Near-optimal Linear Predictive Clustering in Non-separable Spaces via MIP and QPBO Reductions

Jiazhou Liang[*1], **Hassan Khurram**[*1], **Scott Sanner**[1 2]

[1] University of Toronto
[2] Vector Institute for Artificial Intelligence
joe.liang@mail.utoronto.ca, hassan.khurram@mail.utoronto.ca, ssanner@mie.utoronto.ca

Abstract

Linear Predictive Clustering (LPC) partitions samples based on shared linear relationships between feature and target variables, with numerous applications including marketing, medicine, and education. Greedy optimization methods, commonly used for LPC, alternate between clustering and linear regression but lack global optimality. While effective for separable clusters, they struggle in *non-separable* settings where clusters overlap in feature space. In an alternative constrained optimization paradigm, previous work formulated LPC as a Mixed-Integer Program (MIP), ensuring global optimality regardless of separability but suffering from poor scalability. This work builds on the constrained optimization paradigm to introduce two novel approaches that improve the efficiency of global optimization for LPC. By leveraging key theoretical properties of separability, we derive near-optimal approximations with provable error bounds, significantly reducing the MIP formulation's complexity and improving scalability. Additionally, we can further approximate LPC as a Quadratic Pseudo-Boolean Optimization (QPBO) problem, achieving substantial computational improvements in some settings. Comparative analyses on synthetic and real-world datasets demonstrate that our methods consistently achieve near-optimal solutions with substantially lower regression errors than greedy optimization while exhibiting superior scalability over existing MIP formulations.

Code — github.com/D3Mlab/LPC-NS
Extended version — arxiv.org/abs/2511.10809

1 Introduction

Linear Predictive Clustering (LPC) clusters samples into K distinct groups, each characterized by a unique set of linear coefficients shared among the samples within the group, thereby capturing clusterwise patterns and variations in linear relationships. The versatility of LPC enables its application across diverse fields. In marketing, LPC facilitates market segmentation by predicting customer purchase behavior (Wedel and Kistemaker 1989). In medicine, LPC supports patient stratification for predictive health analytics, enhancing personalized treatment strategies (Ntani et al. 2020). Similarly, in education, LPC contributes to personalized learning by analyzing student performance and tailoring educational approaches based on outcomes predicted through linear models (Naik et al. 2017).

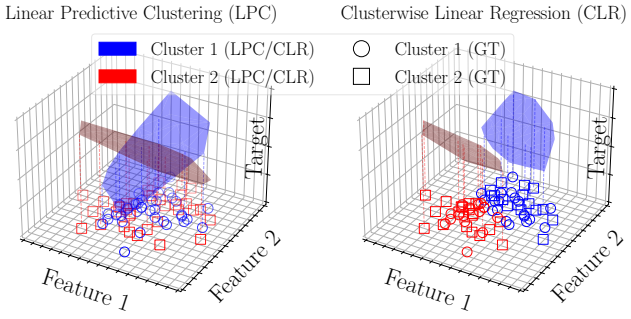


Figure 1: The plot illustrates a case where the feature variables (horizontal axes) from samples belonging to two different ground truth clusters (circles and squares) overlap, making them non-separable in the feature space. However, each cluster follows a distinct linear relationship (hyperplane) with the target variable (vertical axis). This non-separability renders CLR (cf. *right*)—which clusters solely based on feature variables—ineffective: it assigns clusters (shown in blue and red) that mix the ground truth labels. In contrast, LPC (cf. *left*) recovers the ground truth linear predictor assignment.

In a special case of the LPC problem, feature variables may be *non-separable*, meaning that samples from different clusters overlap in the feature space. Specifically, when the means and uncentered covariance of feature variables are highly similar across clusters, clustering based solely on feature variables becomes ineffective. As shown in Fig. 1 (left), although each cluster follows a distinct relationship with the target variable (hyperplanes on the vertical axis), samples from the two ground truth clusters (shown as circles and squares) exhibit significant overlap in their feature space (horizontal axes). This non-separability renders simple *clustering-then-regression* approaches highly impractical. As illustrated in Fig. 1 (right), applying an external clustering method to feature variables in Clusterwise Linear Regression (CLR) produces two partitions (blue and red) as a mixture of samples from two ground truth clusters (circles and squares), which fail to recover the ground truth linear predictor assignment. This further demonstrates the limitations of traditional

[*]These authors contributed equally.

clustering-based approaches in the non-separable setting.

Existing LPC methods for *non-separable feature spaces* fall into two categories. The first, Iterative (Greedy) Optimization (Späth 1979; Manwani and Sastry 2015; Wang 2020), alternates between updating regression coefficients and cluster assignments. While this approach has been widely adopted due to its scalability, it is highly sensitive to initialization, lacks global optimality, and is prone to convergence at suboptimal local minima (Chembu and Sanner 2023). The second, Constrained Optimization (Bertsimas and Shioda 2007; Carbonneau, Caporossi, and Hansen 2012; Zhu, Li, and Kong 2012; Chembu and Sanner 2023), formulates LPC as a Mixed-Integer Programming (MIP) problem, jointly optimizing cluster assignments and regression coefficients. This guarantees global optimality by minimizing total regression loss but suffers from significant scalability limitations due to its computational complexity.

Proposed Contribution. To the best of our knowledge, developing an approach that balances the scalability of greedy optimization with the global optimality of MIP-based formulations for LPC in non-separable feature spaces remains an open challenge. Hence, to address these challenges, we make the following key contributions in this work:

- We build on the global optimization MIP approach to LPC and propose a novel near-optimal approximation of regression coefficients by leveraging fundamental properties of non-separability. This insight allows us to substitute and ablate the regression component from the MIP LPC constrained optimization, thereby significantly reducing the size of the MIP formulation. This leads to our first contribution of LPC-NS-MIP for efficient LPC in non-separable (NS) feature spaces.
- We derive error bounds on LPC-NS-MIP that prove the optimality of our approach under non-separability conditions on the uncentered covariances of each cluster.
- We further transform LPC-NS-MIP into a Quadratic Pseudo-Boolean Optimization (QPBO) formulation. Empirically, LPC-NS-QPBO offers substantial scalability improvements while maintaining near-optimal performance in the setting of two clusters.
- We conduct rigorous experiments demonstrating that LPC-NS-MIP and LPC-NS-QPBO consistently outperform the greedy optimization approach, achieving significantly better LPC objective optimization across datasets with varying noise, dimensionality, and outliers, while scaling more efficiently than existing globally optimal MIP methods.

