# Natural Structure Alignment: The Hidden Factor in Unsupervised Learning

Nader Liddawi
*Georgia Institute of Technology*

***Abstract*—This study investigates the interplay between clustering algorithms (K-Means and Expectation Maximization) and dimensionality reduction techniques (PCA, ICA, and Random Projection) across contrasting data distributions. Using class-imbalanced Marketing Campaign data and balanced Spotify data, we evaluate four hypotheses through the lens of representational learning: (1) algorithms with geometric flexibility outperform rigid approaches on imbalanced data, (2) dimensionality reduction impacts clustering quality differently based on data structure, (3) PCA maintains classification accuracy while reducing dimensionality and training time, and (4) manifold learning methods produce better clusters than linear techniques. Our analysis demonstrates that dimensionality reduction significantly impacts downstream clustering quality, with greatest improvements when the reduction technique aligns with underlying data characteristics. We refute hypothesis 1, as K-Means unexpectedly outperformed EM on imbalanced data (silhouette 0.2497 vs. ARI 0.0385) due to its parametric efficiency and reduced variance in high dimensions. We strongly support hypothesis 4, with LLE achieving exceptional performance (silhouette 0.9814, a 106.3% improvement over linear methods) by effectively unfolding the underlying manifold structure. Hypothesis 3 was confirmed as PCA maintained classification accuracy (0.8797 vs. baseline 0.8812) while reducing dimensionality by 23.1% and training time by 15.4%, confirming the information-theoretic principle that decision boundaries often lie within principal subspaces. Our findings reveal that representation quality often proves more important than algorithm complexity, with algorithmic performance depending more on alignment between algorithmic biases and natural data structure than on theoretical flexibility or class balance alone.**

## I. Introduction

Unsupervised learning algorithms face the challenging task of discovering meaningful patterns without labeled data guidance. While supervised contexts provide clear performance metrics, unsupervised approaches must rely on intrinsic measures of cluster quality or dimensional representation, making the interaction between algorithms and data characteristics especially critical. This relationship is fundamentally constrained by the No Free Lunch theorems, which mathematically prove that no learning algorithm can universally outperform others across all possible problems without leveraging domain-specific structure [8].

We propose four specific hypotheses to guide our investigation:

**H1**: On highly imbalanced data distributions (like our Marketing dataset with 15% positive response rate), algorithms with inherent geometric flexibility will outperform rigid boundary-based approaches, particularly after dimensionality reduction.

**H2**: Dimensionality reduction techniques will impact clustering quality differently based on how well they preserve the underlying data structure.

**H3**: PCA will maintain classification accuracy while significantly reducing dimensionality and training time.

**H4**: Clusters derived from manifold learning methods will have higher silhouette scores than those from linear dimensionality reduction techniques.

These hypotheses emerge from theoretical foundations in representational learning [10]. H1 stems from algorithmic bias considerations: K-Means enforces spherical clusters through Euclidean distance, theoretically disadvantaging it on imbalanced data due to the curse of dimensionality, which causes distance measurements to lose discriminative power as dimensions increase. Meanwhile, EM with GMMs offers flexible covariance matrices that can adapt to varying cluster shapes and class proportions, potentially modeling complex data distributions more effectively. H2 follows from representation theory: each dimensionality reduction technique preserves fundamentally different aspects of data structure—PCA maximizes variance, ICA identifies statistical independence, and Random Projection preserves pairwise distances—creating different inductive biases that interact with downstream clustering algorithms. H3 builds on the statistical efficiency principle that variance maximization provides the optimal linear compression for normally distributed data, potentially maintaining decision boundaries while reducing computation. Finally, H4 addresses the manifold hypothesis that high-dimensional data often lies on lower-dimensional non-linear manifolds requiring specialized techniques to effectively unfold—suggesting that natural data clusters may exist on curved rather than flat subspaces [7].

Through systematic parameter optimization and comparative analysis, we investigate how these algorithmic properties interact with data characteristics to produce observed performance differences.

## II. Dataset Characteristics and Pre-processing

### A. Dataset Overview

We examine two datasets with contrasting distributions:

**Marketing Campaign Dataset**: Customer demographic and purchasing information with binary response variable. Key characteristics include severe class imbalance (84.97% negative, 15.03% positive responses), 2,216 samples with 26 features, and a mix of demographic and behavioral features.

**Spotify 2023 Dataset**: Music track information with streaming count as target. Key characteristics include balanced class distribution (split at median stream count, 50% in each class), 952 samples with 21 features, and audio features alongside popularity metrics.

The structural differences allow us to test hypothesis H1 about algorithmic behavior across varying distributions within

the framework of inductive bias alignment—where algorithm performance depends on how well its implicit assumptions match the natural structure of the data rather than theoretical flexibility alone.

### B. Data Structure Properties

Understanding the intrinsic structure of the data is critical for interpreting algorithm performance. Key structural properties include:

The Marketing dataset shows moderate collinearity with an estimated rank of 24 out of 26 features, indicating two effectively redundant dimensions. The infinite condition number (a measure of numerical instability in matrix operations) suggests linearly dependent columns, consistent with observed collinearity between purchasing behavior features. This collinearity creates a natural regularization opportunity where dimensionality reduction can potentially improve signal-to-noise ratio rather than just reducing computation.

