

# Pseudo-Boolean Polynomial Method for Interpretable Dimensionality Reduction: A Paradigm Shift from Abstract to Meaningful Feature Extraction

T.M. Chikake<sup>1</sup>, B.I. Goldengorin<sup>1,2</sup>, P.M. Pardalos<sup>3,4</sup>

<sup>1</sup>Department of Discrete Mathematics, Phystech School of Applied Mathematics and Informatics, Moscow Institute of Physics and Technology, Institutsky lane 9, Dolgoprudny, 141700, Russia;

<sup>2</sup>The Scientific and Educational Mathematical Center "Sofia Kovalevskaya Northwestern Center for Mathematical Research" in Pskov State University, Sovetskaya Ulitsa, 21, Pskov, 180000, Russia;

<sup>3</sup>Center for Applied Optimization, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA;

<sup>4</sup>LATNA, National Research University Higher School of Economics, 20 Myasnitskaya Street, Moscow 101000, Russia

## Abstract

We present a general-purpose, training-free framework for dimensionality reduction and clustering based on per-sample pseudo-Boolean polynomials (PBP). The method constructs compact, interpretable features without model fitting and is evaluated under a standardized protocol that compares PBP to PCA, t-SNE, and UMAP using identical inputs and metrics: clustering alignment (V-measure, Adjusted Rand Index), cluster geometry (Silhouette coefficient, Calinski-Harabasz index, Davies-Bouldin index), and supervised probes (linear separability and boundary complexity (1-NN error)). Across 11 diverse datasets spanning tabular, signal, and ecological domains, PBP leads on linear separability in 5/11 datasets and achieves lower boundary complexity in 2/11 datasets, while remaining competitive on clustering metrics. We report best-performing aggregation and sorting configurations per dataset and provide guidance on when PBP should be preferred for interpretable analysis and reproducible evaluation.

**Keywords:** dimensionality reduction, pseudo-Boolean polynomials, clustering, interpretable features, sample independence, feature selection

**Citation:** Chikake TM, Goldengorin BI, Pardalos PM. Pseudo-Boolean polynomial method for Interpretable dimensionality reduction. *Computer Optics* 2025; 49(6): 1191-1201. DOI: 10.18287/COJ1815.

## Introduction

Dimensionality reduction techniques play a crucial role in modern data science, enabling the analysis of high-dimensional data through low-dimensional representations that preserve meaningful relationships. Traditional approaches such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) have become standard tools in exploratory data analysis and machine learning pipelines. However, these methods often suffer from limitations including lack of interpretability, sensitivity to initialization, and population distribution dependencies.

In our 11-dataset evaluation, PBP leads on linear separability in 5/11 datasets and achieves lower boundary complexity in 2/11 datasets while remaining competitive on clustering metrics.

## 1. Related Work

PCA [23] remains the most widely used linear dimensionality reduction method, transforming data to a new coordinate system where the greatest variance lies on the first coordinate. However, PCA's linear nature limits its ability to capture complex non-linear relationships in data.

t-SNE [32] addresses this limitation through non-linear dimensionality reduction, converting similarities between data points to joint probabilities and minimizing the Kullback-Leibler divergence. However, t-SNE's non-convex cost function leads to different results with different initializations, limiting reproducibility and quality of data analysis.

UMAP [25] has emerged as a powerful alternative, combining the benefits of non-linear dimensionality reduction with improved computational efficiency. UMAP constructs a high dimensional graph representation of the data and optimizes the low-dimensional representation to be as similar as possible.

Beyond these classics, kernel and manifold methods such as kernel PCA [28] and survey treatments [27] position dimensionality reduction within broader representation learning trends. Recent advances in deep representation learning (e.g., autoencoders and contrastive objectives) have further expanded the toolbox; for an overview see [2, 16]. Our goal in this paper is not to compete with end-to-end deep learners but to provide a deterministic, training-free alternative with strong interpretability and standardized evaluation.

Background on pseudo-Boolean polynomials spans the multilinear representation and reduction/equivalence properties [8, 9, 10], quadratization and structural results [11], and penalty-based aggregation [12] with applications to clustering and industrial cell formation [13].

### PBPs in optimization tasks.

Pseudo-Boolean polynomials are a canonical vehicle for discrete optimization, with objective functions represented in multilinear form and optimized under Boolean constraints [7]. Classical results show that higher-order terms can be reduced to quadratic energies via quadratization while preserving minimizers [11, 3], enabling the use of efficient max-flow/min-cut based solvers [14] and related graph-cut methods [9]. In computer vision and machine learning, PBPs have modeled segmentation, labeling [5, 6], and higher-order regularization [12]. In operations research, penalty-based PBPs provide compact formulations for facility location and p-median style problems [21, 22, 23, 24] with powerful reduction and equivalence properties [8, 9, 10, 12, 13], including applications to financial market graph analysis where p-median models enable exact solutions for markets with up to 1,000 stocks [25]. Our use departs from optimization-driven fitting: we leverage the deterministic PBP structure to construct per-sample, interpretable features without solving a global optimization problem.

Despite their widespread adoption, traditional dimensionality reduction methods suffer from several limitations. The reduced features often lack clear interpretation, making it difficult to understand the underlying data structure. Methods like UMAP and t-SNE rely on global optimization, making them sensitive to the overall data distribution and outliers. Non-convex optimization in t-SNE and UMAP leads to different results with different initializations. Traditional methods do not automatically identify and remove noise features.

Pseudo-Boolean polynomials have been extensively studied in optimization theory [4], particularly in the context of p-median problems [10]. The penalty-based formulation described in [1] provides the mathematical foundation for our dimensionality reduction approach.

The key insight from this work is that pseudo-Boolean polynomials can be uniquely represented as multilinear polynomials, enabling compact representations of large problems through reduction and equivalence properties. These properties form the basis of our dimensionality reduction method.