2 Related Work

LPC under Constrained Optimization was first introduced by Bertsimas and Shioda (2007) as CRIO, which formulates the problem as a MIP problem to identify optimal cluster assignments that minimize the within-cluster regression Sum of Absolute Errors (SAE). This formulation guarantees deterministic cluster assignments with global optimality. Zhu, Li, and Kong (2012) introduces a more outlier-robust variation, and Carbonneau, Caporossi, and Hansen (2012) extends CRIO with a Sum of Squared Errors (SSE) objectives. We extend and compare to these global optimization methods.

LPC under Greedy Optimization was first proposed by Späth (1979) and later extended by Chembu and Sanner (2023), drawing inspiration from the Majorization-Minimization (MM) algorithm (Hunter and Lange 2004). These methods employ an iterative approach, where cluster assignments are fixed to compute cluster-specific regression lines, followed by updating cluster assignments to minimize the clusterwise regression loss. Cluster updates are performed either by pairwise sample exchanges (Späth 1979) or by reassigning all samples (Chembu and Sanner 2023), repeating the process until convergence. Greedy methods are fast, but prone to local optima as we show empirically.

Clusterwise Linear Regression (CLR) CLR (Bradley and Mangasarian 2000; Hota et al. 2009; Manwani and Sastry 2015; Feng et al. 2016; Corizzo et al. 2019) fundamentally differs from LPC in its approach to cluster assignments. While LPC assigns data to clusters arbitrarily, focusing solely on the objective of linear regression, CLR often uses external clustering definitions informed by prior knowledge. For instance, K-plane regression (Manwani and Sastry 2015; da Silva and de Carvalho 2017; Wang 2020) extends the K-means algorithm (Lloyd 1982) into a regression context by leveraging Mean Square Sum of Clusters (MSSC) error.

This approach struggles with non-separable feature spaces as feature variables cannot form distinct clusters. As shown in Fig. 1 (right), the distribution of cluster feature variables may conflict with MSSC objectives that require some degree of separability. Since CLR cannot address this non-separable case, it does not address the main focus of our work.

Latent Class Regression (LCR) Unlike LPC, which aims to identify deterministic cluster assignments with minimum regression loss, Latent Class Regression (LCR) (Wedel and DeSarbo 1994) can be viewed as a generalization of CLR that employs an explicit discrete Finite Mixture Model (FMM) for each latent class (Quandt 1972; DeSarbo and Cron 1988; Cosslett and Lee 1985; Hamilton 1990). However, in non-separable settings, where multiple latent classes share the same modality, LCR fails to distinguish clusters effectively for the same reasons that CLR failed in Fig. 1 (right). Since LCR cannot handle the non-separable case, it does not address the main focus of our work.

3 Methodology

3.1 Preliminary Definitions and MIP formulation with Global Optimality

We first define the objective of LPC and introduce the state-of-the-art MIP formulations (Bertsimas and Shioda 2007).

Let $\mathbf{X} \in \mathbb{R}^{N\times(D+1)}$ denote D features variables with bias term for a dataset with N samples, and let $\mathbf{y} \in \mathbb{R}^{N\times 1}$ represent the corresponding target feature. Cluster assignments are encoded by $\mathbf{Z}_k \in \mathbb{Z}^{N\times N}$, a binary diagonal matrix where $\mathbf{Z}_{k,i,i} = 1$ if sample i is assigned to cluster k and 0 otherwise. The regression coefficients and bias for cluster k are given by $\mathbf{w}_k \in \mathbb{R}^{D+1}$.

The Regularized Linear Least Squares (RLS) estimates a single set of coefficients $\mathbf{w}$ minimizing the SSE between $\mathbf{y}$ and its prediction $\mathbf{X}\mathbf{w}$ with ℓ_2 regularization (a.k.a. Ridge Regression). LPC extends from this objective by partitioning $\mathbf{X}$ into K disjoint clusters, where K is predefined based on prior knowledge. The objective function of LPC is formulated as:

$$\min_{\mathbf{Z},\mathbf{w}} \sum_{k=1}^{K} \left(\|\mathbf{Z}_k\mathbf{y} - \mathbf{Z}_k\mathbf{X}\mathbf{w}_k\|_2^2 + \lambda\|\mathbf{w}_k\|_2^2 \right). \tag{1}$$

Since $\mathbf{Z}_{k,i,i} = 1$ only if sample i is assigned to cluster k, the residual $\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_k$ is nonzero only for assigned samples. Consequently, the SSE for $\mathbf{w}_k$ is computed exclusively over samples in cluster k. LPC aims to determine the optimal $\mathbf{Z}_k$ and $\mathbf{w}_k$ such that the total SSE across clusters is minimized.

Equation (1) can be solved using MIP with optimization variables $\mathbf{Z}$ and $\mathbf{w}$ (Bertsimas and Shioda 2007; Zhu, Li, and Kong 2012; Carbonneau, Caporossi, and Hansen 2012), ensuring a globally optimized objective. The MIP formulation is expressed as:

$$\min_{\mathbf{Z},\mathbf{w}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \delta_{i,k}^2 + \lambda\|\mathbf{w}_k\|_2^2 \right) \tag{2}$$

$$\text{s.t. } \delta_i \le (\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_k) + M(1 - \mathbf{Z}_{k,i,i}) \quad \forall i \in \{1,\dots,N\}; k \in \{1,\dots,K\} \tag{3}$$

$$\delta_i \ge (\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_k) - M(1 - \mathbf{Z}_{k,i,i}) \quad \forall i \in \{1,\dots,N\}; k \in \{1,\dots,K\} \tag{4}$$

$$\textstyle\sum_{k=1}^{K} \mathbf{Z}_{k,i,i} = 1, \quad \forall i \in \{1,\dots,N\} \tag{5}$$

$$\delta_i \in \mathbb{R}; \quad \mathbf{Z}_{k,i,i} \in \{0,1\}; \quad W_{jk} \in \mathbb{R} \quad \forall i \in \{1,\dots,N\} j \in \{1,\dots,D+1\}; k \in \{1,\dots,K\}$$

Here, δ is a continuous auxiliary variable used to linearize the clusterwise regression error, i.e., $\mathbf{Z}_{k,i,i} = 1 \rightarrow \delta_i = (\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_k)$, via the Big-M constraint (Dantzig 2016), where M is a large positive scalar. Constraints (3) and (4) ensure that the residual of sample i with respect to $\mathbf{w}_k$ contributes to the objective only when i is assigned to cluster k. Constraint (5) enforces that each sample i is assigned to exactly one cluster, ensuring that the K clusters are mutually exclusive and exhaustively cover all samples.