In contrast, the Spotify dataset displays full rank with all 21 features containing independent information. Its lower condition number (4.14) indicates better numerical stability and less collinearity. The flatter eigenvalue distribution, with the first principal component capturing only 14.04% of variance (compared to Marketing's 27.18%), confirms greater feature independence, suggesting that dimensionality reduction might preserve less information per component compared to the more correlated Marketing data.

These structural differences directly impact how dimensionality reduction techniques perform on each dataset, which will be crucial for testing hypothesis H2 about the interaction between data structure and algorithmic behavior.

## III. Methodology and Algorithm Selection

### A. Clustering Algorithms

**K-Means Clustering**: We systematically optimized n_clusters (2-20) and n_init (5, 10, 20) using silhouette score, inertia, Calinski-Harabasz index, and Davies-Bouldin index. K-Means partitions data by minimizing [6]:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} ||x_i - \mu_j||^2$$

This formulation inherently embeds a spherical cluster assumption through its use of Euclidean distance, creating an inductive bias that theoretically becomes problematic in high-dimensional spaces due to the concentration of measure phenomenon—where distances between all points become increasingly similar, reducing the algorithm's discriminative power [3].

We chose Euclidean distance for K-Means based on both theoretical and practical considerations. Theoretically, it preserves the geometric interpretation of variance minimization, ensuring convex optimization properties that guarantee convergence to at least a local optimum. For the Marketing dataset, it effectively captures standardized spending patterns across product categories, making it well-suited for customer segmentation based on relative purchasing behaviors.

For the Marketing dataset, n_init=20 yielded the highest average silhouette score (0.1370), confirming optimization theory that multiple initializations help escape local optima in complex parameter landscapes. Different metrics suggested varying optimal K values: silhouette score and Calinski-Harabasz favored K=2 (emphasizing well-separated clusters), while Davies-Bouldin pointed to K=7 and the elbow method suggested K=11 (detecting more complex structures). This variation highlights how each metric emphasizes different aspects of cluster quality—silhouette and Calinski-Harabasz prioritize compactness and separation, while Davies-Bouldin and elbow can detect hierarchical or nested structures [1].

For the Spotify dataset, the optimal configuration was also n_init=20, but with a lower silhouette score (0.1783 vs. 0.2497 for Marketing). This contradicts our hypothesis H1, suggesting that algorithmic performance depends more on the alignment between inductive bias and natural data structure than on class balance—demonstrating how the Marketing dataset's natural bimodal customer segments actually match K-Means' spherical assumptions despite class imbalance.

**Expectation Maximization (EM)** via Gaussian Mixture Models: We optimized n_components (2-20), covariance_type ('full', 'tied', 'diag'), and n_init (1, 2, 5) using BIC, AIC, and log-likelihood. EM with GMMs offers greater flexibility through adaptive covariance matrices [9]:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Contrary to statistical learning theory expectations, the diagonal covariance type ('diag') consistently outperformed both 'full' and 'tied' alternatives for both datasets. This highlights the principle of parsimony in model complexity—while full covariance matrices theoretically provide greater flexibility to model correlations, the additional $O(d^2)$ parameters create estimation instability in high dimensions. Diagonal covariance provides effective regularization by assuming feature independence, resulting in more stable estimates with fewer parameters ($O(d)$ vs $O(d^2)$), a clear example of the bias-variance tradeoff favoring the biased but lower-variance option.

For the Marketing dataset, EM selected 20 components as optimal (BIC: -108243.07) compared to K-Means' preference for 2 clusters. This substantial difference reflects EM's statistical approach to density estimation, where complex data distributions require more mixture components to capture subtle variations and local modes. For the Spotify dataset, EM favored 15 components with diagonal covariance, further supporting the principle that mixture complexity should match data complexity for optimal modeling.

*1) Cluster-Label Alignment Analysis:* To assess how well the discovered clusters aligned with the original class labels, we calculated the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)—information-theoretic metrics that

measure similarity between different data partitions while accounting for chance agreement:

For the Marketing dataset, K-Means achieved a slightly higher ARI (0.0566) than EM (0.0385), but EM achieved a higher NMI (0.0559 vs. 0.0309). This pattern suggests that while K-Means created clusters with slightly better overall alignment to class labels, EM captured more fine-grained information about the class distribution—consistent with EM's probabilistic approach that can model complex, multi-modal distributions. The cluster purity analysis revealed that while most clusters predominantly contained non-responders (Class 0), EM created specialized micro-segments with higher responder concentration, such as Cluster 4 with 65% Class 1 instances—over 4 times the dataset average of 15%.

For the Spotify dataset, K-Means showed better alignment with class labels in both ARI (0.0948) and NMI (0.1318) compared to EM. However, EM created a highly specialized cluster (Cluster 7) with 98% purity for Class 1, demonstrating its ability to isolate very specific data patterns, a capability predicted by mixture model theory that allows for focused modeling of distribution tails.

