SORT 44 (1) January-June 2020, 67-98 DOI: 10.2436/20.8080.02.95

Integer constraints for enhancing interpretability in linear regression

Emilio Carrizosa[1], Alba V. Olivares-Nadal[2] and Pepa Ramírez-Cobo[3]

Abstract

One of the main challenges researchers face is to identify the most relevant features in a prediction model. As a consequence, many regularized methods seeking sparsity have flourished. Although sparse, their solutions may not be interpretable in the presence of spurious coefficients and correlated features. In this paper we aim to enhance interpretability in linear regression in presence of multicollinearity by: (i) forcing the sign of the estimated coefficients to be consistent with the sign of the correlations between predictors, and (ii) avoiding spurious coefficients so that only significant features are represented in the model. This will be addressed by modelling constraints and adding them to an optimization problem expressing some estimation procedure such as ordinary least squares or the lasso. The so-obtained constrained regression models will become Mixed Integer Quadratic Problems. The numerical experiments carried out on real and simulated datasets show that tightening the search space of some standard linear regression models by adding the constraints modelling (i) and/or (ii) help to improve the sparsity and interpretability of the solutions with competitive predictive quality.

MSC: 62J05, 90C11.

Keywords: Linear regression, Multicollinearity, Sparsity, Cardinality constraint, Mixed Integer Non Linear Programming.

1 Introduction

A plethora of real world data involve multiple features interacting between them. As a consequence, one of the most common research challenges is trying to predict a variable by making use of attributes that are deterministic or easier to access. A widely studied tool to achieve this is the linear regression model

$$\mathbf{Y} = \beta_0 + \boldsymbol{\beta}\mathbf{X} + \mathbf{a} \tag{1}$$

[1] Institute of Mathematics of the University of Seville (IMUS).

[2] The University of Chicago Booth School of Business, 5751 S. Woodlawn Ave., Chicago, Illinois 60637. Email: alba.nadal@chicagobooth.edu

[3] Department of Statistics and Operational Research, Universidad de Cádiz.

Received: January 2019
Accepted: April 2020

where $\mathbf{Y} = (y_1, \ldots, y_K)'$ contains the K realizations of the random variable to be predicted, $\mathbf{X} \in \mathbb{R}^{K \times N}$ contains the observations of the attributes $X^1, \ldots, X^N$ that influence on $\mathbf{Y}$, and $\mathbf{a} \in \mathbb{R}^K$ denotes the error term. In practice, the coefficients $\boldsymbol{\beta}$ need to be estimated and thus, the user needs to select an estimation method, which is usually derived from solving an optimization problem of the form:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}} & f(\boldsymbol{\beta}) \\ \text{s.t} & \boldsymbol{\beta} \in \mathcal{B} \end{array} \tag{2}$$

where $\mathcal{B}$ denotes the feasible region.

Since the data collection technologies are improving altogether with communication systems and computers' memories and processors, the dimension of the data sets to be handled is increasing drastically. As a consequence, nowadays researchers aim to return an interpretable output which explains the main interactions between the features that conform the pile of data and the dependent variables. Usually, this is understood as a problem of choosing the most relevant features for prediction (Friedman, Hastie and Tibshirani, 2001; Cai, Tsay and Chen, 2009; Hastie, Tibshirani and Wainwright, 2015). Sparse methods will yield solutions $\boldsymbol{\beta}$ in (2) with a large number of zero coefficients, in which only the most significant features are associated with the non-zeroes (Tibshirani, 1996; Hastie et al., 2015). Although sparsity may be a desirable property for our solution, we should take into account that other characteristics need to be sought in order to obtain a more interpretable output. First, correlated variables can provide highly variable estimated coefficients that make it difficult to understand the impact of a feature on the predictive variable. Second, spurious coefficients complicate the judgement of whether a feature is truly relevant for prediction or not. We will explain these two issues with further detail in what follows and motivate why we aim to alleviate them in this paper while still seeking for a sparse solution.

It is known that ordinary least squares (OLS) provides solutions that may be highly dense. A good representative of a possibly sparse estimation method in the form (2) is the lasso (Tibshirani, 1996), which adds a ℓ_1-norm penalization term to the OLS objective:

$$\begin{array}{l} f(\boldsymbol{\beta}) = \|\mathbf{y} - \beta_0 - \boldsymbol{\beta}\mathbf{X}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ \mathcal{B} = \mathbb{R}^N. \end{array} \tag{3}$$

The lasso encompasses OLS when the penalty parameter λ equals zero, but when λ increases the solution becomes more sparse. The lasso is computationally feasible and, under certain technical conditions on the data matrix $\mathbf{X}$, it enjoys good statistical properties, see Friedman et al. (2001); Bühlmann and van de Geer (2011). However the lasso also presents certain shortcomings well-documented in the literature (for a brief review, see Bertsimas et al. (2016) and the references therein). In particular, it is known that estimation through OLS or the lasso may be quite unstable in the presence of strong

collinearity on the data (Silvey, 1969; Sengupta and Bhimasankaram, 1997; Hesterberg et al., 2008). On one hand, the presence of correlated variables may yield a high variability in the estimated coefficients, complicating thus the interpretation of the results (Farrar and Glauber, 1967; Watson and Teelucksingh, 2002; Montgomery, Peck and Vining, 2012). On the other hand, a consequence of collinearity that leads to problems for interpreting the effect of the regressors is that two variables that are highly positively (negatively) correlated may have associated estimated coefficients with different (same) signs. This problem can be illustrated through the following numerical example in Hesterberg et al. (2008). Consider the `diabetes` database (Efron and Hastie, 2003), which consists of the measures of 10 variables (age, sex, body mass index, average blood pressure and six different blood serums) on 442 patients. The top panel of Figure 1 depicts the path of solutions of the lasso for this database; that is to say, the estimates of the coefficients $\boldsymbol{\beta}$ obtained are depicted against the different values of the penalty λ. As noted by Hesterberg et al. (2008), features *tc* and *ldl* (bottom left panel), have a correlation of 0.89. However, their estimated coefficients take opposite signs, which is in contradiction with their dependence degree. Similarly, the coefficients for variables *hdl* and *tch* (bottom right panel), which show a correlation of -0.73, have the same sign when estimated by the OLS (the case $\lambda = 0$ in Figure 1). Hence it seems that the coefficients of highly correlated variables may take values that compensate each other. Finally, an additional inconsistency is that the sign of the estimated coefficient of *hdl* (squared blue line, left bottom panel) varies depending on the level of sparsity required.

The negative effects of collinearity have been differently addressed in the literature. On one hand, some authors (Chatterjee and Hadi, 2015; Montgomery et al., 2012) suggest to remove variables that are highly correlated or unimportant, often carrying out significancy tests to determine if a variable can be discarded. However, the results of these tests may be misleading in the presence of strong collinearity (Watson and Teelucksingh, 2002). In this line, the recent paper by Bertsimas and King (2015) proposes to tighten the estimation procedure (2) by adding constraints that explicitly forbid the coefficients of variables with a high pairwise correlation to be simultaneously non-zero. Nonetheless, as it will be seen in Section 3.3.1, these approaches may be detrimental if highly correlated features own a strong predictive power. On the other hand, some authors encourage highly correlated predictors to be altogether in the model. The graph-guided fused lasso (GFlasso hereafter, proposed in Kim and Xing, 2009) encourages two highly correlated variables to have similar estimated coefficients by adding the penalization $\gamma|\beta_i - sign(\rho_{ij})\beta_j|$ to the lasso objective function, where ρ_{ij} denotes the correlation between X^i and X^j. The SRIG method (Sparse Regression Incorporating Graphical Structure Among Predictors, introduced in Yu and Liu, 2016), determines the value of the coefficient β_j not only by feature X^j but also by all features X^i such that ρ_{ij} is large in absolute value. Under certain technical conditions, the SRIG is endowed with nice properties that, for instance, ensure the recovery of the original model. However, as it will be shown in Section 3.2, these conditions may not be fulfilled in real databases, yielding outputs that may not improve the performance of the

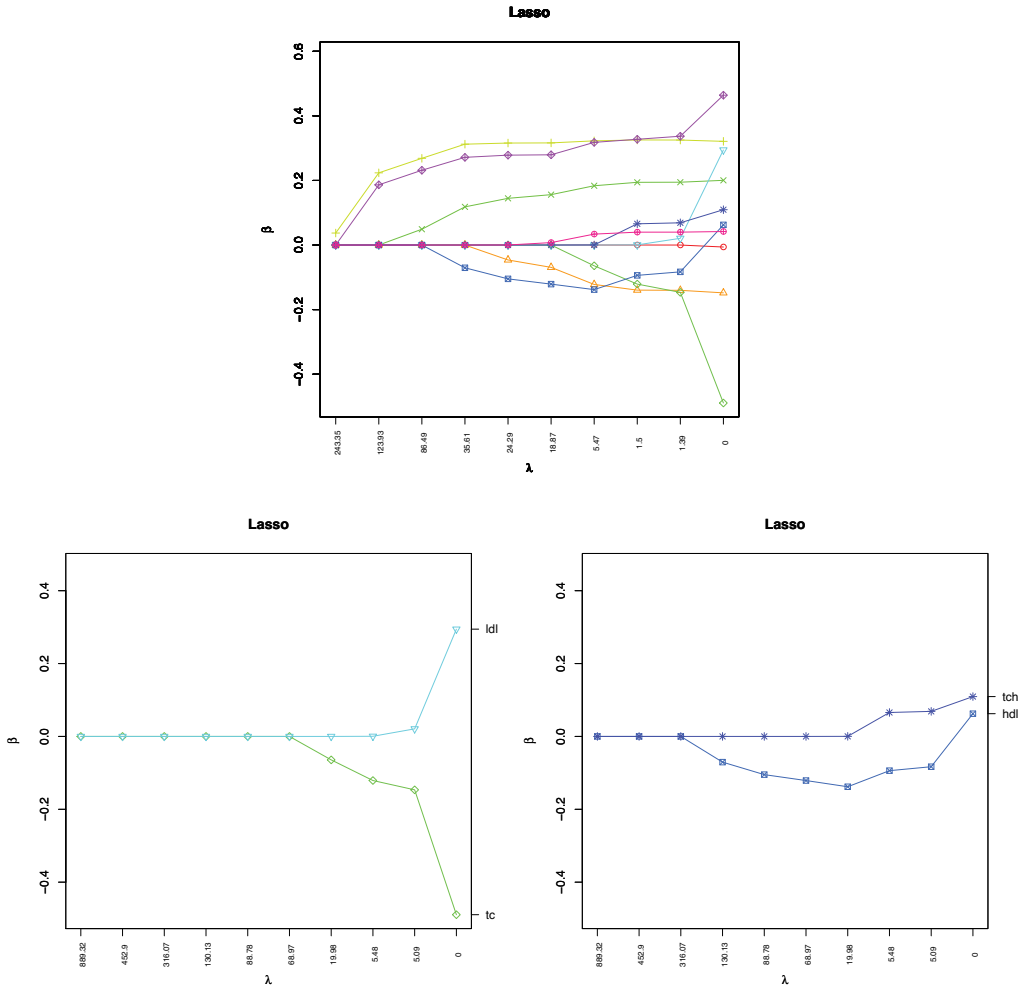

Figure 1: *Top: path of solutions of the lasso for the diabetes database (the size of the coefficients β are depicted against the values of the penalty λ). Bottom: problematic paths.*

current methodologies. Our approach is less restrictive than the previous one, since it neither encourages the removal of variables nor the presence of groups of correlated features. Instead, we propose a constraint (called *sign coherence constraint*) that aims to elude the signs' inconsistency phenomenon related to collinearity shown in Figure 1, avoiding that the coefficients of highly correlated variables compensate each other. This constraint provides more flexible models, as for some cases two highly correlated variables may appear altogether in the output, and for some other cases one feature of a highly correlated pair may be removed. This will be illustrated in Figure 5, Section 3.2. Our approach is not the first one to restrict the sign of the estimated coefficients in linear regression. For instance, Meinshausen (2013) makes use of non-negative least squares to recover the real sparsity pattern in high-dimensional data under certain conditions. Also, the LARS algorithm (Efron, Hastie, Johnstone and Tibshirani, 2004) emulates the lasso

solution by requiring the sign of the coefficients to match the sign of their correlation with the residuals. Another well-known example is the non-negative garotte (Breiman, 1995), which performs subset selection while forcing the signs of the coefficients to match the signs of the OLS estimates.