3.2 Optimization of Regression Coefficients $\mathbf{w}$

We observed that jointly optimizing $\mathbf{Z}$ and $\mathbf{w}$ in (2) significantly limits the scalability of LPC. To improve efficiency, we aim to reduce the number of optimization variables in this formulation.

Given a known $\mathbf{Z}_k$, the corresponding regression coefficients $\mathbf{w}$ can be optimized implicitly via the closed-form solution as follows:

$$\mathbf{w}_k = \underbrace{\left(\mathbf{X}^\top\mathbf{Z}_k\mathbf{X} + \lambda\mathbf{I}\right)^{-1}}_{\mathbf{A}_k} \mathbf{X}^\top\mathbf{Z}_k\mathbf{y}, \tag{6}$$

Substituting (6) into (1), the optimal $\mathbf{w}$ can be directly computed using this implicit formulation, eliminating the need for two sets of optimization variables in the MIP formulation.

However, the matrix inversion in $\mathbf{A}_k$ depends on the optimization variable $\mathbf{Z}_k$, which is infeasible to formulate in an MIP problem. The term $\mathbf{X}^\top\mathbf{Z}_k\mathbf{X}$ in $\mathbf{A}_k$ represents the clusterwise uncentered covariance of the feature variables $\mathbf{X}$, computed using only the samples assigned to cluster k.

The definition of *non-separable* $\mathbf{X}$ suggests that the uncentered covariance structure $\mathbf{X}^\top\mathbf{X}$ across different subsets of samples often remains similar, cf. Fig. 1(left), where the non-separable setting is precisely the case where the clusters have *matching* uncentered covariance. Consequently, while computing the matrix inverse for each cluster in $\mathbf{A}_k$ is infeasible to formulate as a MIP problem, fortuitously, it may actually be unnecessary!

To address this, we propose an approximation for $\mathbf{A}_k$, denoted as $\mathbf{A}^*$, defined as:

$$\mathbf{w}_k^* = \underbrace{\left(\alpha_k\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}}_{\mathbf{A}^*} \mathbf{X}^\top\mathbf{Z}_k\mathbf{y}, \tag{7}$$

where α_k is a hyperparameter reflecting the relative size of cluster k. Under balanced cluster sizes, this reduces to $\alpha_k = 1/K$. For imbalanced clusters, a multi-hop tuning procedure yields a more suitable value (cf. App.D for details).

Eliminating $\mathbf{Z}_k$ from $\mathbf{A}_k$ enables precomputing $\mathbf{A}^*$, reducing the optimization of $\mathbf{w}$ to a matrix multiplication, leading to a simplified MIP formulation. Specifically, substituting the $\mathbf{w}_k^*$ back into (1), we derive a quadratic form solely in terms of the cluster assignment variables $\mathbf{Z}_k$:

$$\min_{\mathbf{Z}} \sum_{k=1}^{K} \left(\|\mathbf{Z}_k\mathbf{y} - \mathbf{Z}_k\mathbf{X}\mathbf{A}^*\mathbf{X}^\top\mathbf{Z}_k\mathbf{y}\|_2^2 + \lambda\|\mathbf{A}^*\mathbf{X}^\top\mathbf{Z}_k\mathbf{y}\|_2^2 \right) \tag{8}$$

Here, the only variable requiring optimization is the binary indicator $\mathbf{Z}_k$, while the regression coefficients $\mathbf{w}$ are implicitly optimized as given in (7). This transformation successfully reduces the MIQP formulation of LPC into a higher-order Pseudo-Boolean Optimization (PBO) problem, which we reduce to a Quadratic PBO (QPBO) form in Sec. 3.4.

With this critical insight, we have shown that this leads to a substantial reduction in the complexity of the MIP encoding due to the ability to directly solve for w_k^* and ablate these optimization variables without sacrificing optimality under the separability assumption. However, before we proceed further under this assumption, we first investigate settings where this approximation is sensible.

3.3 Approximation Error of $\mathbf{w}^*$

Although the non-separable nature of $\mathbf{X}$ provides insight into the approximation of $\mathbf{A}^*$, it is crucial to both theoretically and empirically assess the bound of the approximation error in the LPC objective introduced by approximating $\mathbf{w}$ as $\mathbf{w}^*$ in (7).

Let $\mathcal{E}_{\text{obj}}$ denote the sum of differences between the globally optimal linear regression error objective in cluster k (O_k) and the linear regression error objective using $\mathbf{w}^*$ from (7) (O_k^*), i.e., $\mathcal{E}_{\text{obj}} = \sum_{k=1}^{K} (O_k^* - O_k)$. This term quantifies the approximation error introduced in the LPC objective by using $\mathbf{w}^*$ instead of the globally optimal solution. Assuming

$\lambda > 0$, the upper bound of $\mathcal{E}_{\text{obj}}$ is as follows (cf. App. A for the detailed error derivation):

$$\mathcal{E}_{\text{obj}} \leq \sum_{k=1}^{K} \left(\frac{3\|\mathbf{y}_k\|_2^2 \|\mathbf{X}_k\|_2^2 \|\alpha_k \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}_k \mathbf{X}\|_2}{2\lambda^2} \right), \tag{9}$$

In (9), the terms $\|\mathbf{y}_k\|_2^2$ and $\|\mathbf{X}_k\|_2^2$ correspond to the norms of the target and feature variables of samples within each cluster. Except for degenerate cases of purely null data, these values cannot be zero for non-empty clusters.