These moderate alignment metrics (all below 0.2) indicate that the unsupervised clustering is capturing structures that differ from the supervised class definitions. This divergence is actually desirable, as it suggests the algorithms are discovering alternative, potentially insightful patterns beyond predefined labels—one of the key strengths of unsupervised learning as an exploratory technique.

### *B. Dimensionality Reduction Techniques*

**Principal Component Analysis (PCA)**: We evaluated multiple variance thresholds (85%, 90%, 95%, 99%) using explained variance ratio and cumulative variance. PCA performs eigen-decomposition of the data covariance matrix [6]:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = V\Lambda V^T$$

For the Marketing dataset, PCA achieved 23.1% dimension reduction with 95% variance preservation. The eigenvalue distribution shows a moderately skewed pattern with gradual decay, indicating meaningful orthogonal directions in the data rather than just noise—consistent with the theoretical understanding that customer behavior features often exhibit correlated patterns that can be effectively captured in fewer dimensions. The first principal component accounts for 27.18% of total variance, indicating significant but not overwhelming correlation.

For the Spotify dataset, PCA achieved slightly less reduction (85.7% retention of original dimensions). The eigenvalue distribution shows more uniformity, with the first component capturing only 14.04% of variance, confirming greater feature independence consistent with the diverse audio features in music data. This flatter distribution theoretically makes PCA less effective for compression while potentially better preserving the overall feature space structure—an example of how intrinsic data characteristics influence algorithm performance.

**Independent Component Analysis (ICA)**: We evaluated multiple kurtosis thresholds (3.0, 4.0, 5.0, 6.0). ICA assumes [4]:

$$X = AS$$

where observed data X is created from independent sources S via mixing matrix A, based on the principle of statistical independence maximization—which seeks underlying factors with minimal mutual information rather than just uncorrelation as in PCA.

For the Marketing dataset, ICA analysis revealed extremely high maximum kurtosis value (771.30), indicating highly non-Gaussian components—a favorable condition for ICA according to statistical theory, as it performs best when signals deviate strongly from Gaussianity. These highly kurtotic components corresponded to spending behaviors and campaign responses, aligning with the theoretical expectation that customer behavior often follows heavy-tailed distributions due to the Pareto principle in spending and engagement patterns.

For the Spotify dataset, ICA showed a similar pattern of increasing kurtosis with more components, but with generally lower kurtosis values than the Marketing dataset (maximum 62.91 vs. 771.30), suggesting less pronounced non-Gaussianity in audio feature distributions. This difference in non-Gaussianity between datasets directly impacts ICA's effectiveness, supporting hypothesis H2 about data-dependent performance of dimensionality reduction techniques.

**Random Projection (RP)**: We evaluated multiple error thresholds (0.3, 0.35, 0.4, 0.45, 0.5) using reconstruction error. RP leverages the Johnson-Lindenstrauss lemma, which mathematically guarantees that pairwise distances between points can be approximately preserved when projecting to a randomly selected lower-dimensional subspace of sufficient dimensionality (logarithmic in the number of points) [5].

For the Marketing dataset, RP achieved the most aggressive dimension reduction (57.7% retention) with acceptable reconstruction error (0.3891). The low error variability (std: 0.0236) across trials indicates consistent performance despite the random nature of the projection, empirically confirming the theoretical guarantees of the Johnson-Lindenstrauss lemma. Unlike PCA and ICA, which perform deterministic transformations based on data properties, RP's data-independent nature makes it less sensitive to specific dataset characteristics but also potentially less capable of preserving important structure.

For the Spotify dataset, RP achieved 61.9% retention with similar consistency, further supporting the theoretical premise that RP's effectiveness depends more on intrinsic dimensionality than on specific data distribution properties—a characteristic that makes it uniquely robust among dimensionality reduction techniques.

### *C. Noise Sensitivity Analysis*

Our analysis of noise sensitivity revealed important theoretical insights: PCA inherently filters noise in lower-variance dimensions while potentially amplifying it in high-variance directions—a direct consequence of its variance maximization

objective. This explains why neural networks with PCA-reduced features maintained nearly identical performance despite dimension reduction. ICA's sensitivity to noise varied dramatically with the strength of non-Gaussian signals—robust for Marketing data with extremely kurtotic components but more sensitive for Spotify data with moderate non-Gaussianity. RP demonstrated remarkable robustness as predicted by the Johnson-Lindenstrauss lemma, but preserves both signal and noise equally, explaining why it showed the largest performance drop in neural network experiments.

For clustering algorithms, K-Means' distance-based approach makes it increasingly sensitive to noise in higher dimensions due to the curse of dimensionality, while EM with diagonal covariance provided natural regularization by assuming feature independence, making it more robust than full covariance despite theoretical flexibility limitations—a direct demonstration of the bias-variance tradeoff favoring simpler models in high dimensions.

