International Journal of Operations Research

International Journal of Operations Research Vol. 9, No. 3, 162–172 (2012)

Clusterwise Linear Regression with the Least Sum of Absolute Deviations – An MIP Approach

Zhen Zhu[1], Yan Li[2] and Nan Kong[2*]

[1]The School of Industrial Engineering, Purdue University, 315 N. Grant Street, West Lafayette, IN, 47907

[2]The Weldon School of Biomedical Engineering, Purdue University, 206 S. Martin Jischke Drive, West Lafayette, IN, 47907

Received April 2012; Revised September 2012; Accepted September 2012

Abstract — In this paper, we study the application of mixed-integer programming (MIP) to the clusterwise linear regression (CLR) problem with the least sum of absolute deviations, which is a type of CLR problem that has received both theoretical and practical interests in recent years. We formulate the problem with a big-M formulation and investigate related issues, including the integration of outlier detection into CLR analysis. To improve the global optimization solution, we explore the resolution of breaking the solution symmetry that is prevailing in conventional formulations of many clustering analysis problems. Our numerical studies on randomly generated problem instances and two real data sets offer insights into the computational performance of solving the MIP formulations.

Keywords — Integer programming, cluster analysis, linear regression, outlier detection

1. INTRODUCTION

Mathematical programming (MP) approaches have received considerable attention in statistical analysis since the seminal work by Charnes, Cooper and Ferguson (1955). Only to cite a few works in the past decades, Arthanari and Dodge (1981), Bertsimas and Shioda (2007), and Agullo (2001) in regression analysis; Aronson and Klein (1989), and Stanfel (1981,1986) in cluster analysis; and Nguyen and Welsch (2010a, 2010b), and Zioutas and Avramidis (2005) in outlier detection. However, most MP models are designed for one data analysis task. On the other hand, most real-world data require the combination of several tasks, e.g., clusterwise regression analysis that integrates cluster and regression analysis.

In marketing research and practice, benefit segmentation is extensively used to provide indications for marketing (Beane and Ennis 1987, Wind 1978). The benefit segmentation problem can be described with clusterwise regression model (Punj and Stewart 1983, Wedel and Kistemaker 1989). Clusterwise regression has also been discussed in response based segmentation of customers, regions, subjects, strategies or investors (Carbonneau *et al.*, 2011; Hennig, 2000; Lauet *et al.*, 1999; Spath, 1979). As a motivating example to this investigation of clusterwise regression analysis, we consider a service spending segmentation problem in analyzing publicly funded traumatic brain injury (TBI) inpatient rehabilitation service payment claim data. This segmentation problem is important to publicly funded health insurance policy makers. In this problem, a public insurer intends to model the interaction of service duration and spending based on the data collected from a cohort of TBI patients who receive inpatient rehabilitation services. If the TBI patients have homogeneous duration elasticity, the elasticity can simply be estimated by a regression of the spending on the duration. However, in the real world, the patients are heterogeneous on the service duration, depending on their injury severity, acute hospitalization condition, and rehabilitation service usage behavior. If one ignores the duration elasticity, the estimated spending elasticity would certainly be biased and inaccurate. Therefore, for the public insurer, the task is to identify mutually exclusive segments that partition the patients on the basis of duration elasticity and conduct a regression within each cluster. This leads to a clusterwise regression problem.

In summary, the analysis of real data set often involves simultaneous application of several related statistical models. In the case where one realizes through data collection that a set of linear relationships may well explain the data set, it may be required to apply clustering and linear regression models simultaneously. The traditional regression model assumes the regression coefficients to be identical for all subjects in the sample. This homogeneity assumption is sometimes unrealistic. On the other hand, the clusterwise regression model assumes identical coefficients for members within a cluster, which is more suitable for many real-world data modeling applications.

To integrate cluster analysis into a regression framework, clusterwise linear regression has been investigated extensively. This type of regression is based on the concept of fitting multiple lines to mutually exclusive subsets of the data. Therefore,

* Corresponding author's email: nkong@purdue.edu

it is a clustering problem with the objective of finding a pre-determined number of lines that best fit the data in some form of minimization. Spath (1979, 1982, 1986) proposed the so-called exchange algorithms which use the QR-decomposition technique to minimize the sum of square errors in the integrated model. The exchange algorithms were further adapted by Meier (Meier, 1987) to minimize the sum of absolute deviations. Although these algorithms are easy to implement, they are not exact solution methods and their performance is sensitive to the initial partition and outliers. More recently, metaheuristics were developed to solve the CLR problem, including genetic algorithms (Aurifeille, 2000; Aurifeille and Medlin, 2001), variable neighbor search (Caporossi and Hansen, 2005), and simulated annealing (Desarbo, Oliver and Rangaswamy 1989). Although the results of these metaheuristics are encouraging, they are sensitive to algorithm parameters.

In this paper, we focus on the MP formulation of clusterwise linear regression (CLR) analysis. MP based solution methods offer a generic set of methods for identifying global optimal solutions to certain types of CLR problems. Lau, Leung and Tse (1999) stated that the CLR problem is a hard combinatorial optimization problem and conjectured that it is NP-hard given its similarity with the set covering problem. This well explains why the focus of the literature before 2000 had been the development of heuristic methods. However, applying MP to the CLR problem is of theoretical and practical interest. Such interest is sustainable as continuing innovation on MP based solution methods is more feasible and desirable than that on the heuristics. First, global optimization solutions lead to better CLR models than those derived from random local optima identified by the heuristics. Second, several decades' development on MP-based solution for clustering and regression models lay solid foundation on developing efficient methods for solving MP formulations of CLR models. Meanwhile, rapid development of commercial mixed-integer programming solvers in recent years makes efficient solution of industry-size CLR instances become increasingly possible. Third, a global optimization algorithm can serve as a building block for developing efficient heuristics by finding optimal solutions for subsets of subjects, which can subsequently be used to analyze the entire set of subjects. Finally, globally optimal solutions can be useful to further research on quality measures of CLR models.