The term $\left\|\alpha_k \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}_k \mathbf{X}\right\|_2$ in (9) quantifies the difference in uncentered covariance between all samples ($\mathbf{X}^\top \mathbf{X}$) and those assigned to cluster k (i.e., $\mathbf{X}^\top \mathbf{Z}_k \mathbf{X}$).

When the feature variables are perfectly non-separable (i.e., identical across clusters) and cluster sizes are balanced (N/K samples per cluster with $\alpha_k = 1/K$) , as illustrated in the first plot of Tab. 1, $\left\|\alpha_k \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}_k \mathbf{X}\right\|_2 = 0, \quad \forall k \in \{1, \ldots, K\}$. Thus, the approximation error satisfies $\mathcal{E}_{\text{obj}} \leq 0$, ensuring perfect recovery of the globally optimal objective. For imbalanced cluster settings, α_k can be iteratively tuned to adapt to varying cluster sizes (cf. App. D for details).

Note that this 0-bound assumes that the Z_k obtained from optimizing (8) is aligned with the ground truth cluster assignment. Nevertheless, (8) is inherently robust to suboptimal Z_k since it produces poor regression coefficients with high cluster-wise SSE, which are penalized in the objective.

Tab. 1 empirically evaluates the deviation from the globally optimal objective when using $\mathbf{w}^*$, denoted as $\mathcal{E}_{\text{obj}}^{\mathbf{w}^*}$, across four synthetic datasets with $K = 2, N = 50$ (cf. Sec. 4.2). Each dataset maintains fixed parameters and orthogonal regression coefficients $\mathbf{w}_k$ between clusters while varying $\mathbf{X}$, representing different levels of separability:

$$\mathcal{E}_{\text{sep}} = \sum_{k=1}^{K} \left\|\alpha_k \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}_k \mathbf{X}\right\|_2 . \tag{10}$$

When $\mathbf{X}$ is perfectly non-separable ($\mathcal{E}_{\text{sep}} = 0$), the approximation error $\mathcal{E}_{\text{obj}}^{\mathbf{w}^*}$ is zero. More generally, in non-separable cases (second column of Tab. 1), $\mathbf{w}^*$ remains a near-optimal approximation of $\mathbf{w}$.

Two additional cases on the right of Tab. 1 illustrate: *semi-separable*, where two clusters overlap in X but differ in uncentered covariances, and *separable*, where clusters are distinct in X and have different covariance. In both cases, as feature variables become more separable, the approximation error of $\mathbf{w}^*$ increases, resulting in a larger optimality gap. However, these separable cases already have well-established solutions, such as CLR, achieving near-optimal solutions in the separable setting; thus, we remark that we specifically focus on the more challenging *non-separable* setting.

3.4 QPBO formulation of Cluster Assignment Z

Although (8) can technically be rewritten as a cubic PBO problem (cf. App. B), no modern constrained optimization solvers currently provide scalable solutions due to the cubic growth of optimization variables $\mathbf{Z}$. Consequently, in practice, a MIP formulation incorporating a continuous auxiliary variable $\boldsymbol{\delta}$ (cf. (2)) offers superior scalability compared to direct PBO formulations.

However, LPC-NS-MIP can be approximated as a Quadratic Pseudo-Boolean Optimization (QPBO) problem, termed LPC-NS-QPBO. We first reformulate and simplify (8) as a maximization problem after expanding the square (cf. App. B for a detailed derivation):

$$\max_{\mathbf{Z}} \sum_{k=1}^{K} \left[\mathbf{y}^\top \mathbf{Z}_k \mathbf{X}^\top A^* [2\mathbf{I} - \underbrace{(\mathbf{X}^\top \mathbf{Z}_k \mathbf{X} + \lambda \mathbf{I})}_{\mathbf{A}_k^{-1}} A^*] \mathbf{X}^\top \mathbf{Z}_k \mathbf{y} \right] \tag{11}$$

From (6), we recognize the form of $\mathbf{A}_k^{-1}$ and then approximate it as A^{*-1} under the non-separable assumption using (7). Then we simplify $[2\mathbf{I} - A^{*-1}A^*] = [2\mathbf{I} - \mathbf{I}] = \mathbf{I}$, yielding the final reduced formulation for LPC-NS-QPBO:

$$\max_{\mathbf{Z}} \textstyle\sum_{k=1}^{K} \left[\mathbf{y}^\top \mathbf{Z}_k \mathbf{X}^\top A^* \mathbf{X}^\top \mathbf{Z}_k \mathbf{y} \right] \tag{12}$$

$$\text{s.t.} \textstyle\sum_{k=1}^{K} \mathbf{Z}_{k,i,i} = 1, \mathbf{Z}_{k,i,i} \in \{0,1\}; \quad \forall i \in \{1, \ldots, N\}; k \in \{1, \ldots, K\} \tag{13}$$

LPC-NS-QPBO effectively reduces LPC-NS-MIP into an approximate QPBO formulation with binary optimization variable $\mathbf{Z}$. Moreover, since $\mathbf{X}^\top A^* \mathbf{X}^\top$ does not involve any optimization variables, it can be precomputed, further simplifying the optimization process.

3.5 Refitting Regression Coefficients w using Optimal Clustering Assignment Z

After obtaining the optimal $\mathbf{Z}$ using LPC-NS-QPBO or LPC-NS-MIP, optimal $\mathbf{w}_k$ can be recovered using (6). Thus, this refitting process, denoted as $\mathbf{w}_k^{\text{refit}}$ for all $k \in \{1, \ldots, K\}$, is given by:

$$\mathbf{w}_k^{\text{refit}} = \left(\mathbf{X}^\top \mathbf{Z}_k \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{Z}_k \mathbf{y}. \tag{14}$$

This step ensures that the refitted regression coefficients $\mathbf{w}_k^{\text{refit}}$ are optimally aligned with the final cluster label assignments.