## 2. Methods

### 2.1. Mathematical foundations and formulation

A pseudo-Boolean function is  $f: \mathbf{B}^n \rightarrow \mathbb{R}$  with  $\mathbf{B} = \{0,1\}$ . We adopt the penalty-based aggregation of [1], and use the multilinear representation [4]:

$$f(\mathbf{y}) = \sum_{S \subseteq I} c_S \prod_{i \in S} y_i, \quad (1)$$

constructed per-sample from an input matrix  $C \in \mathbb{R}^{m \times n}$  via column sorting, difference encoding, and term aggregation. This yields compact, per-sample PBP vectors suitable for clustering and probing.

Throughout, we write  $C = (c_{ij}) \in \mathbb{R}^{m \times n}$  for the per-sample construction matrix with measurement index set  $I = \{1, \dots, m\}$  and feature index set  $J = \{1, \dots, n\}$ . We standardize to  $c_{ij}$  (no comma) for matrix entries and reserve  $c_S$  for polynomial coefficients associated with monomial index sets  $S \subseteq I$ .

We distinguish between: (i) the measurement dimension  $m = |I|$  governing the maximum monomial order, and (ii) the visualization target dimension  $p \in \{1,2,3\}$  used only for plotting when desired. Dimensionality reduction in PBP arises from sparsity of the aggregated coefficient vector (at most  $2^m - 1$  non-zero terms and typically far fewer), not from choosing  $p = |I|$ ; for visualization, we project to  $p \leq 3$  without altering the Euclidean geometry used for clustering.

### 2.2. Determinism and sample independence

We formalize two properties used throughout.

**Determinism.** Fixing the configuration (aggregation rule, sorting scheme, and encoding), the PBP vector  $\phi(x; C) \in \mathbb{R}^d$  computed for a sample  $x$  is a deterministic function of  $x$  and  $C$ ; no randomized step is used.

**Sample independence.** For each sample  $x$ , the construction of  $\phi(x; C)$  depends only on  $x$  (the sample's measurement vector) and fixed configuration parameters; it does not depend on the values of other samples. Consequently, adding, removing, or re-ordering other samples does not change  $\phi(x; C)$ .

Sketch justification: the per-sample column sorting, difference encoding, and aggregation are applied to the sample's own measurement vector with no statistics computed over the population. Hence both properties follow by construction.

### 2.3. Mathematical foundations and uniqueness conditions

Following the theoretical framework established by [1] and [10], we provide formal statements of the key properties and uniqueness conditions underlying our PBP construction.

#### 2.3.1. Uniqueness of Multilinear Representation

**Theorem 1 (Uniqueness of Multilinear Form [4]).** Any pseudo-Boolean function  $f: \mathbf{B}^m \rightarrow \mathbb{R}$  can be uniquely represented as a multilinear polynomial of the form:

$$f(\mathbf{y}) = \sum_{S \subseteq I} c_S \prod_{i \in S} y_i, \quad (2)$$

where  $\mathbf{y}$  related to  $S \subseteq I = \{1,2, \dots, m\}$  and  $c_S \in \mathbb{R}$  are uniquely determined coefficients.

**Proof sketch:** The uniqueness follows from the fact that  $y_i^2 = y_i$  for Boolean variables, ensuring that each variable appears at most once in any monomial. The coefficients  $c_S$  are uniquely determined by evaluating  $f$  on all  $2^m$  possible Boolean vectors and solving the resulting linear system.

### 2.3.2. Reduction, Truncation, and Equivalence Properties

**Definition 1 (Similar Monomials).** Two monomials  $c_S \prod_{i \in S} y_i$  and  $c_{S'} \prod_{i \in S'} y_i$  are called similar if  $S = S'$  [1, 10].

**Theorem 2 (Reduction Property).** Similar monomials can be combined through aggregation: if  $S = S'$ , then  $c_S \prod_{i \in S} y_i + c_{S'} \prod_{i \in S'} y_i = (c_S + c_{S'}) \prod_{i \in S} y_i$ .

*Example.*  $4 y_1 y_2$  and  $2 y_1 y_2$  are similar and aggregate to  $6 y_1 y_2$ .

**Theorem 2.1 (Truncation Property of PBP).** To minimize Eq. (1) with  $|S| = p$ , the highest degree of any monomial in Eq. (1) is at most  $m - p$ . In this paper we adopt this truncation property as an assumption for analysis and configuration selection.

**Theorem 3 (Equivalence Property).** Two input matrices  $C_1$  and  $C_2$  yield identical pseudo-Boolean polynomials if and only if their *reduced monomial coefficient maps* coincide (same monomial supports with identical coefficients after aggregation) [1, 10]. Equality of sorted-difference multisets is a sufficient (but not necessary) condition for this equality.

### 2.3.3. Invariance Properties

**Theorem 4 (Equivariance under Row Relabeling).** The PBP construction is equivariant to row permutations: if  $C'$  is obtained from  $C$  by permuting rows, then the resulting polynomial is identical *up to a consistent renaming of variables*  $y_i$ . Consequently, the aggregated coefficient multiset is invariant under row relabeling.

**Theorem 5 (Sample Independence).** For each sample  $x$ , the construction of  $\phi(x; C)$  depends only on  $x$  and fixed configuration parameters, not on other samples in the dataset. This independence ensures that adding, removing, or re-ordering other samples does not change  $\phi(x; C)$ .

**Proof sketch:** The per-sample column sorting, difference encoding, and aggregation are applied to the sample's own measurement vector with no statistics computed over the population. Hence both properties follow by construction.

### 2.3.4. Computational Complexity Guarantees

**Theorem 6 (Polynomial Complexity).** The integer-encoded PBP construction achieves  $O(mn \log(mn) + m^2 + n \cdot m \log m)$  time complexity, where  $m$  is the measurement dimension and  $n$  is the feature dimension.

**Proof sketch:** The complexity breakdown is:  $O(n \cdot m \log m)$  for column-wise sorting,  $O(m^2)$  for binary encoding operations, and  $O(mn \log(mn))$  for monomial aggregation. With typical  $m \in [2, 4]$ , the effective complexity is  $O(n \cdot m \log m)$ .

