On linear regression models with hierarchical categorical variables

Preprint · April 2020

4 authors:

Emilio Carrizosa
Universidad de Sevilla
236 PUBLICATIONS **2,876** CITATIONS
SEE PROFILE

Laust Hvas Mortensen
Statistics Denmark
190 PUBLICATIONS **3,693** CITATIONS
SEE PROFILE

Dolores Romero Morales
Copenhagen Business School
94 PUBLICATIONS **1,566** CITATIONS
SEE PROFILE

M. Remedios Sillero-Denamiel
Universidad Pablo de Olavide
10 PUBLICATIONS **86** CITATIONS
SEE PROFILE

On linear regression models with hierarchical categorical variables

Emilio Carrizosa[1], Laust Hvas Mortensen[2], Dolores Romero Morales[3], and M. Remedios Sillero-Denamiel[1]

[1]Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain
{ecarrizosa, rsillero}@us.es

[2]Statistics Denmark and University of Copenhagen, Copenhagen, Denmark
LHM@dst.dk

[3]Copenhagen Business School, Frederiksberg, Denmark
drm.eco@cbs.dk

October 27, 2020

Abstract

In this paper, we study linear regression models built on categorical predictor variables that have a hierarchical structure. For such variables, the categories are arranged as a directed tree, where the categories in the leaf nodes give the highest granularity in the representation of the variable. Instead of taking the fully detailed model, the user can go upstream the tree and use a less complex, and thus more interpretable, model with fewer coefficients to be estimated and interpreted, hopefully without damaging the accuracy. We study the mathematical optimization problem that trades off the accuracy of the linear regression model and its complexity, measured as a cost function of the level of granularity of the representation of the hierarchical categorical variables. We show that finding non-dominated solutions for this problem boils down to solving a Mixed Integer Quadratic Problem with Linear Constraints. We illustrate our approach in a real-world cancer trial dataset, as well as in a simulated one, where our methodology finds a much less complex model with a very mild worsening of the accuracy.

Keywords: Machine Learning; Hierarchical Categorical Variables; Linear Regression Models; Accuracy vs. Model Complexity; Mixed Integer Quadratic Problem with Linear Con-

straints

1 Introduction

Interpreting and visualizing information extracted from complex data is at the core of Data Science [4, 9, 18, 27, 31, 39]. Mathematical Optimization is an important tool to build in an efficient manner data analysis models that can achieve a high accuracy [8, 15, 19, 20, 21], while being able to incorporate desirable properties such as being parsimonious, [3, 5, 6, 10, 14], or tackling multiple objectives, such as a *bias-variance tradeoff* [25]. In this paper, we study the mathematical optimization problem that trades off, in linear regression models, accuracy and model complexity, in the presence of categorical variables that have a hierarchical structure. In the literature, this kind of variable appears in different fields of research, such as nested spatial data in Spatial Statistics [22], behavioral data in Retail Business Analytics [23], or, economic activity in Official Statistics [17, 26].

In linear regression models, we are given a training sample of size n, and the response variable is to be predicted as a linear function of the predictor variables. Throughout this work, continuous as well as categorical predictors are considered, where the latter case is tackled by using dummy variables. In the numerical section, we use a real-life dataset concerning cancer trials, called `cancer-reg` [32], with individuals from the United States of America (U.S.). This dataset aims to look for relationships between the socioeconomic status in U.S. and the mean per capita cancer mortality (target variable). It has a sample of size $n = 3047$ with 32 predictor variables, where *income* (median income per capita binned by deciles) is one of the predictor variables having 10 categories. With the usual representation of categorical variables through dummies, all individuals in the same category of *income* share the same coefficient in the linear regression model.

In this paper, we are interested in linear regression models built on continuous predictors, dummy variables, as well as categorical predictor variables with a hierarchical structure. Let $\mathcal{J}'$ be the set of continuous and dummy predictor variables, whereas $\mathcal{J}$ the set of hierarchical categorical predictor variables. Then, consider the random vector $(\mathbf{X}', \mathbf{X}, Y)$, where $\mathbf{X}'$ denotes the vector of the predictor variables in $\mathcal{J}'$, $\mathbf{X}$ denotes the vector of categorical predictor variables in $\mathcal{J}$, and Y denotes the response variable. In the dataset `cancer-reg`, *geography* is a categorical variable with a hierarchical structure. According to the *U.S. Department of Commerce Economics and Statistics Administration* and the *U.S. Census Bureau*, *geography* can be coded using the states (51 in total), which is the highest level of granularity for which information is available in the data set. This means that 51 coefficients need to be estimated for this variable, where individuals in the same state share the same coefficient in the linear regression model. The variable *geography* can alternatively be coded using the subregions, such as *East-South Central*, *Middle Atlantic* and *New England*, where each state belongs to exactly one of the 9 subregions. Replacing the dummy variables associated with the states by the ones associated with subregions, yields a lower level of granularity for the variable *geography*, where, instead of 51, only 9 coefficients need to be estimated and interpreted. The subregions can be further combined into 4 regions, namely *West*, *South*, *Mid-West* and *North-East*, where only 4 coefficients would be associated to the *geography* variable in the linear regression model. Using these regions, one has the least granular representation of *geography*. This paper is devoted to trading off accuracy of the linear regression model and its complexity, measured as a cost function of the level of granularity used to represent each of the hierarchical categorical variables.

The categories of hierarchical categorical variable $j \in \mathcal{J}$, can be arranged as a directed tree $\mathcal{T}_j$, i.e., a directed graph with a root node, $r(\mathcal{T}_j)$, and a unique path from each node to $r(\mathcal{T}_j)$. In addition, let $\mathcal{V}(\mathcal{T}_j)$ denote the set of nodes in the tree and $\mathcal{L}(\mathcal{T}_j) \subset \mathcal{V}(\mathcal{T}_j)$ the set of leaf nodes. See Figure 1 for the tree associated with the categories of *geography*, where the leaf nodes correspond to the states, going upstream we find the subregions and then the regions, which, in turn, are directly connected with the root node of the tree. Let $(\mathbf{x}'_i, \mathbf{x}_i, y_i)$ be the vector associated with individual i, with $\mathbf{x}'_i = (x'_{ij'})$ and $\mathbf{x}_i = (x_{ijv})$, where x_{ijv} is equal to 1 if individual i belongs to category $v \in \mathcal{V}(\mathcal{T}_j)$ of variable $j \in \mathcal{J}$. If we were to use the most granular representation of the hierarchical categorical variables, we would need to use the categories associated with the leaf nodes $l \in \mathcal{L}(\mathcal{T}_j)$, i.e.,

$$y_i = \beta'_0 + \sum_{j' \in \mathcal{J}'} \beta'_{j'} x'_{ij'} + \sum_{j \in \mathcal{J}} \sum_{l \in \mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}, \qquad (1)$$

where β'_0 is the independent term, $\beta'_{j'}$ is the coefficient of variable $j' \in \mathcal{J}'$, whereas β_{jl} is the coefficient of category $l \in \mathcal{L}(\mathcal{T}_j)$ of hierarchical categorical variable $j \in \mathcal{J}$. In the ordinary least squares (OLS) paradigm, the coefficients are obtained by minimizing the mean squared error

