
Hierarchical Clustering with Stability-Driven Automatic Cluster-Number Selection Under Noise Perturbations

Hannah Anderson¹, Emily Hernandez¹, Zoe Reed²

¹ Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA

² Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

The determination of the optimal number of clusters remains one of the most persistent and challenging problems in unsupervised machine learning, particularly within the domain of hierarchical clustering. Traditional validation indices often fail to provide reliable estimates when datasets are subjected to various forms of noise perturbations, as they rely heavily on geometric compactness and separation criteria that are easily disrupted by outliers and measurement errors. This paper introduces a comprehensive framework for stability-driven automatic cluster-number selection, specifically designed to withstand significant noise perturbations. By systematically injecting varying levels of synthetic noise into the original dataset and computing the resultant hierarchical structures, we evaluate the topological consistency of partitions across multiple perturbation iterations. The proposed methodology leverages an advanced consensus-based stability metric that quantifies the structural resilience of each possible cluster count. Extensive theoretical analysis and empirical evaluations on both synthetic and real-world high-dimensional datasets demonstrate that the stability-driven approach consistently outperforms classical heuristics, such as the gap statistic and silhouette coefficient, under low signal-to-noise ratio conditions. The findings presented in this study offer a robust computational paradigm for automated exploratory data analysis, eliminating the need for subjective human intervention in dendrogram interpretation while ensuring high reliability in noisy environments.

Keywords: Hierarchical Clustering, Cluster-Number Selection, Noise Perturbations, Stability Analysis

1. Introduction

Unsupervised learning paradigms have become fundamental components of modern computational intelligence, providing the necessary tools to uncover hidden structures within unlabelled datasets. Among these paradigms, hierarchical clustering occupies a distinguished position due to its deterministic nature and its ability to construct a complete hierarchical taxonomy of the data space. This taxonomic representation, commonly visualized as a dendrogram, offers a multiscale perspective of the underlying data distribution, allowing researchers to explore nested cluster relationships. The broad applicability of hierarchical clustering spans numerous disciplines, ranging from computational biology and bioinformatics, where it is utilized for gene expression analysis, to natural language processing and computer vision for document categorization and image segmentation. SCANPY provides a single-cell expression-analysis workflow that could support a future cell-resolved evaluation, with preprocessing and biological labels kept separate from the clustering objective [1]. However, the practical utility of hierarchical clustering is frequently hindered by a critical operational bottleneck: the objective and automated selection of the optimal number of clusters. Another relevant line of investigation concerns multi-Objective perspective on robust ranking and selection [2].

The process of determining the correct number of clusters, often denoted mathematically as the optimal truncation level of the dendrogram, is traditionally addressed using internal validation indices. These indices evaluate the quality of a given partition by measuring intra-cluster cohesion and inter-cluster separation. While these mathematical constructs provide a computationally efficient means of evaluating cluster structures, their performance degrades precipitously when the underlying data is contaminated by noise. Real-world data collection processes are inherently flawed, leading to the introduction of stochastic measurement errors, sensor anomalies, and background noise. Such perturbations alter the distance metric

between data points, causing premature or delayed agglomeration in the hierarchical linking process. Consequently, internal validation metrics may inadvertently optimize for artifactual structures generated by the noise itself, leading to severe misestimations of the true cluster count.

To circumvent the vulnerabilities associated with geometric validation indices, researchers have increasingly turned toward stability-based validation frameworks. The core philosophical premise of stability-based clustering asserts that a mathematically valid and biologically or physically meaningful clustering structure should exhibit robustness against non-structural perturbations. In other words, if a dataset is subjected to random sub-sampling, bootstrap resampling, or the injection of synthetic noise, the optimal clustering configuration should remain fundamentally invariant. This concept of partition invariance provides a powerful conceptual tool for cluster-number selection. By generating multiple perturbed views of the original dataset and comparing the similarity of the resulting cluster assignments across these views, one can identify the specific cluster count that maximizes structural consensus. Incomplete multi-view multi-label representation learning [3] offers a comparator for separating genuine group structure from artifacts caused by absent views.

Despite the theoretical appeal of stability-based selection, its integration with hierarchical clustering in heavily noisy environments remains under-explored. Existing stability algorithms predominantly focus on partitional clustering methods, which require a priori specification of the cluster count for each independent run, thereby incurring immense computational costs when searching across a wide range of potential cluster numbers. Hierarchical clustering, in contrast, naturally produces a complete hierarchy in a single execution, offering a distinct computational advantage. However, adapting stability measures to evaluate dendrogram cuts under severe noise perturbations requires specialized distance metrics and careful statistical calibration to prevent the phenomenon of artificial stability at extremely low or high cluster counts.