In this paper, we focus on the CLR problem with the least sum of absolute deviations (LSAD) over clusters. In other words, we identify a fixed number of clusters and for each cluster we estimate the parameters in linear regression. Our objective is to identify the clusters such that we minimize the sum of absolute deviation of each subject to its corresponding linear regression model. I term this problem the *CLR-LSAD problem* and use this acronym in the remainder of the paper. Considering regression alone, the estimation technique using the least absolute deviation (LAD) is more applicable to a board range of cases than the estimation technique using the least square error (LSE) as the latter one has to assume that the regression residual is normally distributed. Narula and Wellington (1982) concluded that 1) the LAD estimator is more robust, i.e., less sensitive than the LSE estimator to the presence of outliers; and 2) the LAD estimator tends to be superior to the LSE estimator in many non-Gaussian cases, especially those when the residual follows distributions with long tails. Bassett and Koenker (1978) showed that the LAD estimator has a strictly smaller confidence ellipsoid than the LSE estimator in cases where the residual follows a distribution whose sample median is a more efficient estimator of location than the sample mean.

In this paper, we first present a big-M formulation for the CLR-LSAD problem, which is a mixed-integer programming (MIP) formulation. We next develop an MIP formulation for integrating outlier detection into the CLR analysis framework. To speed up the global optimization solution, we introduce a class of constraints that are used to break the solution symmetry among clusters. Finally, we use both randomly generated and real-world data to conduct a thorough investigation on the factors that affect the computational performance of solving CLR-LSAD directly via standard branch and bound. Our main contribution in this paper is using an MIP framework to integrate clustering, regression, and outlier detection. Our numerical studies gained insights into the computational aspect of applying MP-based methods to CLR-LSAD and CLR in general.

The remainder of the paper is organized as follows. In Section 2, we present several MIP formulations for the integrated CLR analysis framework. In Section 3, we introduce a set of symmetry-breaking constraints to speed up the global optimization method. In Section 4, we report our numerical studies on how the solution of CLR-LSAD is affected by instance characteristics and the symmetry-breaking constraints. We offer concluding remarks and outline future research in Section 5.

2. AN MIP APPROACH TO CLUSTERWISE LINEAR REGRESSION ANALYSIS

2.1 Model Description

We take a sample of n subjects from a studied population, namely $\left\{x_1^i,\ldots,x_m^i,y^i\right\}$, $i=1,\ldots,n$. Each subject consists of m independent variables (e.g., education, age, gender, etc.) and one dependent variable (e.g., income). We are asked to divide the samples into K mutually exclusive segments, each of which forms a linear regression model. Our task is to cluster the subjects and conduct linear regression in each cluster. Our objective is to minimize the sum of the absolute deviation for each subject to its corresponding linear regression model.

Suppose each of the K clusters is associated with the linear regression model $f_k\left(x\right)=\alpha_0^k+\sum_{j=1}^{m}\alpha_j^k x_j$, where $x=(x_1,\ldots,x_m)$. We first formulate the above CLR problem with a nonlinear MIP formulation. Then we present a linearized reformulation and prove its equivalence. Let binary decision variable ζ_i^k indicate whether subject i belongs to cluster k, $k=1,\ldots,K$. If subject i, $i=1,\ldots,n$. is in cluster k, then $\zeta_i^k=1$; otherwise, $\zeta_i^k=0$. Hence, we have $\sum_{k=1}^{K}\zeta_i^k=1$. Thus the specification of subject i is $\hat{y}^i=\alpha_0^k+\sum_{j=1}^{m}\alpha_j^k x_j^i$ if $\zeta_i^k=1$. The original nonlinear MIP formulation is presented as:

$$\text{(P0):}\quad \min_{\alpha,\zeta}\sum_{k=1}^{K}\sum_{i=1}^{n}\mid\alpha_0^k+\sum_{j=1}^{m}\alpha_j^k x_j^i-y^i\mid\cdot\zeta_i^k \tag{1}$$

$$\text{s.t.}\quad \sum_{k=1}^{K}\zeta_i^k=1,\quad i=1,\ldots,n; \tag{2}$$

$$\zeta_i^k\in\left\{0,1\right\},\left(\alpha_0^k,\alpha_1^k,\ldots,\alpha_m^k\right)\in\Re^{m+1},i=1,\ldots,n,k=1,\ldots,K. \tag{3}$$

Note that both ζ_i^k and α_j^k, i = 1, …, n, j = 1,…,m, and k = 1,…,K, are decision variables in (P0). Constraints (2) indicate that each subject should belong to exactly one cluster. For the subjects that belong to the same cluster, say k, the optimal solution on α_0^k and α_i^k imply the linear regression model.

Several heuristics have been developed to solve (P0) (Desarbo *et al.*, 1989; Spath 1979, 1982), but these methods do not guarantee global convergence and their computational performances are influenced by the initial solutions. Lau *et al.* (1999) applied an MP approach and proposed a nonlinear MIP formulation for the CLR problem with generic likelihood measures. Furthermore, to address the potential regression model overfitting issue, we impose lower bounds on the cardinality of each cluster in (P0) with the following additional constraints:

$$\sum_{i=1}^{n}\zeta_i^k\geq c,k=1,\ldots,K. \tag{4}$$

2.2 A Big-M MIP Reformulation

It is easy to see that (P0) can be naturally reformulated as a quadratic MIP. However, there are computational challenges to solve quadratic MIPs exactly with proof of optimality. To resolve these challenges, we propose a big-M MIP reformulation as:

$$\text{(P1):}\quad \min_{\alpha,\zeta,z^+,z^-}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\left(z_i^k\right)^+-\left(z_i^k\right)^-\right) \tag{5}$$

$$\text{s.t.}\quad \sum_{k=1}^{K}\zeta_i^k=1,\quad i=1,\ldots,n; \tag{6}$$

$$\alpha_0^k+\sum_{j=1}^{m}\alpha_j^k x_j^i-y^i+\left(z_i^k\right)^+-\left(z_i^k\right)^-\geq -M\left(1-\zeta_i^k\right),i=1,\ldots,n,\ k=1,\ldots,K; \tag{7}$$

$$\alpha_0^k+\sum_{j=1}^{m}\alpha_j^k x_j^i-y^i+\left(z_i^k\right)^+-\left(z_i^k\right)^-\leq M\left(1-\zeta_i^k\right),i=1,\ldots,n,\ k=1,\ldots,K; \tag{8}$$

$$\zeta_i^k\in\left\{0,1\right\},\left(z_i^k\right)^+,\left(z_i^k\right)^-\geq 0,\left(\alpha_0^k,\alpha_1^k,\ldots,\alpha_m^k\right)\in\Re^{m+1},i=1,\ldots,n,\ k=1,\ldots,K. \tag{9}$$

