Novel mixed integer optimization sparse regression approach in chemometrics

D. Bertsimas\textsuperscript{a}, D. Lahlou Kitane\textsuperscript{a,\ast}, N. Azami\textsuperscript{b,\textsuperscript{c}}, F.R. Doucet\textsuperscript{c}

\textsuperscript{a} Operations Research Center, MIT, United States
\textsuperscript{b} Mohammed VI Polytechnic University, Morocco
\textsuperscript{c} ELEMISSION Inc., Canada

ARTICLE INFO

Article history:
Received 27 May 2020
Received in revised form
25 August 2020
Accepted 27 August 2020
Available online 2 September 2020

Keywords:
Mixed integer optimization
LIBS
Spectroscopy
Quadratic optimization
LASSO
Partial least squares

ABSTRACT

Sparse mathematical modelling plays an increasingly important role in chemometrics due to its interpretability and prediction power. While many sparse techniques used in chemometrics rely on $L_1$ penalization to create sparser models, Mixed Integer Optimization (MIO) achieves sparsity by imposing constraints directly in the model. In this paper, we develop an intuitive and flexible robust sparse regression framework using MIO. We use constraints and penalization to achieve sparsity and robustness respectively. We test and compare results with those obtained using other techniques generating sparser models such as LASSO and sparse PLS. We also use PLS as a baseline to compare predictive performance. We use a LIBS data set of certified reference materials (CRM) of various mineral ores to illustrate the framework using different objective functions. The MIO framework proposed improves accuracy, sparsity and robustness vs. LASSO and SPLS. MIO achieves an average $R^2$ higher than other methods on average by at least 10.6%. Robust MIO approach also improves interpretability. It also uses 4.3 variables on average while LASSO and SPLS use 16.1 and 805.8 respectively. We also illustrate how interpretability can help build better models through examples derived from the data sets used. When adding noise to the signal, MIO achieves an $R^2$ of 0.69 on average when all models have negative $R^2$ values. The MIO framework proposed is versatile and could be used in other chemometrics applications.

\ast Corresponding author. Massachusetts Institute of Technology, 77 Massachusetts Avenue, Bldg. E40-103, Cambridge, MA, 02139, USA.
E-mail address: driss@mit.edu (D. Lahlou Kitane).

1. Introduction

With the development of rapid data acquisition technologies, researchers and practitioners are often facing the problem of building regression models based on a limited number of...
observations compared to the number of variables acquired. This issue arises in many fields including genomics, finance and chemometrics. One of the main tools to overcome this difficulty is to use dimension reduction techniques. Such techniques aim at projecting the input data into a space of relatively low dimension while capturing the most relevant information to build a model predicting the outcome of new inputs. Principal Component Analysis (PCA) [1] and Partial Least Square (PLS) [2] have been widely used in this context. Both techniques aim at finding independent vectors (called principal components (PC) for PCA and Latent Variables (LV) for PLS) defining a space on which data is to be projected. One of the major issues of these approaches is that PCs and LVs are usually linear combinations of all variables which hinders their interpretability.

To improve interpretability, prediction power and reduce costs of equipment and experimentation, a limitation in the number of variables used to build regression models is desirable. Dimension reduction techniques such as PLS and PCA could be modified in this regard as these techniques can imbed in their formulation limitations of the number of variables used to build sparse regression models. Two main families of methods are used to achieve sparsity; namely Subset Selection methods and Shrinkage methods.