(MSE). The corresponding OLS model reads as follows

$$\mathrm{MSE}^*((\mathcal{T}_j)_{j\in\mathcal{J}}) = \min_{\beta'_0,(\beta'_{j'})_{j'\in\mathcal{J}'},(\beta_{jl})_{l\in\mathcal{L}(\mathcal{T}_j),j\in\mathcal{J}}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'}\beta'_{j'}x'_{ij'} - \sum_{j\in\mathcal{J}}\sum_{l\in\mathcal{L}(\mathcal{T}_j)}\beta_{jl}x_{ijl})^2. \quad (2)$$

In the dataset `cancer-reg` used in the numerical section, we have 51 coefficients associated with the hierarchical categorical variable *geography* in (1). These dummies together with the rest of variables yield an in-sample MSE of 0.4065. The question arises as to whether that level of granularity is necessary, or whether we can merge categories at the bottom of the tree into a broader category upstream in the tree. In the dataset `cancer-reg`, we can eliminate the state information for all the individuals of same subregion, respectively from the same region, and report the subregion, respectively the region. We have done this for the states in the subregions *Middle Atlantic* and *New England*, yielding the subtree in Figure 3 (a) of the tree in Figure 1, where the node *Middle Atlantic* receives all individuals in its descendants and, therefore, they share the same coefficient in the linear regression model. The same occurs for *New England* node. With this representation, the in-sample MSE increases from 0.4065 to 0.408. This mild worsening in accuracy corresponds to a reduction from 51 to 44 in the number of coefficients to be estimated and interpreted for the *geography* variable.

Reducing the granularity of the representation of hierarchical categorical variables has several advantages. First, and as illustrated above, it is a step towards enhancing the interpretability of the linear regression model, where fewer coefficients need to be estimated and interpreted [13]. Second, if the samples of individuals associated with categories are homogeneous enough, a very granular representation would yield an overparametrized model. Instead, we could merge these categories into a broader one upstream the tree, thus having more observations to estimate fewer coefficients. The homogeneity together with the increase in sample size ensure lower errors in the estimation of the coefficients of the broader categories [28]. Third, and again if the samples of individuals associated with categories are homogeneous enough, a very granular representation will yield higher data gathering costs [11, 38], if, for instance, the surveying costs are asymmetric. Indeed, we would need to ensure a large enough sample for each category in the representation, even though the cost of surveying may be high for some of these categories. By merging homogeneous categories into a broader one upstream the tree, we can sample from a larger subpopulation lowering the data gathering costs. Finally, it is important when having data privacy considerations, [29, 30], since it is well-known that more detailed information is linked to confidentiality concerns [2].

In this paper, we study the mathematical optimization problem that trades off accuracy of the linear regression model and its complexity, by choosing the granularity of the representation of the hierarchical categorical variables. This boils down to choosing for each hierarchical categorical variable $j \in \mathcal{J}$, a subtree $\mathcal{S}_j$ of $\mathcal{T}_j$, with the same root as $\mathcal{T}_j$, that is, $r(\mathcal{S}_j) = r(\mathcal{T}_j)$. While the accuracy of the reduced linear regression model will be measured by its MSE, its complexity will be measured by

$$\mathrm{C}((\mathcal{S}_j)_{j\in\mathcal{J}}) = \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{S}_j)} c_{jl}, \tag{3}$$

where c_{jv} represents the cost associated to node $v \in \mathcal{V}(\mathcal{T}_j)$. With this, our problem reads as follows:

$$\min_{(\mathcal{S}_j)_{j\in\mathcal{J}}} \ (\mathrm{MSE}^*((\mathcal{S}_j)_{j\in\mathcal{J}}), \mathrm{C}((\mathcal{S}_j)_{j\in\mathcal{J}})), \tag{4}$$

where $\mathrm{MSE}^*((\mathcal{S}_j)_{j\in\mathcal{J}})$ is defined as in (2) with the leaf nodes of $\mathcal{S}_j$, $\mathcal{L}(\mathcal{S}_j)$, replacing $\mathcal{L}(\mathcal{T}_j)$. Note that Problem (4) performs akin to the pruning of a regression tree [34, 36], but here we have one tree per hierarchical categorical predictor in the dataset. In fact, to trade off the accuracy and the complexity of the linear regression model, we perform the pruning of all trees at simultaneously.

The remainder of this paper is structured as follows. In Section 2, we study the *constrained* problem, yielding a non-dominated solution to Problem (4), in which we minimize the mean squared error of the reduced model subject to a complexity constraint, where a threshold is imposed on the cost of granularity. For this problem, we provide a Mixed Integer Quadratic Problem with Linear Constraints formulation. Section 3 illustrates our approach in the `cancer-reg` dataset, as well as in a simulated one, where the entire set of non-dominated solutions to Problem (4) is obtained solving the constrained problem for the different values of the threshold. To end, some conclusions and lines for future research are provided in Section 4.

2 The constrained problem

To obtain a non-dominated solution to Problem (4), we propose the following constrained problem

$$\begin{aligned} \min_{(\mathcal{S}_j)_{j\in\mathcal{J}}} \quad & \mathrm{MSE}^*((\mathcal{S}_j)_{j\in\mathcal{J}}) \\ \text{s.t.} \quad & \mathrm{C}((\mathcal{S}_j)_{j\in\mathcal{J}}) \leq c, \end{aligned} \tag{5}$$

where c is a threshold on the complexity of the model defined by (3), in which the hierarchical categorical variables are represented by the leaf nodes of $\mathcal{S}_j$, $\mathcal{L}(\mathcal{S}_j)$. In this section, we formulate this constrained problem as a Mixed Integer Quadratic Problem with Linear Constraints, and discuss the choice of values for threshold c.

We first need to introduce some notation. Recall that given a tree $\mathcal{T}$, $\mathcal{V}(\mathcal{T})$ denotes the set of nodes in the tree, $\mathcal{L}(\mathcal{T})$ the set of leaf nodes and $r(\mathcal{T})$ the root node. Given $j \in \mathcal{J}$ and $l \in \mathcal{L}(\mathcal{T}_j)$, let $\mathcal{P}_{jl}$ be the categories associated with the unique path in $\mathcal{T}_j$ from its root node $r(\mathcal{T}_j)$ to l, and let $\mathcal{P}(\mathcal{T}_j)$ be the collection of all these paths

$$\mathcal{P}(\mathcal{T}_j) = \{\mathcal{P}_{jl} \mid l \in \mathcal{L}(\mathcal{T}_j)\}.$$