Here M is some large positive number. With constraints (7) and (8), we linearize the products of cluster indicator variables and continuously valued residuals. To use the big-M reformulation in actually solving the CLR-LSAD problem, the value of M must be specified sufficiently large to ensure the reformulation equivalence. It is clear that the smallest value for

M is the maximum distance among all subjects to any clusterwise regression line, which is not attainable beforehand. In this paper, we set M to be the maximum distance among all pairs of subjects, which can be computed in $O(n^2)$. Next in Theorem 1, we establish the equivalence between the MIP reformulation and the original nonlinear MIP formulation.

Theorem 1 *The two formulations (P0) and (P1) are equivalent. That is, solving them will obtain the same optimal objective function value and the optimal solution in terms of* α *and* ζ.

Proof Let $(\tilde{\alpha},\tilde{\zeta},(\tilde{z})^+,(\tilde{z})^-)$ be an optimal solution to (P1), where $\tilde{\alpha}$, $\tilde{\zeta}$, $(\tilde{z})^+$, and $(\tilde{z})^-$ are all row vectors of proper dimensions that piece together the respective decision variables, e.g. $\tilde{\zeta}=(\tilde{\zeta}_1^1,\ldots,\tilde{\zeta}_n^1,\tilde{\zeta}_1^2,\ldots,\tilde{\zeta}_1^K,\ldots,\tilde{\zeta}_n^K))$. It is easy to see that $(\tilde{\alpha},\tilde{\zeta})$ is a feasible solution to (P0). We then consider two cases for each pair $(i,k), i=1,\ldots n,$ and $k=1,\ldots,K$. First, we assume $\tilde{\zeta}_i^K=1$, it is clear that we have $(\tilde{z}_i^K)^+-(\tilde{z}_i^K)^-=y^i-\tilde{\alpha}_0^K+\sum_{j=1}^m\tilde{\alpha}_j^K x_j^i$ in (P1). Furthermore, to realize the optimality, we have $(\tilde{z}_i^K)^+(\tilde{z}_i^K)^-=0$ and their values are determined according to the sign of $y^i-\tilde{\alpha}_0^K+\sum_{j=1}^m\tilde{\alpha}_j^K x_j^i$. In addition, with respect to the pair (i,k), the objective function value in (P1) associated with $(\tilde{\alpha},\tilde{\zeta})$ is equal to that in (P0) for $(\tilde{\alpha},\tilde{\zeta},(\tilde{z})^+,(\tilde{z})^-)$ In the second case, we assume $\tilde{\zeta}_i^K=0$, then there is no restriction on $(\tilde{z}_i^K)^+$ and $(\tilde{z}_i^K)^-$ except for nonnegativity. Hence, to realize the optimality, we have $(\tilde{z}_i^K)^+=(\tilde{z}_i^K)^-=0.$ Meanwhile, $|\tilde{\alpha}_0^k+\sum_{j=1}^m\tilde{\alpha}_j^k x_j^i-y^i|\cdot\tilde{\zeta}_i^K=0$ in (P0). This implies that the two objective function values are also identical in the second case with respect to the pair (i,k). Either of the two cases applies for each (i,k) pair, $i=1,\ldots,n$ and $k=1,\ldots,K$. Then the two cases together imply that solving (P0) yields a larger optimal objective function value than solving (P1). Next we prove the statement of the opposite direction to complete the proof.

Let $(\tilde{\alpha},\tilde{\zeta})$ be an optimal solution to (P0). We construct a feasible solution to (P1) based on $(\tilde{\alpha},\tilde{\zeta})$ as follows. First, we keep $(\hat{\alpha},\hat{\zeta})$ for the α and ζ components. Then for each pair (i,k), $i=1,\ldots,n$ and $k=1,\ldots,K$, we need to specify $(\hat{z}_i^K)^+$ and $(\hat{z}_i^K)^-$ to satisfy the constraints (7) and (8) with respect to (i,k). For the specification, we consider two cases, $\hat{\zeta}_i^K=1$ and $\hat{\zeta}_i^K=0$. If $\hat{\zeta}_i^K=1$, we can set either $(\hat{z}_i^K)^+$ or $(\hat{z}_i^K)^-$ to be $|\hat{\alpha}_0^k+\sum_{j=1}^m\hat{\alpha}_j^k x_j^i-y^i|$, depending on the sign of $\hat{\alpha}_0^k+\sum_{j=1}^m\hat{\alpha}_j^k x_j^i-y^i$, and the other decision variable to be 0; If $\hat{\zeta}_i^K=0$, we can set $(\hat{z}_i^K)^+=(\hat{z}_i^K)^-=0$ It is clear that with the specification, $(\hat{\alpha},\hat{\zeta},(\hat{z}_i^K)^+,(\hat{z}_i^K)^-)$ satisfies the constraints (7) and (8) with respect to (i,k). Then the two cases together imply that solving (P1) yields a larger optimal objective function value than solving (P0) with the above specification on $(\hat{z}^+,\hat{z}^-)$. Therefore, the result of the theorem follows. ■

Solving (P1) yields an optimal solution on ζ and α. With the optimal ζ_i^k, we can specify which cluster the subject I belongs to. With the optimal solution on α_0^k and α_0^k, we can further specify the regression model. For our preliminary numerical studies presented in this paper, we solve (P1) directly using standard branch and bound. Note that potentially large values of M may lead to substantial computational burden and numerical instability (Lustig, 1990). We leave innovative computational considerations to our future research.