Subset selection methods aim at selecting a limited number of variables that are used to build a regression model that would minimize the error function considered in the regression. Finding a best subset minimizing the error (also called the Best Subset Selection problem (BSSP) [3]) is an NP-hard problem [4]. Indeed, the number of subsets of variables among n variables is 2^n which leads to strong combinatorial difficulties. Furnival and Wilson proposed in 1974 [5] the leaps-and-bound procedure to solve the Best Subset Selection problem using branch-and-bound approach, but this method received little attention in the chemometrics community due to its lack of scalability as the maximum number of variables it can handle in a tractable manner is around forty [6]. Sizeable efforts were then conducted to approach optimal solutions of BSSP. Although not solving the BSSP to optimality, other techniques such as Forward and Backward-Stepwise Selection [7], successive projection algorithms [8], simulated annealing [9], artificial neural networks [10] and genetic algorithms [11] proved to find solutions that improved the state-of-the-art in various fields. A vast literature has been developed using these techniques in chemometrics [12–14].

Shrinkage methods have been widely used to generate sparser models using penalization terms in the objective functions. L1 penalization has been used to generate sparsity in least squares regression (LASSO) [16]. Principal Component Analysis (Sparse PCA [17]) and Partial Least Square techniques (SPLS [18,19]). Sparse PCA algorithm uses L1 regularization term in the successive iterations to generate sparse PCs. An idea that was used to develop several versions of sparse PLS by modifying the PLS procedure to generate sparse LVs [18–20].

The formulation proposed is based on Mixed Integer Optimization (MIO). Unlike most variable selection techniques, the methodology proposed solves the BSSP to optimality retrieving a subset minimizing the error function considered in a tractable manner. One of the drawbacks of Best Subset Selection is the variance of its solution. Thus, we also use Tikhonov regularization term [23] to achieve robustness with mathematical guarantees. MIO techniques were regarded as intractable due to the high combinatorial complexity of such techniques. However, advances in computation power and MIO solving approaches both led to a dramatic improvement in the tractability of MIO problems either by the development of ad-hoc approaches to specific class of problems or the exponential improvement of generic solvers capabilities. Indeed, MIO solvers such as CPLEX and Gurobi experienced massive gains in speed between 1990 and 2016 becoming 1,250,000 faster in the same period. Tremendous improvement in hardware since the 1990 made them more than 2 trillion times faster [24–26]. By construction, this approach can also reduce significantly data pre-processing and performs regression and variables selection in one step.

We summarize our contribution in this paper below:

- We propose a MIO formulation to solve the calibration problem. The method proposed solves the BSSP optimality in a tractable manner for up to 100,000+ variables in some cases including the most widely used loss function such as least square with L1 and L2 penalization functions. In more general cases, commercial solvers can solve problems with 1000+ variables. To the best of our knowledge, this is the first time a MIO methodology is proposed in a chemometrics paper.
- We use a data set derived from various CRM mineral ores types to illustrate accuracy, robustness and sparsity compared to other sparse and robust methods. We use one of the chemical elements studied (K) to illustrate a possible use of sparsity and interpretability.
- We compare and discuss results and propose further development of the present work in the case of chemometrics.

The structure of the paper is as follows. We first provide background on MIO, BSSP and Robust Optimization in Section 2. We then present sparse regression methods using MIO in Section 3. We then propose a MIO formulation and methodology to use this technique in Spectroscopy context in Section 4. Sparse MIO regression is finally compared with other techniques on the data set proposed and results are discussed in Section 5.

2. Background

In this section, we first give a definition of BSSP that we use. We introduce Mixed Integer Optimization and show how best subset selection could be formulated as a quadratic optimization problem when the loss function considered is least squares. We finally present some of the core principles of robust optimization and the underlying strategy to tackle noise.

2.1. Best subset selection problem

Given inputs data \(X \in \mathbb{R}^{n \times p}\), a response \(Y \in \mathbb{R}^{n \times m}\) and an integer \(k\), the BSSP is finding a subset of variables of cardinality \(k\) minimizing the least square error. This could be formulated as follows:

\[
\min_\beta \| Y - X\beta \|_2^2
\]

s.t. \(\| \beta \|_0 \leq k\),

where \(\| \cdot \|_0\) represents the 0-norm i.e., the number of non-zero parameters.

This form of the problem has been widely studied and considered intractable for \(p \geq 40\). Subsequently, a vast literature and approaches have been developed for variables selection.