The goal is to choose for each hierarchical categorical variable $j \in \mathcal{J}$, a subtree $\mathcal{S}_j$ of $\mathcal{T}_j$ with the same root as $\mathcal{T}_j$. This is equivalent to choosing a subset of nodes of $\mathcal{T}_j$ such that only one node per path $\mathcal{P}_{jl}$ is included. This subset of nodes will define the leaf nodes of $\mathcal{S}_j$, $\mathcal{L}(\mathcal{S}_j)$. We introduce $\mathbf{z} = (z_{jv})$, such that $z_{jv} = 1$ if the node associated with category v of the hierarchical categorical variable j is selected, and $z_{jv} = 0$ otherwise. If the node v is selected, it will receive all individuals in its descendant nodes and they will share the same coefficient in the reduced regression model. Therefore, we solve the following Mixed Integer Quadratic Problem with Linear Constraints:

$$\min_{\mathbf{z},\beta'_0,(\beta'_{j'})_{j'\in\mathcal{J}'},(\beta_{jv})_{v\in\mathcal{V}(\mathcal{T}_j),j\in\mathcal{J}}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} x'_{ij'}\beta'_{j'} - \sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} z_{jv}x_{ijv}\beta_{jv})^2 \tag{6}$$

$$\text{s.t.} \quad \sum_{v\in\mathcal{P}_{jl}} z_{jv} = 1, \;\; l \in \mathcal{L}(\mathcal{T}_j), \;\; j \in \mathcal{J}, \tag{7}$$

$$\sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} c_{jv}z_{jv} \leq c, \tag{8}$$

$$z_{jv} \in \{0,1\}, \;\; \forall v \in \mathcal{V}(\mathcal{T}_j), \;\; j \in \mathcal{J}, \tag{9}$$

$$\beta'_0,\, \beta'_{j'},\, \beta_{jv} \in \mathbb{R},\, \forall j' \in \mathcal{J}',\, \forall v \in \mathcal{V}(\mathcal{T}_j),\;\; j \in \mathcal{J}, \tag{10}$$

where the linear constraints (7) imply that only one node is selected per path. Constraint (8) imposes the threshold c on the complexity of the models where predictor variable $j \in \mathcal{J}$ is represented with categories v, for which the associated costs are $c_{jv}z_{jv}$. Constraints (9) and (10) ensure that the decision variables are well defined. Finally, the objective function is the MSE of linear models, where semi-continuous variables $z_{jv}x_{ijv}$ are the observed values for all i.

Since the objective function (6) has semi-continuous variables, $z_{jv}\beta_{jv}$, a smooth formulation can be obtained using big M constraints:

$$\begin{aligned}
\min_{\mathbf{z},\beta'_0,(\beta'_{j'})_{j'\in\mathcal{J}'},(\beta_{jv})_{v\in\mathcal{V}(\mathcal{T}_j),j\in\mathcal{J}}} \quad & \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} x'_{ij'}\beta'_{j'} - \sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} x_{ijv}\tilde{\beta}_{jv})^2 && (11)\\
\text{s.t.} \quad & \sum_{v\in\mathcal{P}_{jl}} z_{jv} = 1,\;\; l \in \mathcal{L}(\mathcal{T}_j),\;\; j \in \mathcal{J}, && (12)\\
& \sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} c_{jv}z_{jv} \le c, && (13)\\
& -Mz_{jv} \le \tilde{\beta}_{jv} \le Mz_{jv},\;\; \forall v \in \mathcal{V}(\mathcal{T}_j),\;\; j \in \mathcal{J}, && (14)\\
& z_{jv} \in \{0,1\},\;\; \forall v \in \mathcal{V}(\mathcal{T}_j),\;\; j \in \mathcal{J}, && (15)\\
& \beta'_0,\, \beta'_{j'},\, \tilde{\beta}_{jv} \in \mathbb{R},\, \forall j' \in \mathcal{J}',\, \forall v \in \mathcal{V}(\mathcal{T}_j),\;\; j \in \mathcal{J}. && (16)
\end{aligned}$$

This is the formulation that will be used in the numerical section. Note that we can sharpen the value of M by imposing an upper bound on the coefficients of the categories of hierarchical variables. This can be seen as a regularization, thus preventing overfitting and allowing for sparser models [12]. Other types of regularization can be easily incorporated into our model, such as those in [35, 40].

We now discuss the choice of values for threshold c. It is easy to show that if c_{jv} are integer numbers, the threshold c is also integer. Moreover, it is easy to define lower ($c^{\min} := |\mathcal{J}|$) and upper ($c^{\max} := \mathrm{C}((\mathcal{T}_j)_{j\in\mathcal{J}})$) bounds on its value. By varying the threshold value c among this finite set of values it can take, we obtain the entire set of non-dominated solutions to Problem (4). In addition, as we will see in Section 3, the set of non-dominated solutions contains those associated with some well-known goodness-of-fit metrics in the literature, the AIC [1] and BIC [33].

Likewise, Problem (4) can be modeled by controlling the MSE while the complexity of the model is optimized:

$$
\begin{aligned}
\min_{(\mathcal{S}_j)_{j\in\mathcal{J}}} \quad & \mathrm{C}((\mathcal{S}_j)_{j\in\mathcal{J}}) \\
\text{s.t.} \quad & \mathrm{MSE}^*((\mathcal{S}_j)_{j\in\mathcal{J}}) \leq f,
\end{aligned} \tag{17}
$$

where f is the threshold value on the MSE of the reduced model. The advantage of constraining the $\mathrm{MSE}^*((\mathcal{S}_j)_{j\in\mathcal{J}})$ is to have a full control on the accuracy of the model and to allow the user to define meaningful values of f, [7]. Therefore, this option is recommended when Problem (4) is going to be solved once, where using a similar formulation to (11)-(16). As seen in the introduction, a lower bound on f is

$$
f^{\min} := \mathrm{MSE}^*((\mathcal{T}_j)_{j\in\mathcal{J}}), \tag{18}
$$

which is the MSE that we achieve for the highest level of granularity. An upper bound on f is found by removing all the variables $j \in \mathcal{J}$. This corresponds to

$$
f^{\max} := \min_{\beta'_0,(\beta'_{j'})_{j'\in\mathcal{J}'}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'}\beta'_{j'}x'_{ij'})^2 \tag{19}
$$

where we consider the subtree with only the root node, i.e., $\mathcal{S}_j = \{r(\mathcal{T}_j)\}\ \forall j \in \mathcal{J}$. In this case, by varying the threshold value f in a thin grid in $[f^{\min}, f^{\max}]$, we also obtain a collection of non-dominated solutions to Problem (4).