2.3 Outlier Detection in CLR Analysis

An implicit assumption in (P1) is that the data set does not contain outliers. Unfortunately, real-world data sets often contain outliers, for which solving (P1) can lead to bias and inaccurate clustering and regression lines. In this section, we extend the big-M reformulation in Section 2.2 to identify outliers together with conducting CLR analysis. We present an

MIP formulation that embeds outlier detection into (P1) as:

$$\text{(P2)} \quad \min_{\alpha,\zeta,z^+,z^-} \sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\left(\left(z_i^k\right)^+ + \left(z_i^k\right)^-\right) + T\cdot\eta_i\right\} \tag{10}$$

$$\text{s.t.} \quad \sum_{k=1}^{K}\zeta_i^k + \eta_i = 1, \quad i = 1,\ldots,n; \tag{11}$$

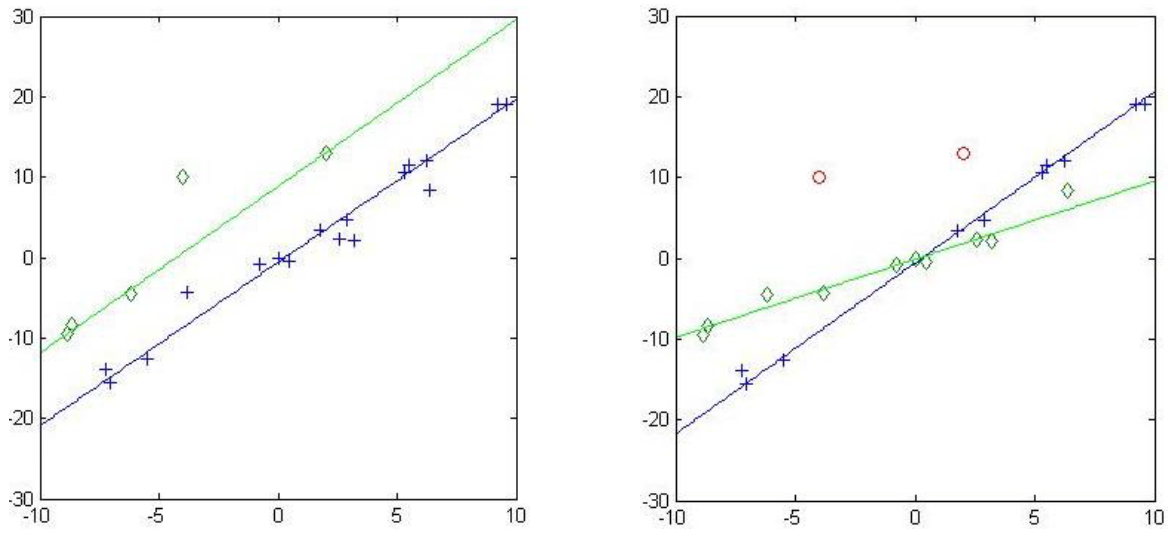
$$\alpha_0^k + \sum_{j=1}^{m}\alpha_j^k x_j^i - y^i + \left(z_i^k\right)^+ - \left(z_i^k\right)^- \geq -M\left(1-\zeta_i^k\right), i = 1,\ldots,n, k = 1,\ldots,K; \tag{12}$$

$$\alpha_0^k + \sum_{j=1}^{m}\alpha_j^k x_j^i - y^i + \left(z_i^k\right)^+ - \left(z_i^k\right)^- \leq M\left(1-\zeta_i^k\right), i = 1,\ldots,n, k = 1,\ldots,K; \tag{13}$$

$$\eta_i, \zeta_i^k \in \{0,1\}, \left(z_i^k\right)^+, \left(z_i^k\right)^- \geq 0, \left(\alpha_0^k, \alpha_1^k, \ldots, \alpha_m^k\right) \in \Re^{m+1}, i = 1,\ldots,n, k = 1,\ldots,K. \tag{14}$$

In the above formulation, for each $i = 1,\ldots,n,$ the additional binary decision variable η_i indicates whether subject *i* is regarded as an outlier. The penalty scalar *T* quantifies the threshold level for each subject being regarded as an outlier to any cluster. For a subject *i,* if the smallest residual among those to the various regression lines is still less preferred to paying the penalty, then the subject should be regarded as an outlier, i.e., $\eta_i = 1.$ Typically, one has some prior knowledge to determine the value of *T*.

In the remainder of this subsection, we use a numerical example to illustrate the effect of embedding outlier detection to CLR analysis. We consider the two underlying linear models, $y = x$ and $y = 2x$. We randomly generate ten subjects associated with each linear model. We arbitrarily select two outliers, each of which has an absolute deviation of at least 5 to either linear model. We specify $T = 3.$ The results are illustrated in Figure 1. Without outlier detection, the CLR analysis (i.e., solving (P1)) outputs two linear models. With outlier detection, the CLR analysis (i.e., solving (P2)) outputs drastically different linear models and identifies all the outliers.



(a) Without outlier detection (b) With outlier detection

Figure 1. An illustration of the importance of embedding outlier detection in CLR analysis. Note that the cardinality of each cluster must be greater than 5, with which we can exclude the possibility of grouping outliers into extra clusters.

3. DEALING WITH SOLUTION SYMMETRY

Many MIP formulations suffer from symmetry in their solution space, i.e. the formulation does not exclude alternative solutions that exist with the same objective function value and thus many of them may be checked during the exploration of the branch-and-bound tree. As a result, unnecessarily duplicate search may be conducted and thus computation time can increase significantly. Solution symmetry is a prevailing challenge in conventional formulations of many clustering problems.