2.2. Mixed Integer Optimization

Given positive integers \(n\) and \(m\), \(X \subset \mathbb{R}^p\), \(Z \subset \mathbb{N}\) and \(f_i, (g_i)_{i \in \{1,...,m\}}\) and \((h_j)_{j \in \{1,...,p\}}\) real functions, a general formulation of a MIO problem is:
\[
\min_{x,z} f(x,z) \\
\text{s.t. } g_i(x,z) \leq 0 \forall i \in \{1, \ldots, m\}, \quad x \in X, \quad z \in Z.
\]

Cases in which \( f \) and \( (g_i)_{i \in \{1, \ldots, m\}} \) are smooth (or in particular quadratic) and convex and in which \( X \) is a compact and convex set are of particular interest as tractable methods have been recently developed to solve this category of problems as we develop in subsection 2.4.

2.3. Robust optimization

Robust optimization is a dynamic field in Operations Research and proved to be effective in numerous practical applications [27]. It aims at addressing uncertainty when solving an optimization problem. Consider the following optimization problem:

\[
\min_{\beta} f(X, \beta) \quad \text{(II)} \\
\text{s.t. } g(X, \beta) \leq 0,
\]

where \( X \) represents the parameters of the problem while \( \beta \) represents the variables. \( g \) is a vector-valued function and \( 0 \) represents the zero vector. Suppose now that \( X \) is not known with certainty (for instance in our case spectra signal and chemical composition) and that it is rather considered in an uncertainty set \( U \). Robust optimization aims at finding a solution to problem (II) that is close to optimal while adapting optimally to the uncertainty set considered. To illustrate this approach, consider on a plane, two given data points \( x_1 \) and \( x_2 \), the line going through these two points \( a^T x = b \) and the half plane delimited by inequality \( a^T x \geq b \). A deterministic approach would consider only one corresponding inequality in the constraints system. Consider now that \( x_1 \) and \( x_2 \) are known with uncertainty and that \( x_i \in U_i \subset \mathbb{R}^2 \) for \( i \in \{1, 2\} \). In this case, we would like, for any points \((x_1, x_2) \in U_1 \times U_2\), to account for the inequality \( a^T x \geq b \) such that the line \( a^T x = b \) goes through \( x_1 \) and \( x_2 \) (see Fig. 1 for an example in which \( U_1 \) and \( U_2 \) are balls centered in the data).

The set of inequalities resulting from robust modelling is generally infinite. However, the problem can still be solved considering an adversarial setting in which the optimization is made against the worst-case scenario. This is formulated as follows:

\[
\min_{\beta} \max_{X \in U} f(X, \beta) \quad \text{(II')} \\
\text{s.t. } g(X, \beta) \leq 0, \quad \forall X \in U.
\]

We consider the minimization over the worst-case scenario in the uncertainty set given \( \beta \). In our case, consider that the input data \( X \) is subject to noise \( \epsilon \) and we note \( X = \tilde{X} + \epsilon \) where \( \tilde{X} \) is the denoised signal (which is unknown). Consider also that \( \|\epsilon\|_q \leq \gamma \). With \( q \in \{1, 2, \infty\} \). In this case, we can consider \( U = \{X : \tilde{X} = X + \epsilon, \|\epsilon\|_q \leq \gamma\} \). Problem (II') could be written in this case:

\[
\min_{\beta} \max_{x/\|x\|_q \leq \gamma} \|Y - (X + \epsilon)\|_2 \\
\text{s.t. } \|\beta\|_0 \leq k.
\]

It turns out that this problem is equivalent to [28]:

\[
\min_{\beta} \|Y - X\beta\|_2 + \gamma \|\beta\|_q \\
\text{s.t. } \|\beta\|_0 \leq k,
\]

with \( 1/q + 1/q^* = 1 \). When \( q = 2 \), we find ridge regression term (note that solution is still sparse because of the constraint \( \|\beta\|_0 \leq k \)). When \( q = \infty \), \( q^* = 1 \) and we retrieve a \( L_1 \) penalization term. Regularization and robustness are intimately related concepts as shown in [28,29]. Finding optimal value of \( \gamma \) and the best regularization term is a matter of cross validation. We note here that a wider range of options is available for uncertainty sets than those presented.