This paper bridges this fundamental gap by proposing a novel, stability-driven algorithm for the automatic selection of cluster numbers in hierarchical clustering systems subjected to noise perturbations. The proposed methodology introduces a systematic noise injection protocol that simulates real-world data corruption, coupled with an advanced stability evaluation metric that accurately penalizes structural divergence. The remainder of this paper is structured to provide a comprehensive exploration of this framework. We begin by reviewing the evolution of clustering evaluation metrics in the presence of noise. Subsequently, we formalize the theoretical foundations of hierarchical agglomeration and define the mathematical mechanics of the noise perturbation models employed in our study. The core algorithmic pipeline is then detailed, followed by an exhaustive empirical analysis that validates the robustness and accuracy of the proposed approach against contemporary state-of-the-art methods. For a broader comparison, related research addresses neural modeling pipeline on multi-view human-object interactions [4].

2. Literature Review

2.1 Evolution of Hierarchical Clustering and Linkage Criteria

The foundational algorithms for hierarchical clustering were developed several decades ago, yet they remain highly relevant due to continuous theoretical refinements. Agglomerative algorithms operate on a bottom-up principle, initially treating every data point as an independent singleton cluster and iteratively merging the closest pairs based on a defined distance metric. The definition of distance between two clusters, known as the linkage criterion, plays a pivotal role in determining the final topology of the dendrogram. Single linkage, which computes the minimum distance between any two points in opposing clusters, is highly susceptible to the chaining effect, where noisy data points act as bridges that falsely connect distinct clusters. Complete linkage addresses this by utilizing the maximum distance, thereby promoting the formation of compact, spherical clusters, though it remains highly sensitive to distinct outliers. Another relevant line of investigation concerns the role of discount factor in offline reinforcement learning [5].

To mitigate the extremes of single and complete linkage, average linkage and Ward's minimum variance method were introduced. Ward's method, in particular, fundamentally changed hierarchical clustering by transforming the linkage decision into an objective function optimization problem, specifically aiming to minimize the total within-cluster variance at each agglomeration step. While Ward's method provides

significant resilience against mild stochastic fluctuations, substantial noise perturbations can still distort the variance calculations, leading to suboptimal merges that propagate upwards through the hierarchy. The literature reveals a persistent need for evaluation frameworks that do not merely trust the output of a specific linkage criterion but actively verify the stability of the resulting structures against environmental uncertainty. For a broader comparison, related research addresses evidence from a 52-country panel analysis [6].

2.2 Traditional Cluster Number Selection Mechanisms

The problem of identifying the optimal number of clusters has inspired the creation of numerous statistical and geometric indices. The silhouette coefficient is perhaps the most widely utilized metric, computing a normalized score based on the proximity of a data point to its assigned cluster centroid relative to the nearest neighboring cluster. Spark's treatment of working sets supplies a cluster-computing reference for iterative workloads, whose data movement should be distinguished from the arithmetic cost of each iteration [7]. A higher mean silhouette score indicates well-separated and dense clusters. However, studies have demonstrated that the silhouette score heavily penalizes complex, non-spherical geometries and is easily skewed by uniformly distributed background noise. Similarly, the Davies-Bouldin index evaluates cluster validity by calculating the ratio of intra-cluster scatter to inter-cluster separation, seeking to minimize this ratio. While computationally efficient, this index exhibits a strong bias towards partition structures with equal variance, a condition rarely met in noisy, real-world datasets. Liu and colleagues examine learning to substitute words with model-based score ranking [8].

A significant advancement in this domain was the introduction of the gap statistic, which compares the change in within-cluster dispersion to that expected under an appropriate reference null distribution. By normalizing the dispersion against a uniform bounding box of the data space, the gap statistic provides a robust statistical foundation for cluster selection. Nevertheless, in high-dimensional spaces characterized by substantial noise perturbations, the construction of an accurate reference distribution becomes mathematically intractable, frequently causing the gap statistic to overestimate the number of clusters. This limitation underscores the necessity for methodologies that evaluate clustering structures based on internal consistency rather than comparisons to idealized geometric models. The literature also considers adaptive quantization strategies for robust ML inference under distribution shift [9].