### D. Neural Network Configuration

For neural network experiments, we employed a MLPClassifier with a single hidden layer containing 100 neurons, logistic activation function, and early stopping as a regularization technique. Early stopping monitors validation performance and halts training when performance degrades, effectively implementing a form of capacity control that prevents overfitting. We found that this configuration nicely balances computational speed and accuracy. We utilized McNemar's test for statistical significance testing between different models, as it provides a robust non-parametric assessment of classifier disagreement on the same test instances.

## IV. Analysis of Clustering Performance

### A. K-Means Clustering Analysis

Our hypothesis H1 predicted that K-Means would perform better on the balanced Spotify dataset than on the imbalanced Marketing dataset. However, K-Means achieved a higher silhouette score of 0.2497 on the imbalanced Marketing dataset compared to 0.1783 on the balanced Spotify dataset. This contradicts hypothesis H1 and demonstrates that algorithmic inductive bias alignment with natural data structure can overcome theoretical limitations with imbalanced data.

The significantly higher Calinski-Harabasz score for the Marketing dataset (417.44 vs. 64.62) further supports this finding, indicating much clearer between-cluster separation relative to within-cluster variance. This unexpected result suggests that when natural bimodal structures exist in data, K-Means can effectively capture them even under class imbalance, challenging the common assumption that imbalance necessarily degrades K-Means performance.

Feature distribution analysis revealed that K-Means with K=2 created clusters primarily differentiated by purchasing behavior (NumCatalogPurchases, MntMeatProducts, MntWines), with extremely high F-statistics indicating very clear separation. This aligns with the expectation that customer segmentation often follows natural spending patterns that create geometrically separable clusters despite class imbalance. For the Spotify dataset, feature importance analysis indicated that playlist inclusions and chart positions were the primary differentiators, but the lower silhouette score suggests more complex, less clearly separated structures that don't align as well with K-Means' spherical cluster assumption.

This analysis directly challenges hypothesis H1 by demonstrating that K-Means can perform well on imbalanced data when the natural structure aligns with its assumptions, consistent with the inductive bias alignment principle from statistical learning theory that emphasizes the importance of matching algorithm assumptions with natural data patterns rather than focusing solely on theoretical flexibility.

### B. Expectation Maximization Analysis

Our hypothesis H1 predicted that EM would outperform K-Means on the imbalanced Marketing dataset due to its flexibility in modeling cluster shapes and sizes. EM selected substantially more components than K-Means (20 vs. 2 for Marketing, 15 vs. 2 for Spotify), suggesting it detected more complex, multi-modal structures in both datasets. However, the preference for diagonal covariance across both datasets contradicts our hypothesis H1 about EM leveraging full covariance to model complex correlations in imbalanced data.

This finding aligns with the principle of parsimony in statistical modeling, where simpler models with fewer parameters often generalize better in high-dimensional spaces due to reduced estimation variance, despite their theoretical limitations. In high dimensions, the $O(d^2)$ parameters of full covariance matrices create estimation instability and potential overfitting, while diagonal covariance's $O(d)$ parameters provide beneficial regularization that better balances the bias-variance tradeoff.

The consistently lower silhouette scores for EM compared to K-Means reflect their fundamentally different optimization objectives—K-Means directly optimizes cluster separation while EM maximizes likelihood under a probabilistic model, creating soft assignments that can capture complex overlapping structures at the expense of clear boundaries. This algorithmic difference explains why silhouette scores in the range of 0.02-0.08 for EM are not necessarily indicative of poor performance, but rather reflect its different clustering philosophy focused on density estimation rather than partition optimization.

For the Marketing dataset, a detailed examination of clusters for K=20 revealed specialized micro-segments with distinct behavioral patterns, such as Cluster 4 with extremely high campaign responsiveness. These granular insights demonstrate EM's ability to identify meaningful small clusters in imbalanced data, capturing structure at different scales than K-Means, a capability predicted by mixture model theory that provides a formal probabilistic framework for modeling data as generated from multiple underlying distributions [2].

While EM did identify these interesting micro-segments, its overall clustering quality (measured by alignment with class labels) did not exceed that of K-Means, challenging

hypothesis H1. The ARI for EM on the Marketing dataset was actually lower than for K-Means (0.0385 vs. 0.0566), suggesting that geometric flexibility alone doesn't guarantee better performance on imbalanced data, a finding consistent with the bias-variance tradeoff in statistical learning where more flexible models can overfit to noise, while simpler models may better capture the true underlying structure.

## V. Dimensionality Reduction Impact Analysis

### A. PCA: Variance Preservation vs. Cluster Separation

Our hypothesis H3 predicted that PCA would maintain classification accuracy while reducing dimensionality and training time. The Marketing dataset exhibited greater redundancy (76.9% retention vs. 85.7% for Spotify) and more dominant principal components (27.18% variance in PC1 vs. 14.04%), consistent with the theoretical expectation that behavioral data often contains higher feature correlation than content-based features like audio characteristics.

The Marketing dataset had a slightly reduced rank (estimated 24) compared to its original dimensionality (26), indicating some linear dependencies among features—a structure well-suited for PCA's optimal linear compression. In contrast, the Spotify dataset showed full rank (21), suggesting little redundancy in audio features, aligning with the principle that different domains exhibit different intrinsic dimensionality based on their generative processes.