2.4. Solving MIO problems

The choice of a method to solve MIO problems depends on the formulation of the problem and the structure of the data. The main methods to solve MIO include branch-and-bound [30], Extended Cutting Plane [15], Extended Supporting Hyperplane [31], Generalized Benders Decomposition [32], Big-M method and outer approximation [34,35]. The later two techniques are of particular interest as they recently have been successfully implemented to solve the BSSP.

Using Big-M method [33], the constraint \( \|\beta\|_0 \leq k \) is replaced by the introduction of binary variables \( z_i \) and the use of a number of linear constraints. Problem (P) can for example be written as follows:

\[
\min_{\beta, z} \|Y - X\beta\|_2 \\
- Mz_1 \leq \beta_i \leq Mz_1, \quad i \in \{1, \ldots, p\}, \\
z_i \in \{0, 1\}, \quad i \in \{1, \ldots, p\}, \\
\sum_{i=1}^{p} z_i \leq k,
\]

where \( M \) is a constant positive number large enough. Using the notations of this formulation, variable \( i \) is used in the solution if \( z_i =\)

---

**Fig. 1.** Illustration of half plane determined by one inequality (deterministic approach) vs. the area determined by a set of inequalities resulting from the uncertainty on the position of \( x_1 \in U_1 \) and \( x_2 \in U_2 \) (Robust Modelling).
1. In this case, $\beta_i$ could be anywhere between $-M$ and $M$. If $z_i = 0$ then $\beta_i = 0$. The choice of $M$ is critical both in terms of computation time and accuracy of the results of the problem. Upper bounds for $M$ could be found in [33]. For $n$ in 100s and $p$ in 1000s, the problem could be solved in minutes [33] for synthetic data.

Outer-approximation optimization is one technique that aims at solving nonlinear optimization problems by iteratively approaching an optimal solution by relaxing the original problem to a linear approximation and adding constraints successively. When considering the BSSP, problem (P) can be written:

$$\min_{p, z} \left\| Y - XZ\beta \right\|_2^2$$

s.t. $\sum_{i=1}^p z_i \leq k,$

$z_i \in \{0, 1\}, i \in \{1, ..., p\},$

where $Z$ is the diagonal matrix with diagonal $z$. The BSSP can be solved for $n$ and $p$ in the 100000s in seconds for synthetic data using outer-approximation technique [22]. Further work reduces the computation time to make it faster than glmnet [36]. This work also enables a wider choice of loss functions including. Other works use $L_0$ as regularization term to approximate optimal solution [21,37].

Further development of outer approximation optimization solvers might enable the use of a wider variety of loss functions.

3. Methods and materials

We use a general MIO formulation to address regression problems in chemometrics. We then propose a formulation that we use to build regression models for the data set proposed. We also describe methods to solve this formulation.

3.1. General formulation

In this method, we consider each chemical element separately. $Y$ is then considered in $\mathbb{R}^n$. One objective of this method is to solve the BSSP for functions of the form $\mathcal{J}(Y - X\beta) + \gamma \pi(\beta)$ where $\mathcal{J}$ is a convex loss function, $\pi$ a convex penalization function and $\gamma$ a non-negative real number. The convexity condition allows the use of optimization algorithms guaranteeing the convergence towards an optimal solution while the penalization term ensures robustness as described in subsection 2.3. One objective of this framework is also to offer flexibility in the choice of the loss function. The general formulation could be written:

$$\min_{p} \mathcal{J}(Y - X\beta) + \gamma \pi(\beta)$$

s.t. $\left\| \beta \right\|_0 \leq k,$

3.2. Formulation

Note that when no regularization term is considered, we retrieve the formulation (P) of the BSSP.