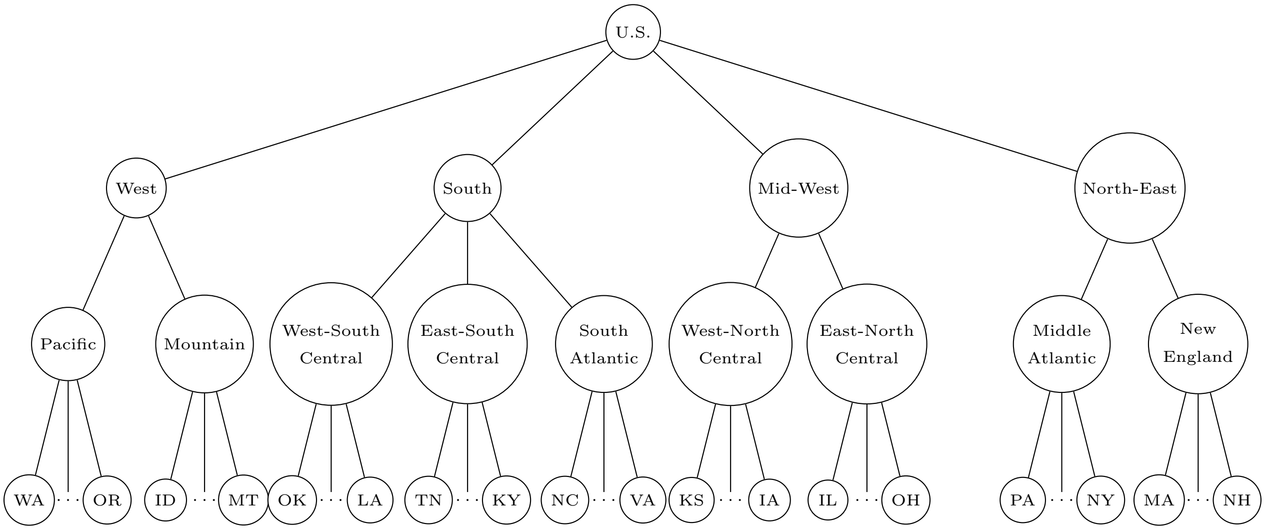
3 Numerical experiments

In this section, we illustrate our approach using both real-life and simulated datasets. Our aim is to depict the tradeoff between the accuracy of the reduced model and its complexity, measured by the number of coefficients to be estimated for the hierarchical categorical variables, which corresponds to $c_{jv} = 1$ in $\mathrm{C}((\mathcal{S}_j)_{j\in\mathcal{J}})$. Throughout this section, we assume that continuous predictors have been standardized. To solve Problem (11)-(16) we use Gurobi [24], where M is set to 1000. The experiments have been run on Intel(R) Core(TM) i7-7500U CPU at 2.70 GHz 2.90 GHz with 8.0 GB of RAM.

3.1 Cancer trials dataset: a real-life dataset

Consider again the real-life dataset `cancer-reg` introduced in Section 1. It has one hierarchical predictor variable ($|\mathcal{J}| = 1$) and 31 non-hierarchical predictor variables ($|\mathcal{J}'| = 31$). This

Figure 1: Tree representation of the variable *geography* in the `cancer-reg` dataset



database was collected from the American Community Survey (`census.gov`), `clinicaltrials.gov` and `cancer.gov` sources. As commented in Section 1, the hierarchical categorical variable is *geography*, which contains the state name, and can be coded according to Figure 1.

We solve Problem (11)-(16) for 51 values of c in the set $\{1, \ldots, 51\}$. Figure 2 reports the pareto frontier for the MSE and the number of coefficients to be estimated in the reduced model for the hierarchical categorical variables. Figure 3 plots the selected subtree $\mathcal{S}_1^*$ for three of the solutions in Figure 2. In particular, Figure 2 (a) is the representation associated with the AIC metric, whereas Figure 2 (c) the BIC one.

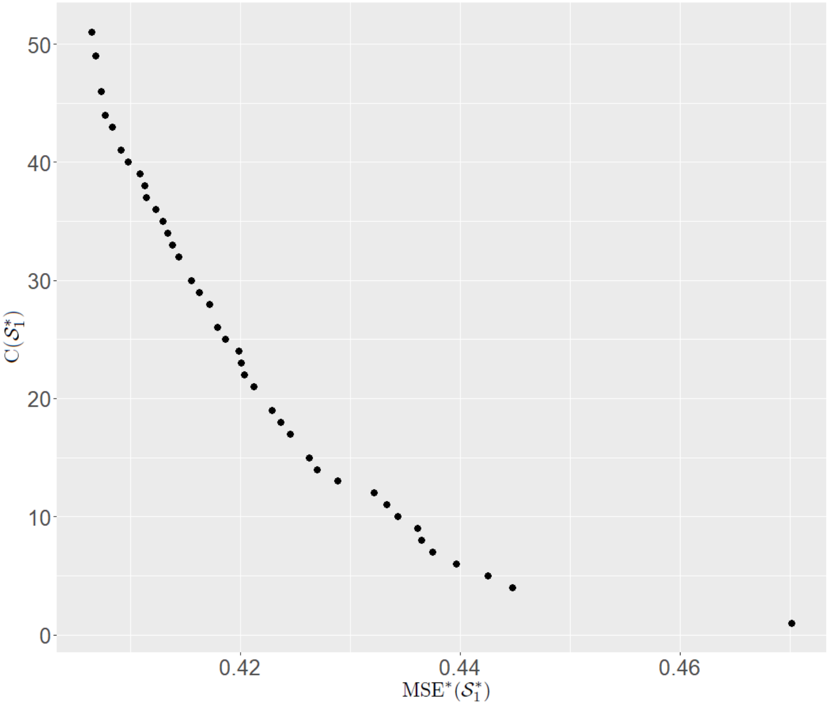
3.2 The simulated data

In this section we illustrate our approach on simulated data. The data generating model is

$$y_i = \sum_{j \in \mathcal{J}} \sum_{l \in \mathcal{L}(\mathcal{T}_j)} \beta_{jl}^{\mathrm{S}} x_{ijl} + \varepsilon_i, \; i = 1, \ldots, n, \tag{20}$$

where $|\mathcal{J}| = 2$ and $\mathcal{J}' = \emptyset$. The values of the coefficients β_{jl}^{S}, $l \in \mathcal{L}(\mathcal{T}_j)$, are given in Figure 4. Note that the first two leaf nodes of $\mathcal{T}_1$ have the same coefficient, and the same holds for the other two leaf nodes. Therefore, the tree can be pruned to avoid unnecessary splits, yielding the subtree in Figure 5. The same holds for $\mathcal{T}_2$. The error is taken $\varepsilon_i \sim N(0, \sigma^2)$ for different values of σ^2 given below. We have $n = 3000$ individuals, evenly distributed across the different

Figure 2: Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model for the hierarchical categorical variable *geography*