To conclude, there are further remedies to alleviate collinearity issues, consisting of harvesting more observations (Sengupta and Bhimasankaram, 1997; Montgomery et al., 2012) or applying methods that decorrelate the data (Cao, Guo and Bouman, 2010; Massy, 1965). Nevertheless, the later approaches imply the transformation of the variables and thus complicate the interpretation of the final models with respect to the original features. More recently, optimization approaches bounding the Variance Inflation Factor (VIF) and condition number of the correlation matrix have also been proposed (Tamura et al., 2019; Jou, Huang and Cho, 2014; Tamura et al., 2017).

On top of the unreliable interpretation of coefficients in presence of high correlations, the lasso also suffers a drawback, mitigated in this paper: the presence of spurious coefficients. For $\lambda > 0$ the ℓ_1-norm performs a shrinkage of the coefficients in the lasso solution that eventually attains sparsity as a side effect. However, as will be shown in the numerical section, the solutions of the lasso may be still dense for large datasets due to these spurious coefficients. In this paper we avoid this negative effect of shrinkage by defining a novel constraint (called *significance constraint*) that forces the estimated coefficients to be either zero or larger than a fixed value (to be tuned).

In summary, in this paper we model two novel constraints which will tighten the search space for $\boldsymbol{\beta}$ in Problem (2). As a result, the interpretability of the solutions is improved since (i) the signs of the coefficients are coherent with the sign of the correlations between highly or moderately correlated predictors, and (ii) the shrinkage is combatted while avoiding spurious coefficients, which may lead to the annihilation of some coefficients, thus increasing the sparsity. As will be shown in the numerical experiments, such better interpretability is obtained without damaging the predictive power of the model. When discerning the suitability of these constraints for a particular database, the user should realize that constraints modelling (i) become inactive if no highly correlated predictors are found, while constraints expressing (ii) do if all the variables have non-spurious estimated coefficients. Hence, the user might want to analyse the correlations before adding the sign coherence constraints. However, we do recommend adding the significance constraint if a regularized method is used for estimation.

The resulting optimization problems will belong to the class of Mixed Integer Quadratic Programs (MIQP), which have recently proven very suitable in different statistics problem as linear regression (Tamura et al., 2019; Bertsimas and King, 2015), time series (Carrizosa, Olivares-Nadal and Ramírez-Cobo, 2016), classification (Carrizosa, Nogales-Gómez and Morales, 2016; Carrizosa, Nogales-Gómez and Morales, 2017), or dimensionality reduction (Carrizosa and Guerrero, 2014). Indeed, Bertsimas and King (2015); Bertsimas et al. (2016) use MIQP theory to solve (in tractable way) the best

subset selection problem (Miller, 2002):

$$\begin{aligned} &f(\boldsymbol{\beta}) = \|\mathbf{y} - \beta_0 - \boldsymbol{\beta}\mathbf{X}\|_2^2 \\ &\mathcal{B} = \left\{\boldsymbol{\beta} \in \mathbb{R}^N : \|\boldsymbol{\beta}\|_0 \leq V_T\right\}, \end{aligned} \tag{4}$$

where the ℓ_0-*norm* is the cardinality function $\|\boldsymbol{\beta}\|_0 = \#(j : \beta_j \neq 0)$ and

$$\|\boldsymbol{\beta}\|_0 \leq V_T, \tag{5}$$

denotes the cardinality constraint which leads to attain the desired level of sparsity given by the value V_T. In this work, the two novel constraints (*sign coherence* and *significance* constraints) will be combined with the cardinality constraint (5), so that sparsity is also achieved in addition to a better interpretability.

The paper is structured as follows. In the next section we model the new constraints to be added to Problem (2) in order to enhance interpretability through mathematical programming. The numerical experiments are carried out in Section 3, where the estimation methods under comparison and the design of experiments are also discussed. The last section is devoted to concluding remarks and extensions.

2 Mathematical model formulation

In this paper, it is our aim to enhance the interpretability of the outputs by replacing any estimation procedure in the form (2) by:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & f(\boldsymbol{\beta}) \\ \text{s.t} \quad & \boldsymbol{\beta} \in \mathcal{B} \cap \mathcal{S} \end{aligned} \tag{6}$$

where $\mathcal{S}$ will gather the proposed constraints. Tightening an estimation procedure by adding constraints, i.e., solving (6) instead of (2), has already been considered in the literature in order to improve the performance of linear regression estimation methods like (2), see for example Bertsimas and King (2015).

In this section we model the tightening set $\mathcal{S}$ by defining constraints that can be added to a classic (possibly sparse) linear regression estimation method (2), in order to enhance the interpretability of the outcome as well as improving its sparsity. As commented in the previous section, the first novel constraint, called *sign coherence constraint*, imposes coherence between the signs of the estimated coefficients and the signs of large pairwise correlations between predictors. The second novel constraint, the so-called *significance constraint*, allows only for truly significant features to be considered in the model. The idea is that the user should feel free to add any of these constraints, when compatible, to her selected estimation method given by (2), yielding (6).

2.1 The sign coherence constraint

The presence of correlated variables in the data is demonstrated to lead to undesired consequences, such as a high variability on the estimated coefficients and the sign inconsistencies explained in the introduction; see e.g. Bartholomew et al. (2008). As it was commented, the traditional procedure to avoid these undesired behaviour consists of removing highly correlated variables. In particular, Bertsimas and King (2015) forbids two highly correlated variables to be simultaneously non-zero. Specifically, the following pairwise correlation constraints are modelled

$$\gamma_i + \gamma_j \leq 1 \quad \forall (i,j) \in \Omega_\eta, \tag{7}$$

where $\Omega_\eta = \{(i,j) : |\rho_{ij}| \geq \eta\}$ is the set of pairs of features considered to be highly correlated, and γ_i, γ_j are defined as

$$\gamma_j \quad = \quad \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0. \end{cases} \tag{8}$$

From now on, the constraint defined by Bertsimas and King (2015) and stated as (7) will be called *correlation constraint*.

In contrast to Bertsimas and King (2015), we propose here a less restrictive approach that allows two highly correlated variables to be in the model at the same time, but forbids misleading interpretations and misrepresentative coefficients. Our aim is to avoid sign inconsistencies while allowing the model to include two correlated variables if they contribute to improve or maintain the prediction quality. Therefore, we propose to model constraints that avoid the *compensation* of coefficients for correlated variables. Under the light of the example illustrated in Figure 1, these are the requirements we aim to gather when modelling the *sign coherence* constraint:

1. The coefficients of two features that are moderately or highly positively correlated must have the same sign.
2. The coefficients of two features that are moderately or highly negatively correlated must have opposite signs.

In order to model these constraints, we introduce the following binary variables:

$$\nu_j^+ = \begin{cases} 1 & \text{if } \beta_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\nu_j^- = \begin{cases} 1 & \text{if } \beta_j < 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the previous requirements 1-2 can be easily formulated as constraints as:

$$\nu_i^+ + \nu_j^- \leq 1 \quad \forall (i,j) \in \Omega_\alpha^+ \tag{9}$$
$$\nu_i^- + \nu_j^+ \leq 1 \quad \forall (i,j) \in \Omega_\alpha^+ \tag{10}$$
$$\nu_i^+ + \nu_j^+ \leq 1 \quad \forall (i,j) \in \Omega_\alpha^- \tag{11}$$
$$\nu_i^- + \nu_j^- \leq 1 \quad \forall (i,j) \in \Omega_\alpha^- \tag{12}$$

where Ω_α^+ and Ω_α^- are the sets of pairs of features that are moderately or highly correlated, expressed as $\Omega_\alpha^+ = \{(i,j) : \rho_{ij} \geq \alpha\}$ and $\Omega_\alpha^- = \{(i,j) : \rho_{ij} \leq -\alpha\}$. That is to say, constraints (9)-(10) mean that, if two variables i, j are highly positively correlated (i.e. $(i,j) \in \Omega_\alpha^+$), then we do not allow one of the coefficients to be positive and the other negative. Similarly, constraints (9)-(10) imply that, if two variables i, j are highly negatively correlated (i.e. $(i,j) \in \Omega_\alpha^-$), we forbid their coefficients to be both positive or both negative.

Note that variables ν_j^+, ν_j^- are linked with γ_j, defined in Equation (8), as follows:

$$\gamma_j = \nu_j^+ + \nu_j^-,$$

and thus the cardinality constraint (5) can be also written as:

$$\sum_{j=1}^{N} (\nu_j^+ + \nu_j^-) \leq V_T.$$

In order to illustrate the impact of these constraints we compare the path of solutions depicted in Figure 1 for the lasso applied to the `diabetes` dataset, against the lasso tightened with the sign coherence constraints, as depicted in Figure 2. As it can be observed, the use of constraints (9)-(12) to tighten the feasible region of the lasso might avoid the sign of a coefficient to vary depending on the level of sparsity required, easing the interpretation of the impact of the predictors over the response variable.

2.2 The significance constraint

In this section we formulate a novel constraint that helps combatting the negative effects of shrinkage of the lasso while discarding spurious coefficients. The idea is to allow only for *significant* variables to be represented in the model and to improve the sparsity of the output. Intuitively, *large* coefficients are identified with the significance of a feature once the data are normalized. Following this reasoning we propose to establish a threshold of *significance* that a feature must be able to exceed to be allowed in the model. We model the significance constraint as follows:

$$|\beta_j| \in \{0\} \cup [\epsilon, +\infty) \qquad j = 1, \ldots, N \tag{13}$$