$$\min_{p} \left\| Y - X\beta \right\|_2^2 + \gamma \left\| \beta \right\|_2^2$$

s.t. $\left\| \beta \right\|_0 \leq k.$
5. Results

We test predictive accuracy, sparsity and robustness using the mineral ore samples presented in the previous section. We randomly choose forty samples for training and validation and the remaining 12 are used for tests. We use and compare four methods:

- **LASSO**: We use glmnet [38] on R environment to solve the problem;
- **MIO**: We use Formulation 1 and solve it using SubsetSelection [36] on Julia 0.6 environment. We pick parameters $\gamma$ and $k$ through cross validation using grid exploration;
- **PLS**: As we are looking at each chemical element separately, a univariate version of PLS is adopted. We use here PLS package v.2.7-1 [39] on R environment;
- **SPLS**: We use SPLS on R environment. We cross validate the number of LVs and parameter $\eta$ through grid search. Values of $\eta$ considered are between 0.8 and 0.99 with a step of 0.01. As the results depends on the algorithm starting point, we run the algorithm 30 times for each couple of number of LVs and $\eta$ value considered.

We report mean square errors and coefficients of determination in Table 1 and Fig. 2. We also report the number of selected variables for sparse methods to compare models' sparsity in Table 2.

### 5.1. Predictive power

With the exception of $K$, Sparse MIO (SMIO) is either superior to all methods or close to the best predictive method with a difference in $R^2$ [40] less than 2%. It proved also to be the only one capable of building a significant model for detecting $Li$, a trace element in the set of samples studied; SPLS and PLS building models with negative $R^2$ while glmnet is predicting a constant. Among sparse methods, we also notice that SPLS is more prone to overfitting compared to SMIO or glmnet. This could be caused by the multiplicity of

---

**Table 1**

<table>
<thead>
<tr>
<th>Element</th>
<th>MSE SMIO</th>
<th>MSE GLMNET</th>
<th>MSE SPLS</th>
<th>MSE PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>8.95E-4</td>
<td>5.57E-3</td>
<td>1.29E-3</td>
<td>3.33E-3</td>
</tr>
<tr>
<td>S</td>
<td>5.37E-5</td>
<td>8.15E-5</td>
<td>1.35E-3</td>
<td>2.85E-4</td>
</tr>
<tr>
<td>Mg</td>
<td>3.97E-5</td>
<td>3.35E-5</td>
<td>2.52E-4</td>
<td>8.86E-5</td>
</tr>
<tr>
<td>K</td>
<td>8.32E-5</td>
<td>4.17E-5</td>
<td>6.3E-5</td>
<td>4.75E-5</td>
</tr>
<tr>
<td>Cu</td>
<td>2.94E-5</td>
<td>3.85E-5</td>
<td>3.61E-5</td>
<td>1.92E-4</td>
</tr>
<tr>
<td>Ca</td>
<td>1.84E-5</td>
<td>2.28E-5</td>
<td>1.63E-5</td>
<td>2.55E-5</td>
</tr>
<tr>
<td>Na</td>
<td>7.83E-6</td>
<td>1.72E-5</td>
<td>5.75E-5</td>
<td>3.82E-5</td>
</tr>
<tr>
<td>Ni</td>
<td>2.56E-6</td>
<td>9.06E-6</td>
<td>3.52E-6</td>
<td>1.37E-6</td>
</tr>
<tr>
<td>P</td>
<td>6.01E-7</td>
<td>3.78E-5</td>
<td>1.29E-6</td>
<td>4.76E-7</td>
</tr>
<tr>
<td>Li</td>
<td>3.49E-11</td>
<td>9.93E-11</td>
<td>1.64E-10</td>
<td>1.82E-10</td>
</tr>
</tbody>
</table>