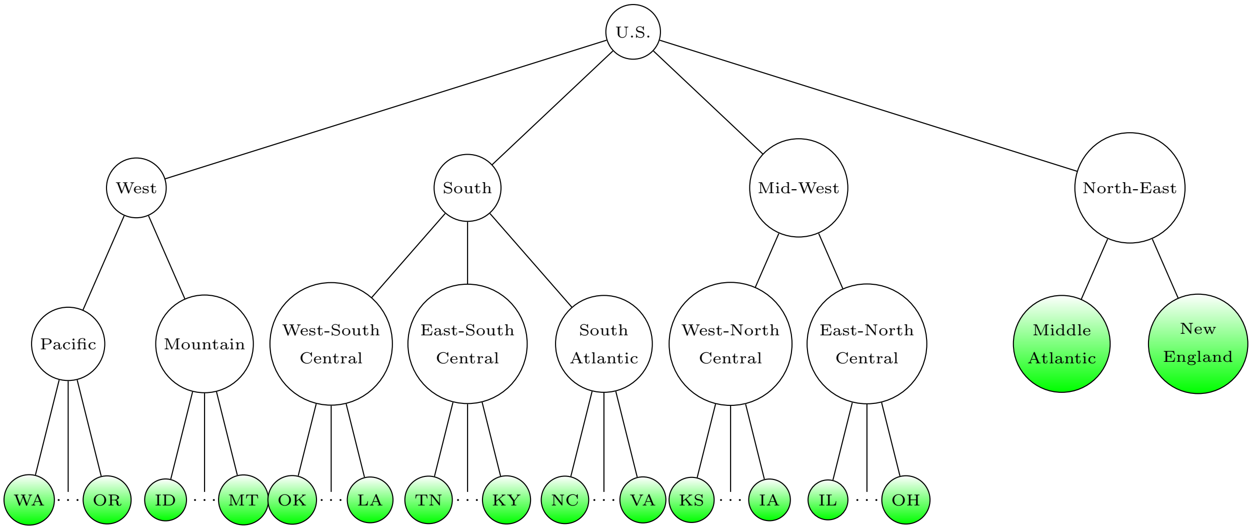


combinations of categories $l_1 \in \mathcal{L}(\mathcal{T}_1)$ and $l_2 \in \mathcal{L}(\mathcal{T}_2)$. The purpose of this section is twofold. First, we illustrate how our approach is able to recover the pruned tree underlying our simulated data. Second, we carry out an *out-of-sample* study.

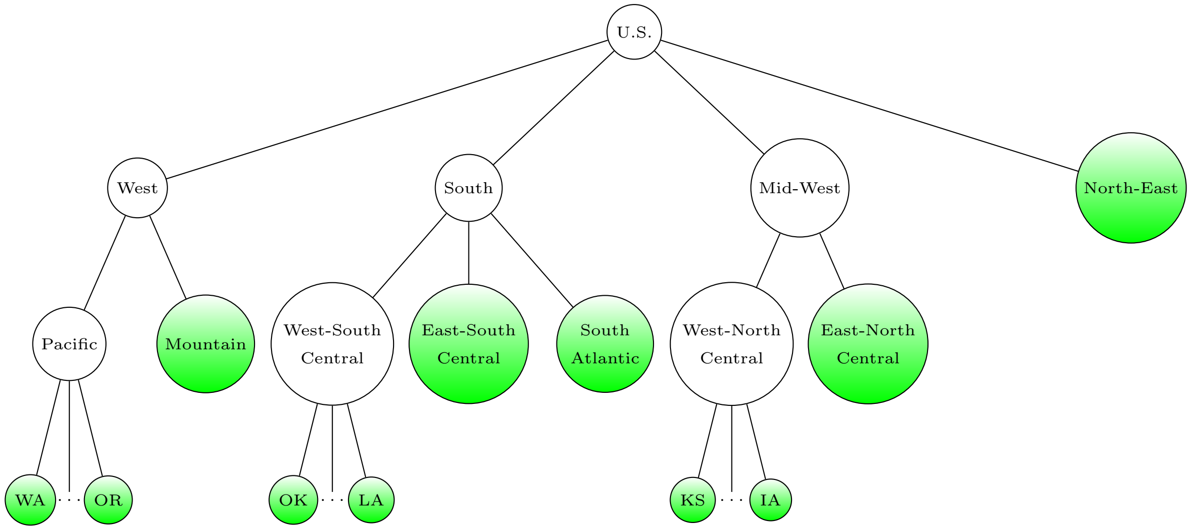
Let us consider $\sigma^2 = 0.04$ and solve Problem (11)-(16) for 10 values of c in the set $\{2, \dots, 11\}$. Figure 6 shows the pareto frontier for the number of coefficients to be estimated in the reduced model versus the MSE. For small values of MSE, the chosen nodes are the 8 green leaf nodes in Figure 5, which implies that our methodology is able to successfully detect the pruned tree underlying each hierarchical categorical variable in our data. Similar conclusions can be drawn when $\sigma^2 = 0.16$ and $\sigma^2 = 0.36$.

To end the numerical section, we provide an estimation for the MSE and the complexity of the reduced model using a 10-fold cross validation approach, showing that our procedure works properly with the available (*in-sample*) individuals, but also for future (*out-of-sample*) individuals. For each fold, the *in-sample* set is used to solve Problem (11)-(16) and get $\mathcal{S}_j^*$, $j \in \mathcal{J}$. Once the subtrees are found, and thus the reduced linear regression model, we calculate

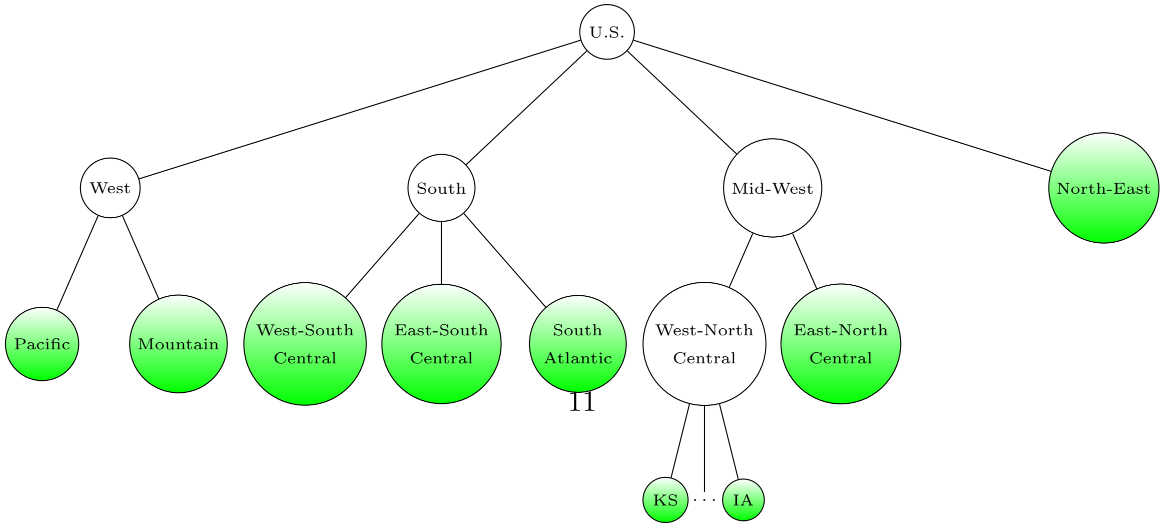
Figure 3: Less granular representations for *geography* variable in the `cancer-reg` dataset



