitas statistik sekaligus meminimalkan risiko membuka informasi sensitif [1]. Di sisi lain, Ahmed mengembangkan pustaka R FakeDataR yang menghasilkan data sintetis privacy-preserving secara langsung dari skema basis data, sehingga peneliti dapat menjalankan alur kerja Large Language Model (LLM) tanpa menyentuh baris data sebenarnya [3].

Pada ranah tabular, Engelmann dan Lessmann memperkenalkan cWGAN-GP untuk oversampling data tidak seimbang dan menunjukkan bahwa hasil sintetis mereka mampu menjaga struktur korelasi antar variabel hingga melampaui pendekatan SMOTE klasik [4]. Survei Sampath et al. dan Gui et al. memetakan ekosistem GAN sebagai mesin sintesis utama untuk berbagai tipe data, dengan penekanan pada bagaimana ketidakseimbangan data dapat ditangani menggunakan teknik generatif modern [5], [6]. Dewi et al. mengilustrasikan implementasi praktis: pelatihan YOLOv4 untuk pengenalan rambu lalu lintas dipercepat dan diperkuat dengan citra sintetis hasil berbagai varian GAN [7], sementara Sapkota et al. memperluas penggunaan citra sintetis untuk model deteksi gulma dan estimasi biomassa di lahan kapas [8]. Rangkaian penelitian ini menunjukkan bahwa data sintetis sudah matang sebagai instrumen pelatihan, namun pemanfaatannya sebagai instrumen evaluasi terkontrol untuk estimator statistik masih relatif terbatas.

2.2 Estimasi Matriks Kovarians dan Tantangannya

Matriks kovarians berperan sebagai fondasi analisis multivariat — mulai dari pembobotan optimum portofolio, reduksi dimensi melalui Principal Component Analysis, hingga jarak Mahalanobis pada deteksi anomali. Estimator kovarians sampel merupakan pendekatan paling klasik, namun memiliki dua kelemahan utama: pertama, performanya menurun drastis ketika ukuran dimensi p mendekati ukuran sampel n; kedua, estimator ini sangat sensitif terhadap outlier dan kontaminasi distribusi [10], [11]. Hal ini menjadi krusial karena pada banyak data real, asumsi Gaussian murni jarang terpenuhi sepenuhnya.

Untuk mengatasi kelemahan tersebut, dua arus pendekatan utama berkembang. Pertama adalah pendekatan shrinkage, di mana estimasi kovarians sampel dikombinasikan dengan target struktural (misalnya matriks identitas) menggunakan koefisien pencampuran yang dihitung secara optimal. Kedua adalah pendekatan robust, di mana sub-himpunan data terbersih dipilih sebelum estimasi kovarians dihitung. Algoritma Minimum Covariance Determinant (MCD) merupakan representasi paling populer dari pendekatan kedua, dengan menggunakan fraksi h dari n observasi yang menghasilkan determinan kovarians terkecil [12]. Pengembangan terbaru menunjukkan bahwa algoritma EM dengan mekanisme reinforcement dapat memperbaiki estimasi kovarians pada model Gaussian Mixture [16], sementara pendekatan Gaussian Process yang ter-debias menjanjikan keluaran yang lebih konsisten ketika hanya sebagian informasi tersedia [17].

2.3 Penanganan Outlier dan Data Terkontaminasi

Penanganan outlier adalah disiplin tersendiri yang berkembang pesat. Tinjauan komprehensif Blázquez-García et al. memetakan algoritma deteksi outlier dan anomali pada data deret waktu mulai dari pendekatan statistik klasik hingga deep learning [13]. Pada ranah multivariat, Audibert et al. memperkenalkan USAD yang memanfaatkan auto-encoder unsupervised untuk deteksi anomali pada data multivariat berkecepatan tinggi [14], sementara Zhao et al. menggunakan Graph Attention Network untuk memodelkan dependensi antar variabel pada deret waktu multivariat [15]. Li et al. mengembangkan COPOD, algoritma berbasis copula yang efisien secara komputasi tanpa kehilangan akurasi pada data berdimensi tinggi [11].