4 Experiments

4.1 Research Questions

To evaluate the effectiveness of our proposed methods, LPC-NS-MIP and LPC-NS-QPBO, against existing approaches – Greedy Optimization (Greedy) and globally optimal methods (GlobalOpt) – we conduct experiments focusing on three key aspects: (a) deviation of each method from the GlobalOpt LPC objective, (b) recovery performance of regression coefficients $\mathbf{w}^{\text{gt}}$ and clustering assignments $\mathbf{Z}^{\text{gt}}$ in the synthetic datasets with known ground truth, and (c) optimization times vs. data sample size. The research questions (RQs) guiding our evaluation are as follows:

- **RQ1: Time vs. Error Scalability Trade-off Analysis.** How do objective error and optimization times change as the number of samples increases?
- **RQ2: Performance Analysis and Algorithm Properties on Synthetic Datasets.** For synthetic datasets with known

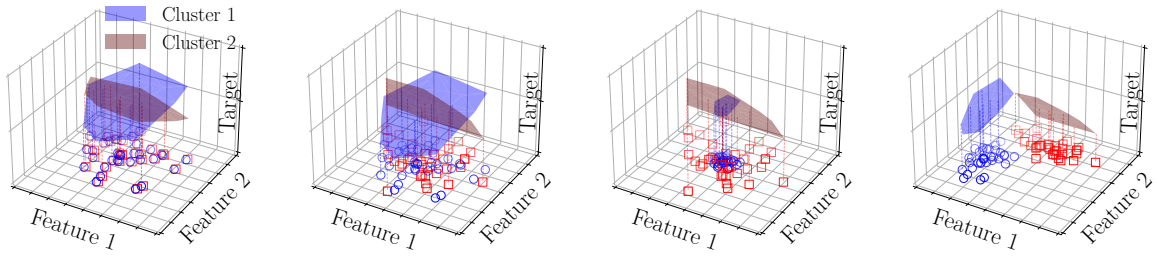
Types	Identical	Non-Separable	Semi-Separable	Separable
Samples				
$\mathcal{E}_{\text{sep}}$	0.0	60.4043	193.791	666.579
$\mathcal{E}_{\text{obj}}^{\mathbf{w}^*}$	0.01 ± 0.01	0.0208 ± 0.0054	2.7080 ± 0.4973	2.6250 ± 0.4868
$\mathcal{E}_{\text{obj}}^{\text{CLR}}$	18.4862 ± 0.4441	72.2481 ± 0.8360	9.4433 ± 0.3633	0.5996 ± 0.0316

Table 1: Samples from two clusters exhibits two distinct linear relationships ($\mathbf{w}_1 = [2, 4, -5]$ and $\mathbf{w}_2 = [-2, -4, 5]$) under four types of feature variables in 2-dimensional space ($\mathbf{X} \in \mathbb{R}^{N\times 2}$), ordered by ascending values of $\mathcal{E}_{\text{sep}} = \sum_{k=1}^{2} \left\|\alpha_k \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}_k \mathbf{X}\right\|_2$. The difference between the globally optimal objective (O) and optimal objective via $\mathbf{w}^*$ approximation ($O^{\mathbf{w}^*}$) is computed as $\mathcal{E}_{\text{obj}}^{\mathbf{w}^*} = O^{\mathbf{w}^*} - O$ with 95% confidence interval. The clusterwise regression (CLR) objective (O^{CLR}) and its deviation from O, denoted as $\mathcal{E}_{\text{obj}}^{\text{CLR}}$, is obtained by K-Means clustering on $\mathbf{X}$ and Ridge regression to each cluster.

ground truth, how do different methods compare in terms of objective and accuracy? We evaluate these metrics for different K while varying feature dimensionality, noise levels, and proportion of target variable outliers.

- **RQ3: Performance Analysis on Real Datasets.** How do different methods compare in terms of objective error on real datasets, specifically curated to contain two distinct clusters with differing regression coefficients?

4.2 Experimental Setup

Synthetic Datasets (SD) To simulate a non-separable observed feature space $\mathbf{X}$ for K clusters, we generate $\frac{N}{K}$ samples per cluster. For each sample $i \in \{1, \ldots, N\}$, the feature variables $\mathbf{X}_i \in \mathbb{R}^D$ follow the same Gaussian distribution ($\mu = 1, \sigma = 2$) across all clusters. We focus solely on generating the challenging case of non-separable $\mathbf{X}$ since CLR already performs well in separable cases (cf. Tab. 1). We define K sets of coefficient $\mathbf{w}$, with $\mathbf{w}_k$ serving as the ground truth for cluster k. The target variable $\mathbf{y}_i$ then is computed based on $\mathbf{w}_k$, $\mathbf{X}_i$, and Gaussian noise.

Following previous work on LPC (Bertsimas and Shioda 2007; Chembu and Sanner 2023; Zhu, Li, and Kong 2012), we focus on cases where $K = 2, 3$. SD1 consists of $K = 2$ clusters where the two regression lines are orthogonal. SD2 consists of $K = 3$ regression lines that represent positive, negative, and uncorrelated relationships, as illustrated in Fig. 1 (cf. App. C for detail setup).

We need to avoid using a large number of clusters K since for two key reasons: (1) GlobalOpt becomes computationally infeasible for large K values using available MIP solvers, and (2) as K increases, multiple distinct cluster assignments may yield the same error for the defined clustering objective, leading to spurious misalignments with ground truth labels. Our experimental setup ensures each cluster reflects an identifiable and examinable linear relationship under the LPC objective. In App. F.2, we further explain the rationale of this design choice and empirically verify these behaviors in the setting with $K = 5$.

Real-world Datasets We selected 10 datasets from the UCI Machine Learning repository (Kelly, Longjohn, and Nottingham 2023). Each dataset was partitioned into K clusters based on the value of its categorical feature, and coefficients $\mathbf{w}_k$ were fitted on samples in each cluster separately.

We ensured each dataset had at least two clusters suitable for LPC. Specifically, we identified the two clusters with the largest l_2 distance in their fitted linear models, defined by $\max_{i,j} \|\mathbf{w}_i - \mathbf{w}_j\|_2$ for $i, j \in \{1, \ldots, K\}$. App. G further analyzes the l_2 distance across all possible pairs of clusters and finds that, in most datasets, only two clusters exhibit significant differences in their fitted linear models. For datasets with multiple categorical features, the feature with the largest $\|\mathbf{w}_i - \mathbf{w}_j\|_2$ was chosen. Datasets also required a lower cluster-wise SSE than standard regression.