### 2.3.5. Dimensionality Reduction Guarantees

**Theorem 7 (Compression Property).** The PBP representation contains at most  $2^{m-1} - 1$  non-zero monomials, achieving substantial compression compared to the original  $m \times n$  matrix [1, 10].

**Theorem 8 (Preservation of Structural Information).** The PBP construction preserves essential structural relationships while eliminating redundant information, enabling effective clustering and classification in the reduced space.

### 2.3.6. Assumptions and Conditions

Our theoretical framework relies on the following assumptions:

1. **Non-negativity:** All entries of the input matrix  $C$  are non-negative and finite.
2. **Finite Dimensionality:** The measurement dimension  $m$  is bounded and typically small ( $m \leq 4$  for practical applications).
3. **Deterministic Configuration:** Aggregation rules, sorting schemes, and encoding methods are fixed and deterministic.
4. **Boolean Domain:** Variables  $y_i$  are restricted to the Boolean domain  $\{0, 1\}$ .
5. **Truncation Property:** When minimizing Eq. (1) with  $|S| = p$ , we assume truncation of monomial degrees at  $m - p$  as in Theorem 2.1.

These assumptions ensure the mathematical rigor and computational tractability of our approach while maintaining practical applicability across diverse data analysis contexts.

## 2.4. Complexity and practical efficiency

End-to-end construction runs in  $O(mn \log(mn) + m^2 + n \cdot m \log m)$ :  $O(n \cdot m \log m)$  for sorting,  $O(m^2)$  for binary encoding, and  $O(mn \log(mn))$  for aggregation. With typical  $m \in [2, 4]$ , runtime is effectively  $O(n \cdot m \log m)$  and scales linearly with sample size. The deterministic, single-pass pipeline ensures reproducibility.

## 2.5. Detailed Algorithmic Framework

To achieve practical implementation of pseudo-Boolean polynomials for large-scale data analysis, we present a domain-agnostic computational mechanism that leverages integer encoding of Boolean variables, column-wise sorting, coefficient differencing, and aggregation of identical monomials including their truncations.

### 2.5.1. Algorithmic Framework

Let  $C \in \mathbb{R}^{m \times n}$  be the input matrix. The computational mechanism proceeds through four fundamental steps:

**Step 1: Permutation Encoding.** For each column  $j$  of matrix  $C$ , we determine a permutation  $\pi_j$  that sorts the elements in ascending order. Let  $\Pi \in \mathbb{N}^{m \times n}$  be the permutation matrix whose  $j$ -th column is  $\pi_j$ . This permutation matrix serves as the foundation for encoding Boolean variables.

**Step 2: Boolean Variable Representation.** We introduce Boolean variables  $y = \{y_{ik} | 1 \leq i \leq m, 1 \leq k \leq m\}$ , where each  $y_{ik}$  is associated with row  $i$  and position  $k$  in the permutation matrix. These variables are encoded into integers using a binary scheme: for position  $v$  in a column permutation, we define  $\text{bin\_encoder}(v, \ell)$  which converts index  $v$  to an  $\ell$ -bit binary representation (with  $\ell = m$ ) using little-endian bit ordering.

**Step 3: Coefficient Matrix Generation.** We form a coefficient matrix  $\text{Coeffs} \in \mathbb{R}^{(m-1) \times n}$  based on the sorted matrix  $C_\pi$  (obtained by sorting columns of  $C$  according to  $\Pi$ ). For each column  $j$ , coefficients are computed as differences between consecutive sorted values, effectively capturing relative changes across permuted rows.

**Step 4: Polynomial Construction and Reduction.** We construct the pseudo-Boolean polynomial  $P(y)$  as a sum of terms, each being a product of a coefficient from  $\text{Coeffs}$  and a combination of Boolean variables encoded from  $\Pi$ . The polynomial is reduced by combining terms with identical variable combinations and summing their coefficients.

### 2.5.2. Detailed Algorithm

**Algorithm 1:** CreatePBP: Integer-Encoded Construction from Matrix  $C$

---

**Input:**  $C \in \mathbb{R}^{m \times n}$   
**Output:** Reduced terms as pairs  $(y, \text{coeff})$  with degree  $\text{popcount}(y)$

```

Π ← argsort(C); // Column-wise indices
Csorted ← take along axis(C, Π);
coeffs ← stack([Csorted, 0]) - stack([0, Csorted]);
Drop last row from coeffs;
Y ← ∅;
for  $i = 1$  to  $m - 1$  do
     $b_i$  ← ToInt(BitArray(m) with bit at  $\Pi[i]$  set); // Per column
     $Y[i]$  ←  $\sum_{j=1}^i b_j$ ; // Cumulative logical OR
Group identical  $Y$  values and sum their corresponding coefficients;
return grouped terms with degrees via popcount

```

---

### 2.5.3. Correctness Rationale

The algorithm's correctness stems from the mathematical foundations of penalty-based pseudo-Boolean formulations. Column-wise permutations induce an ordered chain of fallback options, while differencing between consecutive sorted entries yields incremental penalties as defined in the penalty-based formulation. One-hot integer encodings of permutation positions correspond to Boolean variables  $y_i$ , and cumulative sums implement inclusion of prior options (logical OR in the bit domain). Grouping identical monomials across columns produces the reduced multilinear representation consistent with the mathematical framework.

### 2.5.4. Computational Properties

The integer encoding mechanism exhibits four key computational properties:

**Compression:** Aggregation reduces the number of distinct monomials relative to raw matrix size, often substantially. The degree of compression depends on the structural regularity within the input data.

**Equivalence Detection:** Different matrices can reduce to identical pseudo-Boolean polynomials when their sorted differences and induced monomials coincide, providing a principled notion of structural equivalence.

**Degree Computation:** The polynomial degree equals the maximum popcount (number of set bits) over encoded monomials. This degree metric informs structural complexity and guides algorithm selection for downstream optimization.

**Dependency Analysis:** Bitwise AND operations between encoded monomials reveal shared variables and hierarchical relations, useful for transformations such as quadratization and structural analysis.