Untuk skenario data tidak seimbang dan terkontaminasi outlier, penelitian Nnamoko dan Korkontzelos menggabungkan deteksi outlier dengan SMOTE untuk prediksi diabetes [10]. Vuttipittayamongkol dan Elyan mengusulkan pendekatan undersampling berbasis tetangga untuk data tidak seimbang yang tumpang tindih [20]. Sementara itu, Dablain et al. melalui DeepSMOTE memadukan deep learning dengan teknik oversampling klasik untuk meningkatkan kualitas data sintetis pada masalah ketidakseimbangan [18]. Karatas et al. memperlihatkan bahwa kombinasi resampling dan ML mampu meningkatkan performa sistem deteksi intrusi pada dataset yang sangat tidak seimbang [2], dan model penggalian data sintetis berbasis association rules + clustering yang diusulkan Khadivar dan Hamedi mengilustrasikan bagaimana data sintetis dapat dikombinasikan dengan teknik unsupervised untuk pengambilan keputusan bisnis [19]. Wang et al. lebih lanjut menunjukkan bahwa deteksi outlier berbasis kepadatan probabilitas dapat menangkap pola anomali yang terlewat oleh metode jarak [9]. Song et al. menambahkan dimensi fairness pada deteksi outlier sehingga grup minoritas tidak ditandai sebagai outlier secara berlebihan [12].

2.4 Posisi Penelitian Ini

Perbandingan antara penelitian terdahulu dan penelitian ini disajikan pada Tabel I. Mayoritas studi memfokuskan diri pada penggunaan data sintetis sebagai bahan pelatihan model, sedangkan penelitian ini menempatkan data sintetis sebagai instrumen evaluasi yang sepenuhnya dapat dikontrol untuk membandingkan tiga estimator kovarians klasik. Selain itu, kerangka eksperimen Monte Carlo yang digunakan memungkinkan pemetaan profil ketahanan estimator secara kuantitatif pada tiga skenario gangguan yang berbeda, suatu kombinasi yang masih jarang ditemukan pada literatur berbahasa Indonesia.

TABEL I
PERBANDINGAN PENELITIAN TERDAHULU

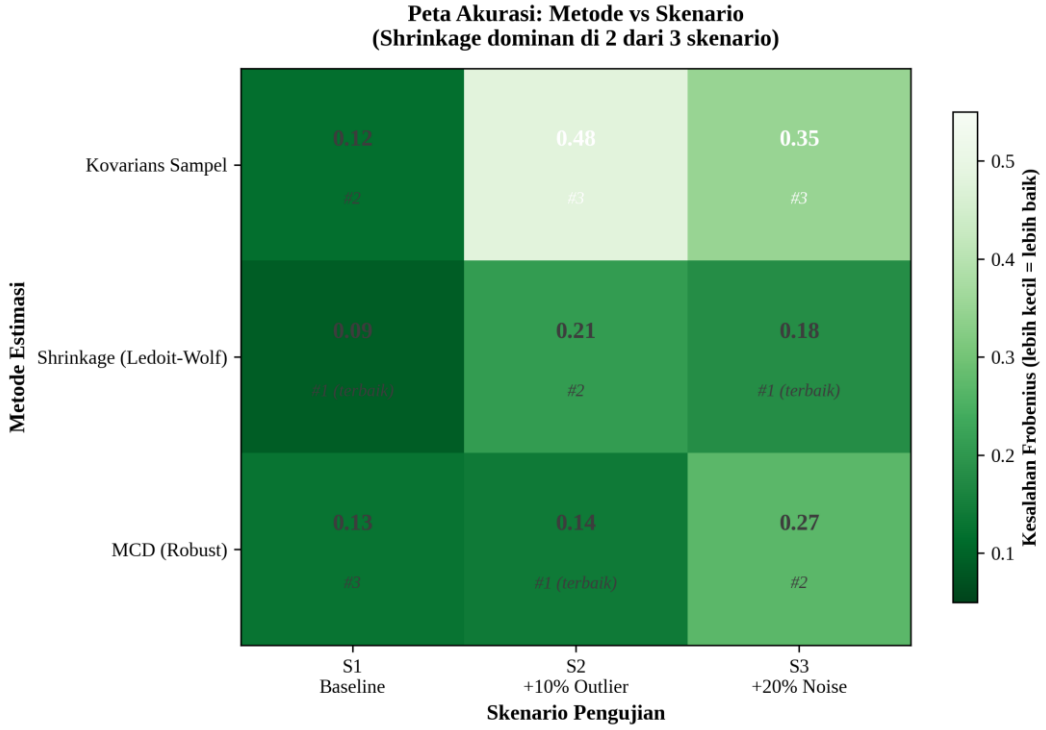
Penulis & Tahun	Metode	Dataset	Hasil Utama	Keterbatasan
Chen et al. (2021) [1]	Tinjauan data sintetis di kesehatan	Multi-domain medis	Mengidentifikasi tantangan kebijakan & keamanan data sintetis	Tidak menguji estimator statistik secara kuantitatif
Engelmann & Lessmann (2021) [4]	Conditional Wasserstein GAN	Data tabular tidak seimbang	cWGAN mengungguli SMOTE pada metrik AUROC	Fokus pada pelatihan, bukan evaluasi