A well known example is the node coloring problem (Campelo *et al.*, 2008; Mehrotra and Trick, 1996; Mendez-Diaz and Zabala, 2006), also known as the graph or vertex coloring problem. A set of symmetric solutions can be obtained by clusterwise permuting a feasible solution in (P0) and (P1). For example, suppose there are two clusters and four subjects, if a feasible solution is subjects 1 and 2 in cluster 1 and subjects 3 and 4 in cluster 2, then exchanging the cluster indices leads to an alternative feasible solution with the same objective function value. In this section, we present two ideas to eliminate the symmetry between the clusters which exists in (P0) and (P1).

First, we present an asymmetric representative formulation (ARF) for the CLR-LSAD problem, inspired by the ideas in Campelo *et al.* (2008) and Jans and Desrosiers (2010). Campelo *et al.* (2008) first introduced an ARF for the node coloring problem. Jans and Desrosiers (2010) generalized the idea to model a variety of clustering problems. The decision variables in these ARFs indicate whether a subject belongs to a specific cluster, but the cluster is identified by the lowest indexed subject. Let $\nu_i^h = 1$ if subject $i = 1,\ldots,n$ is in the same cluster with subject $h = 1,\ldots,n$ and subject h is the lowest numbered in that cluster; otherwise, let $\nu_i^h = 0$. We use the above defined asymmetric decision variables to present the ARF with the big-M notation as:

$$\min_{\alpha,\nu,z^+,z^-} \sum_{h=1}^{n}\sum_{i=h}^{n}\left(\left(z_i^h\right)^+ + \left(z_i^h\right)^-\right) \tag{15}$$

$$\text{s.t.} \quad \sum_{h=1}^{i}\nu_i^h = 1, i = 1,\ldots,n; \tag{16}$$

$$\nu_i^h \le \nu_h^h, i,h = 1,\ldots,n, i \ge h; \tag{17}$$

$$\sum_{h=1}^{n}\nu_h^h = K; \tag{18}$$

$$\alpha_0^{h,i} + \sum_{j=1}^{m}\alpha_j^{h,i}x_j^i - y^i + \left(z_i^h\right)^+ - \left(z_i^h\right)^- \ge -M\left(1-\upsilon_i^h\right), i,h = 1,\ldots,n, i \ge h; \tag{19}$$

$$\alpha_0^{h,i} + \sum_{j=1}^{m}\alpha_j^{h,i}x_j^i - y^i + \left(z_i^h\right)^+ - \left(z_i^h\right)^- \le M\left(1-\upsilon_i^h\right), i,h = 1,\ldots,n, i \ge h; \tag{20}$$

$$\upsilon_i^h \in \{0,1\}, \left(z_i^h\right)^+, \left(z_i^h\right)^- \ge 0, \left(\alpha_0^{h,i}, \alpha_1^{h,i}, \ldots, \alpha_m^{h,i}\right) \in \Re^{m+1}, i,h = 1,\ldots,n, i \ge h. \tag{21}$$

In the above ARF, we use constraints (16) – (18) to replace constraints (6) in (P1). It is easy to see the equivalence with the definition of asymmetric decision variables. We again introduce the big-M notation to realize the linearity in the above ARF. In a similar manner to the proof for Theorem 1, we can verify its correctness and equivalence to the original nonlinear ARF. Note that the difference from (P1) is that we need to index each regression line coefficient by both *i* and *h* in the above formulation since we do not have a cluster indicator from 1 to *K*.

In the above formulation, the numbers of variables υ, z^+, z^- are all $n(n+1)/2$. Additionally, we need to define $n(n+1)/2$ variables α_i for each subject dimension $j = 0,1,\ldots,m$. Hence, more decision variables are required compared to (P1), especially when *K* is small. Meanwhile, more constraints are also likely to be needed. As a result, the computational effort of solving the above formulation can be enormous, which is suggested by our preliminary computational experiments as well. In summary, although we can ensure symmetric solutions not to be explored in the branch-and-bound tree, large LP subproblems must be solved at each tree node. Furthermore, our experiments suggest that the computational time for our problem depends on the order of the input data, which is similar to the past experience in treating ARFs for other clustering problems (Jans and Desrosiers, 2010).

With these unsatisfactory features, we take a direct approach by introducing the following set of constraints to greatly alleviate the symmetry in (P1):

$$\zeta_1^1 = 1; \tag{22}$$

$$\sum_{s=k+1}^{K}\zeta_i^s \le \sum_{j=1}^{i-1}\zeta_j^k, \quad i = 2,\ldots,n-1, k = 2,\ldots,K-1. \tag{23}$$

As in many other clustering problems, symmetric solutions are primarily due to clusterwise permutation. Note that one can obtain $K!$ symmetric solutions from clusterwise permutation of each feasible solution, where *K* is the number of clusters.

The above constraints force that exactly one such symmetric solution is checked in the branch-and-bound tree. To be more specific, constraint (22) forces subject 1 to be assigned to cluster 1 and constraints (23) restricts the smallest indexed subject that is unassigned to be either the first subject assigned to k if no subjects are yet assigned to cluster k, or assigned to a cluster with a smaller index than k. In other words, constraints (23) ensure that such a subject cannot be assigned to a cluster indexed larger than k.

Theorem 2 *For any set of symmetric solutions to (P1) that are caused by clusterwise permutation, there is exactly one solution from the set that can satisfy constraints (22) and (23).*

Proof For each symmetric solution, uniquely identified by some ζ, we define $l_k(\zeta)$ to be the smallest index in cluster k, and $L(\zeta) = \{l_1(\zeta),\ldots,l_K(\zeta)\}$ to be the set that contains the smallest indices for clusters $k = 1,\ldots,K$. It is clear that the sets $L(\zeta)$ are identical for all symmetric solutions ζ. Constraint (22) specifies $l_1(\zeta)$ to be 1, and constraints (23) specify that $l_1(\zeta) < \ldots < l_K(\zeta)$. Since all the elements in $L(\zeta)$ are unique, there is one and only one ascending order among $l_i(\zeta), i = ,\ldots,K$. Hence there is a unique solution ζ that satisfies constraints (22) and (23).