(a) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.408$ and $c = 44$

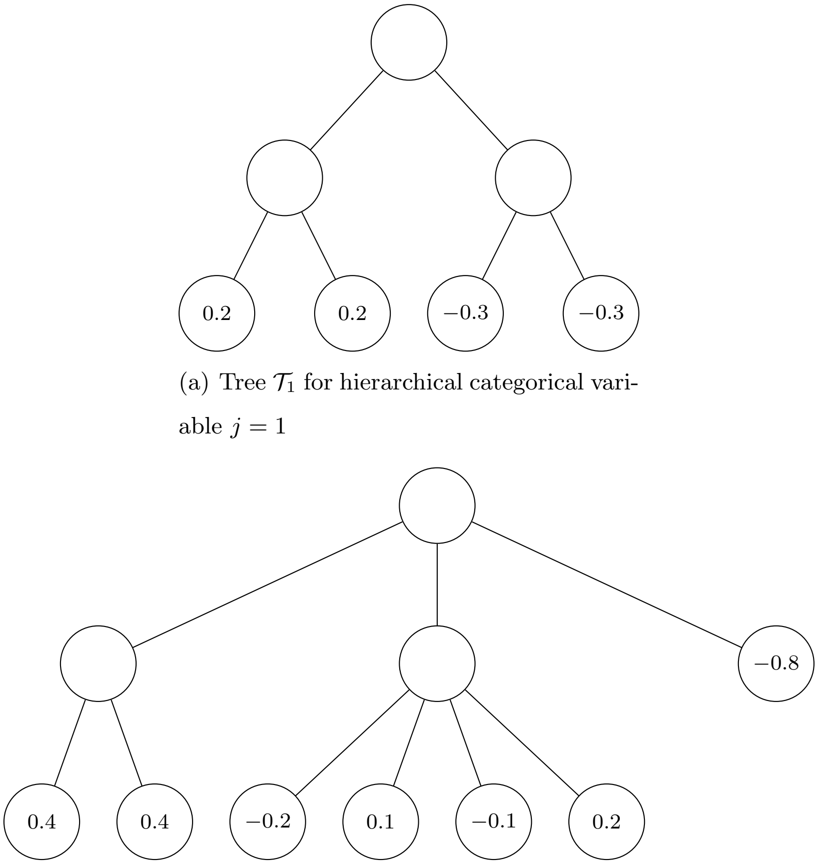


(b) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.421$ and $c = 21$



(c) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.427$ and $c = 14$

Figure 4: Trees associated with the two hierarchical categorical variables together with β^{S}_{jl}, $l \in \mathcal{L}(\mathcal{T}_j)$



(a) Tree $\mathcal{T}_1$ for hierarchical categorical variable $j = 1$

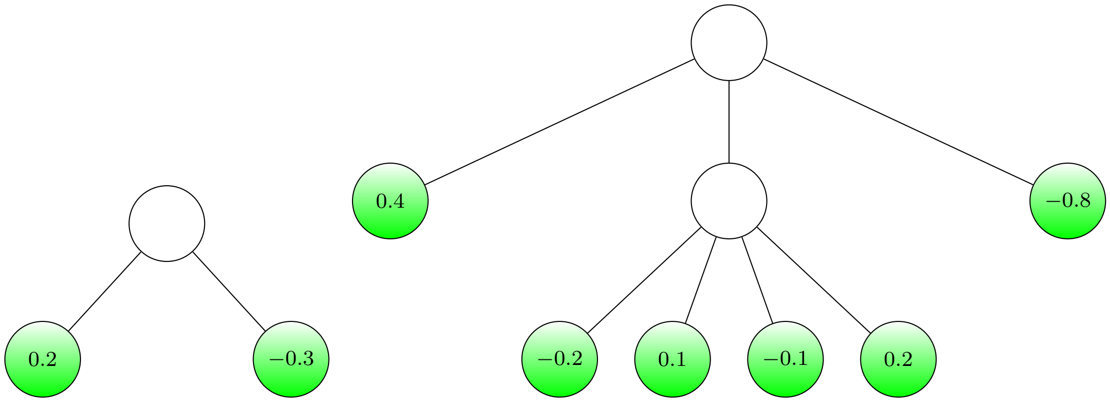
(b) Tree $\mathcal{T}_2$ for hierarchical categorical variable $j = 2$

its *in-sample* and *out-of-sample* MSE, which are plotted in Figures 9-11 for the different values of σ^2. As can be observed, the *in-sample* MSE values (red lines) are only slightly smaller than the *out-of-sample* values (blue lines). Then, in view of results, we can conclude that our methodology generalizes well.

4 Conclusions and extensions

In this paper a new methodology to deal with hierarchical categorical variables, i.e., categorical variables that can be measured at different levels of granularity, has been developed. Through a Mixed Integer Quadratic Problem with Linear Constraints, we study the tradeoff between

Figure 5: Pruned tree and less granular representation of the two hierarchical categorical variables of Figure 4



(a) Pruned tree for the first hierarchical categorical variable

(b) Pruned tree for the second hierarchical categorical variable

Figure 6: Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model when $\sigma^2 = 0.04$

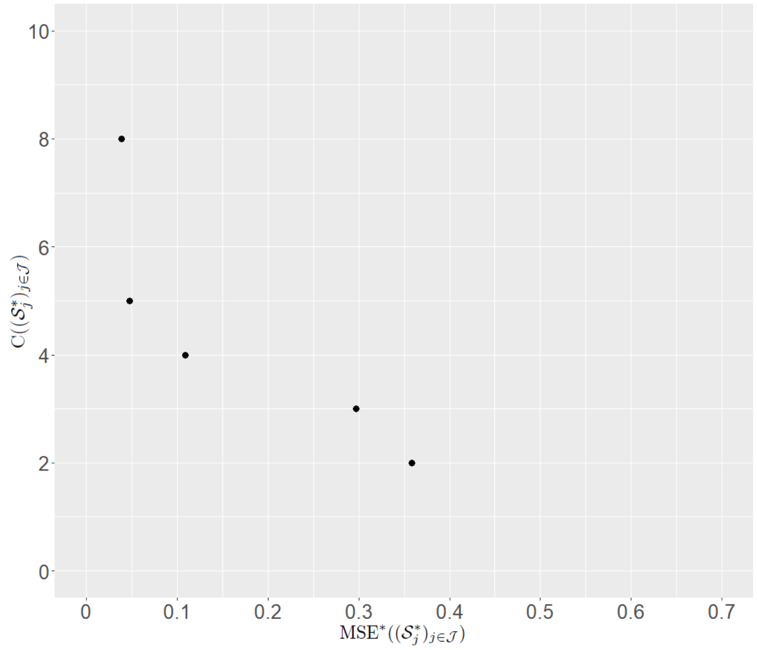
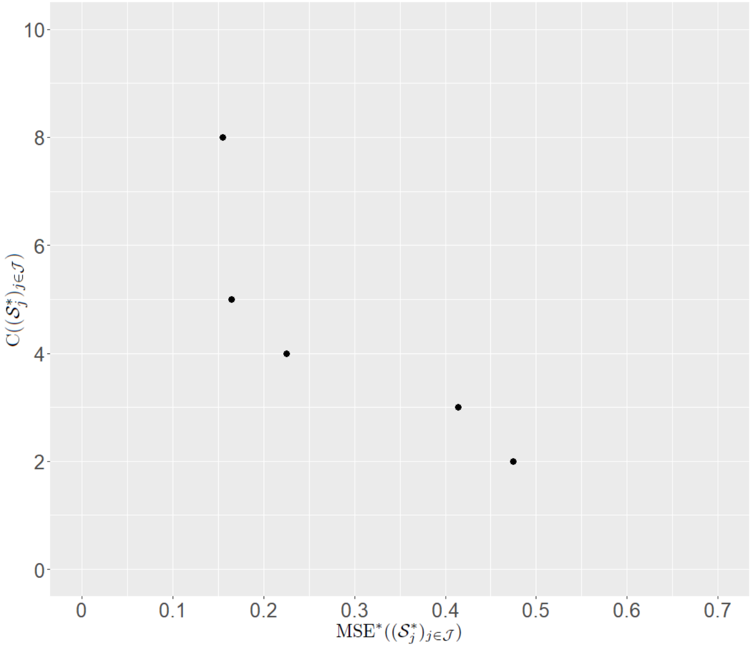