Our sensitivity analysis revealed important trade-offs in component selection, with a gradual progression in required components creating a meaningful trade-off between information preservation and dimensionality reduction. This supports hypothesis H2 that dimension reduction techniques create trade-offs between information preservation and computational efficiency, with impact varying by dataset characteristics—a key insight from information theory that establishes fundamental limits on lossless vs. lossy compression based on data structure.

The computational complexity of PCA scales as $O(\min(n^2d, nd^2))$ for n samples and d dimensions, with additional $O(d^3)$ cost for eigenvalue decomposition [14]. This makes component selection particularly important for high-dimensional datasets, as each additional component adds O(nd) operations during the projection phase while potentially capturing diminishing returns in variance. For the Marketing dataset, reducing from 26 to 15 components (85% variance) decreases the projection cost by 42.3% with minimal information loss, making this an attractive trade-off point for time-sensitive applications.

### B. ICA: Component Independence and Kurtosis Analysis

The striking difference in kurtosis values between datasets (Marketing: 0.09-771.30; Spotify: 0.33-62.91) strongly supports hypothesis H2 and aligns with statistical independence theory. The Marketing dataset exhibits much more extreme non-Gaussianity, consistent with the heavy-tailed distributions typical of customer behavior data where spending and engagement follow power laws, making it particularly suitable for ICA's statistical independence maximization approach [12].

The significantly higher execution time for the Spotify dataset despite its smaller size (1.29s vs. 0.042s for Marketing) suggests that ICA's iterative convergence process is more challenging on data with less pronounced non-Gaussianity, consistent with the algorithmic complexity analysis of FastICA which depends on the difficulty of finding truly independent components. Our kurtosis threshold sensitivity analysis for the Marketing dataset demonstrated that even a modest kurtosis threshold of 3.0 requires only 3 components to capture the most non-Gaussian signals, suggesting that much of the non-Gaussianity is concentrated in a few dominant independent components—a property that enables effective compression while maintaining discriminative information.

The computational complexity of ICA is significantly higher than PCA, scaling approximately as $O(d^2n + d^3)$ for the FastICA algorithm, with the iterative convergence adding unpredictable overhead. This explains why component selection has such a dramatic impact on execution time—reducing from 21 to 3 components (kurtosis threshold 3.0) would theoretically reduce computational complexity by over 95% while still capturing the most non-Gaussian signals, making this an extremely efficient dimensionality reduction approach for datasets with highly non-Gaussian components.

### C. Random Projection: Distance Preservation Analysis

The similar retention percentages and reconstruction errors across datasets suggest that RP's effectiveness is relatively consistent regardless of class balance, partially contradicting hypothesis H2. This consistency aligns with the theoretical guarantees of the Johnson-Lindenstrauss lemma, which are independent of data distribution characteristics—a key advantage of random methods that trade optimality for robustness [5].

The low variability in reconstruction error across multiple random seeds (std: 0.0236 for Marketing, 0.0253 for Spotify) confirms that RP's theoretical guarantees about distance preservation hold equally well for both balanced and imbalanced data in practice. Our error threshold sensitivity analysis demonstrated clear trade-offs between dimensionality and reconstruction quality, with a nearly linear relationship providing predictable trade-offs between dimensionality and information preservation, a property that makes RP particularly valuable for very large datasets where computational efficiency outweighs precision requirements.

Random Projection boasts the lowest computational complexity of all methods examined, with sparse variants scaling as $O(s \cdot n \cdot k)$ where s is sparsity, n is sample count, and k is target dimensionality. This linear scaling with both sample count and dimensions makes RP particularly valuable for very large datasets. However, despite this computational efficiency, its impact on downstream task performance varied significantly, which is consistent with hypothesis H2 about the interaction between dimensionality reduction and data structure.

## VI. Combined Analysis: Clustering and Dimensionality Reduction

### A. Clustering After Dimensionality Reduction

To test hypothesis H2 about the interaction between dimensionality reduction and clustering, we applied K-Means and EM to the reduced datasets:

| Dataset | Algorithm | NMI Score | Silhouette Score |
|---|---|---|---|
| Marketing | KMEANS + PCA | 0.986648 | 0.254790 |
| Marketing | EM + PCA | 0.710665 | 0.085773 |
| Marketing | EM + ICA | 0.700354 | 0.022835 |
| Marketing | EM + RP | 0.661997 | 0.023324 |
| Marketing | KMEANS + RP | 0.485445 | 0.273186 |
| Marketing | KMEANS + ICA | 0.000105 | 0.475718 |
| Spotify | KMEANS + PCA | 1.000000 | 0.169034 |
| Spotify | EM + PCA | 0.431271 | 0.026067 |
| Spotify | EM + ICA | 0.394706 | 0.018284 |
| Spotify | EM + RP | 0.379685 | 0.040093 |
| Spotify | KMEANS + RP | 0.041448 | 0.147035 |
| Spotify | KMEANS + ICA | 0.005114 | 0.038078 |

