stats

MDPI

Article

Robust Variable Selection with Optimality Guarantees for High-Dimensional Logistic Regression

Luca Insolia [1,2,*], Ana Kenney [3], Martina Calovi [4] and Francesca Chiaromonte [2,3]

[1] Faculty of Sciences, Scuola Normale Superiore, 56126 Pisa, Italy
[2] Institute of Economics & EMbeDS, Sant'Anna School of Advanced Studies, 56127 Pisa, Italy; fxc11@psu.edu
[3] Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA; ajk5910@psu.edu
[4] Department of Geography, Norwegian University of Science and Technology, 7491 Trondheim, Norway; martina.calovi@ntnu.no
* Correspondence: luca.insolia@santannapisa.it

Abstract: High-dimensional classification studies have become widespread across various domains. The large dimensionality, coupled with the possible presence of data contamination, motivates the use of robust, sparse estimation methods to improve model interpretability and ensure the majority of observations agree with the underlying parametric model. In this study, we propose a robust and sparse estimator for logistic regression models, which simultaneously tackles the presence of outliers and/or irrelevant features. Specifically, we propose the use of L_0-constraints and mixed-integer conic programming techniques to solve the underlying double combinatorial problem in a framework that allows one to pursue optimality guarantees. We use our proposal to investigate the main drivers of honey bee (*Apis mellifera*) loss through the annual winter loss survey data collected by the Pennsylvania State Beekeepers Association. Previous studies mainly focused on predictive performance, however our approach produces a more interpretable classification model and provides evidence for several outlying observations within the survey data. We compare our proposal with existing heuristic methods and non-robust procedures, demonstrating its effectiveness. In addition to the application to honey bee loss, we present a simulation study where our proposal outperforms other methods across most performance measures and settings.

Keywords: classification; logistic slippage model; mixed-integer conic programming; model selection; honey bee loss; outlier detection; robust estimation

Citation: Insolia, L.; Kenney, A.; Calovi, M.; Chiaromonte, F. Robust Variable Selection with Optimality Guarantees for High-Dimensional Logistic Regression. *Stats* **2021**, *4*, 665–681. https://doi.org/10.3390/stats4030040

Academic Editor: Wei Zhu

Received: 29 July 2021
Accepted: 26 August 2021
Published: 31 August 2021

1. Introduction

Logistic regression is widely used to solve classification tasks and provides a probabilistic relation between a set of covariates (i.e., features, variables or predictors) and a binary or multi-class response [1,2]. The use of the logistic function can be traced back to the early 19th century, when it was employed to describe population growth [3]. However, despite its popularity, the classical logistic regression framework based on maximum likelihood (ML) estimation can suffer from several drawbacks. In this work, we specifically focus on two key challenges: high dimensionality and data contamination. The large dimensionality might lead to overfitting or even singularity of the estimates if the sample size is smaller than the number of features, and this motivates the use of penalized estimation techniques. Importantly, penalized methods can also promote sparsity of the estimates in order to improve the interpretability of the model [4]. On the other hand, the presence of outliers might disrupt classical and non-robust estimation methods leading to biased estimates and poor predictions. In particular, since the log-odds ratio depends linearly on the set of covariates included in the model, an *adversarial contamination* of the latter might create *bad leverage values* that break down ML-based approaches [5]. This motivates the development of robust estimation techniques. Notably, penalized estimation

and robustness with respect to the presence of outliers are very closely related topics [6–8], and they have recently also been combined for logistic regression settings [9,10].

In this work, we provide a provably optimal approach to perform simultaneous feature selection and estimation, as well as outlier detection and exclusion for logistic regression problems. Here optimality refers to the fact the the global optimum of the underlying "double" combinatorial problem is indeed achievable and, even if the algorithm is stopped before convergence, one can obtain optimality guarantees by monitoring the gap between the best feasible solution and the problem relaxation [11,12]. Specifically, we consider an L_0 sparsity assumption on the coefficients [13] and a logistic slippage model for the outlying observations [14]. We further build upon the work in [7] and rely on L_0-constraints to detect outlying cases and select relevant features. This requires us to solve a double combinatorial problem, across both the units and the covariates. Importantly, the underlying optimization can be effectively tackled with state-of-the-art *mixed-integer conic programming* solvers. These target a global optimum and, unlike existing heuristic methods, provide optimality guarantees even if the algorithm is stopped before convergence.

We use our proposal to investigate the main drivers of honey bee (*Apis mellifera*) loss during winter (overwintering), which represents the most critical part of the year in several areas [15–17]. In particular, we use survey data collected by the Pennsylvania State Beekeepers Association, which include information related to honey bee survival, stressors and management practices, as well as bio-climatic indexes, topography and land use information [18]. Previous studies mainly focused on predictive performance and relied on statistical learning tools such as random forest, which capture relevance but not effect signs for each feature, and do not account for the possible impact of outlying cases—making results harder to interpret and potentially less robust. In our analysis, based on a logistic regression model, we are able to exclude redundant features from the fit while accounting for potential data contamination through an estimation approach that simultaneously addresses sparsity and statistical robustness. This provides important insights on the main drivers of honey bee loss during overwintering—such as the exposure to pesticides, as well as the average temperature of the driest quarter and the precipitation level during the warmest quarter. We also show that the data set does indeed contain outlying observations.