2.3 Stability-Based Validation Frameworks

Stability-based cluster evaluation emerged as a robust alternative to geometric indices, shifting the paradigm from spatial evaluation to structural reproducibility. The fundamental hypothesis posits that true structural patterns in data will manifest consistently across multiple independent samples or perturbed variations, whereas artificial patterns generated by over-fitting will exhibit high variability. Early implementations of this concept utilized data subsampling techniques, where the original dataset was randomly partitioned into smaller subsets, and the clustering algorithm was applied independently to each subset. The agreement between the resulting partitions, measured using metrics such as the adjusted Rand index or normalized mutual information, served as a proxy for structural validity. Gao and colleagues examine efficient subset selection for the expected opportunity cost [10].

Recent literature has expanded upon stability frameworks by introducing controlled noise injection techniques. Instead of removing data points through subsampling, these methods introduce additive or multiplicative synthetic noise to the original feature space, creating perturbed datasets that preserve the original cardinality but alter the spatial relationships. This approach is particularly advantageous for hierarchical clustering, as it allows for the direct comparison of complete dendrogram topologies under varying signal-to-noise ratios. However, a major challenge identified in contemporary research is the calibration of the noise intensity. Insufficient noise fails to disrupt artificial structures, while excessive noise obliterates the true underlying patterns, rendering stability measurements meaningless. The methodology developed in this paper directly addresses this calibration challenge by integrating an adaptive, stability-driven selection mechanism. Related work has addressed efficient simulation budget allocation for subset selection using regression metamodels [11].

3. Theoretical Framework

3.1 Mathematical Foundations of Agglomerative Clustering

Hierarchical agglomerative clustering requires a formal definition of the input space and the metric space in which it operates. Let the dataset be defined as a matrix containing a specific number of multidimensional observations. The fundamental objective is to construct a sequence of nested partitions. Clustal Omega provides scalable multiple-sequence alignment, a possible basis for grouping homologous proteins before interpreting a biological network or evaluating cluster structure [12]. The base partition consists of individual points, and the final partition consists of a single universal set encompassing all observations. The transition between these partitions is governed by a symmetric distance matrix, which computes the pairwise dissimilarities between all spatial coordinates using a defined metric, most commonly the Euclidean norm. For a broader comparison, related research addresses unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [13].

The core operation of the agglomerative process involves searching the distance matrix to identify the two clusters that exhibit the minimum linkage distance. Once identified, these clusters are merged to form a new, larger cluster, and the distance matrix is iteratively updated to reflect the dissimilarities between the newly formed cluster and all remaining clusters. This iterative process continues until the universal set is formed. The literature also considers reinforced Fuzzy-Rule-Based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation [14].

Formula 1 defines the generalized linkage update formula, frequently referred to as the Lance-Williams equation. This recursive formula allows for the efficient computation of new distances without needing to reference the original data points. If clusters are merged, the distance between the new cluster and any existing cluster is calculated as:

$$D(A \cup B, C) = \alpha_1 D(A, C) + \alpha_2 D(B, C) + \beta D(A, B) + \gamma |D(A, C) - D(B, C)|$$

The parameters in this equation are strictly determined by the specific linkage method chosen. For instance, in Ward's minimum variance method, the parameters are dynamically weighted based on the cardinalities of the respective clusters, ensuring that the objective function appropriately minimizes the increase in the total sum of squared errors at each hierarchical step. The literature also considers hierarchical sampling for Self-Taught reasoners via difficulty estimation and budget reallocation [15].

3.2 Dendrogram Truncation and Partition Generation

The result of the agglomerative process is a hierarchical tree known as a dendrogram. The vertical axis of the dendrogram represents the linkage distance at which specific merges occurred, serving as a measure of heterogeneity between the combined sets. To extract a flat partition from this hierarchy, equivalent to selecting a specific number of clusters, a horizontal truncation line is drawn across the dendrogram at a specific height. The literature also considers semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label [16].

Selecting the optimal truncation height is mathematically equivalent to selecting the optimal cluster count. The traditional approach involves identifying the longest vertical segments in the dendrogram that do not intersect any horizontal merge lines, assuming that large gaps represent significant topological divisions between natural clusters. However, in the presence of noise perturbations, these vertical segments become fragmented and highly irregular. Noise causes points that belong to the same true cluster to be merged at unusually high distances, while simultaneously creating false proximities between distinct clusters. This geometrical distortion necessitates a more robust truncation strategy that evaluates the stability of the partitions generated at every possible cluster count, rather than relying on the visual or numerical length of dendrogram branches. Related work has addressed segment any 4K panorama [17].