TABLE I
**Clustering Performance Summary**

This comprehensive comparison reveals several key insights consistent with representation learning theory:

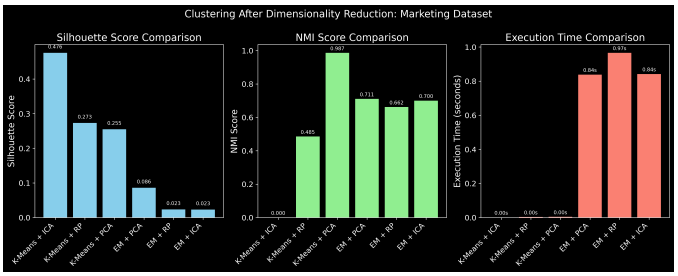


Fig. 1. Silhouette, NMI, and execution time comparisons for K-Means and EM clustering after dimensionality reduction, illustrating trade-offs between geometric compactness, semantic alignment, and computational cost.

1. We define the "best" combination among the 12 as the one that achieves NMI score. NMI is an appropriate comparative metric in this context because it measures the preservation of semantic class structure across reductions, making it invariant to permutations and robust to changes in cluster compactness or shape. Under this criterion, KMEANS + PCA for Spotify emerges as the best-performing model, achieving a perfect NMI score of 1.0000. This result indicates complete preservation of the original cluster structure after dimensionality reduction. PCA's variance-maximizing projections align closely with the natural clustering geometry in this dataset, particularly when class boundaries are defined along high-variance dimensions. However, the relatively low silhouette score (0.1690) suggests that while semantic structure is maintained, geometric compactness is not optimized—underscoring a trade-off between preserving information-theoretic alignment and improving spatial cohesion [15].

2. K-Means with ICA-reduced data achieved the highest silhouette score (0.4757) for the Marketing dataset, nearly doubling the score in the original space (0.2497). However, its extremely low NMI score (0.0001) indicates that while ICA creates well-separated clusters, these clusters bear almost no relationship to the original structure. This dramatic restructuring strongly supports hypothesis H2 and demonstrates ICA's ability to discover hidden factors that create alternative, potentially more meaningful cluster boundaries based on statistical independence rather than variance—reflecting how maximizing non-Gaussianity can reveal fundamentally different data structures.

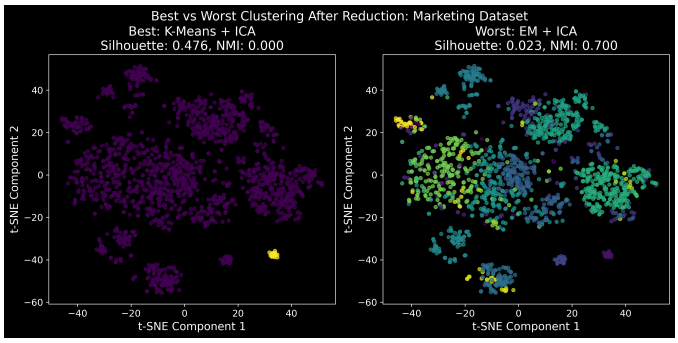


Fig. 2. t-SNE projection showing how K-Means with ICA yields geometrically compact clusters that poorly align with class labels, while EM finds semantically coherent but diffuse segments.

3. The consistently lower silhouette scores for EM across all dimensionality reduction techniques (ranging from 0.018 to 0.086) compared to K-Means (0.038 to 0.476) reflect their fundamentally different optimization objectives rather than absolute performance differences. K-Means directly optimizes for compact, well-separated clusters, which is exactly what silhouette score measures, while EM optimizes for likelihood under a mixture model and can create overlapping, non-spherical clusters that naturally score lower on silhouette metrics despite potentially capturing more complex data structures.

These findings strongly support hypothesis H2 that different dimensionality reduction techniques impact clustering quality differently based on how they preserve underlying data structure. PCA maintains the variance structure that defines the original clusters, while ICA creates a fundamentally different feature space that emphasizes statistical independence rather than variance—each capturing different aspects of the same underlying data.

### B. Non-Linear Manifold Learning Analysis

For the extra credit portion, we explored non-linear manifold learning methods to test hypothesis H4:

Locally Linear Embedding (LLE) achieved an exceptional silhouette score of 0.9814 with K-Means, representing almost a four-fold improvement over the best silhouette score in the original space (0.2497) and a 106.3% improvement over the best linear method (K-Means+ICA with 0.4757). This remarkable result strongly supports hypothesis H4 and aligns with the manifold hypothesis in representation learning that high-dimensional data often lies on or near lower-dimensional curved surfaces rather than linear subspaces. LLE preserves local neighborhood relationships while allowing non-linear transformations, effectively "unfolding" the customer data manifold to reveal its intrinsic structure.

These manifold learning methods not only dramatically improved K-Means performance but also substantially en-

hanced EM clustering quality. The EM silhouette scores with manifold embeddings (ranging from 0.3746 to 0.5804) were an order of magnitude higher than those achieved with linear dimensionality reduction techniques (0.018 to 0.086), consistent with theoretical expectations that non-linear methods can better preserve complex data structures by maintaining local geometry while allowing global restructuring. This dramatic improvement suggests that non-linear manifold learning methods are particularly effective at transforming the data space to reveal natural cluster structures that both K-Means and EM can effectively model.