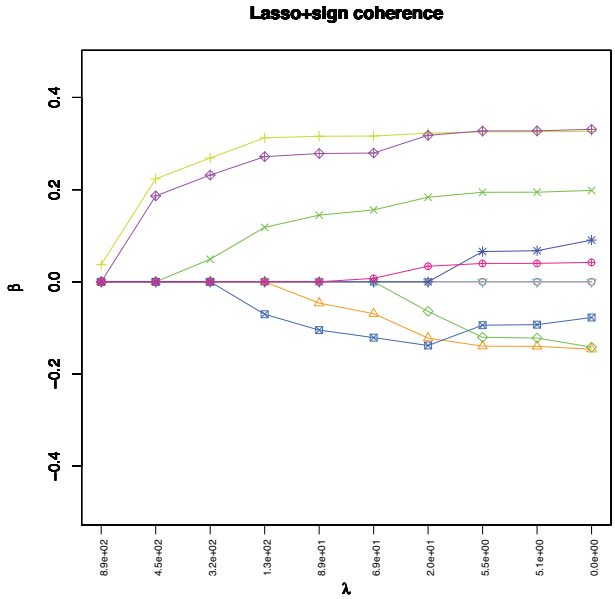

Figure 2: *Path of solutions of the tightened lasso by the sign coherence constraint for the diabetes database.*

where ϵ is called here the significance threshold, to be fixed by the user or to be tuned (see Section 3.1). In (13), a forbidden region in $(0,\epsilon)$ is defined with the aim to avoid shrinkage and to forbid spurious coefficients in the solution. This constraint was already used by Carrizosa et al. (2016) to discover potential causalities in multivariate time series. Note that using the binary variables ν_j^+ and ν_j^- defined in Section 2.1, constraints (13) can be expressed in a more manageable form via two sets of linear constraints

$$\begin{aligned} \beta_j &\geq \epsilon\nu_j^+ - \nu_j^- M \qquad \forall j = 1,\ldots,N \\ \beta_j &\leq -\epsilon\nu_j^- + \nu_j^+ M \quad \forall j = 1,\ldots,N \end{aligned} \tag{14}$$

where M is a *large* constant. This big M, often appearing when modelling problems with integer variables, is large enough so it does not exclude reasonable values of the parameters β_j (see, e.g. Camm, Raturi and Tsubakitani, 1990). In order to clarify the effect of the significance constraint (13), consider the heat map given by Figure 3. The left panel depicts the values of the estimated coefficients for the lasso, in the first column, and the lasso with the significance constraint (taking $\epsilon = 0.3$), in the second column, for the `golf2009` database (Winner, 2016). The right panel represents the values of the estimated coefficients for the OLS and its counterpart tightened with the significance constraint (taking $\epsilon = 0.05$) for the `compact` database (Torgo, 2016). Such datasets will be described with further details in Section 3. The colour represents the sign of the coefficients β_j (blue for negative, red for positive) and the intensity is related to the magnitude of such coefficients. In Figure 3 it can be observed that adding significance constraints establishes a clearer cut between zero and non-zero coefficients. Also note

that the tightened approach is not equivalent to making zero all the coefficients estimated by the lasso to be smaller in absolute value than our threshold ϵ. For instance, in the left panel β_4, estimated to be 0.143, is enlarged to $\epsilon = 0.3$, while β_3 is enlarged to -0.436 despite being estimated by a value of -0.311, which was already larger in absolute value than ϵ. Finally, it should be noted that the solutions under the significance constraint lead to an improvement of 10.58% over the out-of-sample mean squared error (MSE hereafter) of the lasso for the considered database. Moreover, on the right panel we observe that the tightened OLS shrinks coefficient $\beta_{14} = -0.024$ to zero, while enlarges coefficient $\beta_4 = 0.012$ to $\epsilon = 0.05$, despite being smaller in absolute value. Also, adding the significance constraints yields to an improvement of 1.72% over the out-of-sample MSE of the OLS.

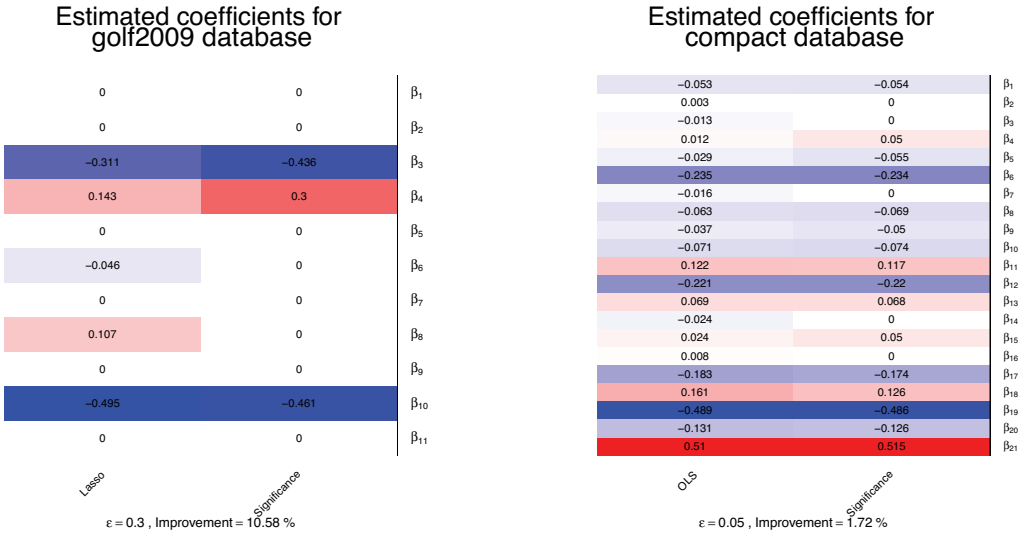

Figure 3: *Heatmaps representing the coefficients $\boldsymbol{\beta}$ estimated by the lasso (left panel) and the OLS (right panel) and their respective counterparts tightened by adding significance constraint* (13) *for the* `golf2009` *and* `compact` *datasets.*

3 Numerical illustrations

In this section we describe and undertake the numerical experiments performed to compare two benchmark estimation methods in linear regression of the form (2) against their tightened versions (6) derived from reducing the search of the coefficients $\boldsymbol{\beta} \in \mathcal{B} = \mathbb{R}^N$ to the set $\mathcal{B} \cap \mathcal{S}$, where $\mathcal{S}$ is defined through some constraints. Specifically, in the next section we outline the design of the experiments, Section 3.2 shows the results for real databases, Section 3.3.1 replicates the simulated study in Yu and Liu (2016), while in Section 3.3.2 the databases are generated following the simulations in Bertsimas and King (2015).

3.1 Design of experiments

In order to assess the impact of the novel constraints over the estimated coefficients $\boldsymbol{\beta}$ and the predictive quality of the solutions, we will analyse the differences in the performance between Problems (2) and (6). As baseline estimation methods (i.e. Problem (2)), we consider the lasso (Problem (3)) and the OLS (Problem (3) with $\lambda = 0$). Their tightened versions (Problem (6)) consider the same objective functions but reduce the search space of the coefficients $\boldsymbol{\beta}$ to the so-called tightened regions $\mathcal{B} \cap \mathcal{S}$.

In our numerical setting, we consider various tightening sets $\mathcal{S}$, whose related problems, taking form (6), are explicitly formulated in Appendix A. In order to analyse the effect of the first novel constraint proposed in this paper, the sign coherence constraint, we consider the set $\mathcal{S}_1 = \{(9)-(12)\}$. To compare our approach with the recent constraints by Bertsimas and King (2015), the correlation constraint, we will also test the performance of the set $\mathcal{S}_2 = \{(7)\}$. Both sets will be considered in the first part of Section 3.2 where the sign coherence constraint is analysed. Then, to clarify the performance of the new significance constraint, the tightening set $\mathcal{S}_3 = \{(14)\}$ will be considered in the second part of Section 3.2. Finally, we will analyse the global performance of our novel constraints when the cardinality constraint is also imposed (that is, $\mathcal{S}_4 = \{(5),(9)-(12),(14)\}$) in comparison to the tightening set of Bertsimas and King (2015), for which $\mathcal{S}_5 = \{(5),(7)\}$. In these cases we also show the predictive quality and number of non-zero coefficients for the elastic net (Enet hereafter) (Zou and Hastie, 2005), the SRIG method in Yu and Liu (2016) and the GFlasso in Kim and Xing (2009). The Enet, which trades off between lasso and ridge regression, is known to avoid erratic paths of correlated variables in the lasso (Hastie et al., 2015). In fact, for non-trivial values of the parameters, the Enet problem has a unique solution, no matter the correlations between the regressors. This shall be addressed in the last part of Section 3.2 as well as Sections 3.3.1 and 3.3.2. The Enet method was run using `R cran` package `glmnet`, and the SRIG method was run using the `R` packages recommended by the authors in Yu and Liu (2016). All the tightened procedures and the GFlasso were easily coded in the algebraic language `AMPL` (Fourer, Gay and Kernighan, 2002), but the latter was solved using `Knitro` solver. As Problems (2) and (6) are MIQPs with quadratic convex objective function and linear constraints, they were solved using `CPLEX`. For the interested reader, the code is included in Appendix D of the Supplementary Material. Even though MIQP problems may be hard to solve, the current solvers already incorporate a plethora of heuristics that turn them into highly efficient optimizers. For instance, `CPLEX` incorporates various preprocessing steps whose aim is to reduce the size of the problem and improve its formulation (Savelsbergh, 1994; Atamurk, Nemhauser and Savelsbergh, 2000). On the other hand, many other techniques and local search heuristics are implemented and implicitly run during the process (see, for instance, Danna, Rothberg and LePape, 2005; Fischetti and Lodi, 2005; Rothberg, 2007). As done in Bertsimas and King (2015), a time limit of 20 seconds was imposed to solve each MIQP for $K \leq N$, although this limit was reached

only for the largest datasets and in most cases cases the optimal solution was attained in a few seconds. For the case with $K > N$, a time limit of 40 seconds was imposed instead.

In order to make a fair comparison against existing procedures, the experiments developed here closely follow those in Bertsimas and King (2015). First, unless otherwise specified, the datasets are normalized and divided in train, test and validation sets (50%, 25% and 25% of the data, respectively). All the problems are solved in the training set, and the solution that minimizes the MSE in the test set is chosen. Two criteria are used to compare the methods, namely, the MSE and the sparsity. All the MSEs reported in this paper correspond to the values obtained in the validation sets and are normalized by dividing by the MSE of the OLS solution; that is to say, when any method attains a MSE greater than 1 their prediction power is estimated to be worse than that of the OLS, while for smaller values the accuracy has improved.

The sparsity of the solution of the unconstrained lasso increases as its regularization parameter $\lambda \in \mathbb{R}^+$ does. The critical values of λ are easily computed using any implementation of the LAR algorithm in various standard statistical packages. In particular, in this paper the lasso set of solutions was obtained by using the `lars()` function of `R-cran` package `lars` (Hastie and Efron, 2013).

For the tightened MIQPs (6), the pairwise correlation considered to generate the sets Ω_α^+ and Ω_α^- in constraints (9)-(12) is fixed to $\alpha = 0.6$. Following Bertsimas and King (2015), the maximum pairwise correlation allowed is $\eta = 0.8$; that is to say, the set Ω_η in (7) is defined here as $\Omega_\eta = \{(i,j) : |\rho_{ij}| \geq 0.8\}$.

The significance parameter ϵ in constraints (13) is tuned by chosing amongst the ten values $\{0.05, 0.06, 0.08, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3\}$ so as to minimize the MSE in the test set. The parameter V_T controlling the sparsity is chosen sequentially in $\{1, \ldots, N\}$. However, in order to restrict the search only to likely values of V_T, a stopping criterion is imposed: when no more features are added to the model (i.e., when the constraint (5) becomes inactive), no larger values of V_T are considered. On top of this, to further improve the speed of the tightened procedures, we have restricted the size of the parameters grids for large instances, as recommended in Tibshirani et al. (2005). In particular, for large simulated datasets we have required our output to have a maximum of 25% of non-zeroes over N, the number of predictors.

3.2 Real datasets

In this section we show the results obtained for some real datasets, which are easily reachable on internet and well referenced in the literature (Bertsimas and King, 2015). Further details about the data sets and their sources are displayed in Table 1. The columns provide information about the name, number of observations (K), the number of covariates (N), and data source.

Table 1: *Real data sets specifications and sources.*

	K	N	Source
`cpu`	105	6	Lichman (2016)
`yacht`	154	6	Lichman (2016)
`whitewine`	2499	11	Lichman (2016)
`redwine`	800	11	Lichman (2016)
`golf2008`	78	6	Winner (2016)
`golf2009`	73	11	Winner (2016)
`compact`	4096	21	Torgo (2016)

The median MSE and number of non-zeroes attained by the different estimation procedures are displayed in Tables 2-4, where the first column of results corresponds to the normalized MSE and number of non-zero coefficients (NZ), and each row shows the results of a real dataset. To obtain such results, the databases were randomly divided ten times in training, test and validation sets.