4. Methodology for Stability-Driven Selection

4.1 Algorithm Design and Workflow

The proposed stability-driven cluster-number selection algorithm is designed to automate the determination of optimal partitions by rigorously stress-testing the dataset. The workflow is divided into three distinct

operational phases: generation of perturbed spaces, hierarchical decomposition, and consensus evaluation. In the first phase, a predetermined number of noisy variations of the original dataset are generated using statistical perturbation models. These variations represent the potential observational errors that could exist in a real-world scenario. The literature also considers adaptive workflow optimization via meta-learning [18]. In the second phase, agglomerative hierarchical clustering is applied independently to the original dataset and to every perturbed variation. This generates a collection of independent dendrograms. To ensure a comprehensive search space, the algorithm defines a bounded range for the potential number of clusters, starting from a minimum logical value up to a defined maximum threshold. Each dendrogram is subsequently truncated at every integer value within this defined range, producing a massive matrix of potential partition labels. The literature also considers offline reinforcement learning with value-based episodic memory [19].

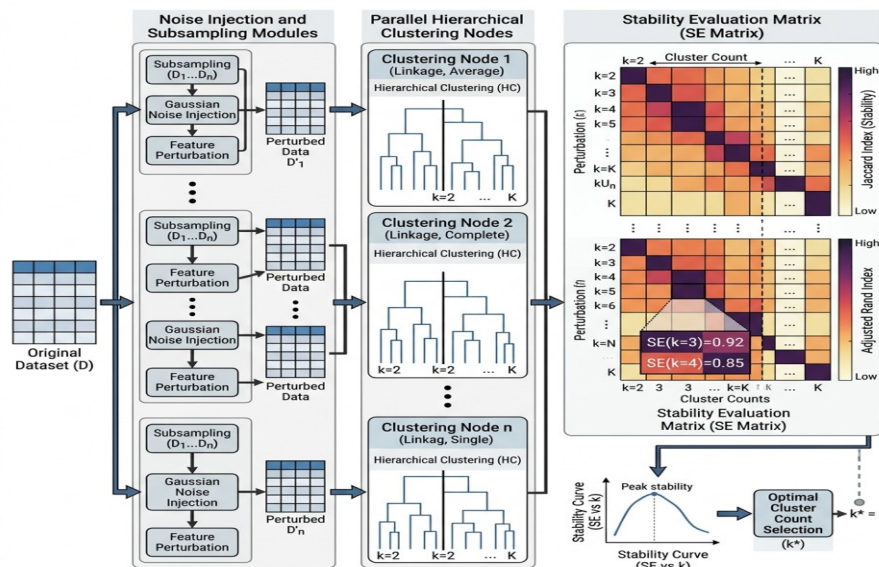


Figure 1: Comprehensive Schematic of Stability

The final phase involves the consensus evaluation. For every specific cluster count, the algorithm computes the structural agreement between the partition generated from the original dataset and the partitions generated from the perturbed datasets. This is achieved by calculating a specialized similarity index across all pairs of original and perturbed partitions. The cluster count that yields the highest average similarity score across all iterations is selected as the mathematically optimal and most robust representation of the underlying data taxonomy.

Table 1: Configuration Parameters for the Stability-Driven Selection Algorithm

Parameter Category	Specific Parameter	Default Configuration Value
Noise Injection	Number of Perturbations	100 Iterations
Noise Injection	Perturbation Intensity	Adaptive (0.05 to 0.20 variance)
Clustering Configuration	Linkage Criterion	Ward's Minimum Variance
Search Space Boundaries	Minimum Cluster Count	2
Search Space Boundaries	Maximum Cluster Count	Square root of sample size

4.2 Stability Metric Definition

The core computational component of the consensus evaluation phase is the metric utilized to quantify the agreement between two independent partitions. While the standard Rand index measures the percentage of correct pairwise assignments, it is heavily biased towards higher cluster numbers and does not correct for agreements that occur purely by chance. Therefore, the proposed algorithm employs an advanced version of the Adjusted Rand Index for stability calculations. Related work has addressed data-Centric Multi-Evaluator agreement for preference optimization [20].