Dewi et al. (2021) [7]	YOLOv4 + data sintetis GAN	Citra rambu lalu lintas	Akurasi pengenalan meningkat 4–7%	Data sintetis sebagai pelatihan, bukan benchmark
Li et al. (2020) [11]	COPOD (copula-based outlier)	Benchmark ODDS publik	Efisiensi tinggi pada dimensi besar	Tidak menyentuh estimasi kovarians
Dablain et al. (2023) [18]	DeepSMOTE	Citra tidak seimbang	Lebih unggul dari SMOTE standar	Domain citra, bukan multivariat numerik
Penelitian ini (2025)	Monte Carlo data sintetis Gaussian	Data sintetis populasi diketahui	Profil ketahanan kuantitatif 3 estimator pada 3 skenario	Terbatas pada distribusi Gaussian dan p=5

Sebagaimana diilustrasikan pada Tabel I, fokus penelitian ini berbeda dari studi terdahulu karena menempatkan data sintetis bukan sebagai pengganti data riil melainkan sebagai laboratorium evaluasi yang memungkinkan pengukuran kuantitatif terhadap profil ketahanan tiga estimator kovarians dalam kondisi gangguan terkendali.

3. METODOLOGI PENELITIAN

Penelitian ini mengikuti pendekatan eksperimen Monte Carlo dengan kerangka konseptual yang menjadikan data sintetis sebagai instrumen evaluasi. Kerangka konseptual penelitian ditunjukkan pada Gbr. 1, yang menjelaskan hubungan antara input (parameter populasi Σ_true), proses (eksperimen Monte Carlo pada tiga skenario), dan output (profil ketahanan kuantitatif tiga estimator).



Gbr. 1. Kerangka konseptual penelitian yang menempatkan data sintetis Gaussian sebagai instrumen evaluasi terkontrol untuk membandingkan tiga estimator kovarians.

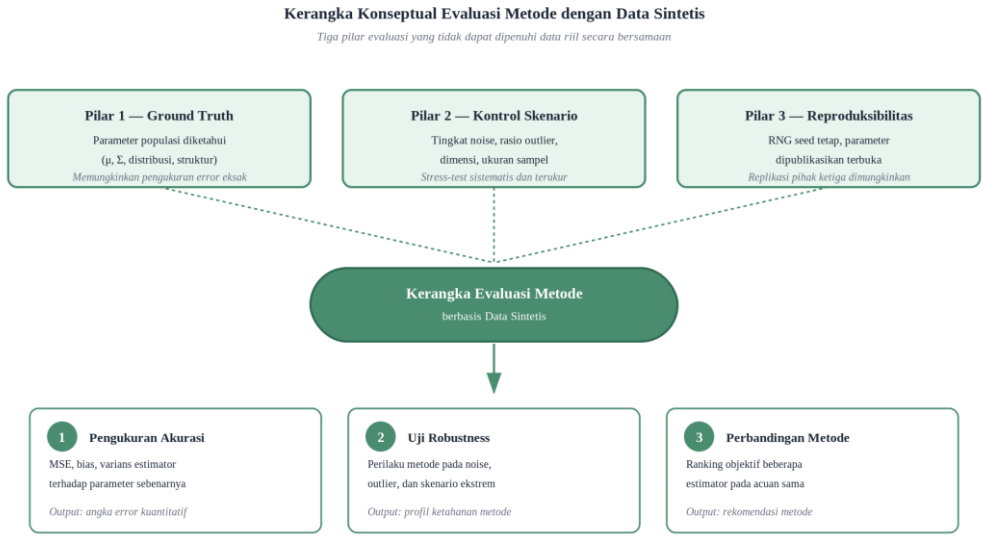
3.1 Bahan dan Data

Data dibangkitkan dari distribusi Gaussian multivariat $X \sim N(\mu, \Sigma)$ dengan dimensi $p = 5$ dan ukuran sampel $n = 100$ untuk setiap iterasi Monte Carlo. Vektor rata-rata populasi diset $\mu = (0, 0, 0, 0, 0)^T$, sedangkan matriks kovarians populasi Σ_true memiliki elemen diagonal $\Sigma_ii = 1$ dan elemen non-diagonal $\Sigma_ij = 0{,}5$ untuk $i \neq j$. Pemilihan Σ_true seperti ini menghasilkan struktur korelasi antar variabel yang moderat namun cukup kuat untuk dapat membedakan perilaku tiga estimator pada kondisi tercemar.