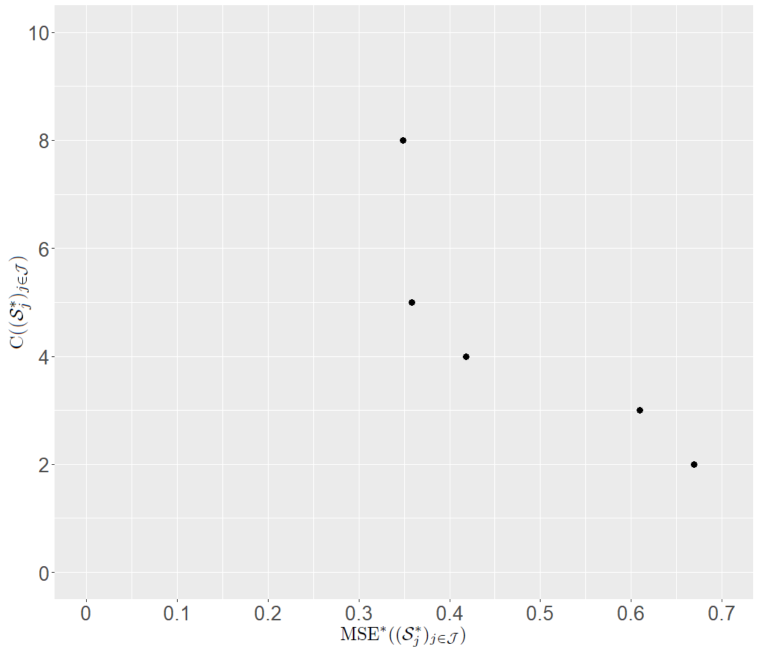
Figure 7: Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model when $\sigma^2 = 0.16$



accuracy and model complexity. Our proposal has been tested on both synthetic and real-life datasets. The numerical section shows that much less granular representations for the hierarchical categorical variables can be found at the expense of slightly damaging the accuracy.

A number of extensions to this work are are worth investigating. Firstly, when the number of categories is large, instead of solving Problem (11)-(16) considering all the categories at once, a sequential pruning can be used instead. The main idea is to consider subtrees in $\mathcal{T}_j$ and try to compress their categories solving Problem (11)-(16) sequentially. Another option to deal with large number of categories is to cluster them based on a dissimilarity, see [13, 16] and references therein. Secondly, our methodology can be extended to generalized linear models [37], where, instead of predicting the response variable as in (1), a non-linear relationship between the response variable and the predictors is through a linkage function. However, this extension makes the problem highly nonlinear and its resolution is very challenging and outside the scope of this paper.

Figure 8: Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model when $\sigma^2 = 0.36$



Acknowledgements

This research is partially supported by research grants and projects MTM2015-65915-R (Ministerio de Economía y Competitividad, Spain), FQM-329 and P18-FR-2369 (Junta de Andalucía, Spain), and EC H2020 MSCA RISE NeEDS Project (Grant agreement ID: 822214). This support is gratefully acknowledged.

References

[1] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.

[2] D. Baena, J. Castro, and A. Frangioni. Stabilized benders methods for large-scale combinatorial optimization, with application to data privacy. Forthcoming in *Management Science*, 2020. https://doi.org/10.1287/mnsc.2019.3341.

[3] S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo. Cost-sensitive Feature

Figure 9: Average MSE (10-fold CV) versus the imposed threshold c when $\sigma^2 = 0.04$

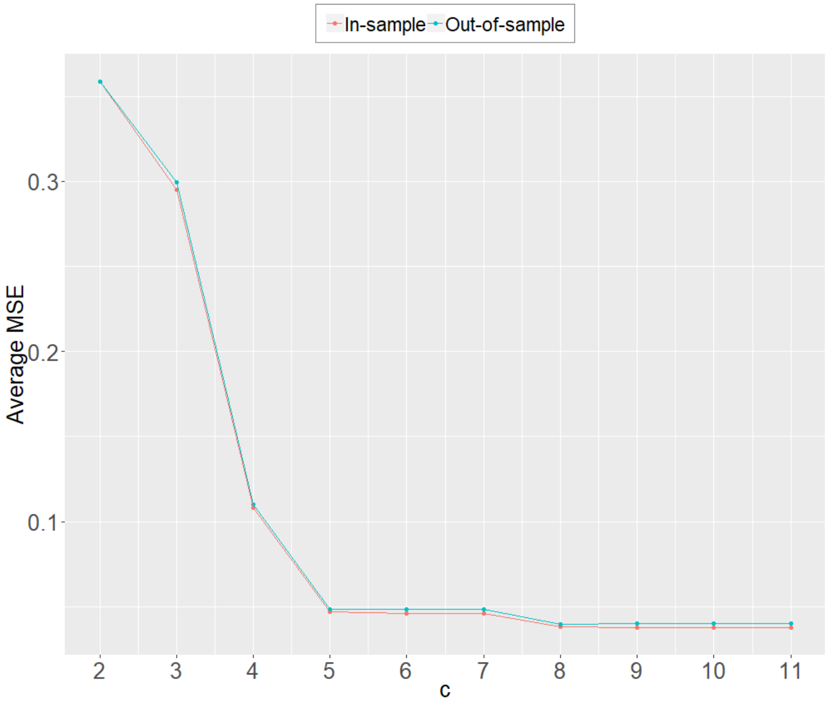


Figure 10: Average MSE (10-fold CV) versus the imposed threshold c when $\sigma^2 = 0.16$