3.2.1 The effect of the sign coherence constraints

The sign coherence constraints described in Section 2.1 and formulated as (9)-(12) are claimed to avoid the inconsistencies shown by Figure 1. Now we analyse the effect of such constraints in the accuracy and sparsity of the obtained solutions. In addition, the results are compared with those under the correlation constraint (7) by Bertsimas and King (2015). The first three rows of Table 2 show the results for the untightened OLS, and the OLS with tightening sets $\mathcal{S}_1$ (the novel *sign coherence constraint*) and $\mathcal{S}_2$ (the *correlation constraint* of Bertsimas and King (2015)), respectively. Analogously, the remaining rows display the performance of the untightened and tightened lasso problems.

Table 2: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline methods and the approaches tightened by the correlation-based constraints.*

	Cpu		**Yacht**		**Whitewine**		**Redwine**		**Golf2008**		**Golf2009**		**Compact**	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
OLS	1.000	6	1.000	6	1.000	11	1.000	11	1.000	6	1.000	11	1.000	21
$\mathcal{S}_1$	1.031	5	0.970	4	1.009	9	0.999	10	1.004	5	0.896	7.5	1.014	15
$\mathcal{S}_2$	1.042	5	0.996	5	1.007	10	1.000	10	0.998	5	0.971	8	1.016	15
lasso	1.000	5	0.959	2.5	1.000	10.5	0.997	9.5	1.000	4	0.798	9.5	1.000	20.5
$\mathcal{S}_1$	1.036	4	0.954	3.5	1.009	9	0.996	10	1	5	0.79	8	1.010	14
$\mathcal{S}_2$	1.049	4.5	0.961	4	1.008	10	0.994	8.5	0.986	4	0.966	7	1.013	14

From Table 2 it can be deduced that both constraints related to multicollinearity yield a similar performance: adding sign coherence constraints slightly improves sparsity for `yacht` and `whitewine` databases, but may also attain slightly more dense solu-

tions (golf2008 and the tightened lasso in redwine database). Regarding the predictive quality, the MSEs are quite similar in most cases. An exception is the golf2009 database, where the tightening set $\mathcal{S}_1$ improves the predictive power of the tightening set $\mathcal{S}_2$ in a 7.5% for the OLS and a 17.6% for the lasso. We conclude that, not only the sign coherence constraint improves the interpretability of the results by avoiding the inconsistencies described in Section 1, but also it does not damage the level of sparsity and predictive power. Indeed, when comparing the novel coherence constraints (9)-(12) against the correlation constraint (7), they give overall the same accuracy and sparsity. We should remark that, in addition, our constraints yield more stable results than the correlation constraint when used in the lasso model. To illustrate this, consider Figure 4, which displays the paths of solutions attained for the baseline and tightened lasso in random shuffles of the golf2009 and yacht databases. The path of solutions attained by the baseline lasso on the golf2009 database (top left panel) shows that coefficients β_{11} and β_6 grow quickly in opposite directions for small values of λ. These coefficients are inflated due to the high pairwise correlation (0.91) between these variables. This phenomenon disappears when we strictly forbid coefficients β_{11} and β_6 to be simultaneously non-zero (central left panel). However, coefficients β_3, β_4 and β_6, which are considerably less correlated ($\rho_{3,4} = 0.05$, $\rho_{3,6} = 0.77$ and $\rho_{4,6} = -0.39$), still show this behaviour. Moreover, coefficient β_{10}, which was significant even for large values of the penalty λ, suddenly disappears as λ approaches zero. On the contrary, sign coherence constraints (bottom left panel) seem to avoid the inflation of coefficients β_3, β_4, β_6 and β_{11}, also leading to smoother paths. A similar behaviour is observed for yacht database, where coefficients β_2, β_3, β_4 and β_5 explode for $\lambda = 0$ in the baseline lasso (right top panel). Since the pairwise correlations between these variables do not exceed the threshold 0.8 imposed by Bertsimas and King (2015) (indeed, the largest correlation coefficient is $\rho_{3,5} = 0.63$), they are not explicitly forbidden simultaneously in the model, thus yielding the same solution path as the lasso when adding the correlation constraint (central right panel). As sign coherence constraints are less restrictive, they allow highly correlated variables to simultaneously appear in the model. This may lead to alternative solutions that may improve the stability of the estimated parameters $\boldsymbol{\beta}$ in the presence of highly correlated variables. In fact, the bottom right panel represents a considerably stable path of solutions along λ, which clearly identify the more significant feature for prediction.

As mentioned in the introduction, our constraints differ from most of the approaches previously considered in the literature: we do no explicitly forbid two highly correlated variables in the model (as recommended in Chatterjee and Hadi (2015); Bertsimas and King (2015)), nor encourage groups of correlated variables to be altogether in or out of the model (Yu and Liu, 2016; Kim and Xing, 2009). An example of this is illustrated on Figure 5, representing the estimated $\boldsymbol{\beta}$ in the 10 shuffles of golf2009 database. The $\boldsymbol{\beta}$ have been estimated by the classic OLS (first row), the lasso (second row), and two tightened OLS approaches: adding the correlation constraint (7) of Bertsimas and King (2015) (tightening set $\mathcal{S}_2$) or adding the sign coherence constraints (9)-(12) (tightening

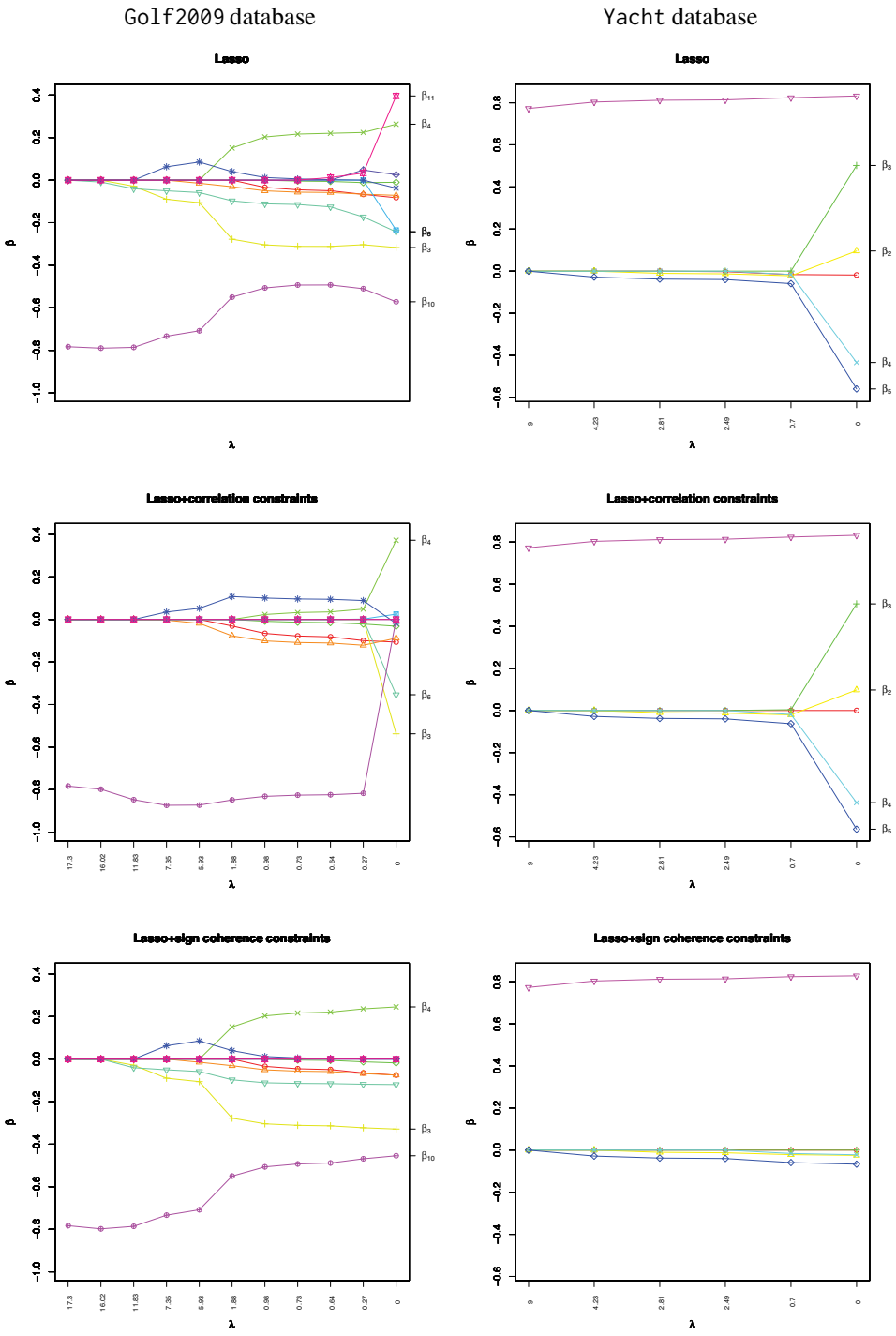

Figure 4: *Paths of solutions of the lasso (top panels), the lasso with the correlation constraint (mid panels) and sign coherence constraints (bottom panels) in shuffle 9 of the* `Golf2009` *database (left panels) and shuffle 7 of the* `Yacht` *database (right panels).*

set $\mathcal{S}_1$). In this heatmap, features X^6 and X^{10}, with correlation 0.91, do not appear simultaneously in the same shuffle when adding the correlation constraints. Nonetheless, their coefficients are non-zero simultaneously in various shuffles when considering sign coherence constraints.

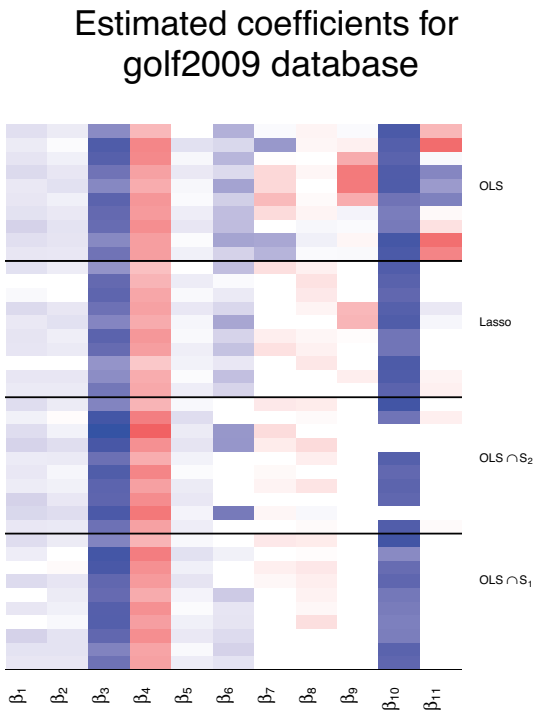

Figure 5: *Heatmap representing the coefficients β_i estimated by OLS, lasso and OLS tightened by adding correlation constraints (tightening set $\mathcal{S}_2$), or sign coherence constraints (tightening set $\mathcal{S}_1$) for* `golf2009` *database.*

The above results were obtained for fixed values of η and α, which were set to 0.6 and 0.8, respectively. These numerical examples followed the experimental design in Bertsimas and King (2015), who fixed the correlation threshold α. However, in Appendix B we explore the sensitivity of the tightening procedures to changes in η and α. We conclude that, in general, the calibration of these parameters seem to yield less sparse solutions with a similar MSE. As a consequence the results shown in our numerical experiments disregard the calibration of the correlation thresholds.