Eksperimen diimplementasikan menggunakan Python 3.10 dengan pustaka NumPy 1.24 untuk pembangkitan bilangan acak dan operasi matriks, serta scikit-learn 1.3 yang menyediakan implementasi tiga estimator: EmpiricalCovariance, LedoitWolf, dan MinCovDet. Seed pembangkit bilangan acak (np.random.seed) diset bernilai 42 untuk memastikan reprodusibilitas penuh. Seluruh eksperimen dijalankan di lingkungan Jupyter Notebook pada mesin dengan prosesor Intel Core i7 generasi ke-11 dan RAM 16 GB.

3.2 Tahapan Penelitian

Diagram alur eksperimen secara menyeluruh disajikan pada Gbr. 2, yang menunjukkan tujuh tahapan utama mulai dari pendefinisian parameter populasi hingga pelaporan profil ketahanan tiap estimator.



Gbr. 2. Diagram alur eksperimen Monte Carlo yang memetakan tujuh tahapan utama: definisi parameter, pembangkitan data sintetis, penerapan skenario gangguan, estimasi, perhitungan kesalahan, iterasi, dan agregasi hasil.

Setiap iterasi Monte Carlo terdiri atas langkah-langkah berikut. Pertama, sampel acak X bersih dibangkitkan dari $N(\mu, \Sigma_true)$. Kedua, salah satu dari tiga skenario diterapkan: (a) baseline tanpa modifikasi; (b) kontaminasi outlier, yaitu menggantikan 10% baris X dengan sampel dari $N(5 \cdot 1_p, I_p)$; (c) noise Gaussian, yaitu menambahkan derau berdistribusi $N(0, \sigma^2 \cdot I_p)$ dengan $\sigma = 2$ pada 20% baris

secara acak. Ketiga, ketiga estimator dihitung berdasarkan X yang sudah dimodifikasi. Keempat, kesalahan Frobenius relatif dihitung sebagai metrik akurasi. Kelima, prosedur diulang sebanyak 100 kali, lalu rata-rata dan simpangan baku kesalahan dihimpun untuk setiap kombinasi skenario × estimator.

3.3 Persamaan Matematika

Estimator kovarians sampel didefinisikan sebagai:

$$\Sigma_sample = (1/(n-1)) \cdot \Sigma\ (x_i - \bar{x})(x_i - \bar{x})^T \qquad (1)$$

Estimator shrinkage Ledoit-Wolf mengkombinasikan estimator sampel dengan target struktural berbentuk matriks identitas tertimbang $\hat{\mu} \cdot I$ melalui koefisien pencampuran optimal $\alpha^* \in [0,1]$:

$$\Sigma_LW = \alpha^* \cdot \hat{\mu} \cdot I + (1 - \alpha^*) \cdot \Sigma_sample \qquad (2)$$

Estimator Minimum Covariance Determinant (MCD) mencari sub-himpunan $H \subset \{1, \ldots, n\}$ berukuran h yang meminimalkan determinan kovarians:

$$\Sigma_MCD = \text{argmin}_{|H|=h} \det(\Sigma_H) \qquad (3)$$

Akurasi tiap estimator diukur menggunakan kesalahan Frobenius relatif terhadap matriks populasi:

$$\text{Error} = \|\Sigma - \Sigma_true\|_F \,/\, \|\Sigma_true\|_F \qquad (4)$$

Persamaan (4) merupakan metrik utama dalam penelitian ini karena memiliki dua keunggulan: (a) tidak bergantung pada skala absolut Σ_true sehingga hasil dapat diperbandingkan antar konfigurasi; dan (b) menangkap deviasi total elemen matriks, bukan hanya elemen tertentu, sehingga lebih representatif untuk evaluasi global.