Finally, the selected groups were merged to construct the final dataset, with the categorical variable used for clustering removed to ensure that the model learns solely from the feature variables. We refer the reader to App. G for further details on each dataset and preprocessing steps.

Implementation GlobalOpt follows the formulation in (2). Greedy optimization follows the algorithm proposed by Chembu and Sanner (2023). Since Greedy is sensitive to initialization, each experimental trial evaluates Greedy Optimization using 20 different initializations with a maximum of 100 iterations and reports the mean performance. We used the industry-standard solver `Gurobi` (Gurobi Optimization, LLC 2024), with an optimality gap fixed at 5%. Although LPC-NS-QPBO is formulated as a QPBO problem,

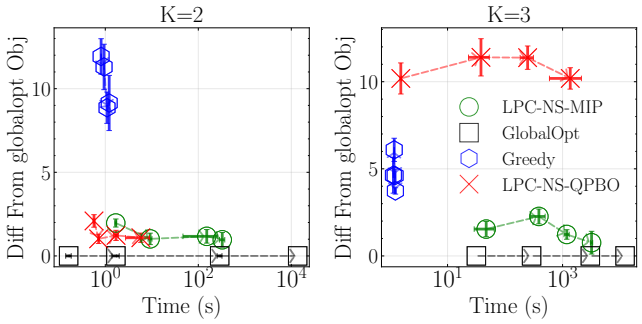


Figure 2: *RQ1* Trade-off between difference from the GlobalOpt objective (y-axis) and optimization time (seconds) (x-axis) across different methods as sample sizes increase. The sample size is limited to 200 for $K = 2$ and 90 for $K = 3$, to allow GlobalOpt to achieve optimality in ≤ 2 hours (cf. Fig. 3). Each data point represents the mean over trials and the solid interval indicates the 95% confidence intervals.

we observed that the MIP solver in Gurobi outperforms the PBO solver SCIP (Bolusani et al. 2024) (cf. App. E for performance evaluations). We rely on Gurobi's M optimization and were unable to outperform it using customized constraints.

4.3 Evaluation Metrics

Difference From Optimal Objective Let O represent the globally optimized LPC objective obtained from GlobalOpt. Let O' denote the optimal LPC objective achieved by each method. For LPC-NS, O' is recalculated using $\mathbf{w}^{\text{refit}}$ and $\mathbf{Z}$ instead of the optimal objective given by constrained optimization using approximated $\mathbf{w}^*$. $O' - O$ quantifies the deviation of each method from GlobalOpt.

Cluster Label Difference Let $\mathbf{Z}^{\text{gt}} \in \mathbb{R}^{K \times N}$ denote the binary matrix representing the ground truth cluster assignments, and let $\mathbf{Z}' \in \mathbb{R}^{K \times N}$ be the predicted cluster assignments obtained using LPC and aligned with $\mathbf{Z}'$ using Hungarian algorithm. The discrepancy between $\mathbf{Z}^{\text{gt}}$ and $\mathbf{Z}'$ is defined as the percentage of samples with misassigned cluster labels across all K clusters: $\text{mismatch}(\mathbf{Z}^{\text{gt}}, \mathbf{Z}') = \frac{1}{N} \sum_{k=1}^{K} \|\mathbf{Z}_k^{\text{gt}} - \mathbf{Z}_k'\|_1$. It represents the percentage of misassigned samples relative to the total number of samples.

Regression Coefficient Difference Let $\mathbf{w}^{\text{gt}} \in \mathbb{R}^{K \times (D+1)}$ be the clusterwise regression coefficients based on $\mathbf{Z}^{\text{gt}}$, and let $\mathbf{w}' \in \mathbb{R}^{K \times (D+1)}$ denote the predicted regression coefficients obtained by different methods. The error between $\mathbf{w}^{\text{gt}}$ and $\mathbf{w}'$ is measured as: $\text{difference}(\mathbf{w}^{\text{gt}}, \mathbf{w}') = \sum_{k=1}^{K} \|\mathbf{w}_k^{\text{gt}} - \mathbf{w}_k'\|_2$.

5 Experimental Results

RQ1 Fig. 2 illustrates the trade-off between deviation from GlobalOpt and computational time across $K = 2$ (sample range $\{50, \ldots, 200\}$) and $K = 3$ (sample range $\{40, \ldots, 60\}$) in a high noise setting ($\sigma = 3.5$). *Sample sizes were limited for GlobalOpt to run to optimality in* ≤ 2 *hours*[1].

[1]Runtime from the start to the completion of the `Gurobi` solver with an Intel(R) Xeon(R) Platinum 8260 CPU 32 threads.

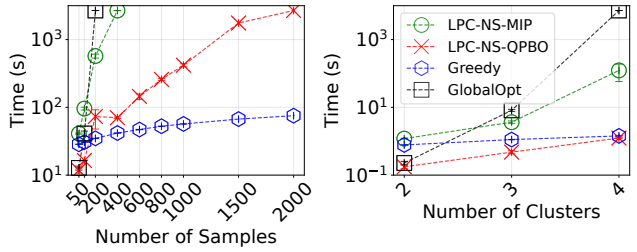


Figure 3: Runtime as the number of samples increases for $K = 2$ (left) and as the number of clusters increases (right). LPC-NS-QPBO exhibits better scalability with $N = 2000$ samples, compared to $N = 200$ in GlobalOpt under 2 hours.

Greedy achieves the shortest optimization time but deviates significantly from the GlobalOpt objective. In contrast, LPC-NS-QPBO maintains a comparable runtime to Greedy for smaller sample sizes while achieving a substantially lower deviation, indicating a near-optimal solution for all samples in $K = 2$ but exhibiting higher error for $K = 3$. (Fig. 3-left) further shows that LPC-NS-QPBO optimizes 2000 samples in $K = 2$ within the same time limit, whereas GlobalOpt reaches its limit at 200 samples. Scalability declined significantly for all methods in $K > 2$ (Fig. 3-right), as samples exhibited less distinct relationships across clusters. Although LPC-NS-MIP requires a longer optimization time, it consistently achieves a more optimized solution than Greedy in both cases while remaining significantly more efficient than GlobalOpt. GlobalOpt attains the optimal objective value but at a higher computational cost, with the efficiency gap widening as the number of samples increases.