3.2.2 The effect of the significance constraint

Now we aim to study the impact of adding significance constraints (13) to the OLS and the lasso. This was briefly analysed in Section 2.2, where heat maps representing the estimated coefficients $\boldsymbol{\beta}$ were represented in Figure 3. We observed that imposing a threshold ϵ to the estimates may lead to more sparse solutions by avoiding spurious

coefficients and discarding unimportant variables. The first two rows of Table 3 display the results for the untightened OLS and its counterpart tightened with $\mathcal{S}_3$ (*significance constraint*). Analogously, the last two rows display the results for the unrestricted and tightened lasso. In this table, we observe that tightening the feasible region of the OLS and the lasso by using the set $\mathcal{S}_3$ always improves the sparsity of the output while usually attaining a competitive predictive quality. For instance, the significance constraints improves the MSE of the OLS and the lasso in a 4% and a 2.6% for `yacht` database, respectively, while reducing in 3.5 and 1.5 the number of non-zeroes. However, an exception is found in `golf2009` dataset, where the novel constraint worsens the accuracy of the OLS and the lasso, although yielding 3 more zeroes in both cases.

Table 3: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline methods and the approach tightened by the significance constraint.*

	Cpu		**Yacht**		**Whitewine**		**Redwine**		**Golf2008**		**Golf2009**		**Compact**	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
OLS	1.000	6	1.000	6	1.000	11	1.000	11	1.000	6	1.000	11	1.000	21
$\mathcal{S}_3$	0.978	4.5	0.960	2.5	0.999	8	1.000	7.5	1.007	5	1.205	8	1.002	15
lasso	1.000	5	0.959	2.5	1.000	10.5	0.997	9.5	1.000	4	0.798	9.5	1.000	20.5
$\mathcal{S}_3$	0.989	4	0.934	1	0.999	8	1.001	7	0.987	3	0.936	6.5	1.003	13

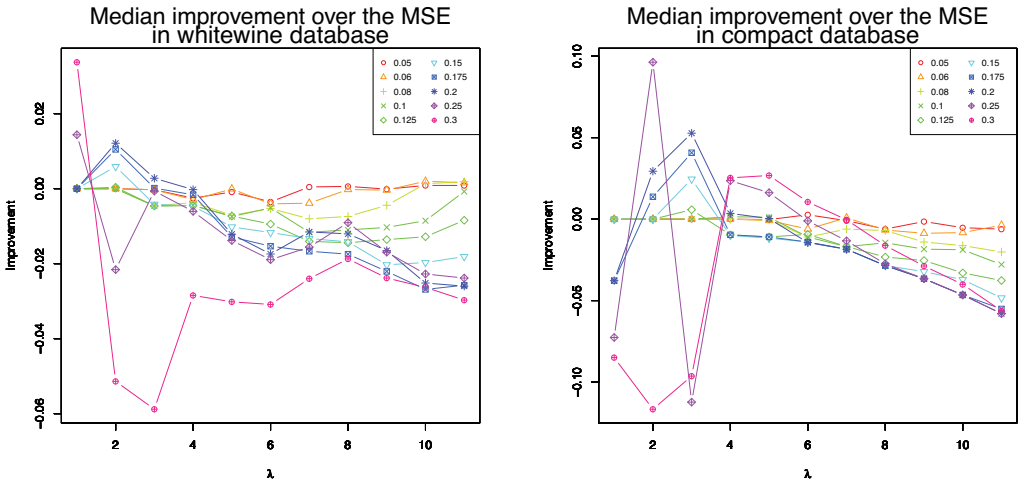

Figure 6: *Improvements over the MSE of the lasso when tightened via the significance constraint.*

Since it is not straightforward to choose a grid of thresholds ϵ to calibrate from, we will try to gain some intuition by studying the improvement of the predictive quality when adding the significance constraints for each value of ϵ considered. To this aim, Figure 6 shows the median improvement on the lasso MSE for each value of the penalty λ and each threshold in the proposed grid for `whitewine` (left panel) and `compact` (right panel) database.

From Figure 6 we observe that the behaviour for the largest ϵ $(0.3, 0.25)$ can be more effective specially for the larges values of λ. That is to say, a large ϵ may help combating the strong shrinkage of the lasso when highly sparse solutions are sought. Still large but more conservative values of the threshold ($\epsilon = 0.2, 0.175, 0.15$) seem to also improve the MSE of the lasso with large λ, also providing less extreme behaviours than the choices $\epsilon = 0.3, 0.25$. Finally, the smallest values of the threshold ($\epsilon = 0.06, 0.05$) may slightly improve the lasso with small λ and the OLS. In conclusion, there is no straightforward a priori choice for the parameter ϵ, which should be calibrated. Nevertheless, if the user is seeking a highly sparse solution (i.e., the user is choosing a high penalty λ) it seems advisable to choose larger values for ϵ in order to combat the shrinkage more effectively. On the other hand, when estimating via OLS (or lasso with small values of λ) the user might want to focus on a grid with a majority of small values of ϵ.

3.2.3 Global performance

Finally, we show in Table 4 the results when our novel constraints are jointly considered in combination to the cardinality constraint, or equivalently, when the set $\mathcal{S}_4$ (*cardinality + sign coherence + significance constraints*) is used to tighten the OLS or lasso approaches. For comparison reasons, the table also shows the results under the tightening set $\mathcal{S}_5$ (*cardinality + correlation constraints*) proposed in Bertsimas and King (2015). Table 4 also displays the predictive quality and number of non-zero coefficients for the Enet, the SRIG method in Yu and Liu (2016) and the GFlasso method in Kim and Xing (2009).

Table 4: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline estimation methods (OLS, lasso, SRIG, Enet and GFlasso), and OLS and lasso tightened with $\mathcal{S}_4$ (cardinality + novel constraints) or $\mathcal{S}_5$ (cardinality + correlation constraint), for the real datasets.*

Method	**Cpu**		**Yacht**		**Whitewine**		**Redwine**		**Golf2008**		**Golf2009**		**Compact**	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
OLS	1.000	6	1.000	6	1.000	11	1.000	11	1.000	6	1.000	11	1.000	21
$\mathcal{S}_4$	0.990	4	0.934	1	1.011	6	1.002	6	0.999	3	0.963	4	1.012	9
$\mathcal{S}_5$	1.002	4	0.934	1	1.008	8	1.000	6	1.006	3	1.024	5	1.016	11
lasso	1.000	5	0.959	2.5	1.000	10.5	0.997	9.5	1.000	4	0.798	9.5	1.000	20.5
$\mathcal{S}_4$	0.993	4	0.934	1	1.013	6	0.998	6	0.982	3	0.971	4.5	1.013	9
$\mathcal{S}_5$	1.049	4	0.934	1	1.008	7.5	0.995	6	0.988	3	1.045	4	1.015	11.5
SRIG	0.988	6	0.942	2	1.000	11	0.999	11	0.983	6	1.016	11	0.999	21
Enet	0.917	5.5	0.948	2	1.000	11	0.998	11	0.994	4	0.805	9.5	1.000	21
GFlasso	0.972	6	0.959	2.5	1.000	11	0.996	11	1.000	4	0.866	9.5	1.001	20

First, we analyse the performance of the baseline estimation procedures (OLS and lasso) against their tightened counterparts. We can conclude that reducing the search space of the coefficients $\boldsymbol{\beta}$ by intersecting with either $\mathcal{S}_4$ or $\mathcal{S}_5$ always improves the sparsity of the solutions. Moreover, the predictive quality is usually similar to that of the OLS and the lasso. In particular, for `yacht` database, both tightened approaches

improve the MSE of the OLS and the lasso estimates by a 6.6% and a 2.6%, respectively. However, these tightened procedures worsens the accuracy of the lasso in the `golf2009` database.

Second, we compare the performance of the two tightening sets. From the table it can be observed that they deliver a similar accuracy-sparsity trade-off for 3 out of the 7 real datasets (`yacht, redwine, golf2008`). For the remaining databases, the approaches attain different trade-offs between sparsity and predictive quality. Indeed, the novel set $\mathcal{S}_4$ maintains the sparsity attained by the set $\mathcal{S}_5$ og Bertsimas and King (2015) on the `cpu` database, while slightly improving the MSE in a 1.2% and a 5.3% for the tightened OLS and lasso, respectively. In contrast, the proposed tightening set provides more sparse solutions, with a similar predictive quality for `whitewine` and `compact` datasets. Finally, $\mathcal{S}_4$ improves the MSE of the OLS and the lasso in a 6% and 7.1% in `golf2009` database, but it provides a slightly more dense solution for the lasso. Note that the best accuracy for the lasso in this database was attained when adding exclusively sign coherence constraints (see Table 2), although the solution provided was more dense. Third, both the SRIG and GFlasso are clearly outperformed by the tightened approaches. On top of this, for the real datasets considered here there is no guarantee that the necessary assumptions to preserve the theoretical properties of the SRIG are fulfilled. In order to compare the performance of our approach against these methods under a more favourable scenario for the later, in the next section we replicate the simulation study of Yu and Liu (2016), hence assuring the non-violation of the conditions for the SRIG method.

3.3 Simulations

In the previous section, the behaviour of the novel constraints for the case of real datasets with a small number of predictors was analysed. In this section we aim to examine the sensitivity of the tightened procedures under various settings. First, we will simulate data as in Yu and Liu (2016) to understand the behaviour of the methodology under different correlation structures and for different sizes of the training sample. Second, we aim to test the proposed methodology for larger datasets and for different correlations intensities. To do so we simulate data as in Bertsimas and King (2015).

3.3.1 Sensitivity to correlation structure and training sample size

In this section we aim to test the proposed methodology for datasets simulated following the three examples and training sizes described in Yu and Liu (2016). Ten instances have been generated for each example and for training and testing sizes of 40, 80 and 120. For both Examples 1 and 2, $\beta_i = 3$ for $i = 1,..,15$, and $\beta_i = 0$ for $i = 16,\ldots,100$. In Example 1, however, the predictors were generated as follows:

$$X^i = Z^j + 0.4\mathbf{a}_X^i, \quad Z^j \sim N(0,1), \quad \mathbf{a}_X^i \sim N(0,1)$$

for $5(j-1)+1 \leq i \leq 5j$ and $j = 1,2,3$. For $i > 15$, $X^i \sim N(0,1)$. In Example 2, the k-th vector of observations $\left(X_k^1, \ldots, X_k^N\right)$ was generated following a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = 0.5^{|i-j|}$. Analogously, in Example 3 the vector of observations was also drawn from a multivariate normal distribution with zero mean but with covariance matrix $\Sigma = (B + \psi I)^{-1}$, where $b_{ij} = 0$ for $i = j$, and $b_{ij} = 0.5\delta_{ij}$, $\delta_{ij} \sim Be(0.05)$ otherwise. Parameter ψ is fixed so that the condition number of Σ^{-1} equals N. The real coefficients are $\boldsymbol{\beta} = \Sigma^{-1}\Sigma_{xy}$, where Σ_{xy} is the cross-covariance vector whose elements equal 10 for the four predictors with the largest degrees and 0 otherwise. These three examples have different structures of correlation and, as it will be seen later, this may influence on the performance of the approaches making use of constraints taking into account pairwise correlations. As an illustration, Figure 8 of the Appendix C of the Supplementary Material displays the pairs of variables appearing in the correlation constraints of Bertsimas and King (2015) and our sign coherence constraints.