The remainder of the paper is organized as follows. Section 2 provides some background on existing penalized and robust estimation methods. Section 3 details our proposal and its algorithmic implementation. This is compared with existing methods through numerical simulations in Section 4. Our analysis of the drivers of honey bee loss is presented in Section 5. Final remarks are provided in Section 6.

2. Background

Let $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T \in \mathbb{R}^{n \times p}$ be an observed design matrix, and $\boldsymbol{y} \in \{-1, 1\}^n$ the corresponding set of binary response classes. The *two-class logistic regression model* assumes that the log-odds ratio is a linear function of the covariates

$$\log\left(\frac{\Pr(y_i = 1|\boldsymbol{x}_i)}{1 - \Pr(y_i = 1|\boldsymbol{x}_i)}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the unknown regression parameters (possibly sparse). We also assume the presence of an intercept term, so that $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_{p-1}\}$ and $\boldsymbol{X}$ contains only 1's in the first column. Thus, for any $\boldsymbol{x}_i \in \mathbb{R}^p$, it follows from (1) that

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} = \frac{1}{1 + \exp(-\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

and

$$\Pr(y_i = -1|\boldsymbol{x}_i) = 1 - \Pr(y_i = 1|\boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}.$$

Hence, in full generality, the logistic model can be expressed as

$$\Pr(y_i|\boldsymbol{x}_i) = \frac{1}{1+\exp(-y_i\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{2}$$

Assuming that $y_i|\boldsymbol{x}_i$, for $i = 1,\ldots,n$, follow independent Bernoulli distributions, the likelihood function associated to (2) is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(y_i = 1|\boldsymbol{x}_i)^{(1+y_i)/2} \Pr(y_i = -1|\boldsymbol{x}_i)^{(1-y_i)/2},$$

which provides the ML estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) \tag{3}$$

where the deviance is defined as

$$d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) = \log\Big(1+\exp(-y_i\boldsymbol{x}_i^T\boldsymbol{\beta})\Big).$$

The optimization problem in (3) is convex, and it admits a unique and finite solution if and only if the points belonging to each class "overlap" to some degree (i.e., the two classes are not linearly separable based on predictors information) [19,20]. Otherwise, there exist infinitely many hyperplanes perfectly separating the data, and the ML estimator is undetermined. Importantly, in this setting, the ML estimator is consistent and asymptotically normal as $n \to \infty$ under weak assumptions [21]. However, unlike ML estimation for linear regression problems, there is no closed-form solution for (3), and iterative methods such as the Newton–Raphson algorithm are commonly employed [22], which can be solved through iteratively reweighted least squares [1,22].

2.1. Penalized Logistic Regression

The ML estimator in (3) does not exist if $p > n$. Moreover, in the presence of strong collinearities in the predictor space, even if $p < n$, the ML estimator might provide unstable estimates or lead to overfitting (i.e., to estimates with low bias and high variance and thus poor predictive power). In order to overcome these limitations, penalized estimation methods based on the L_2-penalty have been considered [23,24]. To promote sparse estimates and improve interpretability, several authors also studied the use of the L_1-norm [4,25]. Although this class of "soft" penalization methods is computationally very efficient due to convexity, it provides biased estimates. Further approaches combine the L_1 and L_2-norms in what is known as the *elastic net* penalty [26]—coupled with an adaptive weighting strategy to regularize the coefficients [27]. Importantly, under suitable assumptions, this guarantees that the resulting estimator satisfies the so-called oracle property, meaning that the probability of selecting the truly active set of covariates (i.e., the ones corresponding to nonzero coefficients) converges to one, and at the same time the coefficient estimates are asymptotically normal with the same means and variance structure as if the set of active features was known a priori [28].

Best subset selection is a traditional "hard" penalization method that approaches the feature selection problem combinatorially [29]. Ideally, one should compare all possible fits of a given size, for all possible sizes—say $1 \leq k_p \leq \min(n,p)$. This was long considered unfeasible for problems of realistic size p even in the linear regression setting [22]. Nevertheless, leveraging recent developments in hardware and *mixed-integer programming* solvers, [30] proposed the use of L_0-constraints on $\boldsymbol{\beta}$ to efficiently and effectively solve the underlying best subset logistic regression problem using *mixed-integer nonlinear programming* techniques. This extends the approach in [11] for linear regression and relies on the L_0 pseudo-norm, which is defined as $\|\boldsymbol{\beta}\|_0 = \sum_j I(\beta_j \neq 0)$, where $I(\cdot)$ is the

indicator function. Notably, oracle properties can be established in this setting under weaker assumptions than other proposals [31].

2.2. Robust Logistic Regression

Outliers may influence the fit, hindering the performance of ML-based estimators and leading to estimation bias and weaker inference [32]. Multiple outliers are particularly problematic and difficult to detect since they can create masking (false negative) and swamping (false positive) effects [33]. Here, as in linear regression, raw (deviance) residuals can be used to build several regression diagnostics [33–35]. Different approaches have been introduced to overcome the limitations of classical ML estimation in low-dimensional settings [5]. For instance, a weighted counterpart of ML estimation was proposed in [36] (see also [37]), robust M-estimators were developed in [35], and ref. [38] introduced an additional correction term that provides a robust class of Fisher-consistent M-estimators—see also [39,40] for bounded influence estimators. Furthermore, an adaptive weighted maximum likelihood where the estimator efficiency is calibrated in a data-driven way was considered in [41]. A distributionally robust approach was proposed in [42], which is similar in spirit to the use of robust optimization in [30] where uncertainty sets have to be taken into account.