#### Implementation note (encoding).

We implement binary encodings using fixed-width integer bit operations for portability and speed in mainstream runtimes; this avoids variability of high-level bit-array libraries and enables constant-time masking and population-count operations [17, 13]. Empirically, this reduces memory footprint and improves throughput in our prototypes.

### 2.6. Distance and clustering

We use Euclidean distance between PBP vectors for similarity, supported by reduction/equivalence properties [10]. Clustering quality is evaluated by: V-measure (V), Adjusted Rand Index (ARI), Silhouette coefficient (Silhouette),

Calinski–Harabasz index (CH), and Davies–Bouldin index (DB). Supervised probes: Linear separability (LinearSep; LinearSVC accuracy under a linear decision function), k-nearest neighbors (k-NN) accuracy with  $k = 5$  (5-NN), logistic-regression margin score (margin), and boundary complexity measured as 1-nearest neighbor (1-NN) error (abbreviated as BoundC in tables). For consistency, we report k-NN as accuracy for any  $k$ , and explicitly state when it is an error metric (only for boundary complexity via 1-NN).

Let  $\mathcal{M}$  denote the ordered set of monomials retained after aggregation and let  $\phi(x), \phi(x') \in \mathbb{R}^{|\mathcal{M}|}$  be the aligned coefficient vectors for two samples. We define the similarity metric as

$$d(x, x') = \| \phi(x) - \phi(x') \|_2, \text{ (Euclidean distance between coefficient vectors).} \tag{3}$$

This is the standard  $\ell_2$  distance on the coefficient space; alignment ensures that coefficients correspond to the same monomial in  $\mathcal{M}$ .

### 2.7. Dimensionality reduction characteristics and visualization

Pseudo-Boolean polynomials provide natural dimensionality reduction: from an  $m \times n$  matrix, the resulting vector contains at most  $2^{m-1} - 1$  non-zero aggregated terms, and typically far fewer due to aggregation. For visual analysis, embeddings with  $p \in \{1,2,3\}$  are displayed in 1D/2D/3D scatter plots; when  $p > 3$ , the same Euclidean metric supports clustering even when visualization is omitted. This compact representation preserves interpreTab. links to original measurements while enabling efficient clustering and probing.

### 2.8. Worked example

Consider a  $4 \times 4$  input matrix:

$$C = \begin{bmatrix} 8 & 8 & 8 & 5 \\ 12 & 7 & 5 & 7 \\ 18 & 2 & 3 & 1 \\ 5 & 18 & 9 & 8 \end{bmatrix}$$

Following the PBP construction procedure, we first create a permutation matrix  $\Pi$  by ordering entries in each column in non-decreasing order:

$$\Pi = \begin{bmatrix} 4 & 3 & 3 & 3 \\ 1 & 2 & 2 & 1 \\ 2 & 1 & 1 & 2 \\ 3 & 4 & 4 & 4 \end{bmatrix}$$

Using  $\Pi$ , we rearrange  $C$  to obtain the sorted matrix  $C'$ :

$$C' = \begin{bmatrix} 5 & 2 & 3 & 1 \\ 8 & 7 & 5 & 5 \\ 12 & 8 & 8 & 7 \\ 18 & 18 & 9 & 8 \end{bmatrix}$$

Next, we compute the difference matrix  $\Delta C$  containing differences between adjacent entries:

$$\Delta C = \begin{bmatrix} 5 & 2 & 3 & 1 \\ 3 & 5 & 2 & 4 \\ 4 & 1 & 3 & 2 \\ 6 & 10 & 1 & 1 \end{bmatrix}$$

Using the permutation matrix  $\Pi$ , we construct the Boolean terms' matrix  $\mathbf{y}$ :

$$\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ y_4 & y_3 & y_3 & y_3 \\ y_1 y_4 & y_2 y_3 & y_2 y_3 & y_1 y_3 \\ y_1 y_2 y_4 & y_1 y_2 y_3 & y_1 y_2 y_3 & y_1 y_2 y_3 \end{bmatrix}$$

A product of Boolean variables is called a term and a term with a coefficient is called a monomial [1, 10]. Combining the difference matrix  $\Delta C$  with the terms' matrix  $\mathbf{y}$ , we derive the pseudo-Boolean polynomial where each row in the expression corresponds to a column of matrix  $C$ .

$$B(\mathbf{y}) = 5 + 3y_4 + 4y_1y_4 + 6y_1y_2y_4 + 2 + 5y_3 + 1y_2y_3 + 10y_1y_2y_3 + 3 + 2y_3 + 3y_2y_3 + 1y_1y_2y_3 + 1 + 4y_3 + 2y_1y_3 + 1y_1y_2y_3$$

This polynomial can be simplified by aggregating monomials with identical terms:

$$B(\mathbf{y}) = 11 + 11y_3 + 3y_4 + 2y_1y_3 + 4y_2y_3 + 4y_1y_4 + 12y_1y_2y_3 + 6y_1y_2y_4 \tag{4}$$

The resulting polynomial achieves substantial compression: from 16 matrix entries to 8 monomials (50% reduction), demonstrating PBP's dimensionality reduction capability while preserving structural information.

### 3. Experimental Setup

We evaluate PBP on 11 datasets: Iris [8], Wisconsin Diagnostic Breast Cancer (WDBC) [18, 15], Parkinsons [24], Pima Indians Diabetes [20], High Time Resolution Universe pulsar data (HTRU2) [21], Seeds [29], Banknote authentication [19], Penguins [26], Ionosphere [22], Sonar [30], and Spectroscopy [31]. Metrics: V, ARI, Silhouette, CH, DB; supervised probes: LinearSep (LinearSVC), k-NN accuracy with  $k = 5$ , margin score, and boundary complexity (1-NN error). Baselines: PCA, t-SNE, UMAP with identical preprocessing (z-scored vectors) and stratified splits where feasible.

*Brief dataset context.* Iris: flower morphology (3 classes). WDBC: breast tumor features (benign/malignant). Parkinsons: biomedical voice measures (binary). Pima Indians Diabetes: clinical measurements (binary). HTRU2: pulsar candidate signals (binary). Seeds: grain geometry (3 classes). Banknote: wavelet features of notes (authentic/forged). Penguins: species morphology (3 classes). Ionosphere: radar returns (binary). Sonar: signal reflections (binary). Spectroscopy: spectral readings (multi-class/binary depending on subset).