As done in Bertsimas and King (2015), the grid of values of the parameter λ to be tuned for the tightened MIQP with lasso objective function is logarithmically generated in the interval $(0, \lambda_{max}]$, where λ_{max} is the penalty provided by the `lars` for which only one coefficient is non-zero. Analogously to the real datasets results, Table 5 shows the median MSE and number of non-zeroes (NZ). In particular, for each example (rows) and training sizes (columns), each row shows the results obtained for an estimation method, namely the OLS, the lasso, and their tightened counterparts, which take the form of Problem (6) with tightening sets $\mathcal{S}_4$ (*cardinality + sign coherence + significance constraints*) and $\mathcal{S}_5$ (*cardinality + correlation constraints*). The last rows for each example display the results for the methods dealing with correlated variables: the Enet, the SRIG, and the GFlasso. To make it easier to discuss these results, in Figure 7 we have represented the MSE against the number of nonzeroes for the three simulated examples of Table 5. The different estimation methods have been assigned different colours, the solid items representing the approaches making use of our proposed tightening set. The diversity of training samples have been represented by unalike symbols.

In the top panel of Figure 7 we can observe that, in Example 1, the MIQPs (i.e., the approaches proposed in Bertsimas and King (2015) and in this paper) tend to attain more sparse solutions than the continuous optimization methods (i.e., the lasso, SRIG and GFlasso). As it can be observed in Figure 8 in the Appendix, the true generating model contains predictors that hold a high pairwise correlation and, therefore, are forbidden to appear simultaneously in the outputs yielded by the tightening set of Bertsimas and King (2015). As a consequence, the approaches making use of this set $\mathcal{S}_5$ are not able to recover the original graph, hence delivering more sparse solutions but with worse predictive quality than our tightening set. In constrast, the sign coherence constraints flexibility allows the simultaneous presence of all the variables in the generating model. As a consequence, $\mathcal{S}_4$ yields outputs with similar performance to that of the SRIG, that is the best approach amongst those based on continuous optimization models.

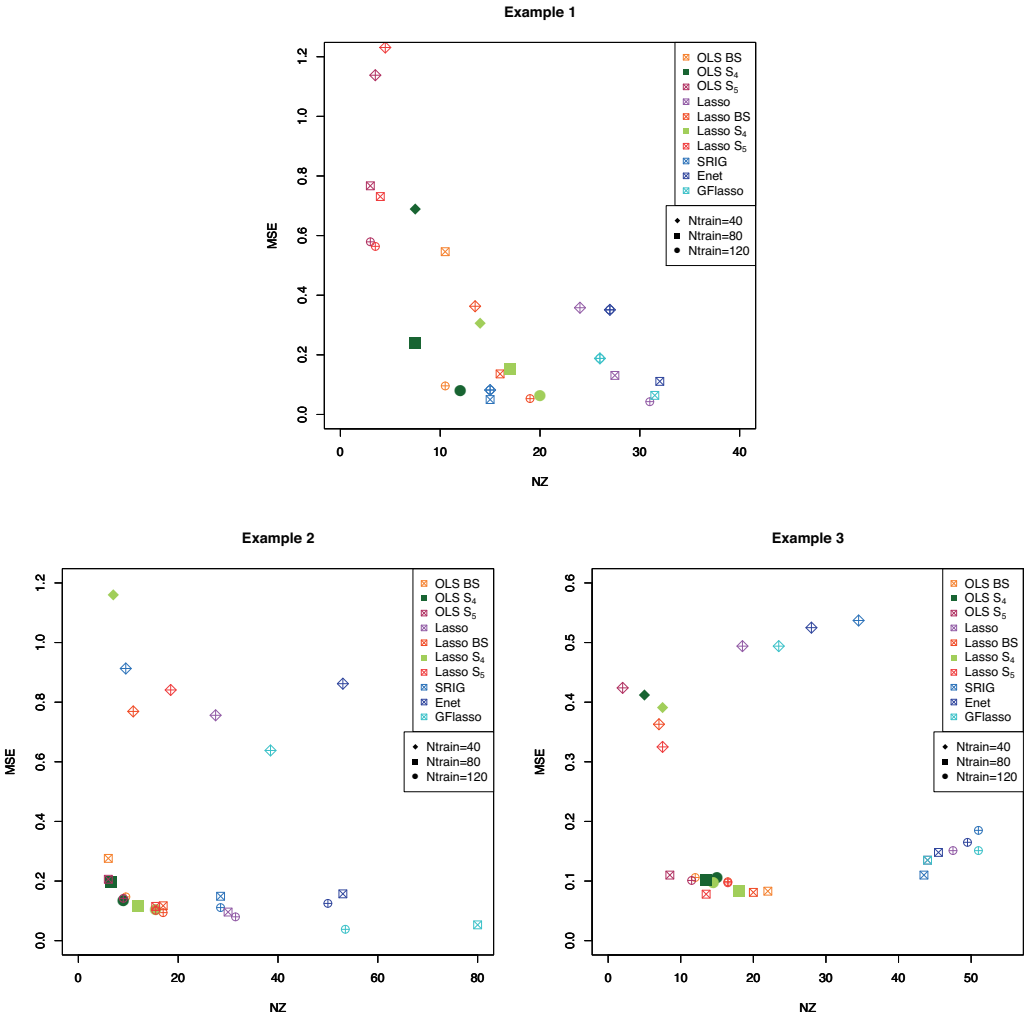

Figure 7: *Median MSE and NZ for each method in the simulated datasets of Yu and Liu (2016).*

In the bottom panels of Figure 7 we observe that the MIQP approaches are clustered together. For both Examples 2 and 3 it is evident that the methods based on integer optimization improve the sparsity of the outputs with no damage to the predictive quality. More specifically, in Example 3 all the methods based on continuous optimization conform a unique cluster with higher density and similar or slightly worse MSE than the MIQP approaches. Nonetheless, in Example 2 the SRIG and the lasso attain more sparse solutions than the rest of continuous optimization approaches, yet still yielding more non-zeroes than the tightened procedures. In this case, all the approaches attain a similar accuracy.

Table 5: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline estimation methods (OLS, lasso, SRIG, Enet and GFlasso), and OLS and lasso tightened with $\mathcal{S}_4$ (cardinality + novel constraints) or $\mathcal{S}_5$ (cardinality + correlation constraint), for the simulated datasets with $N = 100$.*

		Ntrain= 40		Ntrain= 80		Ntrain= 120	
		MSE	NZ	MSE	NZ	MSE	NZ
Example 1	OLS	1.000	100.0	1.000	100.0	1.000	100.0
	$\mathcal{S}_4$	0.689	7.5	0.240	7.5	0.081	12.0
	$\mathcal{S}_5$	1.138	3.5	0.767	3.0	0.579	3.0
	lasso	0.358	24.0	0.131	27.5	0.043	31.0
	$\mathcal{S}_4$	0.306	14.0	0.154	17.0	0.062	20.0
	$\mathcal{S}_5$	1.231	4.5	0.731	4.0	0.564	3.5
	SRIG	0.082	15.0	0.050	15.0	0.082	15.0
	Enet	0.351	27.0	0.111	32.0	0.351	27.0
	GFlasso	0.188	26.0	0.064	31.5	0.188	26.0
Example 2	OLS	1.000	100.0	1.000	100.0	1.000	100.0
	$\mathcal{S}_4$	1.380	5.0	0.197	6.5	0.133	9.0
	$\mathcal{S}_5$	1.560	4.0	0.206	6.0	0.142	9.0
	lasso	0.756	27.5	0.096	30.0	0.080	31.5
	$\mathcal{S}_4$	1.160	7.0	0.115	12.0	0.105	15.5
	$\mathcal{S}_5$	0.841	18.5	0.115	15.5	0.102	15.5
	SRIG	0.913	9.5	0.149	28.5	0.111	28.5
	Enet	0.862	53.0	0.157	53.0	0.125	50.0
	GFlasso	0.638	38.5	0.053	80.0	0.038	53.5
Example 3	OLS	1.000	100.0	1.000	100.0	1.000	100.0
	$\mathcal{S}_4$	0.412	5.0	0.101	13.5	0.106	15.0
	$\mathcal{S}_5$	0.424	2.0	0.110	8.5	0.101	11.5
	lasso	0.494	18.5	0.135	44.0	0.151	47.5
	$\mathcal{S}_4$	0.391	7.5	0.083	18.0	0.097	14.5
	$\mathcal{S}_5$	0.325	7.5	0.078	13.5	0.097	16.5
	SRIG	0.537	34.5	0.110	43.5	0.185	51.0
	Enet	0.525	28.0	0.148	45.5	0.165	49.5
	GFlasso	0.494	23.5	0.135	44.0	0.151	51.0

3.3.2 Scalability and sensitivity to correlation intensity

In this section we aim to test the proposed methodology for larger datasets with diverse correlation intensities. As the overall design of experiments, the synthetic generation of the data is similar to that of Bertsimas and King (2015). The k-th vector of observations $\left(X_k^1,\ldots,X_k^N\right)$ was generated following a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = \rho^{|i-j|}$. In particular, we chose $\rho = -0.9$ and $\rho = -0.5$ so as to test the performance of the constraints under highly and moderate correlations. The regression model is taken in small dimension, but with quite a number of irrelevant covariates. More precisely, the number of features was set to 500, only 10 of

which corresponding to explanatory variables, the remaining 490 being noise. For each value of ρ, 10 instances were generated as follows. The β_i were uniformly generated in the interval $(-2,2)$ for i such that i modulus $N/10 = 0$. The response was generated following (1), with $\beta_0 = 0$ and the error terms i.i.d. following a normal distribution with zero mean and variance as in Bertsimas and King (2015). As done in Section 3.3.1, the sequence of λ has been logarithmically generated. Table 6 shows the median MSE and number of non-zeroes (NZ) of the solutions for the OLS, the lasso and their tightened counterparts, as well as the SRIG, Enet and GFlasso.

Table 6: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline estimation methods (OLS, lasso, SRIG, Enet and GFlasso), and OLS and lasso tightened with $\mathcal{S}_4$ (cardinality + novel constraints) or $\mathcal{S}_5$ (cardinality + correlation constraint), for the simulated datasets with $N = 500$.*

	$\rho = -0.5$		$\rho = -0.9$	
	MSE	NZ	MSE	NZ
OLS	1.000	500	1.000	500
$\mathcal{S}_4$	0.745	6	0.732	6
$\mathcal{S}_5$	0.849	6	1.474	4
lasso	0.535	80.5	0.533	183
$\mathcal{S}_4$	0.526	10.5	0.528	11
$\mathcal{S}_5$	0.501	13	0.807	19
SRIG	0.630	34	0.792	102
Enet	0.535	50	0.534	51.5
GFlasso	0.537	70.5	0.533	54.5

Note that, for the simulated data with moderately correlated features ($\rho = -0.5$), the correlation constraints (7) and the sign coherence constraints (9)-(12) are inactive, since the highest pairwise correlation in absolute value is roughly 0.5. In this case, both sets of tightening constraints help to considerably improve the sparsity of the baseline estimation procedures. Indeed, the density of the OLS is drastically reduced by increasing in 494 the zeroes of the output, while its predictive quality is also substantially improved. However, the novel tightening set $\mathcal{S}_4$ provides an accuracy 12.2% better than the attained with the set $\mathcal{S}_5$ proposed in Bertsimas and King (2015). On the other hand, the benchmark sparse regression method, the lasso, attains a median of 80.5 non-zeroes, while its tightening counterparts produce much more sparse solutions with better MSEs. Although the predictive quality of the outputs of the tightened procedure with $\mathcal{S}_5$ is slightly better than that obtained with $\mathcal{S}_4$, this comes at the price of yielding more dense solutions. More generally, the methods based on MIQP solvers clearly outperform the approaches relying on continuous optimization techniques: the former manage to considerably reduce the sparsity of the later with a slightly better predictive quality. Amongst the later approaches, the GFlasso is outperformed by the Enet which, with a similar accuracy, provides a much more sparse solution. Nonetheless, the most sparse of these methods is the SRIG, although yielding the worst predictive quality.