The *logistic slippage model*, which closely resembles the mean-shift outlier model for linear regression problems [43], was explicitly considered in [14] and leads to the removal of outliers from the fit. However, since the number and position of outlying cases are generally unknown, one should in principle compare the exclusion of $0 \leq k_n \leq n/2$ points from the fit (if one is willing to assume that less than half of the data are in fact contaminated). Building upon high breakdown point estimators and deletion diagnostics, a forward search procedure based on graphical diagnostic tools that is effective in detecting masked multiple outliers and highlights the influence of individual observations on the fit was developed in [44,45]. This approach is robust, computationally cheap and provides a natural order for the observations according to their agreement with the model.

For high-dimensional settings, the authors in [9] focused on the possible contamination of the y labeling and proposed L_1 penalization methods for reducing the influence of outliers and performing feature selection. However, this provides a sub-optimal strategy both for sparse estimation [31] and outlier detection [6]. More recently, the elastic net penalty has been combined with a trimmed loss function which excludes the k_n most influential observations from the fit [10]. This mimics the *least trimmed squares* (LTS) estimator for linear regression [46], and is equivalent to assuming a logistic slippage model. On the other hand, the trimmed loss function is solved through heuristic methods based on resampling, and the elastic net penalty in use is sub-optimal in terms of feature selection.

3. MIProb: Robust Variable Selection under the Logistic Slippage Model

We consider a two-class logistic regression model affected by data contamination (i.e., outliers) and comprising irrelevant covariates. Specifically, we focus on the logistic slippage model, where the number, position and strength of the outliers are unknown [14,43]. The main idea is to enforce integer constraints on the number of outlying cases and relevant features in order to improve the interpretability of the model and its robustness. Now, we introduce a general formulation that in addition to simultaneous feature selection and outlier detection encompasses an optional ridge penalty, which can be useful to tackle strong collinearity structures [26,30], low signal-to-noise ratio regimes [47] and data perturbations [48]. Thus, we propose to solve the following discrete optimization problem:

$$\left[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}\right] = \underset{\boldsymbol{\beta},\boldsymbol{\phi}}{\arg\min} \sum_{i=1}^{n} d(\boldsymbol{x}_i^T \boldsymbol{\beta} + \phi_i, y_i) \tag{4}$$

$$\text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k_p \tag{4a}$$

$$\|\boldsymbol{\phi}\|_0 \leq k_n \tag{4b}$$

$$\|\boldsymbol{\beta}\|_2 \leq l. \tag{4c}$$

Due to the (double) combinatorial nature of the problem, the formulation in (4) is computationally daunting [49]. Nevertheless, nowadays it can be solved effectively and at times also efficiently with specialized solvers. Importantly, it relates to the use of a trimmed loss function as in [10], and it extends the work in [7] for sparse linear regression models affected by data contamination in the form of mean-shift outliers. However, here the use of a nonlinear and nonquadratic objective function complicates the matter and requires special attention.

We also note that (4) can be easily extended to model structured data, such as hierarchical or group structures. For instance, in Section 5 we enforce the so-called *group sparsity constraints* [50] to model categorical features. Moreover, it can be naturally extended to multinomial logistic regression models along lines similar to those in [38].

3.1. Algorithmic Implementation

The optimization problem in (4) can be formulated as a *mixed-integer conic program*. For simplicity, we first consider only the objective function and the L_2 ridge-like penalty. Specifically, including auxiliary variables $t_1, \cdots, t_n$ and r, the objective (4) and the constraint (4c) can be equivalently reformulated as

$$\min_{t,r,\boldsymbol{\beta}} \sum_{i=1}^{n} t_i + \lambda r \tag{5}$$

$$\text{s.t.} \quad t_i \geq \log\left(1 + \exp\left(-y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \phi_i)\right)\right) \tag{5a}$$

$$r \geq \|\boldsymbol{\beta}\|_2. \tag{5b}$$

The constraints in (5a) can be expressed using the exponential cone

$$K_{\text{exp}} = \left\{(x, y, z) \in \mathbb{R}^3 : y \exp(x/y) \leq z\right\},$$

and provide

$$\exp(-t_i) + \exp(u_i - t_i) \leq 1$$

where $u_i = -y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \phi_i)$. Including auxiliary variables z_{i1} and z_{i2} such that $z_{i1} \geq \exp(u_i - t_i)$ and $z_{i2} \geq \exp(-t_i)$, it follows that (5a) is equivalent to

$$\begin{cases} (u_i - t_i, 1, z_{i1}) \in K_{\text{exp}} \\ (-t_i, 1, z_{i2}) \in K_{\text{exp}} \\ z_{i1} + z_{i2} \leq 1. \end{cases}$$