RQ2 We evaluate deviations from the GlobalOpt objective and the performance of recovering $\mathbf{w}^{\text{gt}}$ and $\mathbf{Z}^{\text{gt}}$ across different methods under varying conditions. We analyze the impact of varying the level of Gaussian noise in the target variable (Fig. 5), feature dimensionality (Fig. 4), and the proportion of outliers (Fig. 6) in the target variable. LPC-NS-MIP consistently outperforms Greedy for both $K = 2$ and $K = 3$ (cf. Appendix F), achieving a significantly smaller gap from the globally optimal objective. LPC-NS-QPBO also outperforms Greedy for $K = 2$ but performs suboptimally for $K = 3$. Additionally, both proposed methods achieve higher accuracy in recovering ground truth $\mathbf{Z}$ and $\mathbf{w}$ for $K = 2$, while LPC-NS-MIP maintains its superiority in $K = 3$.

The performance of Greedy deteriorates in high-dimensional settings (Fig. 4), whereas LPC-NS-QPBO remains robust, maintaining near-optimal objective values. In addition, LPC-NS-MIP continues to outperform Greedy as the proportion of outliers increases.

RQ3 To assess performance in real-world scenarios, Tab. 2 and App. G.3 report the performance of different methods[2] across 10 real-world datasets, along with the separability measure $\mathcal{E}_{\text{sep}}$. LPC-NS-MIP and LPC-NS-QPBO outper-

[2]Due to scalability constraints, GlobalOpt was not evaluated as it is infeasible for most datasets within a reasonable runtime.

Dataset	SSE			Separability
Dataset (UCI Repository ID)	LPC-NS-QPBO	LPC-NS-MIP	Greedy	$\mathcal{E}_{\text{sep}}$
Stock Portfolio Performance (390)	16.1798	**12.4745**	17.1278 ± 1.6503	0.0000
Servo (87)	12.4865	**2.5305**	4.0431 ± 1.1435	0.0543
Solar Flare (89)	**3.8289**	11.0511	8.3014 ± 1.042	0.5885
Productivity Prediction (597)	41.4977	**38.2234**	43.4662 ± 2.7841	0.5914
Heart Failure Prediction (519)	4.733×10^{-30}	$\mathbf{6.920 \times 10^{-32}}$	25.0644 ± 8.6361	0.7685
Liver Disorders (60)	108.3983	**94.9041**	112.0968 ± 5.4914	0.9784
Student Performance (320)	**4.8812**	5.6557	6.8896 ± 0.4574	2.0107
Parkinsons Telemonitoring (189)	**135.3372**	146.9704	174.6422 ± 12.7622	3.9580
Facebook Metrics (368)	2.882×10^{-20}	$\mathbf{2.0464 \times 10^{-20}}$	$3.1 \times 10^{-5} \pm 1.2 \times 10^{-5}$	5.6163
Infrared Thermography (925)	1.080×10^{-13}	$\mathbf{6.904 \times 10^{-15}}$	0.1029 ± 0.038	7.9983

Table 2: *RQ3* Performance across different methods in 10 real datasets. GlobalOpt is excluded due to infeasible runtime on this scale. Mean and 95% confidence interval in Greedy (20 initializations) is reported.

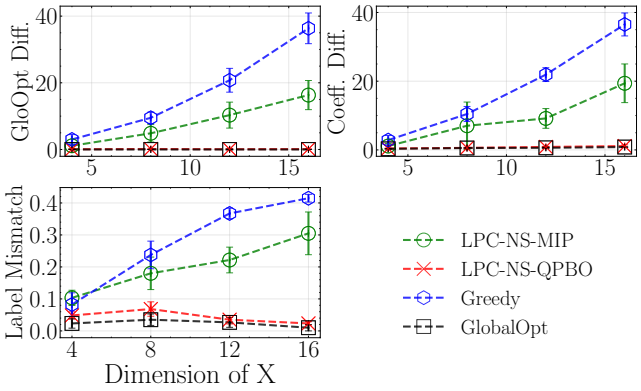


Figure 4: *RQ2 Feature Variables* The performance difference across methods as the number of feature variables increases.

form Greedy. In some separable datasets, LPC-NS surpasses Greedy in cases where relationships are easier to identify (low SSE), where LPC-NS is more robust to noise, or when clusters are highly imbalanced – scenarios in which Greedy fails to recover correct partitions.

We also observed that the performance of LPC-NS-MIP and LPC-NS-QBPO is highly dependent on the dataset. If scalability is a concern, LPC-NS-QBPO is the preferred choice, as it significantly improves scalability with comparable performance in the majority of cases.

6 Conclusion

With the aim of developing scalable and near-optimal LPC optimization methods, we leverage properties of non-separable feature spaces to derive novel MIP and QPBO reductions from an existing globally optimal MIP formulation. Specifically, LPC-NS-MIP offers a more compact and efficient formulation than GlobalOpt while maintaining near-optimal performance in non-separable settings. LPC-NS-QPBO further enhances scalability via a QPBO reduction. Extensive experiments show that LPC-NS-MIP consistently outperforms Greedy in optimizing the objective and recovering hidden cluster assignments and regression coefficients across varying noise, dimensions, and outliers, while LPC-NS-QPBO remains highly competitive for $K = 2$.