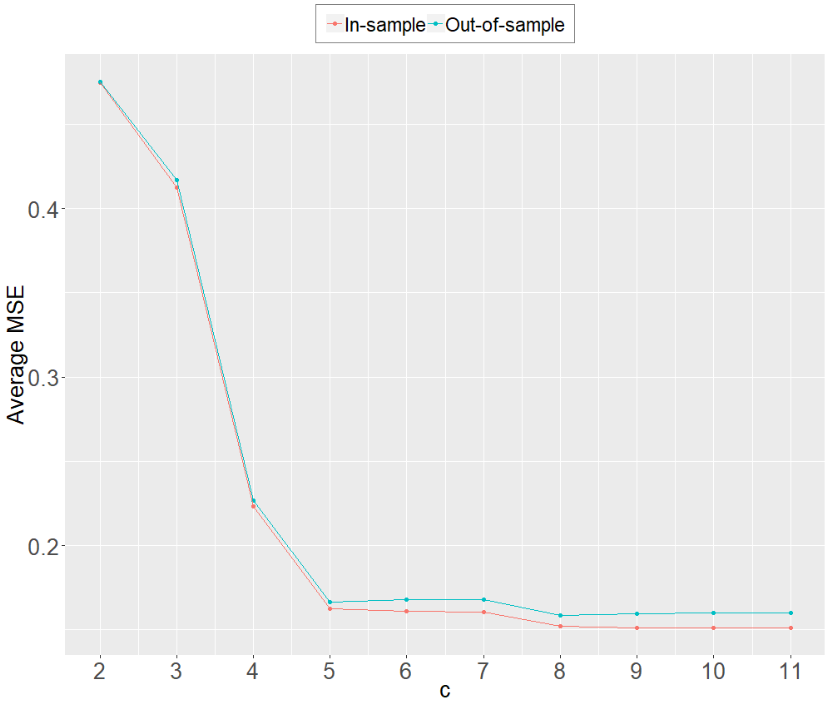
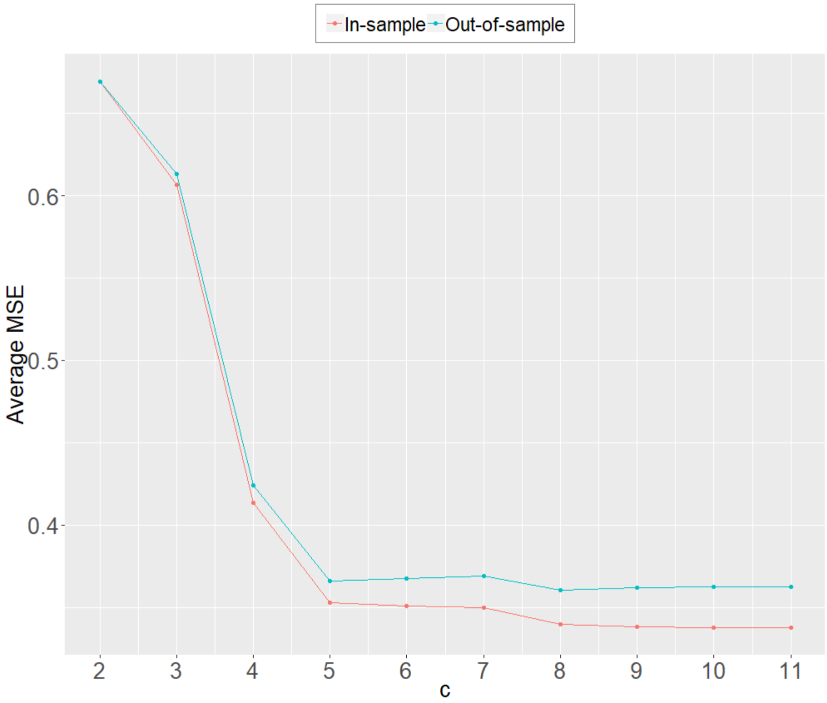


Figure 11: Average MSE (10-fold CV) versus the imposed threshold c when $\sigma^2 = 0.36$



Selection for Support Vector Machines. *Computers & Operations Research*, 106:169 – 178, 2019.

[4] D. Bertsimas, A. O'Hair, S. Relyea, and J. Silberholz. An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer. *Management Science*, 62(5):1511–1531, 2016.

[5] R. Blanquero, E. Carrizosa, A. Jiménez-Cordero, and B. Martín-Barragán. Variable selection in classification for multivariate functional data. *Information Sciences*, 481:445 – 462, 2019.

[6] R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, 284(1):255 – 272, 2020.

[7] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel. A cost-sensitive

constrained lasso. Forthcoming in *Advances in Data Analysis and Classification*, 2020. https://doi.org/10.1007/s11634-020-00389-5.

[8] L. Bottou, F. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018.

[9] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming, Series B*, 169:119–140, 2018.

[10] E. Carrizosa, V. Guerrero, D. Romero Morales, and A. Satorra. Enhancing Interpretability in Factor Analysis by Means of Mathematical Optimization. Forthcoming in *Multivariate Behavioral Research*, 2019. https://doi.org/10.1080/00273171.2019.1677208.

[11] E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Multi-group support vector machines with measurement costs: A biobjective approach. *Discrete Applied Mathematics*, 156:950–966, 2008.

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

[13] E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Clustering categories in support vector machines. *Omega*, 66:28 – 37, 2017.

[14] E. Carrizosa, A. V. Olivares-Nadal, and P. Ramírez-Cobo. A sparsity-controlled vector autoregressive model. *Biostatistics*, 18(2):244–259, 2016.

[15] E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers and Operations Research*, 40(1):150–165, 2013.

[16] P. Cerda, G. Varoquaux, and B. Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8):1477–1494, 2018.

[17] European Commission. *NACE Rev. 2 Statistical classification of economic activites in the European Community.* Luxembourg: Office for Official Publications of the European

Communities, 2008. https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.

[18] X. Fang, O. R. Liu Sheng, and P. Goes. When is the right time to refresh knowledge discovered from data? *Operations Research*, 61(1):32–44, 2013.

[19] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

[20] Z. Fu, B. Golden, S. Lele, S. Raghavan, and E. Wasil. Genetically engineered decision trees: Population diversity produces smarter trees. *Operations Research*, 51(6):894–907, 2003.

[21] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[22] C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

[23] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos. Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100:1 – 16, 2018.

[24] L. Gurobi Optimization. Gurobi optimizer reference manual, 2018.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[26] T. Katz-Gerro and J. López Sintas. Mapping circular economy activities in the European Union: Patterns of implementation and their correlates in small and medium-sized enterprises. *Business Strategy and the Environment*, 28(4):485–496, 2019.

[27] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017.

[28] M. LeBlanc and R. Tibshirani. Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, 7(4):417–433, 1998.

[29] X.-B. Li and S. Sarkar. Against classification attacks: A decision tree pruning approach to privacy protection in data mining. *Operations Research*, 57(6):1496–1509, 2009.

[30] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao. Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4):46–50, 2014.

[31] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466 – 1476, 2007.

[32] N. Rippner. Cancer Trials, 2017. Retrieved from http://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv.

[33] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[34] H. Sherali, A. Hobeika, and C. Jeenanunta. An optimal constrained pruning strategy for decision trees. *INFORMS Journal on Computing*, 21(1):49–61, 2009.

[35] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[36] X. Su, M. Wang, and J. Fan. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13(3):586–598, 2004.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[38] P. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

[39] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[40] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.