Thus, the proposed mixed-integer conic programming formulation for logistic regression in (4) (denoted *MIProb* for simplicity), which provides sparse estimates for $\boldsymbol{\beta}$ and removes outliers through $\boldsymbol{\phi}$, is

$$\min_{\boldsymbol{t},z,r,\boldsymbol{\beta},z^{\beta},\boldsymbol{\phi},z^{\phi}} \sum_{i=1}^{n} t_i + \lambda r \tag{6}$$

$$\text{s.t.} \quad -\mathcal{M}_j^{\beta} z_j^{\beta} \leq \beta_j \leq \mathcal{M}_j^{\beta} z_j^{\beta} \tag{6a}$$

$$-\mathcal{M}_i^{\phi} z_i^{\phi} \leq \phi_i \leq \mathcal{M}_i^{\phi} z_i^{\phi} \tag{6b}$$

$$\sum_{j=1}^{p} z_j^{\beta} \leq k_p \tag{6c}$$

$$\sum_{i=1}^{n} z_i^{\phi} \leq k_n \tag{6d}$$

$$(u_i - t_i, 1, z_{i1}) \in K_{\text{exp}}$$

$$(-t_i, 1, z_{i2}) \in K_{\text{exp}}$$

$$z_{i1} + z_{i2} \leq 1$$

$$r \geq \|\boldsymbol{\beta}\|_2$$

$$z_j^{\beta} \in \{0,1\}, \;\; \beta_j \in \mathbb{R}, \;\; j = 1, \ldots, p$$

$$z_i^{\phi} \in \{0,1\}, \;\; \phi_i \in \mathbb{R}, \;\; i = 1, \ldots, n.$$

The big-$\mathcal{M}$ bounds $\boldsymbol{\mathcal{M}}^{\beta}$ and $\boldsymbol{\mathcal{M}}^{\phi}$ in constraints (6a) and (6b) have p and n entries, respectively, which can be tailored for each β_j and ϕ_i. These should be wide enough to include the true regression coefficients and zero-out the effects of the true outliers, but not so wide as to substantially increase the computational burden.

For instance, an ensemble method based on existing heuristic and robust procedures to create suitable big-$\mathcal{M}$ bounds was considered in [7]. However, a similar approach is challenging in this framework given a "pool" of openly available robust algorithms is not available for logistic regression models—unlike in linear regression. Here, we simply set large, more conservative bounds to maintain accuracy at the cost of computing time. Extensions of additional heuristics to strengthen these bounds are worth further investigation, but beyond the scope of this work. The L_0-norm constraints (6c) and (6d) depend on positive integers k_p and k_n, which control the sparsity level for feature selection and the trimming level for outlier detection, respectively. As with any selection procedure, these tuning parameters are key to retain selection and detection accuracy. However, k_p and k_n can be treated differently. For the former, any deviation from the true sparsity level will result in false negatives/positives. For the latter, a common approach [10,45] is to select an inflated trimming amount (i.e., higher than the true level) to avoid masking and swamping effects, and then refine the solution to recover efficiency.

Importantly, in this work we use existing specialized solvers (see Section 4) but the development of a tailored approach could be beneficial. For instance, outer approximation techniques in mixed-integer nonlinear programming with dynamic constraint generation were combined in [30], as well as the use of first-order methods, which reduce the computational burden compared to general-purpose solvers. Extensions of such approaches to this setting are left for future work.

3.2. Additional Details

In order to achieve good estimates it is essential to tune the sparsity level k_p and the trimming level k_n, as well as the ridge-like tuning parameter λ if present, in a data-driven fashion. For instance, one might consider robust counterparts of information criteria or cross-validation. In our simulation study, we do not include the L_2-constraint and, for a given trimming level k_n, we use a robust version of the *Bayesian information criterion* (BIC) similarly to [7]. In symbols, this is $\text{BIC} = k_p \ln(n - k_n) + \sum_{i=1}^{n} d(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}, y_i)$, where $d(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}, y_i)$ are the final deviances for a given estimator—recall that deviances corresponding to trimmed points are equal to 0. If an intercept term is included in the model, we force its selection as an active feature. Other tuning procedures such as cross-validation benefit

from the use of effective warm-starts to accelerate convergence of the algorithm when solving over several training and testing sets splits—see [7] for additional details.

The *breakdown point* (BdP) is the largest fraction of contamination that an estimator can tolerate before it might provide completely unreliable estimates [51]. It can be formalized either by replacing good units with outliers or adding outliers to an uncontaminated dataset. Using a unit-replacement approach, it has been shown that one can break down (unpenalized) ML estimation by simply removing units belonging to the overlaps among classes [39,52]. Using unit-addition,

the authors in [53] showed that when severe outliers are added to a non-separable dataset, ML estimates do not break down due to "explosion" (to infinity), but they can break down due to "implosion" (to zero). Specifically, the BdP for the ML estimator is equal to $\varepsilon^*_{\text{ML}} = 2(p-1)/\{n+2(p-1)\}$ (which is 0% asymptotically), since the estimates can implode to zero, adding $2(p-1)$ appropriately chosen outliers. Thus, unlike in linear regression, here one has to take into account not only the explosion of the estimates, but also their implosion, which is often more difficult to detect.