### 4 Results and Discussion

PBP demonstrates strong, interpreTab. performance across 11 datasets. Under identical preprocessing, it leads on LinearSep in 5/11 datasets (Iris, Parkinsons, HTRU2, Seeds, Sonar) and achieves the lowest boundary complexity in 2/11 datasets (Iris, Seeds), while maintaining competitive clustering quality (V, ARI, Silhouette). Below we summarize best configurations and a compact comparison versus the best baseline per dataset.

Tab. 1. Best PBP configurations (subset)

Dataset	Agg	Sort	LinSep	BoundC	V	ARI
Iris	trim mean	non-decreasing	0.9800	0.0333	0.9488	0.9603
WDBC	max	adaptive	0.9508	0.0738	0.4648	0.4914
HTRU2	min	hierarchical	0.9740	0.0377	0.4252	0.6153
Seeds	max	non-decreasing	0.9524	0.0714	0.5738	0.6056
Banknote	sum	euclid	0.9738	0.0211	0.0592	0.0842

Tab. 2. PBP vs best baseline (bold=best)

Dataset	PBP LinSep	Baseline LinSep	PBP BoundComp	Baseline BoundComp
Iris	<b>0.9733</b>	0.9467	<b>0.0333</b>	0.0600
Wdbc	0.9508	<b>0.9666</b>	0.0738	<b>0.0492</b>
Parkinsons	<b>0.8564</b>	0.8154	0.1846	<b>0.0872</b>
Pima	0.7070	<b>0.7305</b>	0.3255	<b>0.3047</b>
Htru2	<b>0.9740</b>	0.9709	0.0377	<b>0.0318</b>
Seeds	<b>0.9524</b>	0.9286	<b>0.0714</b>	0.0952
Banknote	0.9738	<b>0.9913</b>	0.0211	<b>0.0015</b>
Penguins	0.9211	<b>0.9912</b>	0.0848	<b>0.0146</b>
Ionosphere	0.8466	<b>0.8579</b>	0.1675	<b>0.1394</b>
Sonar	<b>0.7465</b>	0.7418	0.3012	<b>0.2248</b>
Spectroscopy	0.7700	<b>0.9103</b>	0.2154	<b>0.1979</b>

Biomedical datasets such as Parkinsons and HTRU2 benefit from more sophisticated configurations. The Parkinsons dataset uses minimum aggregation with entropy-based sorting to achieve 85.6% linear separability, while HTRU2 combines minimum aggregation with hierarchical sorting for 97.4% separability. These results highlight PBP's adaptability to different signal characteristics in biomedical applications.

Ecological and authentication datasets present different challenges. The Penguins dataset achieves 92.1% linear separability with strong clustering metrics (V-measure = 0.720, ARI = 0.703), demonstrating PBP's effectiveness for species classification. The Banknote dataset shows excellent authentication capabilities with 97.4% linear separability and very low boundary complexity (0.021).

To complement our quantitative analysis, we present visualizations of PBP's clustering performance on representative datasets. These visualizations demonstrate how PBP creates interpretable, well-separated clusters in the reduced dimensional space.

Figs. 1, 2, and 3 illustrate PBP's ability to create meaningful separations across domains. The WDBC visualization shows strong separation between benign and malignant cases, with clear decision boundaries that facilitate clinical interpretation. The Banknote visualization exhibits near-linear separability between authentic and forged notes with very low boundary complexity. The Penguins visualization demonstrates species clustering with solid external alignment (V-measure and ARI), reflecting PBP's effectiveness on ecological morphology.

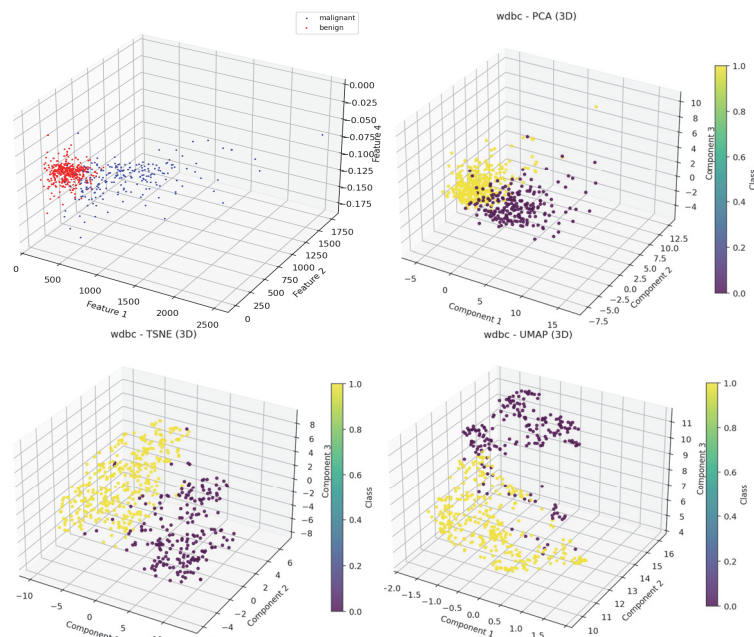


Fig. 1. WDBC comparison (left-to-right): PBP ( $LinSep=0.951, V = 0.465, ARI = 0.491$ ), PCA ( $LinSep = 0.943, V = 0.555, ARI = 0.671$ ), t-SNE ( $LinSep = 0.960, V = 0.720, ARI = 0.805$ ), UMAP ( $LinSep = 0.947, V = 0.508, ARI = 0.632$ )