Remark 1 When $K = 2$, only constraint (22) is required. As indicated earlier, the cardinality of any permutation solution set is $K!$. This implies that only two solutions in each permutation solution set when $K = 2$. It is thus sufficient to specify the cluster indicator for one subject with constraint (22).

4. NUMERICAL STUDIES

In this section, we use both randomly generated data sets and real-world data sets from the literature to test our proposed approach. The randomly generated instances provide us with more flexibility in designing experiments and exploring the performance of our algorithm in different scenarios, and solving the CLR-LSAD problems in real life further demonstrates the usability of our proposed approach. We solve both randomly generated instances and real instances using the CPLEX MIP solver with default settings. We conduct all the computational experiments on a PC with 16GB RAM and a CPU of 3.0GHz.

4.1 Randomly Generated Data

With randomly generated data sets, we intend to investigate 1) the influences of data set characteristics on the computational performance of solving the CLR-LSAD problem directly via a standard MIP solver; and 2) the effectiveness of the symmetry-breaking constraints introduced in Section 3. In particular, we considered planar subjects in all our experiments, i.e., $m = 1$. To construct a set of subjects, we specified K regression line models, i.e., $y = \alpha_0^k + \alpha_1^k x, k = 1,\ldots,K$. For each line k, $k = 1,\ldots,K$, we determined n_k. For each $i = 1,\ldots,n_k$, we first selected x_i^k randomly and determined $\overline{y}_i^k$ accordingly, and then drew a sample ϵ_i^k from a normal distribution $N(0,\sigma)$, where σ was set to be 1, and determine $y_i^k = \overline{y}_i^k + \epsilon_i^k$. After generating n_k subjects for line k, we continued to generate subjects for another line. We continued the generation procedure until $n = \sum_{k=1}^{k} n_k$ subjects had been generated.

For a set of subjects (x_i^k, y_i^k), $k = 1,\ldots,K$ and $i = 1,\ldots,n_k$, we characterized the set by counting the number of subjects that are one standard deviation within any other lines, i.e., a subject (x_i^k, y_i^k) is selected if $| y_i^k - \alpha_0^{k'} - \alpha_1^{k'} x_i^k | \le \sigma$ for some k' with $1 \le k' \ne k \le K$. We term this measure the *overlap count*. We next constructed an instance of the CLR-LSAD problem according to (P1). For each instance, we first solved (P1) and then solved the formulation with addition of constraints (22) and (23) for the comparison purpose. For each instance, we reported the CPU times taken to solve both formulations. We also reported the overlap count of the subject set to assess its influence on the computational performance.

In the first set of experiments, we considered two line models that cross the origin, i.e., $y = a_1 x$ and $y = a_2 x$. We set $y = x$ to be the reference line model, i.e., $a_1 = 1$, and varied a_2 to be 1.4, 1.8, 2.2, 2.6, and 3 to control the overlap count. We let $n_k = 25$ for $k = 1,$ and generate ten instances for each test value of a_2. We present our computational results in Table 1. The first column indicates the two line models we used to construct the CLR problem instances. Within

each instance group, the second row reports the percentage of subjects that are misclassified. The last two rows report the CPU times (in seconds) taken to solve (P1) without and with symmetry-breaking constraints (SBCs), respectively.

Table 1. Set 1 computational results. Reference line model is $y = x$; $K = 2$; $n_1 = n_2 = 25$.

Model	Instance no.	1	2	3	4	5	6	7	8	9	10
	Overlap Count (%)	18	20	22	22	24	24	24	24	24	28
$y = x$	Misclassified (%)	14	20	14	14	16	18	18	26	28	12
$y = 1.4x$	(P1)	338	2145	46	290	256	356	1022	239	973	463
	(P1) + SBC	271	892	22	158	137	172	333	213	409	350
	Overlap Count (%)	6	8	8	10	12	14	14	16	18	20
$y = x$	Misclassified (%)	6	2	6	8	14	6	6	8	18	16
$y = 1.8x$	(P1)	140	15.3	30.8	55.3	25.8	7.3	9.5	14.2	49.8	17.5
	(P1) + SBC	77.2	11.5	15.9	18.2	11.9	3.3	6.4	14.2	13.6	8.2
	Overlap Count (%)	4	6	6	8	8	10	10	10	10	14
$y = x$	Misclassified (%)	6	4	6	8	8	4	8	10	12	10
$y = 2.2x$	(P1)	17	37.5	12.3	9.5	10.4	6.5	30.7	2.4	6.3	6.2
	(P1) + SBC	7.8	13.1	7.0	4.8	3.2	3.1	5.4	2.0	3.1	2.9
	Overlap Count (%)	2	4	4	4	6	6	8	8	8	14
$y = x$	Misclassified (%)	2	0	2	2	4	4	4	6	8	6
$y = 2.6x$	(P1)	6.9	5.3	3.1	5.6	4.3	11.7	2.2	3.4	2.3	4.5
	(P1) + SBC	3.9	0.4	0.6	2.1	1	3.6	1.1	2.9	1.9	2.2
	Overlap Count (%)	2	4	4	4	4	6	8	8	8	12
$y = x$	Misclassified (%)	2	2	2	2	4	12	6	8	10	10
$y = 3.0x$	(P1)	1.4	0.8	0.9	1.6	1.6	1.0	1.8	2.6	1.7	5.6
	(P1) + SBC	0.5	0.5	0.4	0.9	0.6	0.9	1.1	1.2	1.0	3.0

SBC: symmetry-breaking constraints

The results in Table 1 suggest that the overlap count, in general, follows the angle between the two line models. That is, larger angles typically lead to fewer points that are associated with one line but are close to the other line as well. Consequently, larger angles, on average, make the instances easier to solve, and result in fewer misclassified subjects in the optimal solutions. As for the computational improvement with the addition of the symmetry-breaking constraints (see (20) – (21)), the results are encouraging with 2-4 times of speed up. The speed-up ratio seems to be insensitive to the angle between the two line models.