We leave theoretical derivations concerning our MIProb proposal in (6) to future work. However, we note that MIProb clearly represents a trimmed likelihood estimator as a special case, so in this special case it inherits properties such as the high breakdown point [54,55]. Moreover, these results might be combined with the oracle properties for feature selection described in [31] in order to obtain a logistic version of the robustly strong oracle property introduced in [7].

4. Simulation Study

In this section, we use a simulation study to compare the performance of our proposal with state-of-the-art methods. The simulated data is generated as follows. The first column of the $n \times p$ design matrix $\boldsymbol{X}$ comprises all 1's (for the model intercept) and we draw the remaining entries of each row independently from a standard $(p-1)$-variate normal distribution $N(\mathbf{0}, \boldsymbol{I}_{p-1})$. The values of the p-dimensional coefficient vector $\boldsymbol{\beta}$ comprise p_0 non-zero entries (including the intercept) and $p - p_0$ zeros. The response labels $y_i \in \{1, -1\}$, for $i = 1, \ldots, n$, are generated from Bernoulli distributions with probabilities $1/(1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}})$. Next, without loss of generality, we contaminate the first n_0 cases with a logistic slippage model, adding the scalar mean shifts μ_X to the active predictors only (excluding the intercept). In order to generate *bad leverage points*, we also assign opposite signs to the labels of each contaminated unit: $\text{sign}(y_i) = -\text{sign}(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

The simulation scenarios are defined according to the values of the parameters discussed above. Here, we present results for $p_0 = 4$ active predictors with $\beta_j = 3$ (without loss of generality, these correspond to the intercept and the last 3 features), sample size $n = 100$, increasing dimension $p = 20, 50$ (low) and 150 (high), $n_0 = 5$ contaminated units (i.e., 5% contamination), and mean shifts $\mu_X = 10$. Each simulation scenario is replicated q independent times, and random test data, say $(\boldsymbol{y}^*, \boldsymbol{X}^*)$, are generated from the same simulation scheme, but without any form of contamination.

Different estimators are compared based on: (i) the *mean of the negative log-likelihoods* $\text{MNLL}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{x}_i^{*\boldsymbol{T}}\widehat{\boldsymbol{\beta}}, y_i^*)$, i.e., the average of deviances computed on the uncontaminated test set; (ii) the outlier *misclassification rate* $\text{MR}(\widehat{\boldsymbol{\beta}}) = c/n$, where c counts the points of the uncontaminated test set erroneously labeled as outliers; (iii) estimation accuracy in terms of *average mean squared error* $\text{MSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{p}\sum_{j=1}^{p}\text{MSE}(\widehat{\beta}_j)$, where for each $\widehat{\beta}_j$ we decompose $\text{MSE}(\widehat{\beta}_j) = \frac{1}{q}\sum_{i=1}^{q}(\widehat{\beta}_{ji} - \beta_j)^2 = (\overline{\beta}_j - \beta_j)^2 + \frac{1}{q}\sum_{i=1}^{q}(\widehat{\beta}_{ji} - \overline{\beta}_j)^2$ in squared bias and variance (here $\overline{\beta}_j = \frac{1}{q}\sum_{i=1}^{q}\widehat{\beta}_{ji}$) (iv) feature selection accuracy, measured by the *false positive rate* $\text{FPR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1, \ldots, p\} : \widehat{\beta}_j \neq 0 \wedge \beta_j = 0\}|/|\{j \in \{1, \ldots, p\} : \beta_j = 0\}|$ and the *false negative rate* $\text{FNR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1, \ldots, p\} : \widehat{\beta}_j = 0 \wedge \beta_j \neq 0\}|/|\{j \in \{1, \ldots, p\} : \beta_j \neq 0\}|$; (v) outlier detection accuracy, which is similarly measured by $\text{FPR}(\widehat{\boldsymbol{\phi}})$ and $\text{FNR}(\widehat{\boldsymbol{\phi}})$.

We use the *robust oracle* estimator as a benchmark, which is a logistic fit computed only for the active set of features and only on the uncontaminated units (we used our

MIP formulation to compute the robust oracle). The following estimators are compared: (a) *enetLTS* with $\alpha = 1$ (i.e., robust Lasso) [10]; (b) *MIProb*, our robust MIP proposal without a ridge-like constraint (see Section 3); (c) *MIP*, the non-robust MIP implementation performing only feature selection (i.e., as MIProb but using $k_n = 0$); (d) *Lasso*, the non-robust L_1-penalized loss computed through the `glmnet` package in `R` [4]. Robust methods trim the true number of outliers ($k_n = n_0$) , though this does not guarantee exact outlier detection, and only the sparsity level in the feature space is tuned for each method based on (robust) information criteria or cross-validation. However, since enetLTS is a heuristic method that relies on resampling rather than exact trimming, we inflate the trimming proportion to 20% and then take the re-weighted estimates in order to improve its outlier detection performance.