**Fig. 2.** $R^2$ for each chemical element depending on regression method used. Average excludes $Li$.

### Table 2

<table>
<thead>
<tr>
<th>#variables</th>
<th>SMIO</th>
<th>GLMNET</th>
<th>SPLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>5</td>
<td>21</td>
<td>1950</td>
</tr>
<tr>
<td>S</td>
<td>5</td>
<td>20</td>
<td>121</td>
</tr>
<tr>
<td>Mg</td>
<td>5</td>
<td>15</td>
<td>2941</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>19</td>
<td>480</td>
</tr>
<tr>
<td>Cu</td>
<td>4</td>
<td>13</td>
<td>38</td>
</tr>
<tr>
<td>Ca</td>
<td>4</td>
<td>22</td>
<td>124</td>
</tr>
<tr>
<td>Na</td>
<td>5</td>
<td>13</td>
<td>113</td>
</tr>
<tr>
<td>Ni</td>
<td>4</td>
<td>8</td>
<td>1461</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>14</td>
<td>24</td>
</tr>
<tr>
<td>Average</td>
<td>4.3</td>
<td>16.1</td>
<td>805.8</td>
</tr>
</tbody>
</table>

**Fig. 3.** Accuracy improvement for $K$ from initial performance (SMIO) to later one after removal of variables leading to overfitting (SMIO').

---

Fig. 4. $R^2$ evolution for different values of $\frac{\|\text{Noise}\|_2}{\|\text{Signal}\|_2}$
parameters that need to be tuned (including the starting point) which could lead to a higher probability of overfitting.

5.2. Sparsity and interpretability

When comparing the number of variables used, MIO is far sparser than any other sparse technique considered with a number of variables uses between 2 and 5 while glmnet and SPLS use between 8 and 22 and between 24 and 2941 variables respectively. These results on sparsity echo similar findings in recent works in spectroscopy [42,43]. Indeed the proportion of variables selected relative to the total number of variables by SPLS and glmnet approach those found in this work.

We show now how interpretability can help improve models. Consider the case of K in which MIO performed less than the other approaches. The four variables used by the model built using MIO correspond to wavelengths 279.6 nm, 324.6 nm, 334.8 nm and 766.5 nm. We notice in particular the use of 766.5 nm corresponding to the only persistent line of K in the wavelength band captured (220–827 nm) [44]. K signal is then subject to the interference of Mg which is also has a strong line in 765.9 nm. The algorithm is then correcting the interference of Mg using the variable at 279.6 nm which also corresponds to a strong line of Mg (279.8 nm). The other two variables seem to build overfitting behavior. We remove then the bands [323.3nm, 326nm] and [333.4nm, 336.1nm] and report the improvement in accuracy in Fig. 3. The variables used by the new model involve wavelengths corresponding to strong lines of Mg (279.6 nm, 279.8 nm), K (766.5 nm) and Al (396.08 nm). Performing such operations based on coefficients found by glmnet or SPLS is more difficult regarding the number of variables used by these approaches (19 and 480 in the case of K). We also note that these improvements to the model could be performed in the training phase. Indeed, some factors such as a wide gap between the prediction power of the training and the validation sets or a choice a variables that do not correspond to a physical logic could indicate that the model should be modified.

Regarding sparsity, we finally note here that forcing more sparsity in glmnet by increasing $\gamma$ or in SPLS by increasing $\eta$ to the level of MIO deteriorates significantly prediction power as the penalty term becomes more important and then skews the objective function.

5.3. Robustness

We provide now indication on the robustness of the approach proposed. Although the conditions of the experimentation are already noisy, for the sake of illustrating the robustness to noise, we generate synthetic noise (using normal multivariate distribution with zero mean and covariance corresponding to the signal captured) and add it to the spectra initially captured. We report the evolution of $R^2$ when the amplitude of noise is increased in Fig. 4. We notice that the robust optimization approach is significantly more robust for all chemical elements when compared to other techniques. Results suggest that sparsity combined with robust optimization yields provides a strong approach when dealing with noise. We also notice an improvement in the predictive power of K when MIO is used; the random noise added may have disturbed the variables leading to overfitting in this case as discussed earlier in this subsection.
5.4. Manual selection of wavelengths

For illustration purposes, we compare results obtained using the different techniques considered while using only a set of selected regions of the spectra. We consider a region of ±1 nm around the wavelengths corresponding to the chemical elements studied in Table 3.