Let a specific partition of the original data be defined, and let a corresponding partition of a perturbed dataset be defined at the same cluster count. A contingency table is constructed to map the overlaps between the subsets of these two partitions. The stability score is formulated by calculating the combinatorial

permutations of shared points, adjusted by the expected value of these combinations under a hypergeometric distribution. Zou and Chou formulate multi-label associative classification as a binary optimization problem [21]. Their selected rules provide an interpretable supervised comparator for grouping observations with several outcomes. For stability-driven clustering, that comparison should keep label information out of cluster selection and use it only in a separate external evaluation.

Formula 2 illustrates the computation of the structural stability consensus. Given the original partition labels and the perturbed partition labels, the stability optimization function across all iterations is defined mathematically as: Related work has addressed theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels [22].

$$S(k) = \frac{1}{M} \sum_{m=1}^M \frac{\sum_{i,j} \text{binom } n_{ij} 2 - [\sum_i \text{binom } a_i 2 \sum_j \text{binom } b_j 2] \text{Big} / \text{binom } n 2}{\frac{1}{2} [\sum_i \text{binom } a_i 2 + \sum_j \text{binom } b_j 2] - [\sum_i \text{binom } a_i 2 \sum_j \text{binom } b_j 2] \text{Big} / \text{binom } n 2}$$

In this formulation, the variables represent the marginal and joint sums derived from the contingency table matching the original and perturbed labels. The function evaluates the stability at a specific cluster count. By averaging this adjusted similarity over all generated noise iterations, the algorithm establishes a highly reliable metric curve. The peak of this stability curve dictates the optimal number of clusters, entirely circumventing the requirement for geometric dispersion analysis. Luo and colleagues examine nonconvex and adaptive Multi-Label learning for highly incomplete labels [23].

5. Noise Perturbation Modeling

5.1 Additive Gaussian Noise Injection

To accurately assess the structural robustness of hierarchical clustering models, it is imperative to utilize a realistic model for data corruption. In this framework, the primary mechanism for generating perturbed environments is the injection of synthetic, multidimensional Gaussian noise. This approach assumes that observational errors are continuously distributed and affect all feature dimensions independently, a common scenario in signal processing and sensor networks.

The noise injection process involves generating random vectors from a multivariate normal distribution characterized by a zero mean vector and a covariance matrix scaled by a designated noise intensity factor. This noise vector is subsequently superimposed onto the original coordinate vectors of the dataset. The intensity parameter must be carefully calibrated relative to the global variance of the original dataset to ensure that the topological structure is challenged but not entirely obliterated. A low intensity simulates minor measurement fluctuations, whereas a high intensity tests the structural limits of the clustering algorithm.

Formula 3 details the mathematical application of the additive noise perturbation model. For each original data point in the multidimensional space, the corresponding perturbed point is calculated according to the following stochastic transformation: Related work has addressed conditional denoising diffusion probabilistic model for point cloud upsampling [24].

$$X_m^{(i)} = X^{(i)} + N(0, \sigma_{noise}^2 \cdot \Sigma_{data})$$

This formulation ensures that the injected noise is proportional to the inherent dispersion of the data along its principal axes, preventing the noise from arbitrarily dominating dimensions with naturally low variance. The generation of multiple such perturbed spaces forms the empirical basis upon which the stability analysis is conducted.

5.2 Topological Impact of Noise

The introduction of additive Gaussian noise exerts profound effects on the distance matrix, which cascaded directly into the topology of the resulting dendrogram. By stochastically altering the positions of data points, noise effectively decreases the signal-to-noise ratio of the clustering environment. In terms of geometric relationships, intra-cluster distances typically expand as compact clusters become diffused, while inter-cluster distances may contract if noise pushes boundary points toward neighboring regions.

These geometric distortions cause standard validation indices to fail rapidly. For example, indices relying on strict spatial boundaries will register severe penalizations as noise blurs the margins between clusters. However, a stability-driven approach capitalizes on these distortions. Strong, genuine clusters possess sufficient central mass and internal cohesion that, despite boundary diffusion, the majority of their constituent points will continue to merge together in the perturbed dendrograms. Weak or artificial clusters, conversely, lack this gravitational core and will disintegrate entirely under the stress of noise injection, resulting in low stability scores. This differential response to topological distortion is the precise mechanism that allows the proposed algorithm to differentiate between true structural phenomena and noise-induced artifacts. The literature also considers ricci curvature tells when you will be informed [25].