Table 1 provides medians and median absolute deviations (MAD) of simulation results over $q = 30$ replications. Our proposal substantially outperforms competing methods in most criteria. In both low ($p = 20, 50$) and high ($p = 150$) dimensional settings, the MNLL and MR of MIProb are closest to values produced by the oracle. In terms of estimation accuracy, MIProb has the lowest bias, but the non-robust Lasso has distinctly lower variance than all procedures aside from enetLTS when $p = 150$. MIProb has very strong feature selection accuracy with FPR($\widehat{\beta}$) and FNR($\widehat{\beta}$) equal to 0 in the low-dimensional settings ($p = 20, 50$). In the high-dimensional setting, it maintains the lowest false positive rate, but has a higher false negative rate than enetLTS (though still lower than the non-robust methods). This motivates the development of more effective tuning strategies as p increases. On the other hand, enetLTS tends to overselect, since it has FNR($\widehat{\beta}$) $= 0$ in all settings, but the highest FPR($\widehat{\beta}$) across methods. Similar results were found in [7]. Regarding outlier detection, enetLTS and MIProb produce very similar solutions with FPR and FNR almost always 0. Thus, both methods are highly effective at detecting contaminated units.

Table 1. Median (MAD in parenthesis) of MNLL, misclassification rate, variance and squared bias for $\widehat{\beta}$, false positive rate and false negative rate for feature selection and outlier detection based on 30 simulation replicates.

n	p	Method	MNLL	MR	var($\widehat{\beta}$)	bias($\widehat{\beta}$)^2	FPR($\widehat{\beta}$)	FNR($\widehat{\beta}$)	FPR($\widehat{\phi}$)	FNR($\widehat{\phi}$)
100	20	Oracle	0.24(0.08)	0.11(0.04)	0.85(0.00)	0.29(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.50(0.08)	0.28(0.06)	0.03(0.00)	1.11(0.00)	0.18(0.09)	0.00(0.00)	0.005(0.008)	0.00(0.00)
		MIProb	0.27(0.04)	0.11(0.04)	0.01(0.00)	0.46(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIP	0.64(0.07)	0.31(0.04)	0.02(0.00)	1.52(0.00)	0.06(0.09)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.03)	0.29(0.04)	0.005(0.00)	1.52(0.00)	0.00(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
100	50	Oracle	0.22(0.04)	0.09(0.03)	0.10(0.00)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.51(0.10)	0.27(0.13)	0.02(0.00)	0.45(0.00)	0.12(0.06)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIProb	0.26(0.04)	0.10(0.02)	0.01(0.00)	0.19(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIP	0.69(0.09)	0.37(0.06)	0.02(0.00)	0.63(0.00)	0.04(0.03)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.02)	0.29(0.03)	0.002(0.00)	0.63(0.00)	0.00(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
100	150	Oracle	0.21(0.05)	0.09(0.03)	0.07(0.00)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		enetLTS	0.54(0.07)	0.29(0.03)	0.007(0.00)	0.16(0.00)	0.06(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		MIProb	0.34(0.12)	0.16(0.06)	0.03(0.00)	0.08(0.00)	0.00(0.00)	0.25(0.37)	0.00(0.00)	0.00(0.00)
		MIP	0.88(0.12)	0.42(0.03)	0.02(0.00)	0.21(0.00)	0.03(0.00)	0.75(0.00)	0.00(0.00)	1.00(0.00)
		Lasso	0.62(0.05)	0.30(0.06)	0.002(0.00)	0.21(0.00)	0.007(0.01)	0.75(0.00)	0.00(0.00)	1.00(0.00)

Computational Details

In this section, we discuss further computational details and the tuning approaches for each procedure. Our proposal, MIProb is computationally more demanding than the other methods under comparison, including the non-robust MIP. This is natural, given methods like enetLTS are heuristics and avoid directly solving the full combinatorial problem. As discussed in more detail in [11,56], a common challenge with MIP formulations is the weak lower bound produced by the relaxed version of the problem. Thus, while the optimal solution may have already been found, the majority of computing time may be used to verify its optimality. For our settings, we set generous stopping criteria where the algorithm

ends when either a maximum computing time of 40 min (this can be as low as 3 min in other literature [47]) or an optimality gap of 2.5% (i.e., the relative difference between the upper and lower bounds) is met. While this maximum time may be hit, especially under the most challenging scenarios with $p = 150$, the consistent quality of solutions close to the oracle (see Table 1) further supports this observation of weak lower bounds. However, for comparison, enetLTS only takes an average of 14 s. Thus, the use of other warm-starts, heuristics, etc., to improve lower bounds would be very beneficial for MIP-based feature selection and outlier detection approaches.

We also found that the computational burden of MIProb varies vastly based on the tuning parameter k_p. In our numerical experiments, computing time decreases as more features are selected, especially for $k_p > p_0$. For instance, we considered other simulation scenarios not reported here, including one with a lower sample size $n = 50$ and thus a higher contamination percentage. We observed the pattern in Figure 1 where the average computing time is much higher for lower values of k_p, but rapidly decreasing after the "elbow" occurring around $k_p = p_0$. This could be due to the outlier detection portion of the problem being more difficult when some of the relevant features are not included. Recall that our simulations add mean shift contamination only to the relevant features; when some are missing, it is more challenging to detect contaminated units.