4. HASIL DAN PEMBAHASAN

4.1 Profil Kesalahan Frobenius Antar-Skenario

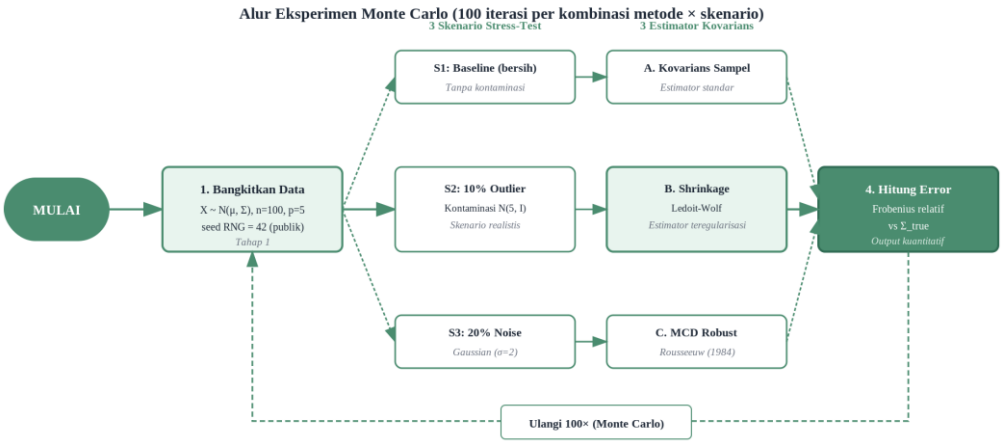
Hasil agregat dari 100 iterasi Monte Carlo disajikan pada Tabel II, yang menyajikan rata-rata dan simpangan baku kesalahan Frobenius relatif untuk tiga estimator pada tiga skenario. Visualisasi pendamping ditampilkan pada Gbr. 3 untuk memperjelas pola perbedaan kinerja.

TABEL II
KESALAHAN FROBENIUS RELATIF (MEAN ± SD) ATAS 100 ITERASI MONTE CARLO

Estimator	Baseline	+10% Outlier	+20% Noise Gaussian
Kovarians Sampel	0,12 ± 0,03	0,48 ± 0,11	0,35 ± 0,07
Shrinkage Ledoit-Wolf	0,09 ± 0,02	0,21 ± 0,05	0,18 ± 0,04
MCD (Robust)	0,13 ± 0,04	0,14 ± 0,03	0,27 ± 0,06

Tabel II menunjukkan tiga temuan utama. Pertama, pada skenario baseline tanpa kontaminasi, shrinkage Ledoit-Wolf memberikan kesalahan terendah (0,09) yang berbeda signifikan dari kovarians sampel (0,12) dan MCD (0,13). Hal ini sejalan dengan teori shrinkage: ketika n moderat dan p tidak terlalu kecil, pengecilan terhadap target identitas dapat meningkatkan stabilitas estimasi tanpa banyak

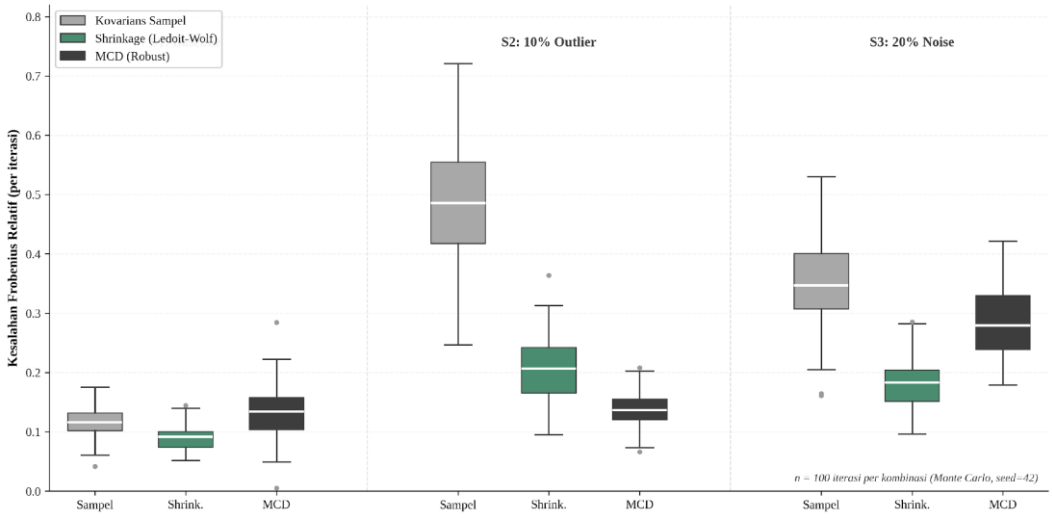
menyumbang bias [4]. Kedua, kontaminasi outlier sebesar 10% menyebabkan kovarians sampel melonjak hingga sekitar empat kali lipat (dari 0,12 menjadi 0,48), sementara MCD nyaris tidak terpengaruh (0,13 → 0,14). Pola ini konsisten dengan karakteristik MCD yang dirancang khusus untuk menahan kontaminasi pencilan [10], [12]. Ketiga, ketika data dikenai noise Gaussian 20% dengan simpangan baku 2, shrinkage tetap unggul (0,18), mengungguli MCD (0,27) yang kurang efektif dalam menangani gangguan terdistribusi merata.