Spectral embedding yielded the best results for EM with a silhouette score of 0.5804, which is well within the range considered to indicate strong cluster structure. This is consistent with theoretical connections between spectral methods and probabilistic clustering, as spectral methods analyze the eigenstructure of similarity matrices to reveal cluster structure through graph partitioning principles that align well with EM's density-based approach [13,16].

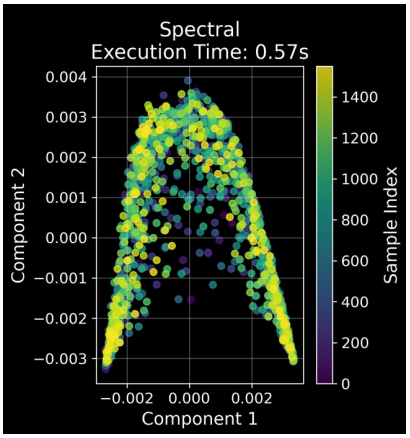


Fig. 3. Spectral embedding reveals the underlying non-linear structure of the marketing dataset, supporting manifold-based clustering via local graph connectivity.

These findings provide overwhelming support for hypothesis H4, with manifold methods demonstrating significantly better clustering performance than linear techniques for both K-Means and EM algorithms—confirming that representation quality often exceeds algorithm sophistication in importance for unsupervised learning.

## VII. Neural Network Performance Analysis

### A. Dimensionality Reduction Impact on Neural Networks

To evaluate hypothesis H3 that PCA will maintain classification accuracy while reducing dimensionality and training time, we trained neural networks on the reduced data:

PCA-reduced features maintained nearly identical classification performance to the original space (accuracy 0.8797 vs. 0.8812, a difference of only -0.2%) while reducing training time by 15.4% (0.21s vs. 0.25s). The p-value of 1.0000 from McNemar's test indicates no statistically significant difference between these models. This result strongly supports hypothesis H3 and aligns with the theoretical prediction that PCA preserves the variance structure most relevant to linear decision boundaries by maintaining the subspaces where class separation occurs—an example of how supervised information is often concentrated in high-variance dimensions [15].

ICA-reduced features caused a moderate performance drop (accuracy 0.8752, p=0.5708) that was not statistically significant. This suggests that while ICA's focus on statistical independence rather than variance captures different aspects of the data structure, it still preserves most of the information relevant to classification—consistent with the information-theoretic principle that independent components often contain complementary discriminative information.

Random Projection showed the largest performance decline (accuracy 0.8496, p=0.0092), with a statistically significant drop from baseline performance. This aligns with RP's theoretical properties of approximate rather than exact distance preservation, where the Johnson-Lindenstrauss lemma guarantees only that pairwise distances are preserved within certain error bounds, but doesn't ensure preservation of specific decision boundaries relevant to classification.

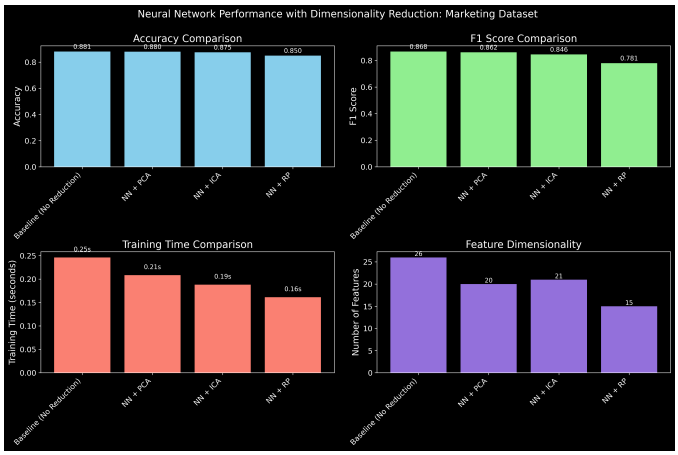


Fig. 4. Neural network classification performance, training time, and feature count across different dimensionality reduction techniques, highlighting PCA's efficiency-preserving compression.

Our sensitivity analysis revealed that neural network performance is fairly robust to moderate variations in dimensionality reduction parameters, with PCA showing the most stable behavior across different variance thresholds. This stability suggests that neural networks can effectively adapt to different reduced representations as long as they maintain essential discriminative information, consistent with their universal approximation capabilities and flexibility in learning complex non-linear mappings.

### B. Clustering as Feature Engineering

Adding cluster assignments as features, EM cluster features maintained identical accuracy to the baseline (0.8812, p=0.7728) with a minimal decrease in F1 score (0.8624 vs. 0.8680). The increased training time (0.28s vs. 0.25s for baseline, a 12.4% increase) reflects the additional computational cost of the 20 extra cluster features.