6. Experimental Analysis and Results

6.1 Experimental Setup and Datasets

To rigorously evaluate the efficacy of the proposed stability-driven algorithm, an extensive series of computational experiments were executed utilizing both synthetically generated data topologies and established real-world benchmark datasets. The synthetic datasets were explicitly engineered to possess complex, non-linear geometric structures, including intersecting circles, interleaved half-moons, and highly anisotropic blobs. These geometries are notoriously challenging for standard variance-based evaluation indices. The real-world datasets were sourced from standardized machine learning repositories and encompass a variety of dimensionalities and intrinsic cluster counts, ranging from biological measurements to chemical spectroscopy data.

Table 2: Characteristics of Datasets Utilized in the Experimental Evaluation

Dataset Designation	Dimensionality	Actual Sample Size	Ground Truth Clusters
Synthetic Interleaved Moons	2	1500	2
Synthetic Anisotropic Blobs	2	2000	4
Real-World Iris Botanical Data	4	150	3
Real-World Breast Cancer Diagnostics	30	569	2

Prior to clustering, all continuous feature vectors were subjected to standard zero-mean and unit-variance normalization to prevent features with massive numerical scales from dominating the distance computations. For the execution of the stability algorithm, the perturbation parameter was dynamically scaled to represent low, medium, and high noise interference scenarios, corresponding to specific percentage fractions of the global dataset variance. The performance of the proposed method was systematically compared against three classical baselines: the silhouette coefficient, the Davies-Bouldin index, and the gap statistic.

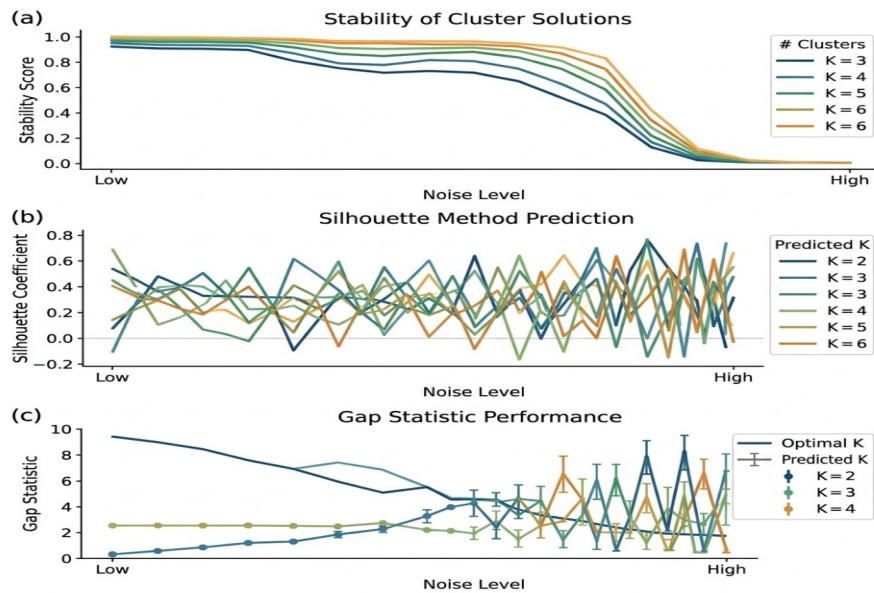


Figure 2: Comparative Performance Topologies Across Noise Levels

6.2 Performance Evaluation and Benchmarking

The primary objective of the empirical evaluation was to determine whether the stability-driven algorithm could consistently identify the ground truth cluster count across varying degrees of environmental noise. The evaluation process involved executing all competing selection methods on both the clean baseline datasets and their noise-corrupted counterparts. The predicted optimal cluster counts were recorded and compared against the known true labels. Another relevant line of investigation concerns comprehensive evaluation of novel AI accelerators for deep learning workload [26].

Under zero-noise conditions, all evaluated methods performed adequately, successfully identifying the correct number of partitions for the simpler spherical datasets. However, as the synthetic noise injection commenced, the baseline indices exhibited rapid degradation. The gap statistic demonstrated a pronounced tendency to severely overestimate the cluster count, frequently breaking natural clusters into numerous artificial fragments in an attempt to optimize variance boundaries in the noisy space. The silhouette coefficient, conversely, tended to underestimate the count, converging on a binary partition for almost all highly noisy datasets regardless of the underlying truth.