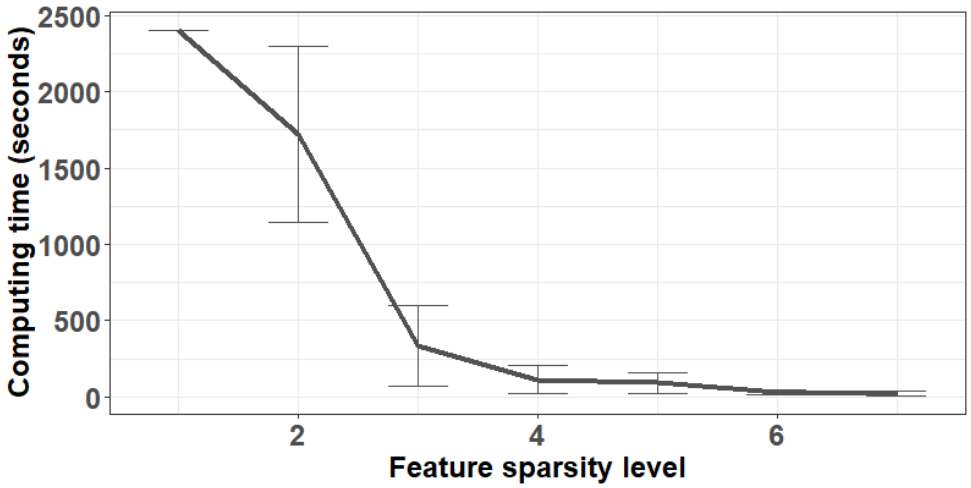

Figure 1. Average computing times across various feature sparsity levels k_p in simulated data following the data generation approach described above with $n = 50$, $p = 7$, $p_0 = 4$, and $k_n = 5$. Bars represent ± 1 standard deviations over 5 simulation replicates.

Regarding tuning, we utilized different approaches for each procedure as appropriate. The oracle operates on uncontaminated units and relevant features only, and requires no tuning. EnetLTS is tuned with cross-validation through the `enetLTS` package in `R` following the default settings with 5 folds [57]. For our proposal, MIProb, we used a robust version of BIC as described in Section 3.2, selecting the k_p corresponding to the minimum. Similarly, MIP is tuned based on the traditional BIC (without trimming incorporated). Finally, the non-robust Lasso is tuned through 10-fold cross-validation in the `glmnet` package. We note that MIProb and MIP are implemented in `Julia 1.3.1` to interact with the `Mosek` solver through its `JuMP` package. MIProb and enetLTS utilize 24 cores per replication through their multi-thread options.

5. Investigating Overwintering Honey Bee Loss in Pennsylvania

Pollinators play a vital role supporting critical natural and agricultural ecosystem functions. Specifically, honey bees (*Apis mellifera*) are of great economic importance and play a primary role in pollination services [58,59]. The added value of honey bees pollination for the crops produced in the United States (in terms of higher yield and quality of the product) is annually estimated around 15–20 billion dollars [58,60], and

according to the Pennsylvania Beekeepers Association, their yearly contribution has an estimated value of 60 million dollars in the state of Pennsylvania alone; see https://pastatebeekeepers.org/pdf/ValueofhoneybeesinPA3.pdf (accessed on 15 July 2021). Yet the decline of the honey bee populations is a widespread phenomenon around the globe [61–65]. Major threats for honey bees include habitat fragmentation and loss, mites [66,67], parasites and diseases [68], pesticides [69], climate change [70], extreme weather conditions, the introduction of alien species [71], as well as the interactions between these factor [72]. Moreover, the overwintering period is often a major contributor to honey bee loss [15,16,73,74]. We thus focus on honey bee winter survival.

In the United States, beekeepers suffered an average 45.5% overwinter colony loss between 2020 and 2021 [75]. This figure was 41.2% in the state of Pennsylvania for the same overwintering period; see https://beeinformed.org/2021/06/21/united-states-honey-bee-colony-losses-2020-2021-preliminary-results/ (accessed on 15 July 2021). In both cases, this was an increase compared to the previous year, when the reported losses were 43.7% and 36.6% for the United States and Pennsylvania, respectively [76]. In recent years the trend of overwintering loss for Pennsylvania is comparable to the one at the national level, making it an interesting case study. Thus, in the following we analyze overwintering survey data for Pennsylvania covering the years 2016–2019.

5.1. Model Formulation and Data

Focusing on the state of Pennsylvania, honey bee winter survival was recently investigated in [18] based on winter loss survey data provided by the Pennsylvania State Beekeepers Association. The data cover three winter periods (2016–2017, 2017–2018, and 2018–2019), and the main goals of the analysis were to assess the importance of weather, topography, land use, and management factors on overwintering mortality, and to predict survival given current weather conditions and projected changes in climate. The authors utilized a random forest classifier to model overwintering survival. Importantly, they also controlled for the treatment of varroa mites (*Varroa destructor)* at both apiary and colony levels, since this represents a key factor in describing honey bee survival—all untreated colonies were excluded from the dataset. Their main findings suggest that growing degree days (see Table 2) and precipitations in the warmest quarter of the preceding year were the most important predictors, followed by precipitations in the wettest quarter, as well as maximum temperature in the warmest month. These results highlight the strong association between weather events and overwintering survival of honey bees.

The data set used in our analysis is extracted from the Supplementary Information published in [18]—see Table 2 for a description of the variables included into our model.

Since observations in the original data set represent colonies that may belong to the same apiary, we aggregated the data to obtain unique apiary information. This is particularly important in order to reduce dependence across observations, and leads to a sample of $n = 257$ apiaries from 1429 colonies (in the absence of publicly available geo-localized information, apiary identification was made possible through the features "bioc02" and "slope").