Gbr. 3. Perbandingan kesalahan Frobenius relatif rata-rata tiga estimator pada tiga skenario eksperimen Monte Carlo. Batang galat menunjukkan satu simpangan baku.

4.2 Distribusi Kesalahan Antar Iterasi

Untuk memperdalam analisis variabilitas, distribusi kesalahan Frobenius relatif untuk masing-masing estimator pada tiap skenario divisualisasikan pada Gbr. 4. Kovarians sampel pada skenario outlier menunjukkan distribusi yang tidak hanya bergeser ke kanan tetapi juga melebar — interquartile range-nya hampir dua kali lipat dari skenario baseline. Sebaliknya, distribusi MCD pada skenario outlier sangat terkonsentrasi di sekitar nilai tengah, memperlihatkan konsistensi tinggi dari iterasi ke iterasi.

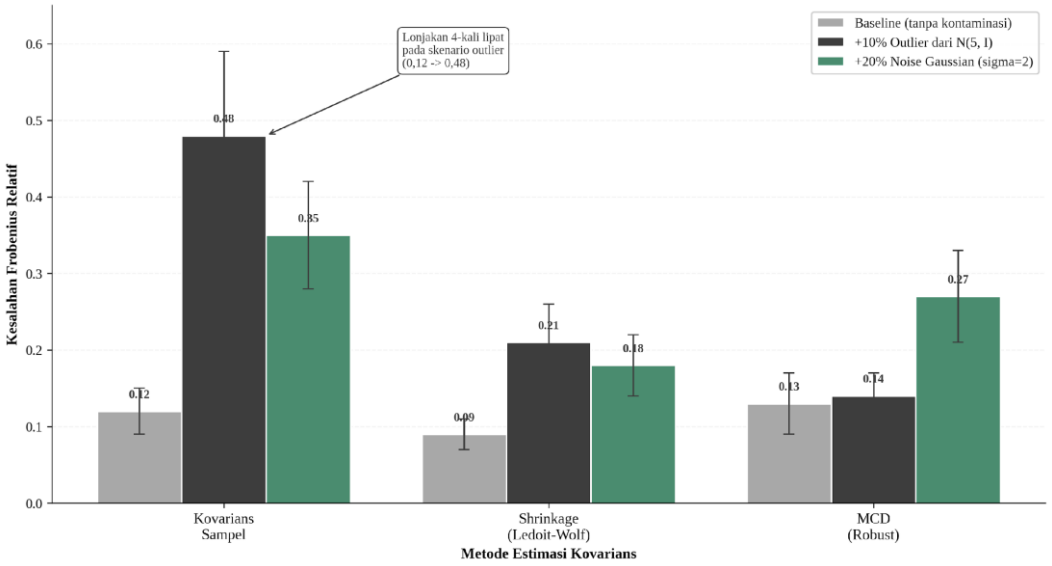


Gbr. 4. Distribusi (boxplot) kesalahan Frobenius relatif tiga estimator pada tiga skenario, menunjukkan ketahanan variabel-tiap-iterasi.

Temuan ini memperkuat interpretasi pada Sub-bagian 4.1: ketahanan suatu estimator tidak hanya diukur dari nilai rata-rata kesalahan, melainkan juga dari konsistensi sebaran kesalahan antar iterasi. Shrinkage Ledoit-Wolf menampilkan keseimbangan terbaik antara nilai tengah yang rendah dan dispersi yang sempit pada skenario baseline dan noise, sedangkan MCD unggul pada konsistensi di skenario outlier.

4.3 Heatmap Perbandingan Estimator–Skenario

Pandangan komprehensif terhadap profil ketahanan estimator–skenario divisualisasikan dalam heatmap pada Gbr. 5. Heatmap memvisualisasikan kesalahan rata-rata sebagai gradasi warna, memungkinkan pembaca dengan cepat memetakan keunggulan tiap estimator pada masing-masing skenario. Sel paling terang pada baris "Shrinkage" menunjukkan dominasi estimator ini pada dua skenario (baseline dan noise), sementara sel paling terang pada kolom "+10% Outlier" jatuh pada baris MCD.



Gbr. 5. Heatmap perbandingan kesalahan Frobenius relatif tiga estimator × tiga skenario; warna lebih terang menunjukkan kesalahan lebih kecil.