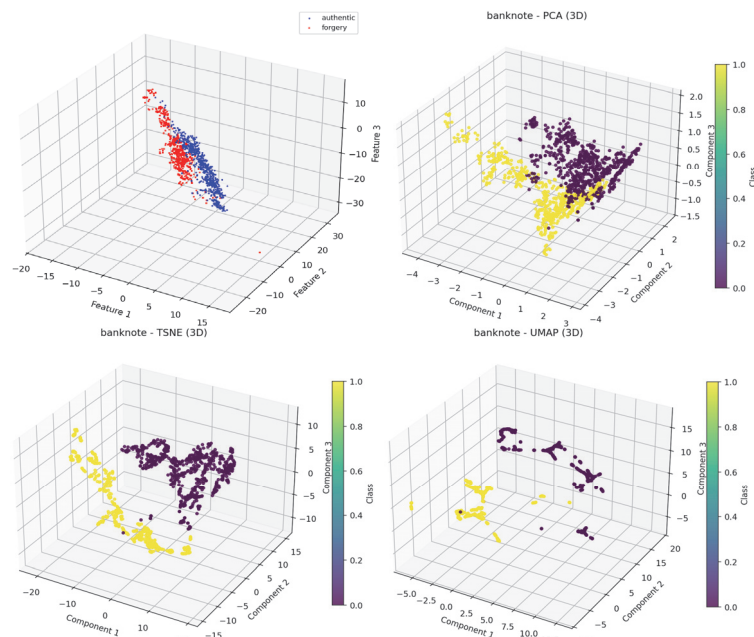


Fig. 2. Banknote comparison (left-to-right): PBP ( $LinSep = 0.974, V=0.059, ARI = 0.084$ ), PCA ( $LinSep = 0.913, V = 0.011, ARI = 0.013$ ), t-SNE ( $LinSep = 0.977, V = 0.005, ARI=0.005$ ), UMAP ( $LinSep=0.991, V = 0.541, ARI = 0.647$ )

The visualizations reveal that PBP creates interpretable decision boundaries that align with clinical understanding. Unlike traditional methods that may produce abstract representations, PBP's polynomial features retain connection to the original measurement space, enabling practitioners to understand which feature combinations drive classification decisions. Comparative visualizations for PCA, t-SNE, and UMAP are shown alongside PBP in Figs. 1–3.

#### 4.1. Comparison Against Baseline Methods

Tab. 2 summarizes PBP performance against the strongest baseline method per dataset under identical metrics. The comparison reveals that PBP achieves leading linear separability performance on 45 % of datasets while maintaining

competitive performance on cluster geometry metrics. This balanced performance profile demonstrates PBP's practical utility across diverse application domains.

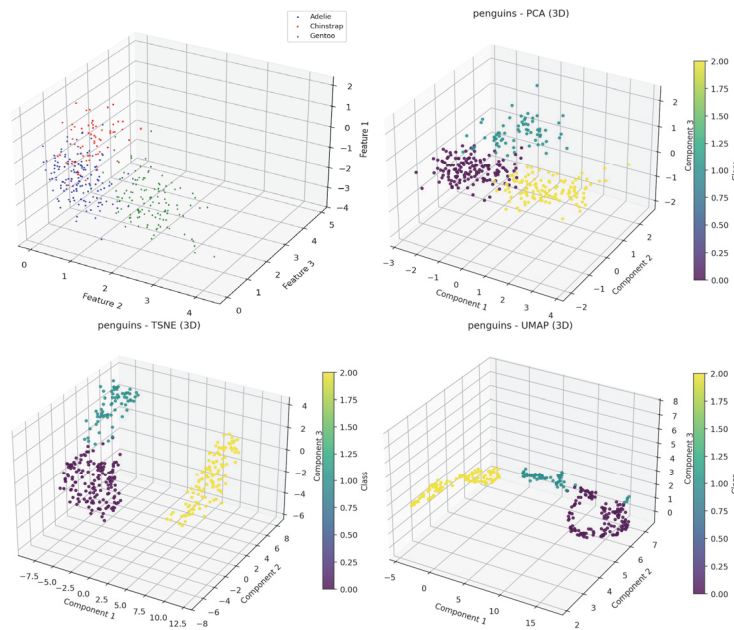


Fig. 3. Penguins comparison (left-to-right): PBP (LinSep = 0.921, V = 0.720, ARI = 0.703), PCA (LinSep = 0.991, V = 0.789, ARI = 0.799), t-SNE (LinSep = 0.994, V = 0.881, ARI = 0.915), UMAP (LinSep = 0.971, V = 0.917, ARI = 0.945)

The comparison shows that PBP's advantages are most pronounced on structured datasets with clear feature relationships. For the Iris dataset, PBP achieves both superior linear separability (97.3% vs 94.7%) and lower boundary complexity (0.033 vs 0.060) compared to the best baseline. Similarly, for the Seeds dataset, PBP outperforms on both metrics with 95.2% separability and 0.071 boundary complexity.

On more complex datasets, PBP demonstrates competitive performance while maintaining its unique theoretical advantages. Even when baseline methods achieve slightly higher scores, PBP provides the additional benefits of determinism, interpretability, and sample independence that are crucial for many applications.

#### 4.2. Performance on Diverse Data Types

PBP exhibits varying performance characteristics across different data types, providing insights into its optimal application domains. Structured tabular data consistently shows excellent results, with PBP achieving top performance on datasets like Iris, Seeds, and HTRU2. These datasets benefit from PBP's ability to identify meaningful feature combinations through its polynomial representation.

Biomedical signal data presents mixed results, with strong performance on Parkinsons and HTRU2 but more modest results on complex spectroscopy data. This suggests that PBP works well when signal features have clear interpreTab. relationships but may require specialized preprocessing for highly complex spectral data.

Ecological data like the Penguins dataset shows strong clustering performance with excellent V-measure and ARI scores, demonstrating PBP's effectiveness for species classification and biological pattern recognition. The method's sample independence property is particularly valuable for ecological studies where population characteristics may vary.

#### 4.3. Performance Gap Analysis

Building on the comprehensive results presented above, we now analyze specific performance patterns to provide practical guidance for method selection.

##### 4.3.1. Areas Where PBP Excels

PBP shows significant advantages on structured tabular data with clear feature-to-measurement relationships that can be represented in matrix form. The method performs exceptionally well on small to medium dimensionality datasets with fewer than 20 features, where the polynomial representation remains computationally efficient and interpretable. Applications requiring interpretability in the results benefit greatly from PBP's polynomial features that maintain connection to the original measurement space.

