Weighted Clusterwise Linear Regression Based on Adaptive Quadratic Form Distance

Jovan Moris D CS23B2058, Dinesh Kumarun K CS23B2057

November 2, 2024

Outline

Introduction

- Introduction to regression and clustering in data analysis.
- Traditional single-model regression approaches may overfit or fail to capture local patterns.
- Objective: Improve clustering regression methods by developing a weighted Clusterwise Linear Regression (WCLR) approach.

Problem Definition

- Clusterwise Regression (CR) seeks to partition data into clusters, each with its own regression model.
- Challenges:
  - Standard CR may lead to data overfitting.
  - Limited capacity to distinguish clusters in the explanatory variable space.
- WCLR: Adds adaptive distance to account for variable correlation and cluster homogeneity.

Objective

- The aim of WCLR is to achieve homogeneous clusters with good regression fitting by minimizing an optimization criterion combining clustering and regression loss.
- This approach utilizes an adaptive quadratic form distance metric to weight variables based on relevance.

Optimization Criterion

Combined Criterion

$$J_{wclr} = \sum_{k=1}^{K} \sum_{i=1}^{n} V_{ik} \left[d^2_{W_k}(\mathbf{x}_i, \mathbf{g}_k) + \alpha(y_i - \tilde{\mathbf{x}}_i^\top \mathbf{b}_k)^2 \right]$$

where $d^2_{W_k}(\mathbf{x}_i, \mathbf{g}_k)$ is the adaptive quadratic distance, and α is a hyperparameter.

- Combines a clustering criterion with regression residual minimization.
- Enables variable weighting through the matrix W_k.

Quadratic Form Distance

Distance Definition

$$d^2_{W_k}(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_i - \mathbf{g}_k)^\top W_k (\mathbf{x}_i - \mathbf{g}_k)$$

where W_k is a positive definite matrix, enabling non-spherical clusters.

- W_k adjusts based on the relevance of explanatory variables.
- Six constraint types explored for W_k matrices, including both local and global adaptive distances.

Optimization Algorithm - Overview

- The algorithm iteratively refines the cluster assignments and regression parameters.
- Steps include: Representation, Weighting, Modeling, and Assignment.
- Each step is designed to optimize the WCLR objective function iteratively until convergence.

Representation Step

- Goal: Update cluster centroids based on current assignments.
- Computation:

$$\hat{\mathbf{g}}_k = \frac{\sum_{i=1}^{n} V_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} V_{ik}}, \quad \text{for each cluster } k.$$

Weighting Step

- Goal: Optimize the weight matrix W_k to reflect variable relevance.
- Computation involves constraints such as:

$$\det(W_k) = 1 \quad \text{or} \quad \sum_{j=1}^{p} w_{jk} = 1.$$

Modeling Step

- Goal: Estimate regression coefficients $\mathbf{b}_k$ for each cluster.
- Solution:
$$\hat{\mathbf{b}}_k = (\tilde{\mathbf{X}}_k^\top \tilde{\mathbf{X}}_k)^{-1} \tilde{\mathbf{X}}_k^\top \mathbf{y}_k,$$
where $\tilde{\mathbf{X}}_k$ is the augmented design matrix.

Assignment Step

- Goal: Reassign data points to clusters to minimize J_{wclr}.
- For each observation i, assign to the cluster k that minimizes:

$$d^2_{W_k}(\mathbf{x}_i, \mathbf{g}_k) + \alpha(y_i - \tilde{\mathbf{x}}_i^\top \mathbf{b}_k)^2.$$

Algorithm Summary

- The algorithm iteratively updates $\mathbf{g}_k$, W_k, $\mathbf{b}_k$, and V_{ik}.
- Convergence: The objective function J_{wclr} is non-increasing, leading to a local optimum.
- Multiple initializations suggested to avoid local minima.

Experimental Evaluation

- Synthetic and benchmark datasets used for evaluation.
- Comparison with standard CR and K-plane methods.

Results - Performance Table

Method	Synthetic Data	Benchmark Data 1	Benchmark Data 2
WCLR	**0.85**	0.78	**0.82**
Standard CR	0.75	0.70	0.76
K-Plane	0.80	**0.79**	0.81

Table: Performance comparison of WCLR with other methods.

Conclusion

- WCLR effectively handles heterogeneous datasets in regression clustering tasks.
- Adaptive quadratic distance enhances clustering by considering variable relevance.
- Future work: Extensions to non-linear models and robustness to outliers.