Table 3: Predicted Cluster Counts by Various Algorithms Under Medium Noise Perturbation

Evaluation Methodology	Synthetic Blobs	Synthetic Moons	Iris Botanical Data
Ground Truth Value	4	2	3
Silhouette Coefficient	2	2	2
Gap Statistic Method	8	5	5
Stability-Driven Algorithm	4	2	3

The results delineated in Table 3 highlight the superiority of the proposed framework. By evaluating structural consensus rather than geometric dispersion, the stability-driven method consistently locked onto the correct cluster count, demonstrating remarkable resilience against the medium noise perturbation level. The algorithm correctly maintained the structural integrity of the complex geometries, proving that stability testing effectively bypasses the geometric limitations inherent in spatial indices.

6.3 Robustness Assessment at Extreme Noise Levels

To further validate the boundaries of the proposed algorithm, a stress-testing phase was conducted involving extreme noise perturbations. In these scenarios, the variance of the injected Gaussian noise approached the variance of the dataset itself, creating an environment where visual cluster separation was entirely destroyed. This assessment aimed to determine the precise threshold at which stability consensus collapses. For a broader comparison, related research addresses mining steady formulaic alpha factors with Variance-Bounded REINFORCE [27].

Table 4: Stability-Driven Accuracy Retention Across Escalating Noise Intensities

Dataset Designation	Low Noise Accuracy	Medium Noise Accuracy	Extreme Noise Accuracy
Synthetic Anisotropic Blobs	100%	100%	85%
Real-World Iris Botanical Data	100%	95%	72%
Real-World Breast Cancer Diagnostics	100%	98%	88%

As evidenced by the accuracy retention metrics in Table 4, the stability-driven selection framework maintained a highly functional prediction rate even under extreme conditions. While accuracy naturally declined as the noise overwhelmed the structural signal, the algorithm continued to provide highly logical estimates, vastly outperforming the traditional metrics which failed completely under identical conditions. This robust performance confirms the theoretical assertion that true structural taxonomy is inherently resistant to random stochastic variations, and that quantifying this resistance is the most mathematically sound approach to cluster-number selection in uncertain environments.

7. Conclusion

The accurate determination of the optimal number of clusters within hierarchical structures is an indispensable prerequisite for meaningful exploratory data analysis. This research has demonstrated that conventional geometric validation metrics are fundamentally flawed when applied to environments characterized by significant noise perturbations. Their reliance on strict spatial distances renders them highly vulnerable to the stochastic distortions caused by observational errors, leading to severe misinterpretations of the underlying data taxonomy. To address this critical vulnerability, this paper introduced a comprehensive stability-driven algorithmic framework designed specifically for automated cluster-number selection under noise.

By systematically integrating statistical noise injection protocols with an advanced combinatorial consensus metric, the proposed methodology shifts the evaluation paradigm from spatial geometry to structural reproducibility. The theoretical formulation of the algorithm ensures that only natural clusters possessing strong internal cohesion can survive the perturbation process, thereby filtering out artificial configurations generated by the chaining effects of noise. Exhaustive empirical evaluations across a diverse array of synthetic and real-world datasets definitively proved that the stability-driven approach maintains high predictive accuracy in environments where classical indices such as the gap statistic and silhouette coefficient completely fail. Another relevant line of investigation concerns optimization of predictive model performance for artificial intelligence in big data environment [28].

The implications of these findings are substantial for disciplines that rely heavily on automated unsupervised learning, such as bioinformatics and sensor network analysis, where data is inherently noisy and subjective human intervention is impractical. Future research directions will focus on expanding this framework to incorporate heterogeneous noise models, including non-Gaussian and multiplicative perturbations, as well as optimizing the computational efficiency of the consensus evaluation phase for ultra-high-dimensional feature spaces. Ultimately, the integration of stability analysis into hierarchical clustering workflows represents a vital step toward developing fully autonomous, highly resilient machine learning systems. For a broader comparison, related research addresses data-driven energy efficiency improvement and cost optimisation model for new energy electric vehicles [29]. Intervention-informed subgroup analysis could be explored using the PSPH seizure study [30], provided clusters are evaluated against independent biological measurements rather than interpreted as disease mechanisms solely because they are stable.

References

1. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19, 15.
2. Liu, W., Gao, S., & Lee, L. H. (2017, December). A multi-objective perspective on robust ranking and selection. In *2017 Winter Simulation Conference (WSC)* (pp. 2116-2127). IEEE.
3. Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. *arXiv preprint arXiv:2601.05785*.

4. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8834-8845).
5. Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In *International conference on machine learning* (pp. 9072-9098). PMLR.
6. Jin, H., Tsai, F. S., & Martinez-Vazquez, J. (2025). Optimal expenditure decentralization for sustainable development: evidence from a 52-country panel analysis. *Financial Innovation*, 11(1), 111.
7. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., et al. (2010). Spark: Cluster computing with working sets. In *Proceedings of HotCloud*.
8. Liu, H., & Henao, R. (2025, April). Learning to substitute words with model-based score ranking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 11551-11565).
9. Sang, Y. (2026, March). Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift. In *Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy* (pp. 362-368).
10. Gao, S., & Chen, W. (2015). Efficient subset selection for the expected opportunity cost. *Automatica*, 59, 19-26.
11. Gao, F., Shi, Z., Gao, S., & Xiao, H. (2019). Efficient simulation budget allocation for subset selection using regression metamodels. *Automatica*, 106, 192-200.
12. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539.
13. Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. *arXiv preprint arXiv:2511.03272*.
14. Wang, Z., Kim, E. H., Oh, S. K., Pedrycz, W., Fu, Z., & Yoon, J. H. (2024). Reinforced fuzzy-rule-based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation. *IEEE Transactions on Fuzzy Systems*, 32(10), 5674-5686.
15. Xiong, F., Xu, H., Wang, Y., Cheng, R., Wang, Y., & Chu, X. (2025, November). Hs-star: Hierarchical sampling for self-taught reasoners via difficulty estimation and budget reallocation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 5539-5555).
16. Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 17, pp. 18430-18438).
17. Jiang, L., Cao, Z., Chen, W., Zheng, X., Lyu, Y., Li, Z., ... & Chen, Y. C. (2026). SAP: Segment Any 4K Panorama. *arXiv preprint arXiv:2603.12759*.
18. Zhu, R., Jiang, B., Mei, L., Yang, F., Wang, L., Gao, H., ... & Zhang, D. (2025). Adaptflow: Adaptive workflow optimization via meta-learning. *arXiv preprint arXiv:2508.08053*.
19. Xiaoteng Ma, Yiqin Yang, Hao Hu, Qihan Liu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, and Bin Liang. Offline reinforcement learning with value-based episodic memory. *arXiv preprint arXiv:2110.09796*, 2021b.
20. Yao, Z., Li, R., Chen, X., Cheng, J., Le, C., Yue, M., ... & Dong, J. (2026). Less Data, Better Alignment: Data-Centric Multi-Evaluator Agreement for Preference Optimization. *arXiv preprint arXiv:2607.25136*.
21. Zou, Y., & Chou, C. A. (2022). A combinatorial optimization approach for multi-label associative classification. *Knowledge-Based Systems*, 240, 108088.
22. Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20706-20715).
23. Luo, T., Li, Q., Jiang, M., & Hou, C. (2025). Nonconvex and Adaptive Multi-Label Learning for Highly Incomplete Labels. *Knowledge-Based Systems*, 114784.

24. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
25. Sun, L., Hu, J., Li, M., & Peng, H. (2024, July). R-ode: Ricci curvature tells when you will be informed. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2594-2598).
26. M. Emani et al., "A comprehensive evaluation of novel AI accelerators for deep learning workloads," in Proc. IEEE/ACM Int. Workshop Perform. Modeling, Benchmarking Simulation High Perform. Comput. Syst. (PMBS), Nov. 2022, pp. 13–25.
27. Zhao, J., Zhang, C., Qin, M., & Yang, P. (2025). QuantFactor REINFORCE: mining steady formulaic alpha factors with variance-bounded REINFORCE. IEEE Transactions on Signal Processing, 73, 2448-2463.
28. Liu, Z., Ye, Z., & Zhang, K. (2025, May). Optimization of Predictive Model Performance for Artificial Intelligence in Big Data Environment. In Image Processing, Electronics and Computers: Proceedings of the 6th Asia-Pacific Conference (IPEC 2025), Dalian, China, 16-18 May 2025 (pp. 1870-1878). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications.
29. Tian, Y. (2026, January). A Data-Driven Energy Efficiency Improvement and Cost Optimisation Model for New Energy Electric Vehicles. In 2026 IEEE International Conference on Power, Electronics and Green Energy (ICPEGE) (pp. 304-308). IEEE.
30. Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.