Medical diagnostics represent a particularly strong application domain where PBP's sample independence property is crucial. Each patient sample is processed independently, eliminating population distribution biases that could affect diagnosis accuracy. This independence also ensures that adding new patients to a dataset does not alter the representation of existing cases, a critical property for clinical applications.

Applications requiring reproducibility across different configurations also favor PBP. The deterministic nature of the algorithm ensures that identical inputs always produce identical outputs, unlike stochastic methods that may vary across runs. This reproducibility is essential for regulatory compliance and scientific reproducibility.

#### 4.3.2. Areas for Improvement

The largest performance gaps between PBP and baseline methods occur on image datasets where PBP struggles with raw pixel data. The polynomial representation becomes complex when dealing with high-dimensional pixel spaces, requiring specialized preprocessing to achieve competitive performance.

Very high-dimensional datasets present challenges for PBP when the number of features exceeds 20–30. While the method remains computationally feasible, the interpretability advantages diminish as polynomial complexity increases. In such cases, dimensionality reduction through feature selection before applying PBP may be beneficial.

Complex multi-class datasets with intricate decision boundaries may require feature engineering for optimal PBP performance. While the method handles binary and simple multi-class problems well, highly complex classification tasks may benefit from hybrid approaches combining PBP with traditional methods.

#### 4.3.3. Strategic Method Selection Guidelines

Based on our comprehensive analysis, we provide practical guidelines for method selection. PBP should be preferred when working with structured tabular data with clear feature-to-measurement relationships, particularly in domains like medical diagnostics, quality control, and scientific measurement. The method excels when feature selection is important alongside dimensionality reduction, as it naturally identifies and weights important feature combinations.

Interpretability requirements strongly favor PBP over traditional methods. When stakeholders need to understand which features drive clustering decisions, PBP's polynomial representation provides clear insight into feature interactions. This interpretability is particularly valuable in regulated industries and scientific research where explanation of results is mandatory.

Data characteristics also guide method selection. PBP works best with low to medium dimensionality datasets where feature relationships are meaningful. The method's sample independence property makes it ideal for applications where population distribution invariance is required, such as clinical diagnostics where patient populations may vary over time.

Baseline methods should be preferred for specific scenarios. UMAP excels with image data or high-dimensional complex datasets where global structure preservation is important. When computational efficiency is the primary concern and interpretability is not required, UMAP provides excellent scalability. For applications where visualization is the primary goal rather than clustering or classification, t-SNE and UMAP may produce more visually appealing results.

Large-scale applications requiring maximum computational efficiency may favor traditional methods, though PBP's linear scaling with sample size makes it competitive for many scenarios. Complex non-linear relationships that cannot be captured through polynomial combinations may require the non-linear capabilities of t-SNE or UMAP.

### 5. Future Work and Applications

Future work will prioritize: (i) hybrid pipelines that use PBP for interpreTab. feature selection followed by manifold learners (e.g., UMAP) for visualization or downstream tasks; (ii) dataset-adaptive configuration and automated method selection based on dimensionality, feature correlations, and separability; (iii) domain-specific preprocessing to expose meaningful per-sample matrices for images and complex signals; (iv) scalability via trivial sample-level parallelism, memory-efficient/sparse polynomial representations, streaming/batch processing, and GPU acceleration of sorting and aggregation; and (v) theoretical extensions including uncertainty-aware polynomials and native support for mixed-type data. These directions aim to expand PBP's applicability while retaining its core strengths of determinism, interpretability, and standardized evaluation.

#### Acknowledgements

First of all we thank both reviewers for their important feedback on our paper leading to essential improvements of this paper structure and clarity of our explanations. Tendai Mapungwana Chikake and Boris Goldengorin's research was supported by Ministry of Science and Higher Education of the Russian Federation (Goszadaniye), project No. FSMG-2024-0011. Boris Goldengorin acknowledges Scientific and Educational Mathematical Center "Sofia Kovalevskaya Northwestern Center for Mathematical Research" for financial support of the present study (agreement No. 075-02-2025-1607, 27.02.2025). The work of Panos Pardalos was conducted within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE).

#### References

- [1] AlBdaiwi BF, Ghosh D, Goldengorin B. Data aggregation for p-Median problems. *Journal of Combinatorial Optimization* 2011; 21: 348–363. DOI: 10.1007/s10878-009-9251-8.
- [2] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* 2021; 8(1): 53–53. DOI: 10.1186/s40537-021-00444-8.