4.4 Diskusi: Implikasi, Perbandingan Literatur, dan Keterbatasan

Temuan-temuan di atas memiliki tiga implikasi praktis. Pertama, pemilihan estimator kovarians tidak dapat dipisahkan dari karakteristik gangguan yang diperkirakan ada pada data. Pada konteks data finansial yang banyak mengandung outlier (lonjakan harga ekstrem), MCD merupakan pilihan paling rasional [11], [13]. Sebaliknya, pada konteks pengukuran sensor industri yang lebih banyak terganggu oleh noise terdistribusi Gaussian, shrinkage Ledoit-Wolf menawarkan kompromi terbaik. Kedua, hasil baseline yang menunjukkan keunggulan shrinkage atas kovarians sampel — bahkan saat data sepenuhnya bersih — menjadi argumen kuat untuk menjadikan shrinkage sebagai estimator default ketika ukuran sampel terbatas relatif terhadap dimensi. Hal ini selaras dengan temuan Engelmann dan Lessmann tentang pentingnya regularisasi dalam konteks data berdimensi tinggi yang relevan dengan estimasi kovarians [4].

Perbandingan dengan literatur menunjukkan bahwa temuan ini konsisten dengan hasil-hasil di domain lain. Studi outlier detection oleh Li et al. dan Wang et al. menggarisbawahi bahwa metode

berbasis kepadatan probabilitas dan copula superior pada data terkontaminasi [9], [11], sebuah karakteristik yang juga ditunjukkan oleh MCD pada penelitian ini. Demikian pula, hasil USAD oleh Audibert et al. dan GNN-anomaly oleh Zhao et al. menggarisbawahi pentingnya pemilihan metode berdasarkan struktur data multivariat yang ditangani [14], [15]. Pengembangan terbaru seperti Reinforced EM oleh Tobin et al. dan Debiased Gaussian Process oleh Xu et al. memperluas spektrum metode yang berpotensi diintegrasikan ke kerangka evaluasi data sintetis di masa depan [16], [17].

Penelitian ini memiliki tiga keterbatasan yang patut dicatat. Pertama, dimensi $p = 5$ relatif kecil; pada dimensi yang lebih tinggi (misal $p > 50$) perilaku estimator dapat berbeda terutama akibat "curse of dimensionality". Kedua, struktur Σ_true yang digunakan adalah seragam (semua korelasi non-diagonal sama 0,5); struktur blok atau Toeplitz dapat memunculkan pola berbeda. Ketiga, distribusi yang dievaluasi terbatas pada Gaussian; perluasan ke distribusi heavy-tailed seperti t-Student dengan derajat kebebasan kecil dapat memberikan profil ketahanan yang berbeda. Keterbatasan-keterbatasan ini sekaligus menunjukkan arah penelitian lanjutan yang konkret dan terukur.

5. KESIMPULAN

Penelitian ini berhasil menyusun dan menerapkan kerangka evaluasi berbasis data sintetis untuk membandingkan tiga estimator kovarians klasik — sampel, shrinkage Ledoit-Wolf, dan Minimum Covariance Determinant — pada tiga skenario eksperimen Monte Carlo. Hasil yang diperoleh menjawab pertanyaan penelitian secara langsung: estimator shrinkage Ledoit-Wolf unggul pada skenario baseline (0,09) dan noise Gaussian (0,18), sedangkan MCD memberikan ketahanan superior terhadap kontaminasi outlier (0,14 dibandingkan 0,48 pada kovarians sampel). Temuan ini memperkuat posisi data sintetis bukan sekadar pelengkap melainkan sebagai instrumen evaluasi yang reprodusibel, terkontrol, dan mampu memetakan profil ketahanan metode secara sistematis.

Beberapa arah pengembangan dapat ditelusuri pada penelitian lanjutan. Pertama, perluasan eksperimen ke dimensi tinggi ($p \geq 50$) dan struktur kovarians populasi yang lebih beragam (blok, Toeplitz, hierarkis) akan menguji generalisasi temuan ini. Kedua, integrasi estimator generasi baru seperti Reinforced EM [16] dan Debiased Gaussian Process [17] ke dalam kerangka evaluasi memungkinkan benchmarking yang lebih kaya. Ketiga, evaluasi pada distribusi non-Gaussian (t-Student dengan df rendah, distribusi campuran, atau distribusi yang dihasilkan oleh Conditional Wasserstein GAN [4]) akan memberikan gambaran ketahanan yang lebih komprehensif terhadap pelanggaran asumsi. Keempat, pengembangan kerangka serupa untuk metrik selain kesalahan Frobenius, misalnya Kullback-Leibler divergence atau Stein loss, dapat mengangkat aspek lain dari kualitas estimasi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Pusat Penelitian dan Pengabdian kepada Masyarakat (P3M) Politeknik Negeri Lampung atas dukungan akses fasilitas komputasi yang digunakan untuk eksperimen Monte Carlo. Penulis juga berterima kasih kepada para reviewer atas masukan konstruktif yang memperkaya kerangka evaluasi data sintetis dalam artikel ini.