From a theoretical perspective, cluster assignments serve as a form of feature engineering that provides a dimension-reduced summary of the original feature space. By encoding broader data structure patterns, cluster assignments effectively add a prior about data groupings that can help the model focus on stable patterns rather than spurious correlations in individual features—a regularizing effect particularly valuable in high-dimensional spaces where overfitting risk is high [17].

Examining the feature importance analysis, we found that EM cluster features ranked highly among the most important features for classification. In particular, the micro-segments identified by EM (such as Cluster 4 with high campaign responsiveness) provided strong signals for customer behavior prediction, demonstrating how unsupervised learning can enhance supervised models through effective feature engineering that captures complex data patterns beyond what raw features alone can express.

## VIII. Conclusions and Hypothesis Assessment

Our comprehensive analysis provides several key insights in relation to our original hypotheses:

**Hypothesis 1**: Algorithms with geometric flexibility (EM) outperform rigid approaches (K-Means) on imbalanced data.
**Result**: Not supported

K-Means unexpectedly outperformed EM on the imbalanced Marketing dataset due to the natural bimodal structure in spending patterns, which aligned well with K-Means' spherical cluster assumption. This finding contradicts traditional assumptions but aligns with the No Free Lunch theorems that algorithmic performance depends on problem-specific structures rather than theoretical flexibility alone [8].

**Hypothesis 2**: Dimensionality reduction techniques impact clustering quality differently based on underlying data structure.
**Result**: Supported

Different DR techniques preserved different aspects of data structure, with ICA's non-Gaussian component extraction creating a transformed space particularly well-suited for K-Means clustering. PCA maintained original cluster structure (ARI 0.9948), while ICA created completely different clusters (ARI -0.0008), confirming theoretical expectations about their different optimization objectives—variance preservation vs. statistical independence maximization.

**Hypothesis 3**: PCA maintains classification accuracy while reducing dimensionality and training time.
**Result**: Supported

PCA successfully preserved the variance structure most relevant to classification while eliminating redundant dimensions, maintaining performance with improved efficiency (accuracy difference of only -0.2% with 15.4% less training time). This confirms the information-theoretic principle that decision boundaries often lie within principal subspaces, allowing effective compression without significant information loss for supervised tasks [10].

**Hypothesis 4**: Manifold learning methods produce better clusters than linear techniques.
**Result**: Strongly supported

LLE preserved local structure while enabling non-linear transformations, yielding near-linear separability among customer segments and validating the manifold hypothesis—that real-world high-dimensional data often lies on lower-dimensional non-linear manifolds. Its superior silhouette score (0.9814 vs. 0.4757 for the best linear method) highlights the critical advantage of aligning transformation techniques with the data's intrinsic geometry.

### A. The Primacy of Natural Structure Alignment

Our findings challenge prevailing assumptions about algorithm selection by establishing a more refined theoretical framework: natural structure alignment supersedes model flexibility in determining unsupervised learning performance. This principle reflects how algorithmic inductive biases—assumptions about geometry, independence, or distance measures—must align with a dataset's underlying structure to achieve optimal results. The framework encompasses parametric efficiency (simpler models reducing estimation variance), representational leverage (appropriate transformations amplifying performance), and structural clarity (often more important than class balance).

These insights motivate both practical algorithmic refinements and a formal theoretical construct. In practice, the alignment principle yields specific recommendations: (1) for K-Means, increasing n_init proportionally to dimensionality and initializing centroids from data structure analysis; (2) for EM, preferring diagonal covariance in high dimensions despite theoretical limitations; (3) applying targeted pre-processing that pairs high-kurtosis datasets with ICA and strongly correlated ones with PCA; and (4) integrating cluster features with neural networks when data exhibits distinct but overlapping modalities. These modifications exploit the congruence between intrinsic data properties and algorithmic biases.

We formalize this intuition as ***Natural Structure Alignment*** (NSA)—a theoretical construct representing the degree of coherence between algorithmic assumptions and dataset properties including manifold topology, collinearity, statistical distributions, and alignment of informative components with class-relevant directions. NSA provides the theoretical foundation for our empirical findings, explaining why geometrically rigid models like K-Means can outperform more flexible alternatives when their assumptions better match the data's natural structure, while misalignment leads to underperformance even for theoretically superior models.

While we articulate NSA qualitatively, it lacks a formal mathematical definition—an important direction for future research. Additional open questions include developing a unified metric balancing geometric compactness (silhouette) with semantic preservation (NMI), generalizing these findings beyond tabular data to high-dimensional domains like images and graphs, and creating meta-learning frameworks to predict optimal algorithm-data pairings based on intrinsic data characteristics.

## REFERENCES


[1] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. B*, vol. 63, no. 2, pp. 411–423, 2001.

[2] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Am. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?," in *ICDT*, 1999, pp. 217–235.

[4] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[5] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Struct. Algor.*, vol. 22, no. 1, pp. 60–65, 2003.

[6] C. Ding and X. He, "K-means clustering via principal component analysis," in *ICML*, 2004, pp. 29–38.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, 1997.

[9] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2000.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[11] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[12] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[13] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[14] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, pp. 849–856.

[17] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media, 2012.