In the second set of experiments, we considered three line models $y = 3x + 4$, $y = -3x + 4$, and $y = 5x + 2$. We let $n_k = 12$ for $k = 1, 2, 3$ and generated five subject sets. We varied *K* to be 2, 3, and 4, and solved (P1) without and with the symmetry-breaking constraints, respectively. We present the computational CPU times (in seconds) for different numbers of clusters and different instances in Table 2. The results in Table 2 suggest that the solution time is significantly affected by the number of clusters. This observation matches our intuition as the instance size increases with the number of clusters. It is encouraging that the benefit of adding the symmetry-breaking constraints seems to become more noticeable as the number of clusters increases. Finally, the comparison between the two tables suggests that the speed-up ratio seems to increase as the number of subjects increases.

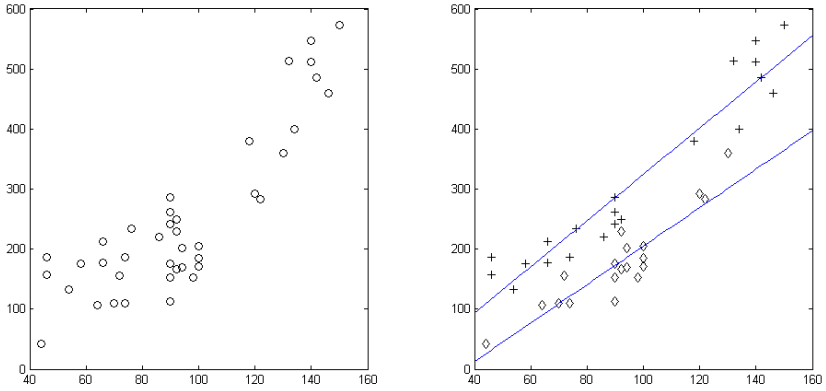
Table 2. Set 2 computational results. There line models are $y = 3x + 4$, $y = -3x + 4$, and $y = 5x + 2$; $n_1 = n_2 = n_3 = 12$.

	Instance no.	1	2	3	4	5
$\mathbf{K = 2}$	(P1)	0.63	0.81	0.86	0.95	1.31
	(P1) +SBC	0.40	0.17	0.31	0.80	1.14
$\mathbf{K = 3}$	(P1)	3.96	3.76	2.29	22.3	2.56
	(P1) +SBC	3.48	2.45	1.98	5.24	2.78
$\mathbf{K = 4}$	(P1)	517	739	226	877	371
	(P1) +SBC	182	465	162	291	181

Zhu, Li and Kong: *Clusterwise Linear Regression with the Least Sum of Absolute Deviations – An MIP Approach*
IJOR Vol. 9, No. 3, 162−172 (2012)

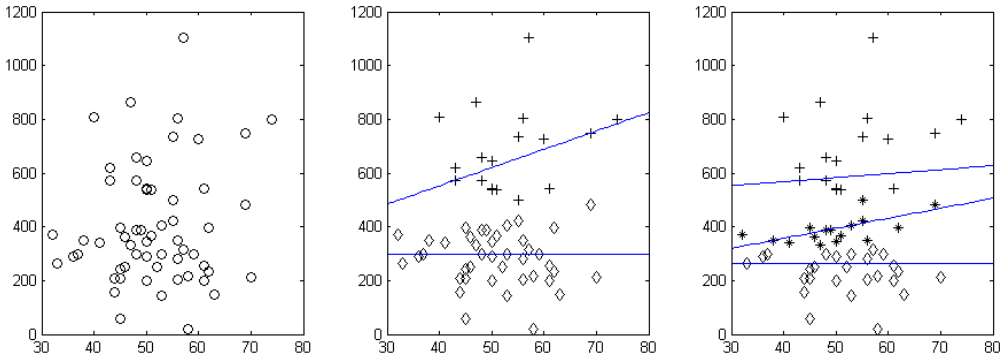
4.2 Real Data

In this section, we test our proposed approach with two real data sets obtained from the literature. The first data set contains the pricing details of 40 houses in a large city in the UK (Miles and Shevlin, 2001). The independent variable is the area of the house in square meters, and the dependent variable is the price of the house in thousands of pounds. The second data set contains the annual salaries of the chief executive officers (CEOs) for 59 small firms in the US (Velleman, 2010). The independent variable is the age of the CEO in years, and the dependent variable is the annual salary of the CEO in thousands of dollars. We show the computational results in Figure 2 and Figure 3.



(a) Original data set (b) Clustering result for K=2

Figure 2. CLR-LSAD analysis with the MIP approach for the house price data set



(a) Original data set (b) Clustering result for K=2 (c) Clustering result for K=3

Figure 3. CLR-LSAD analysis with the MIP approach for the CEO salary data set

Figure 2a is the scatter plot of the data points for the house price data set. Our observation from the scatter plot indicates that some house prices increase with a higher rate than others as the house area increases. Our clustering result shown in Figure 2b confirms our observation and clearly indicates the two distinct linear relationships. The CPU time is 90 seconds in this case. The CEO salary data set shown in Figure 3a does not show obvious linear relationship between the age and salary of CEO. Thus, we cluster the data points with both two clusters and three clusters and present the results in Figure 3b and 3c, respectively. These results offer insight into the interpretation of the relationship between the age and salary of CEO. For this data set, both CPU times are within 10 minutes.

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we investigate the mathematical programming application in CLR analysis. We introduce an MIP reformulation for the problem and extend the reformulation to incorporate outlier detection in CLR analysis. We introduce two symmetry-breaking approaches in response to computational challenges due to the solution symmetry caused by clusterwise permutation. For our preliminary numerical studies, we generate test instances randomly. We use these instances to assess the effects of various data set characteristics on the computational performance of solving our MIP reformulations directly via standard branch and bound. We also use them to verify the effectiveness of a set of symmetry-breaking constraints introduced in this paper. Finally, we demonstrate the usability of our proposed approach to perform clusterwise linear regression analysis for two data sets in real world.