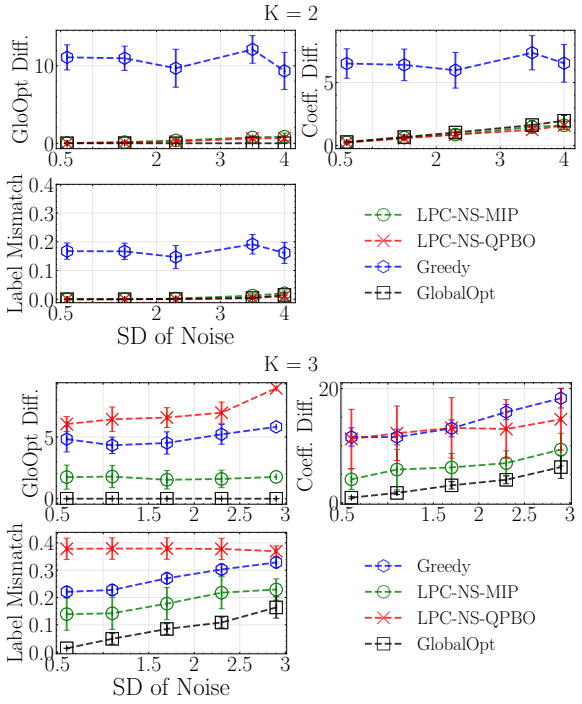


Figure 5: *RQ2 Noise* The performance difference across different methods in an increasing level of Gaussian noise.

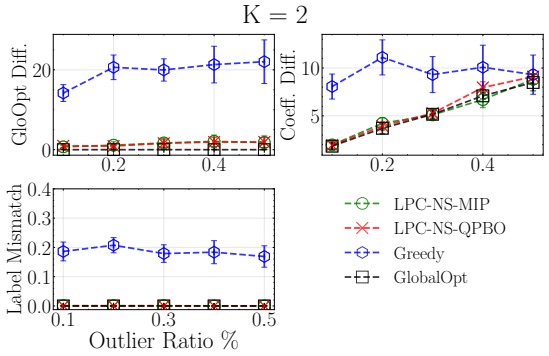


Figure 6: *RQ2 Outliers* The performance of different methods as the proportion of outliers in target variables increases.

References

Bertsimas, D.; and Shioda, R. 2007. Classification and regression via integer optimization. *Operations research*, 55(2): 252–271.

Bolusani, S.; Besançon, M.; Bestuzheva, K.; Chmiela, A.; Dionísio, J.; Donkiewicz, T.; van Doornmalen, J.; Eifler, L.; Ghannam, M.; Gleixner, A.; Graczyk, C.; Halbig, K.; Hedtke, I.; Hoen, A.; Hojny, C.; van der Hulst, R.; Kamp, D.; Koch, T.; Kofler, K.; Lentz, J.; Manns, J.; Mexi, G.; Mühmer, E.; Pfetsch, M. E.; Schlösser, F.; Serrano, F.; Shinano, Y.; Turner, M.; Vigerske, S.; Weninger, D.; and Xu, L. 2024. The SCIP Optimization Suite 9.0. Technical report, Optimization Online.

Bradley, P. S.; and Mangasarian, O. L. 2000. k-Plane Clustering. *J. of Global Optimization*, 16(1): 23–32.

Carbonneau, R. A.; Caporossi, G.; and Hansen, P. 2012. Extensions to the repetitive branch and bound algorithm for globally optimal clusterwise regression. *Computers & operations research*, 39(11): 2748–2762.

Chembu, A.; and Sanner, S. 2023. A Generalized Framework for Predictive Clustering and Optimization. *arXiv preprint arXiv:2305.04364*.

Corizzo, R.; Pio, G.; Ceci, M.; and Malerba, D. 2019. DENCAST: distributed density-based clustering for multi-target regression. *Journal of Big Data*, 6(1): 43.

Cosslett, S. R.; and Lee, L.-F. 1985. Serial correlation in latent discrete variable models. *Journal of Econometrics*, 27(1): 79–97.

da Silva, R. A.; and de Carvalho, F. d. A. 2017. On combining clusterwise linear regression and k-means with automatic weighting of the explanatory variables. In *Artificial Neural Networks and Machine Learning–ICANN 2017: 26th International Conference on Artificial Neural Networks, Alghero, Italy, September 11-14, 2017, Proceedings, Part II 26*, 402–410. Springer.

Dantzig, G. B. 2016. Linear programming and extensions. In *Linear programming and extensions*. Princeton university press.

DeSarbo, W. S.; and Cron, W. L. 1988. A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5: 249–282.

Feng, W.; Lim, C. Y.; Maiti, T.; and Zhang, Z. 2016. Spatial regression and estimation of disease risks: A clustering-based approach. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(6): 417–434.

Gurobi Optimization, LLC. 2024. Gurobi Optimizer Reference Manual.

Hamilton, J. D. 1990. Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1-2): 39–70.

Hota, R. N.; Syed, S.; Bandyopadhyay, S.; and Krishna, P. R. 2009. A Simple and Efficient Lane Detection using Clustering and Weighted Regression. In *COMAD*.

Hunter, D. R.; and Lange, K. 2004. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37.

Kelly, M.; Longjohn, R.; and Nottingham, K. 2023. The UCI machine learning repository. *URL https://archive. ics. uci. edu*.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137.

Manwani, N.; and Sastry, P. 2015. K-plane regression. *Information Sciences*, 292: 39–56.

Naik, P.; Shaikh, R. P.; Diukar, O.; and Dessai, S. 2017. Predicting Student Performance Based On Clustering And Classification. *IOSR Journal of Computer Engineering*, 19: 49–52.

Ntani, G.; Inskip, H.; Osmond, C.; and Coggon, D. 2020. Consequences of ignoring clustering in linear regression. *BMC Medical Research Methodology*, 21.

Quandt, R. E. 1972. A new approach to estimating switching regressions. *Journal of the American statistical association*, 67(338): 306–310.

Späth, H. 1979. Algorithmus 39. Klassenweise lineare Regression. *Computing*, 22: 367–373.

Wang, Y. 2020. *Interpretable machine learning methods with applications to health care*. Ph.D. thesis, Massachusetts Institute of Technology.

Wedel, M.; and DeSarbo, W. S. 1994. A review of recent developments in latent class regression models. *Advanced methods of marketing research*, 352–388.

Wedel, M.; and Kistemaker, C. 1989. Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing*, 6(1): 45–59.

Zhu, Z.; Li, Y.; and Kong, N. 2012. Clusterwise linear regression with the least sum of absolute deviations–an mip approach. *International Journal of Operations Research*, 9(3): 162–172.