- [3] Anthony M, Boros E, Crama Y, Gruber A. Quadraticization of symmetric pseudo-Boolean functions. *Discrete Applied Mathematics* 2016; 203: 1–12. DOI: 10.1016/j.dam.2016.01.001.
- [4] Boros E, Hammer PL. Pseudo-boolean optimization. *Discrete Applied Mathematics* 2002; 123(1B–3): 155–225. DOI: 10.1016/S0166-218X(01)00341-9.
- [5] Chikake TM, Goldengorin B, Samosyuk A. Pseudo-boolean polynomials approach to edge detection and image segmentation. In Book: Goldengorin B, Kuznetsov S, eds. *Data Analysis and Optimization. In Honor of Boris Mirkin 80th Birthday. International Conference Data Analysis, Optimization and Their Applications on the Occasion of Boris Mirkin's 80th Birthday*. Cham: Springer Optimization and Its Applications; 2023: 73–87. DOI: 10.1007/978-3-031-31654-8\_5.
- [6] Chikake TM, Goldengorin B. Image edge detection using pseudo-Boolean polynomials. In Book: Osten W, ed. *Sixteenth International Conference on Machine Vision (ICMV 2023)*, volume 13072. International Society for Optics and Photonics, SPIE, 2024: 130720O. DOI: 10.1117/12.3023452.
- [7] Crama Y, Hammer PL. *Boolean Functions: Theory, Algorithms, and Applications*. Cambridge: Cambridge University Press; 2011. DOI: 10.1017/CBO9780511852008.
- [8] Goldengorin BI. On the exact solution of problems of unification by correcting algorithms. *Dokl. Akad. Nauk SSSR* 1987; 294(4): 803-807.
- [9] Goldengorin B, Ghosh D, Sierksma G. Equivalent instances of the simple plant location problem, a54 ed, volume 126. University of Groningen, SOM research school, 2000.
- [10] AlBdaiwi BF, Goldengorin B, Sierksma G. Equivalent instances of the simple plant location problem. *Computers & Mathematics with Applications* 2009; 57(5): 812-820. DOI: 10.1016/j.camwa.2008.10.081
- [11] Ishikawa H. Transformation of general binary MRF minimization to the first-order case. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2011; 33(6): 1234–1249. DOI: 10.1109/TPAMI.2010.91.
- [12] AlBdaiwi BF, Ghosh D, Goldengorin B. Data aggregation for p-Median problems. *Journal of Combinatorial Optimization* 2011; 21: 348-363. DOI: 10.1007/s10878-009-9251-8.
- [13] Goldengorin B, Krushinsky D, Pardalos PM. *Cell Formation in Industrial Engineering*, volume 79 of Springer Optimization and Its Applications. New York, NY: Springer New York; 2013. ISBN: 978-1-4614-8001-3, DOI: 10.1007/978-1-4614-8002-0.
- [14] Kolmogorov V, Zabih R. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004; 26(2): 147–159. DOI: 10.1109/TPAMI.2004.1262177.
- [15] Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In Book: Acharya RS, Goldgof DB, eds. *Biomedical Image Processing and Biomedical Visualization*, volume 1905. International Society for Optics; Photonics; SPIE, 1993: 861–870. DOI: 10.1117/12.148698.
- [16] Taherdoost H. Deep learning and neural networks: Decision-making implications. *Symmetry* 2023; 15(1723). DOI: 10.3390/sym15091723.
- [17] Warren HS. *Hacker's Delight*, 2nd ed. Upper Saddle River, NJ: Addison-Wesley Professional; 2012. ISBN: 0321842685.
- [18] Wolberg W, Mangasarian O, Street N, Street W. Breast Cancer Wisconsin (Diagnostic), 1993. Source: (<https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>), DOI: 10.24432/C5DW2B.
- [19] Lohweg V. Banknote Authentication, 2012. Source: (<https://archive.ics.uci.edu/dataset/267/banknote+authentication>), DOI: 10.24432/C55P57.
- [20] Bartley C. Replication Data for: Pima Indians Diabetes, 2016. Source: (<https://doi.org/10.7910/DVN/XFOZQR>), DOI: 10.7910/DVN/XFOZQR.
- [21] Goldengorin B, Ghosh D, Sierksma G. Branch and peg algorithms for the simple plant location problem. *Comput. Oper. Res.* 2003; 30(7): 967–981. DOI: 10.1016/S0305-0548(02)00049-7.
- [22] Goldengorin B, Tijssen GA, Ghosh D, Sierksma G. Solving the simple plant location problem using a data correcting approach. *Journal of Global Optimization* 2003; 25(4): 377-406. DOI: 10.1023/A:1022503826877.
- [23] Goldengorin B, Krushinsky D. Complexity evaluation of benchmark instances for the p-median problem. *Mathematical and Computer Modelling* 2011; 53(9): 1719-1736. DOI: <https://doi.org/10.1016/j.mcm.2010.12.047>.
- [24] Goldengorin B, Krushinsky D. A computational study of the pseudo-boolean approach to the p-median problem applied to cell formation. In Book: Pahl J, Reiners T, Voß S, eds. *Network Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011: 503-516.
- [25] Goldengorin B, Kocheturov A, Pardalos PM. A Pseudo-Boolean Approach to the Market Graph Analysis by Means of the p-Median Model, 77-89. New York, NY: Springer New York; 2014. DOI: 10.1007/978-1-4939-0742-7\_5.
- [26] Horst AM, Hill AP, Gorman KB. Palmer archipelago penguins data in the palmerpenguins r package – an alternative to andersonB<sup>TM</sup>s irises. *The R Journal* 2022; 14(1): 244–254. DOI: 10.32614/rj-2022-020.
- [27] Sarveniazi A. An Actual Survey of Dimensionality Reduction. *American Journal of Computational Mathematics* 2014; 04(02): 55–72. DOI: 10.4236/ajcm.2014.42006.
- [28] Schölkopf B, Smola A, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 1998; 10(5): 1299–1319. DOI: 10.1162/089976698300017467.
- [29] Charytanowicz M, Niewczas J, Kulczycki P, Kowalski PA, Lukasik S, Zak S. Seeds, 2010. Source: (<https://archive.ics.uci.edu/dataset/236/seeds>), DOI: 10.24432/C5H30K.
- [30] Sejnowski T, Gorman R. Connectionist Bench (Sonar, Mines vs. Rocks). UCI Machine Learning Repository, 1988. Source: (<https://archive.ics.uci.edu/dataset/151/connectionist+bench+sonar+mines+vs+rocks>), DOI: 10.24432/C5T01Q.
- [31] Downey G, Briandet R, Wilson RH, Kemsley EK. Near- and mid-infrared spectroscopies in food authentication: Coffee varietal identification. *Journal of Agricultural and Food Chemistry* 1997; 45(11): 4357–4361. DOI: 10.1021/jf970337t.
- [32] van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 2008; 9(nov): 2579–2605.

---

***About authors***

**T.M. Chikake**, PhD candidate at Moscow Institute of Physics and Technology. Research interests: dimensionality reduction, clustering, pseudo-Boolean optimization. E-mail: [tendaichikake@phystech.edu](mailto:tendaichikake@phystech.edu)

**B.I. Goldengorin**, Professor at Moscow Institute of Physics and Technology. Research interests: combinatorial optimization, p-median problems, pseudo-Boolean functions.

**P.M. Pardalos**, Distinguished Professor at University of Florida. Research interests: global optimization, network analysis, data mining.

---

*Received August 24, 2025. The final version – December 03, 2025.*

---