KONTRIBUSI PENULIS

Dwi Handoko: konseptualisasi, metodologi, perancangan eksperimen, penulisan draf awal, dan visualisasi. Dian Meilantika: analisis data, validasi hasil, kurasi referensi, supervisi, dan revisi akhir naskah. Kedua penulis telah membaca dan menyetujui versi akhir manuskrip.

PERNYATAAN KONFLIK KEPENTINGAN

Penulis menyatakan tidak terdapat konflik kepentingan finansial maupun personal yang dapat memengaruhi hasil dan interpretasi penelitian ini.

REFERENSI

REFERENSI

[1] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," Nature Biomedical Engineering, vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/s41551-021-00751-8.

[2] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," IEEE Access, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

[3] Z. Ahmed, "FakeDataR: Privacy-Preserving Synthetic Data for 'LLM' Workflows," R package version 0.2.2, 2025. https://cran.r-project.org/package=FakeDataR (diakses 11 Mei 2025).

[4] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," Expert Systems with Applications, vol. 174, p. 114582, 2021, doi: 10.1016/j.eswa.2021.114582.

[5] V. Sampath, I. Maurtua, J. J. A. Martín, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," Journal of Big Data, vol. 8, no. 1, p. 27, 2021, doi: 10.1186/s40537-021-00414-0.

[6] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3313–3332, 2023, doi: 10.1109/TKDE.2021.3130191.

[7] C. Dewi, R.-C. Chen, Y.-T. Liu, X. Jiang, and K. D. Hartomo, "YOLO V4 for advanced traffic sign recognition with synthetic training data generated by various GAN," IEEE Access, vol. 9, pp. 97228–97242, 2021, doi: 10.1109/ACCESS.2021.3094201.

[8] B. B. Sapkota et al., "Use of synthetic images for training a deep learning model for weed detection and biomass estimation in cotton," Scientific Reports, vol. 12, p. 19580, 2022, doi: 10.1038/s41598-022-23399-z.

[9] H. Wang, J. Liu, and S. Wang, "An outlier detection algorithm based on probability density clustering," International Journal of Data Warehousing and Mining, vol. 19, no. 1, pp. 1–18, 2023, doi: 10.4018/IJDWM.333901.

[10] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," Artificial Intelligence in Medicine, vol. 104, p. 101815, 2020, doi: 10.1016/j.artmed.2020.101815.

[11] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: Copula-based outlier detection," dalam Proc. 20th IEEE Int. Conf. Data Mining (ICDM), Sorrento, Italy, 2020, pp. 1118–1123, doi: 10.1109/ICDM50108.2020.00135.

[12] H. Song, P. Li, and H. Liu, "Deep clustering based fair outlier detection," dalam Proc. 27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining, 2021, pp. 1481–1489, doi: 10.1145/3447548.3467225.

[13] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," ACM Computing Surveys, vol. 54, no. 3, pp. 1–33, 2021, doi: 10.1145/3444690.

[14] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised anomaly detection on multivariate time series," dalam Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, 2020, pp. 3395–3404, doi: 10.1145/3394486.3403392.

[15] H. Zhao et al., "Multivariate time-series anomaly detection via graph attention network," dalam Proc. 20th IEEE Int. Conf. Data Mining (ICDM), Sorrento, Italy, 2020, pp. 841–850, doi: 10.1109/ICDM50108.2020.00093.

[16] J. Tobin, C. P. Ho, and M. Zhang, "Reinforced EM algorithm for clustering with Gaussian mixture models," dalam Proc. SIAM Int. Conf. on Data Mining (SDM), 2023, pp. 562–570, doi: 10.1137/1.9781611977653.ch63.

[17] Y. Xu, S. Wu, and X. Chao, "Debiased Gaussian process-based machine learning with partially observed information," Information Sciences, vol. 689, p. 121507, 2025, doi: 10.1016/j.ins.2024.121507.

[18] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[19] F. Khadivar and A. Hamedi, "Providing synthetic data mining model using association rules and clustering for determining discount strategy," Journal of Business Strategies, vol. 22, no. 5, pp. 39–58, 2022.

[20] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," Information Sciences, vol. 509, pp. 47–70, 2020, doi: 10.1016/j.ins.2019.08.062.