Table 7: *Predictive quality (MSE) and sparsity degree (NZ) for the baseline estimation methods (OLS, lasso, SRIG, Enet and GFlasso), and OLS and lasso tightened with $\mathcal{S}_4$ (cardinality + novel constraints) or $\mathcal{S}_5$ (cardinality + correlation constraint), for the simulated datasets with $\rho = -0.9$ and $N = 50, 500, 1000$.*

	$N = 50$		$N = 500$		$N = 1000$	
	MSE	NZ	MSE	NZ	MSE	NZ
OLS	1.000	50	1.000	500	1.000	1000
$\mathcal{S}_4$	0.662	8	0.732	6	1.035	5
$\mathcal{S}_5$	0.634	8	1.474	4	1.310	5.5
lasso	0.671	24.5	0.533	183	0.555	56
$\mathcal{S}_4$	0.637	8.5	0.528	11	0.904	8
$\mathcal{S}_5$	0.595	19	0.807	19	0.406	10
SRIG	0.127	42	0.792	102	0.835	103.5
Enet	0.108	26	0.534	51.5	0.538	49.5
GFlasso	0.693	41	0.533	54.5	0.615	359

The simulated instances with highly correlated features, $\rho = -0.9$, show a similar behaviour. Nevertheless, the lasso provides significantly more dense outputs in this case, with a median of 183 non-zeroes. The tightening set $\mathcal{S}_5$ proposed in Bertsimas and King (2015) worsens its predictive quality although improving substantially its sparsity. In contrast, the novel tightening set $\mathcal{S}_4$ attains around eight more zeroes than the later while maintaining the lasso's accuracy. Regarding the OLS, both tightened estimation methods reduce drastically the density of the solutions, although the $\mathcal{S}_4$ obtains around two more non-zero coefficients than the tightening set $\mathcal{S}_5$. However, the price paid for two extra non-zeroes is shown in the accuracy of the solution: while adding cardinality, sign coherence and significance constraints improves the MSE of the OLS in a 26.8%, adding cardinality and correlation constraints instead worsens the accuracy in 47.4%. Analysing the results for the continuous optimization based methods, we observe that the best performance is yielded by the Enet and GFlasso which, with a similar MSE, significantly enhance the sparsity of the classic lasso. Nonetheless, when tightened with $\mathcal{S}_4$, the later attains outputs with comparable predictive quality and many more zero coefficients (around 40 more).

In order to analyse the scalability of our methodology we have also simulated data as in Bertsimas and King (2015) with 50 and 1000 variables, and a maximum correlation of 0.9. For the later, the MIQP to be solved would have 3001 variables and more than 2000 constraints. As the size of the problem is considerably larger, we have allowed for a time limit of 40 seconds in this case, which was also the time limit considered for high dimensional data where $N > K$. The results are collected in Table 7. As it can be observed, MIQP approaches attain solutions that are considerably more sparse than their continuous counterparts while still delivering a good predictive quality. In particular, SRIG delivers the most dense outputs from the sparse continuous methods, while the GFlasso is outperformed both in terms of accuracy and sparsity. Regarding the MIQP,

they attain different accuracy-sparsity trade-offs. Although the proposed constraints attain better accuracy when combined with the OLS objective, this is not true when the regularization penalty λ is positive. Nonetheless, in this case the tightening set $\mathcal{S}_4$ yields more sparse outputs than $\mathcal{S}_5$.

Summarizing, tightening the search space of the OLS and the lasso provides considerably more sparse solutions, although the sign coherence and significance constraints yield a better accuracy-sparsity trade-off than the correlation constraint. Indeed, the later can substantially worsen the predictive quality of the baseline methods in order to reduce the density of the outputs, while the former entails a more competitive MSE.

4 Concluding remarks

The aim of this paper is to enhance the interpretability in a regression model without worsening its predictive quality. We assume we have a baseline regression estimation procedure based on solving an optimization problem (e.g. OLS or lasso), and then the underlying optimization problems are modified by adding new constraints to those defining the search space. These constraints avoid misleading estimators that may be obtained in the presence of highly correlated variables and detect the most important features for the prediction.

In order to assess the impact of adding the two novel constraints over various estimation procedures, in our numerical experiments we consider the OLS and the lasso. The search space of $\boldsymbol{\beta}$ is reduced in these methods by using the tightening set composed by the sign coherence constraints (9)-(12) and/or the significance constraints (14), possibly in combination with the cardinality constraint (5). The first constraint forces the sign of the coefficients to be coherent with the sign of large and moderately large pairwise correlations between features, while the second avoids spurious coefficients and combats the shrinkage of regularized regression. We compare the performance of our tightening set, including all the proposed constraints, with the recent tightening set by Bertsimas and King (2015), which also defines a MIQP and includes the cardinality constraint (5) and the correlation constraint (7) which explicitly forbids two highly correlated variables to be simultaneously in the regression model. These methods are compared against other approaches also dealing with correlated variables but based on continuous optimization techniques: the Enet (Zou and Hastie, 2005), the SRIG (Yu and Liu, 2016), and the GFlasso (Kim and Xing, 2009). The results show that the novel constraints yield tractable optimization problems, solvable in short time by standard solvers, and may enhance the interpretability while often improving or maintaining the predictive quality and level of sparsity. More specifically, the MIQP approaches attain a different trade off between sparsity and predictive quality than the methods based on continuous optimization, usually yielding more sparse solution with similar or better MSE. Amongst the former methods, the novel constraints tend to improve the predictive quality of the outputs obtained with the tightening set proposed in Bertsimas and King (2015).

Although we have proposed some heuristics to further improve the speed of the tightened procedures, such as reducing the grid of parameters for large datasets, in the future we aim to develop tailored heuristics that improve the computational times of the MIQPs when they particularly model linear regression problems.

References

Atamurk, A., Nemhauser, G. and Savelsbergh, M. (2000). Conflict graphs in solving integer programming problems. *European Journal of Operational Research*, 121, 40–55.

Bartholomew, D. J., Steele, F., Moustaki, I. and Galbraith, J. (2008). *Analysis of Multivariate Social Science Data*. Chapman & Hall.

Bertsimas, D. and King, A. (2015). OR forum – An algorithmic approach to linear regression. *Operations Research*, 64, 2–16.

Bertsimas, D., King, A., Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44, 813–852.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37, 373–384.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Cai, A., Tsay, R. and Chen, R. (2009). Variable selection in linear regression with many predictors. *Journal of Computational and Graphical Statistics*, 18, 573–591.

Camm, J. D., Raturi, A. S. and Tsubakitani, S. (1990). Cutting big *M* down to size. *Interfaces*, 20, 61–66.

Cao, G., Guo, Y. and Bouman, C. A. (2010). High dimensional regression using the sparse matrix transform (SMT). In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 1870–1873. IEEE.

Carrizosa, E. and Guerrero, V. (2014). Biobjective sparse principal component analysis. *Journal of Multivariate Analysis*, 132, 151–159.

Carrizosa, E., Nogales-Gómez, A. and Morales, D. R. (2016). Strongly agree or strongly disagree?: Rating features in support vector machines. *Information Sciences*, 329, 256–273.

Carrizosa, E., Nogales-Gómez, A. and Morales, D. R. (2017). Clustering categories in support vector machines. *Omega*, 66, 28–37.

Carrizosa, E., Olivares-Nadal, A. V. and Ramírez-Cobo, P. (2016). A sparsity-controlled vector autoregressive model. *Biostatistics*, 18, 244–259.

Chatterjee, S. and Hadi, A. S. (2015). *Regression Analysis by Example*. John Wiley & Sons.

Danna, E., Rothberg, E. and Le Pape, C. (2005). Exploring relaxation induced neighborhoods to improve mip solutions. *Mathematical Programming*, 102, 71–91.

Efron, B. and Hastie, T. (2003). LARS software for R and Splus. https://web.stanford.edu/ hastie/ Papers/LARS/.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407–499.

Farrar, D. E. and Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92–107.

Fischetti, M. and Lodi, A. (2005). Local branching. *Mathematical Programming*, 98, 23–47.

Fourer, R., Gay, D. and Kernighan, B. W. (2002). *The AMPL book*. Duxbury Press, Pacific Grove.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Volume 1. Springer series in statistics.

Hastie, T. and Efron, B. (2013). Least Angle Regression, Lasso and Forward Stagewise. http://cran.r-project.org/web/packages/lars/lars.pdf.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press.

Hesterberg, T., Choi, N. H., Meier, L. and Fraley, C. (2008). Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2, 61–93.

Jou, Y.-J., Huang, C.-C. L. and Cho, H.-J. (2014). A VIF-based optimization model to alleviate collinearity problems in multiple linear regression. *Computational Statistics*, 29, 1515–1541.

Kim, S. and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5, e1000587.

Lichman, M. (2016). UCI machine learning repository. http://archive.ics.uci.edu/ml. University of California, Irvine, School of Information and Computer Sciences.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60, 234–256.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7, 1607–1631.

Miller, A. (2002). *Subset Selection in Regression* (2 ed.). Chapman & Hall/CRC.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, Volume 821. John Wiley & Sons.

Rothberg, E. (2007). An evolutionary algorithm for polishing mixed integer programming solutions. *INFORMS Journal on Computing*, 19, 534–541.

Savelsbergh, M. (1994). Preprocessing and probing techniques for mixed integer programming problems. *ORSA Journal on Computing*, 6, 445–454.

Sengupta, D. and Bhimasankaram, P. (1997). On the roles of observations in collinearity in the linear model. *Journal of the American Statistical Association*, 92, 1024–1032.

Silvey, S. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 539–552.

Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K. and Matsui, T. (2017). Best subset selection for eliminating multicollinearity. *Journal of the Operations Research Society of Japan*, 60, 321–336.

Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K. and Matsui, T. (2019). Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. *Journal of Global Optimization*, 73, 431–446.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Torgo, L. (2016). Regression data sets. http://www.dcc.fc.up.pt/ ltorgo/Regression/DataSets.html. University of Porto, Faculty of Sciences.

Watson, P. K. and Teelucksingh, S. S. (2002). *A Practical Introduction to Econometric Methods: Classical and Modern*. University of West Indies Press.

Winner, L. (2016). Miscellaneous data sets. http://www.stat.ufl.edu/ winner/datasets.html. University of Florida.

Yu, G. and Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. *Journal of the American Statistical Association*, 111, 707–720.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

A Mathematical formulation of the methods under comparison

The optimization problems that are solved in the numerical study take the form of Problem (6), whose objective function is that of the lasso or of OLS (i.e., the objective function in Problem (3), where $\lambda = 0$ in the case of the latter). The tightening sets are denoted as $\mathcal{S}_m$, $m = 1,..,5$, and were defined in Section 3.1. For the sake of comprehension, these problems will be explicitly stated now.