We report MSE and $R^2$ for the four techniques considered in Table 4 and Fig. 5 respectively.

We notice that the performance of SMIO is still comparable to the one obtained when the whole spectra were used as inputs. We also note that the predictive power is comparable to the one obtained when using the whole spectra. This indicates that SMIO is effective in finding variables with the highest predictive power. We finally note that we used big-M method to solve the instances relative to this subsection.

5.5. Computation time

We use a personal computer using an Intel Core i7-8750H CPU at 2.20 GHz and 16 GB of RAM. Solving SMIO problem when the entire spectra is fed as input takes less than 2 min for a number of variables equal to 3828 and a number of training samples equal to 30 (10 samples were used for validation of parameters $k$ and $\gamma$). Building a model for a chemical element including cross-validation for five values of $\gamma$ and four values of $k$ takes less than 20 min when we use outer approximation approach [36]. We were unable to run big-M method in this case. Although this method handles this size of problems for randomly generated data, it does not seem to be tractable method when the colinearity of variables is as important as in the LIBS spectra collected. Big-M method takes a fraction of second to solve an instance when wavelengths are selected which corresponds in this study to thirty-four variables.

While PLS builds models in minutes and glmnet in less than 1 min including cross-validation, SPLS take more time to build models because of the multiplicity of parameters to tune.

6. Conclusion

Recent developments in theory and computation power makes robust sparse MIO accessible and tractable in chemometrics. MIO can offer competitive accuracy, higher robustness and consistency compared to other techniques tested in this study and is by far sparser than LASSO and SPLS which improves interpretability. Indeed, MIO achieves an average $R^2$ of 0.88 while PLS, glmnet and SPLS achieve 0.78, 0.77 and 0.68 respectively. In addition, MIO generates models using a significantly lower number of variables which improves interpretability. Indeed, MIO generates models using 4.3 variables while glmnet and SPLS use 16.1 and 805.8 variables respectively. Finally, MIO formulation proved to be the most robust achieving an $R^2$ of 0.69 in average at the highest level of noise considered while all other techniques have a negative $R^2$.

However, MIO technique is more sensitive when it comes to tuning parameters as it requires the tuning of two parameters ($k$ and $\gamma$ in the case of outer-approximation and $k$, $\gamma$ and $M$ in the case of big-M) while only one for PLS (number of LVs) and glmnet. Although computation time is higher than PLS and glmnet, the method is still tractable.

Several developments of MIO technique could be considered. First, Strategies in the training phase including bootstrapping techniques [45] to improve model significance or wavelengths selection to avoid overfitting could be considered. In addition, other loss functions and regularization options could be explored. Indeed, penalization term efficiency depend on the distribution of the noise in the signal. Considering the different sources of noise (spectra continuum, spectrometer white noise, ...), there might be more efficient penalization functions dealing with the real noise. Furthermore, algorithmic advances in speeding up outer approximation optimization could help explore more loss and penalization functions. Advances based on relaxations of outer approximation optimization such as [36] (which deals only with Tikhonov [23] regularization while loss function options are $L_1$, ordinary least square and $L_2$ SVR) could significantly reduce computation time. Second, MIO approach is also applicable to classification [47] and could be used for several fields related to chemometrics including metal and mining grade classification, soil classification etc. Finally, MIO approach could also be considered to develop novel sparse PLS approaches which would improve interpretability of this method.