A number of research directions may be worth undertaking in the future. First, more efficient solution methods, including efficient heuristics, should be explored to improve the computational performance of solving the MIP formulations. For example, we may consider embedding the exact solution to provide promising initial clustering for fast heuristics. We randomly select a subset of subjects with manageable size, and identify an optimal set of multiple regression lines by solving the reformulation with respect to the selected subset. We then associate the rest of the subjects to the derived regression line that offers the least absolute deviation. This method can be used repeatedly with multiple randomly selected initial subsets. We may also develop effective decomposition based methods as these methods have been developed for a variety of clustering analysis problems, e.g. Mulvey and Crowder (1979). Second, more comprehensive numerical studies should be conducted to investigate the computational performance with different data set characteristics (e.g., the values of K, n_k, σ, etc.) and test of effectiveness of the symmetry-breaking constraints. We also plan to test large-scale real-world data sets such as those used in Carbonneau *et al.* (2011).

REFERENCES

1. Agullo, J. (2001). New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis*, 36: 425-439.
2. Aronson, J. and Klein, G. (1989). A clustering-algorithm for computer-assisted process organization. *Decision Sciences*, 20: 730-745.
3. Arthanari, T.S. and Dodge, Y. (1981). *Mathematical Programming in Statistics*. Wiley, New York, NY.
4. Aurifeille, J.M. (2000). A bio-mimetic approach to marketing segmentation: Principles and comparative analysis. *European Journal of Economic and Social Systems*, 14: 93-108.
5. Aurifeille, J.M. and Medlin, C. (2001). A dyadic segmentation approach to business partnerships. *European Journal of Economic and Social Systems*, 15: 3-16.
6. Bassett, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73: 618-622.
7. Beane, T. and Ennis, D. (1987). Market-segmentation - A review. *European Journal of Marketing*, 21: 20-42.
8. Bertsimas, D. and Shioda, R. (2007). Classification and regression via integer optimization. *Operations Research*, 55: 252-271.
9. Campelo, M., Campos, V. and Correa, R. (2008). On the asymmetric representatives formulation for the vertex coloring problem. *Discrete Applied Mathematics*, 156: 1097-1111.
10. Caporossi, G. and Hansen, P. (2005). Variable neighborhood search for least squares clusterwise regression. Les Cahiers du GERAD, Montreal, Canada.
11. Carbonneau, R., Caporossi, G. and Hansen, P. (2011). Globally optimal clusterwise regression by mixed logical-quadratic programming. *European Journal of Operational Research*, 212: 213-222.
12. Charnes, A., Cooper, W. and Ferguson, R. (1955). Optimal estimation of executive compensation by linear programming. *Management Science*, 1: 138-151.
13. Desarbo, W., Oliver, R. and Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear-regression. *Psychometrika*, 54: 707-736.
14. Hennig, C. (2000) Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17: 273-296.
15. Jans, R. and Desrosiers, J. (2010). Binary clustering problems: Symmetric, asymmetric, and decomposition formulations. Les Cahiers du GERAD, Montreal, Canada.
16. Lau, K., Leung, P. and Tse, K. (1999). A mathematical programming approach to clusterwise regression model and its extensions. *European Journal of Operational Research*, 116: 640-652.
17. Lustig, I. (1990). Feasibility issues in a primal dual interior-point method for linear-programming. *Mathematical Programming*, 49: 145-162.
18. Mehrotra, A. and Trick, M.A. (1996). A column generation approach for graph coloring. *INFORMS Journal on Computing*, 8: 344-354.
19. Meier, J. (1987). A fast algorithm for clusterwise linear absolute deviations regression. *OR Spektrum*, 9: 187-189.
20. Mendez-Diaz, I. and Zabala, P. (2006). A Branch-and-Cut algorithm for Graph Coloring. *Discrete Applied Mathematics*, 154: 826-847.
21. Miles, J. and Shevlin, M. (2001). *Applying Regression & Correlation: A Guide for Students and Researchers*. Sage Publications, London.
22. Mulvey, J. and Crowder, H. (1979). Cluster-analysis - application of lagrangian relaxation. *Management Science*, 25: 329-340.
23. Narula, S. and Wellington, J. (1982). The minimum sum of absolute errors regression - a state of the art survey. *International Statistical Review*, 50: 317-326.
24. Nguyen, T. and Welsch, R. (2010a). Outlier detection and least trimmed squares approximation using semi-definite

programming. *Computational Statistics & Data Analysis*, 54: 3212-3226.
25. Nguyen, T. and Welsch, R. (2010b). Outlier detection and robust covariance estimation using mathematical programming. *Advances in Data Analysis and Classification*, 4: 301-334.
26. Punj, G. and Stewart, D. (1983). Cluster-analysis in marketing-research - review and suggestions for application. *Journal of Marketing Research*, 20: 134-148.
27. Spath, H. (1979). Clusterwise linear-regression. *Computing*, 22: 367-373.
28. Spath, H. (1982). A fast algorithm for clusterwise linear-regression. *Computing*, 29: 175-181.
29. Spath, H. (1986). Clusterwise linear least absolute deviations regression. *Computing*, 37: 371-378.
30. Stanfel, L. (1981). A lagrangian treatment of certain non-linear clustering problems. *European Journal of Operational Research*, 7: 121-132.
31. Stanfel, L. (1986). A recursive lagrangian method for clustering problems. *European Journal of Operational Research*, 27: 332-342.
32. Velleman, P.F. (2010). DASL, the Data and Story Library. Available at http://lib.stat.cmu.edu/DASL/DataArchive.html. Accessed on September 2, 2012.
33. Wedel, M. and Kistemaker, C. (1989). Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing*, 6: 45-59.
34. Wind, Y. (1978). Issues and advances in segmentation research. *Journal of Marketing Research*, 15: 317-337.
35. Zioutas, G. and Avramidis, A. (2005) Deleting outliers in robust regression with mixed integer programming. *Acta Mathematicae Applicatae Sinica*, 21: 323-334.