A.1 Lasso regression problem with tightening set $\mathcal{S}_1$

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \beta_0 - \boldsymbol{\beta}\mathbf{X}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ \text{s.t} \quad & \begin{cases} \nu_i^+ + \nu_j^- \leq 1 & \forall (i,j) \in \Omega_\alpha^+ \\ \nu_i^- + \nu_j^+ \leq 1 & \forall (i,j) \in \Omega_\alpha^+ \\ \nu_i^+ + \nu_j^+ \leq 1 & \forall (i,j) \in \Omega_\alpha^- \\ \nu_i^- + \nu_j^- \leq 1 & \forall (i,j) \in \Omega_\alpha^- \\ \nu_j^+, \nu_j^- \in \{0,1\} & \forall j = 1,\ldots,N \end{cases} \end{aligned} \tag{15}$$

A.2 Lasso regression problem with tightening set $\mathcal{S}_2$

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \beta_0 - \boldsymbol{\beta}\mathbf{X}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ \text{s.t} \quad & \begin{cases} \gamma_i + \gamma_j \leq 1 & \forall (i,j) \in \Omega_\eta \\ \gamma_j \in \{0,1\} & \forall j = 1,\ldots,N \end{cases} \end{aligned} \tag{16}$$

A.3 Lasso regression problem with tightening set $\mathcal{S}_3$

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \beta_0 - \boldsymbol{\beta}\mathbf{X}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ \text{s.t} \quad & \begin{cases} \beta_j \geq \epsilon\nu_j^+ - \nu_j^- M & \forall j = 1,\ldots,N \\ \beta_j \leq -\epsilon\nu_j^- + \nu_j^+ M & \forall j = 1,\ldots,N \\ \nu_j^+, \nu_j^- \in \{0,1\} & \forall j = 1,\ldots,N \end{cases} \end{aligned} \tag{17}$$

A.4 Lasso regression problem with tightening set $\mathcal{S}_4$

$$\begin{array}{ll}\min\limits_{\boldsymbol{\beta}} & \|\mathbf{y}-\beta_0-\boldsymbol{\beta}\mathbf{X}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1 \\ \text{s.t} & \left\{\begin{array}{ll} \sum\limits_{j=1}^{N}(\nu_j^+ + \nu_j^-) \leq V_T & \\ \beta_j \geq \epsilon\nu_j^+ - \nu_j^- M & \forall j=1,\ldots,N \\ \beta_j \leq -\epsilon\nu_j^- + \nu_j^+ M & \forall j=1,\ldots,N \\ \nu_i^+ + \nu_j^- \leq 1 & \forall (i,j)\in\Omega_\alpha^+ \\ \nu_i^- + \nu_j^+ \leq 1 & \forall (i,j)\in\Omega_\alpha^+ \\ \nu_i^+ + \nu_j^+ \leq 1 & \forall (i,j)\in\Omega_\alpha^- \\ \nu_i^- + \nu_j^- \leq 1 & \forall (i,j)\in\Omega_\alpha^- \\ \nu_j^+, \nu_j^- \in \{0,1\} & \forall j=1,\ldots,N \end{array}\right. \end{array} \tag{18}$$

A.5 Lasso regression problem with tightening set $\mathcal{S}_5$

$$\begin{array}{ll}\min\limits_{\boldsymbol{\beta}} & \|\mathbf{y}-\beta_0-\boldsymbol{\beta}\mathbf{X}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1 \\ \text{s.t} & \left\{\begin{array}{ll} \sum\limits_{j=1}^{N}\gamma_j \leq V_T & \\ \gamma_i + \gamma_j \leq 1 & \forall (i,j)\in\Omega_\eta \\ \gamma_j \in \{0,1\} & \forall j=1,\ldots,N \end{array}\right. \end{array} \tag{19}$$

B Calibration of correlation thresholds

In this section we calibrate the correlation thresholds η and α in the grid $\{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0, 5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9\}$. Table 8 reports the predictive quality and the number of non-zero coefficients attained by calibrating these parameters. In comparison to results in Table 2, attained for fixed η and α, the calibrateD methods yield more dense outputs with a similar accuracy.

Table 8: *Predictive quality (MSE) and sparsity (NZ) for the approaches tightened with the correlation-based constraints when parameters η and α are calibrated.*

	Cpu		**Yacht**		**Whitewine**		**Redwine**		**Golf2008**		**Golf2009**		**Compact**	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
OLS	1.000	6	1.000	6	1.000	11	1.000	11	1.000	6	1.000	11	1.000	21
$\mathcal{S}_1$	1.000	6	1.000	6	1.000	11	1.000	11	1.000	6	1.000	11	1.000	21
$\mathcal{S}_2$	1.038	5	0.987	5	1.000	10	0.999	7	1.005	3.5	0.889	8	1.016	15
lasso	1.000	5	0.959	2.5	1.000	10.5	0.997	9.5	1.000	4	0.798	9.5	1.000	20.5
$\mathcal{S}_1$	1.000	6	1.000	6	1.000	11	0.999	11	0.999	6	0.871	9.5	1.000	21
$\mathcal{S}_2$	1.042	5	0.969	5	1.000	10	1.000	8	0.998	3.5	0.812	7.5	1.015	15

C Correlated variables in simulated data with $N = 100$

In order to better understand the results of the tightened procedures displayed in Table 5 and Figure 7, in Figure 8 we have represented heatmaps that indicate whether two variables are highly correlated ($|\rho| \geq 0.8$) or moderately correlated ($|\rho| \geq 0.6$) for a random instance of each example of simulated data in Section 3.3.1. Orange colour indicates that the correlation constraint (7) is included in the tightening set $\mathcal{S}_5$ and also that sign coherence constraints (9)-(12) are added to the tightening set $\mathcal{S}_4$. Green colour stands only for the presence of sign coherence constraints (9)-(12) in $\mathcal{S}_4$. Left panels represent the correlations amongst all features, while right panels show the correlations only amongst the predictors truly appearing in the generating model (i.e. $\beta_i \neq 0$). As it can be observed, some features appearing in the generating model of Example 1 have a correlation larger than 0.8 in absolute value, and therefore they are forbidden to appear together in the output model of $\mathcal{S}_5$.

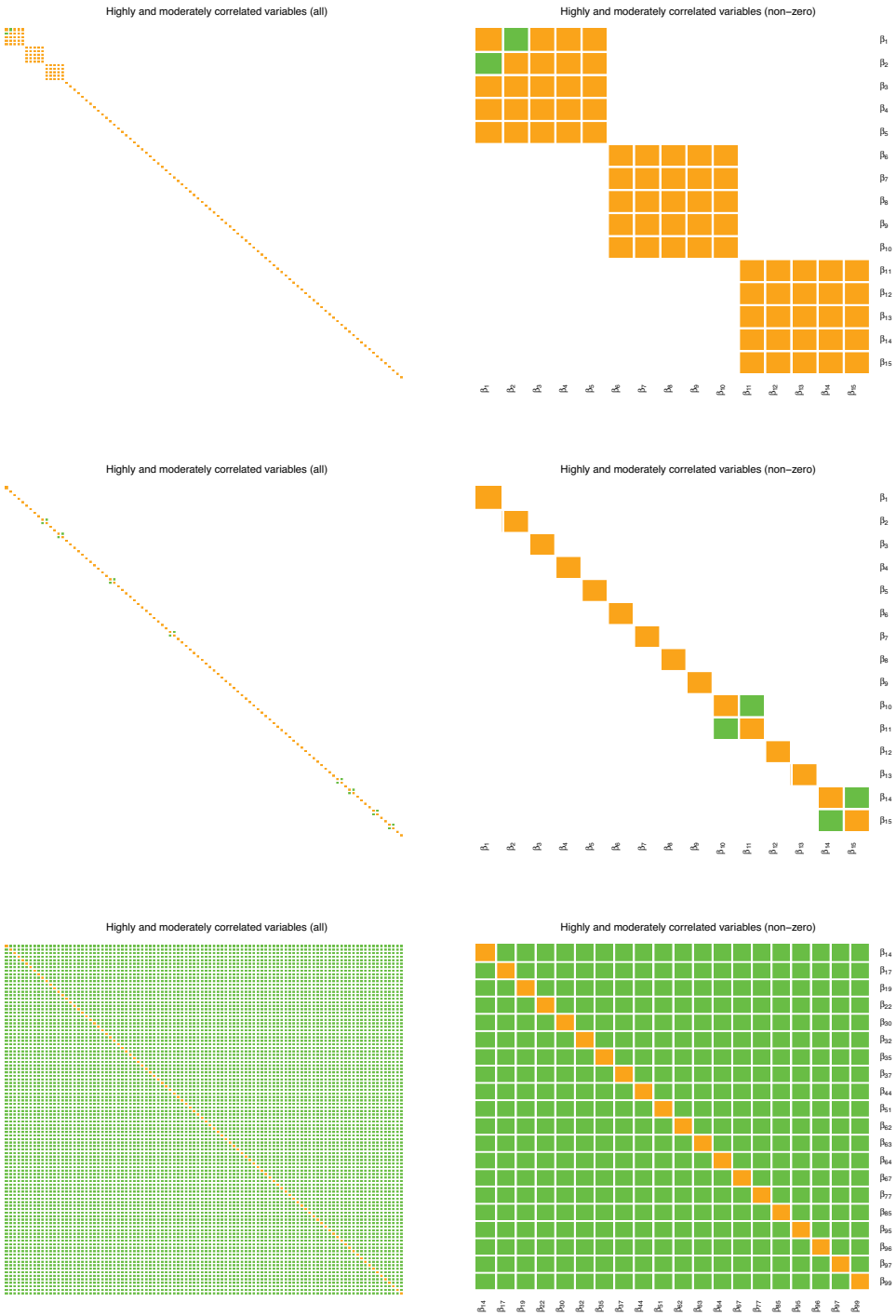

Figure 8: *Features exceeding a pairwise correlation of* 0.8 *in absolute value (orange), hence appearing in the correlation constraint* (7) *and also sign coherence constraints* (9)-(12)*, and features exceeding a pairwise correlation of* 0.6 *in absolute value (green), appearing solely in the sign coherence constraints.*

D AMPL code

```
#DATA PARAMETERS
param N;           #Number of predictors
param K;           #Number of observations
#SETS OF INDICES
set Nvar:=1..N;
set Nobs:=1..K;
#THE MATRIX OF DATA
param X {i in Nobs,j in 1..(N+1)};   # Kx(N+1) matrix of data
                                     # Includes response variable in position N+1
#PARAMETERS OF THE METHODS
param lambda;                        # Lasso penalty
param NZ;                            #Upper-bound on the total number of non-zeroes
param eps;                           #Significance threshold
param M default 100;                 #Upper bound for the coefficients beta
#SETS OF HIGHLY/MODERATELY CORRELATED VARIABLES
set conjcorrpos dimen 2;             # Positively correlated features
set conjcorrneg dimen 2;             # Negatively correlated features
#VARIABLES
var c ;                          # Intercept
var beta {j in Nvar};            # Slopes
var nupos {j in Nvar}, binary;
var nuneg {j in Nvar}, binary;
var v {j in Nvar} >=0 ;          # Auxiliar variables to express the absolute value
#OBJECTIVE FUNCTION
minimize fun:   (1/N)*sum{ i in Nobs} (X [i,p+1] -c-sum{ j in Nvar} (beta[j]*X[i,j]))^ 2
                +(1/N)*lambda*sum{ j in Nvar} v[j];
#CONSTRAINTS

#SPARSITY CONSTRAINT
subject to sparsity:   sum{ j in Nvar} (nuneg[j]+nupos[j])<=NZ;
#SIGNIFICANCE CONSTRAINTS
subject to significancepos {j in Nvar}:   beta[j]>=eps*nupos[j]-nuneg[j]*M;
subject to significanceneg {j in Nvar}:   beta[j]<=-eps*nuneg[j]+nupos[j]*M;
#SIGN COHERENCE CONSTRAINTS
subject to coherencepos1 {(j,r) in conjcorrneg}:   nupos[j]+nupos[r]<=1;
subject to coherencepos2 {(j,r) in conjcorrneg}:   nuneg[j]+nuneg[r]<=1;
subject to coherenceneg1 {(j,r) in conjcorrpos}:   nupos[j]+nuneg[r]<=1;
subject to coherenceneg2 {(j,r) in conjcorrpos}:   nuneg[j]+nupos[r]<=1;
#AUXILIAR CONSTRAINTS
subject to abs1 {j in Nvar}:      v[j]>=beta[j];
subject to abs2 {j in Nvar}:      v[j]>=-beta[j];
subject to sumnusj {j in Nvar}:   nuneg[j]+nupos[j]<=1;
```