CRedit authorship contribution statement


Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank ACA reviewers for their insightful comments that contributed greatly in the quality of this work. We also thank Lütfü Özcan and Kheireddine Rifai (Elemission) for the guidance they have provided. This research has been funded by OCP SA. (www.ocpgroup.ma) with the contribution of Mohammed VI Polytechnic University.

Appendix A. Chemical Composition and References of CRMs

Table A.5

<table>
<thead>
<tr>
<th>Ref</th>
<th>Type</th>
<th>Fe</th>
<th>S</th>
<th>Mg</th>
<th>K</th>
<th>Cu</th>
<th>Ca</th>
<th>Na</th>
<th>Ni</th>
<th>P</th>
<th>Li</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>U bearing sandstone</td>
<td>1.57</td>
<td>0.23</td>
<td>2.59</td>
<td>0.06</td>
<td>0.23</td>
<td>0</td>
<td>0.01</td>
<td>4.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25a</td>
<td>Soil lithogeochem</td>
<td>6.6</td>
<td>0.05</td>
<td>0.33</td>
<td>0.48</td>
<td>0.31</td>
<td>0.13</td>
<td>0</td>
<td>36.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45d</td>
<td>Anomalous Ferruginous Soil</td>
<td>14.52</td>
<td>0.05</td>
<td>0.24</td>
<td>0.41</td>
<td>0.04</td>
<td>0.19</td>
<td>0.1</td>
<td>0.02</td>
<td>0.04</td>
<td>21.49</td>
</tr>
<tr>
<td>45e</td>
<td>Latericite Soil Lithogeochem</td>
<td>24.12</td>
<td>0.05</td>
<td>0.16</td>
<td>0.32</td>
<td>0.08</td>
<td>0.06</td>
<td>0.05</td>
<td>0.03</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>460</td>
<td>REE-Nb laterite ore</td>
<td>18.56</td>
<td>0.03</td>
<td>0.72</td>
<td>1.25</td>
<td>0.7</td>
<td>0.13</td>
<td>0.01</td>
<td>0.19</td>
<td>0.18</td>
<td>18.34</td>
</tr>
<tr>
<td>461</td>
<td>REE-Nb laterite ore</td>
<td>31.52</td>
<td>0.05</td>
<td>1</td>
<td>0.23</td>
<td>0.01</td>
<td>1.21</td>
<td>0.18</td>
<td>0.01</td>
<td>0.34</td>
<td>12.04</td>
</tr>
<tr>
<td>462</td>
<td>REE-Nb laterite ore</td>
<td>33.05</td>
<td>0.05</td>
<td>1.01</td>
<td>0.12</td>
<td>0.01</td>
<td>1.25</td>
<td>0.18</td>
<td>0.01</td>
<td>0.45</td>
<td>10.84</td>
</tr>
<tr>
<td>463</td>
<td>REE-Nb laterite ore</td>
<td>33.33</td>
<td>0.06</td>
<td>1.01</td>
<td>0.11</td>
<td>0.01</td>
<td>1.19</td>
<td>0.17</td>
<td>0.01</td>
<td>0.59</td>
<td>10.5</td>
</tr>
</tbody>
</table>
Table A.5 (continued)

<table>
<thead>
<tr>
<th>Ref</th>
<th>Type</th>
<th>Fe</th>
<th>S</th>
<th>Mg</th>
<th>K</th>
<th>Ca</th>
<th>Na</th>
<th>Ni</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>463</td>
<td>REE-Nb laterite ore</td>
<td>1.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>464</td>
<td>REE-Nb laterite ore</td>
<td>1.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>503b</td>
<td>Porphyry Cu–Au–Mo</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>504b</td>
<td>Porphyry Cu–Au–Mo</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>601</td>
<td>Sulphidation Epithermal Ag–Cu–Au</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>602</td>
<td>Sulphidation Epithermal Ag–Cu–Au</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
</tbody>
</table>

1Oredolomitic, carbonaceous and argillaceous, sandstones and siltstones.

References