We created a binary response taking the proportion of survived colonies per apiary, and assigning the label 1 if such a mean is greater than 0.8, and the label -1 if it is smaller than 0.6. These thresholds are motivated by the "average" winter colony loss rate described above and they allow us to study the most "extreme" behavior (significantly higher or lower losses); they also provide a balanced labeling for the response variable. The remaining observations are completely removed from the data set in use and thus decreasing the sample size to $n = 216$.

Table 2. Description of the features included in our logistic model formulation to describe honey bee overwintering survival. See [18] for details. The bioc# variables refer to bioclimatic variables of the WorldClim database; see https://www.worldclim.org/data/bioclim.html (accessed on 15 July 2021).

	Variable	Description
1	survival	Binary survival response at the apiary level
2	bee2	Winter total precipitation
3	bee4	Winter days with maximum temperature above 16 °C and precipitation below 3 mm
4	gdd	Growing degree days (base 5 °C) as the accumulation of average daily temperatures
5	dd rain	Days between rain events > 0.25 mm
6	bioc02	Mean diurnal temperature range
7	bioc04	Temperature seasonality
8	bioc08	Mean temperature of the wettest Quarter
9	bioc09	Mean temperature of the driest quarter
10	bioc18	Precipitation of the warmest quarter
11	bioc19	Precipitation of the coldest quarter
12	slope	Terrain slope
13	sol rad	Potential incident solar radiation, 21 December
14	pcurv	Profile curvature
15	tcurv	Terrain curvature
16	TWI	Topographic wetness index
17	EW	East/West orientation of slope
18	ITL	Distance-weighted insect toxic load
19	col nov	Number of colonies in November
20	exp 1–2	Beekeeper years of experience between 1 and 2 (binary variable)
21	exp 2–5	Beekeeper years of experience between 2 and 5 (binary variable)
22	exp < 1	Beekeeper years of experience less than 1 (binary variable)
23	exp > 10	Beekeeper years of experience greater than 10 (binary variable)

We compared the same procedures considered in our simulation study (see Section 4) without introducing a ridge-like penalty for any of the methods. Relatedly, we did not use all features in the original study, which presented sizable collinearities. In particular, for each pair of features with an absolute pair-wise correlation above 0.7, we computed the mean absolute correlation of each feature against all the others and removed the one with the largest mean absolute correlation from our pool.

Each column of the design matrix $\boldsymbol{X}$ (excluding the intercept and categorical factors) was standardized to have zero median and median absolute deviation (MAD) equal to the average MAD across columns (standardization does not affect our proposal and each of the other approaches included in our comparison performs its own standardization as needed). Importantly, for MIP and MIProb we introduced *group sparsity constraints* [50] to tackle the categorical feature "beekeepers' experience"; the reference category is "between 5 and 10 years" and all coefficients for the dummy variables are included or excluded from the fit together.

5.2. Results

We randomly split the data into training and test sets, encompassing 100 and 116 points, respectively. For robust methods, we fix the trimming proportion at 10% after exploring a range of values suitable for the nature of the problem, and only tune the sparsity level. Figure 2 compares the *balanced accuracy*, defined as (sensitivity+specificity)/2, on the test set across different methods Here sensitivity is defined as (# true positives)/(# true positives + # true negatives) and specificity is defined as (# true negatives)/(# true negatives + # false positives). While this is a function of the sparsity level imposed on MIP and MIProb, for enetLTS and Lasso we show the mean values across eight repetitions due to the intrinsic randomness induced by cross-validation methods (horizontal dashed lines). Here MIP and MIProb are quite comparable and generally outperform competing methods, although we notice a drop in predictive performance for MIProb if the sparsity level $k_p \geq 9$—which is likely a result of overfitting due to data trimming compared to MIP. Based on these

findings, in the following we present the results based on $k_p = 8$ (including the intercept), where the balanced accuracy for both methods is very close to their maximum.

Table 3 displays the features selected by each method on the training set. We focus on the interpretation of the signs of the estimated coefficients, represented as green (positive) and red (negative) cells, respectively. The estimates provided by MIProb are in line with the findings of the original study [18]. Specifically, MIProb estimates a positive association between honey bee survival and "bee2" (winter total precipitation), "gdd" (growing degree days), "EW" (East/West orientation of slope) and "ITL" (distance-weighted insect toxic load, see [77]). This suggests that the impact of precipitations and the accumulation of average daily temperatures (gdd), which influence the growth of crops, have an overall positive effect on honey bee survival. In contrast, MIProb estimates a negative association between honey bee survival and "bioc09" (mean temperature of the driest quarter), "bioc18" (precipitation of the warmest quarter), "tcurv" (terrain curvature) and "TWI" (topographic wetness index). This highlights once more the major impact of weather predictors, as well as topographic factors and humidity levels. Notably, beekeepers' experience was not selected as a relevant feature by MIProb, which further supports the findings in [18].

Table 3. Features selected by Lasso, MIP, enetLTS and MIProb (robust MIP) on a training set encompassing 100 points. Green and red cells indicate estimated coefficients with positive and negative signs, respectively. White cells indicate non-selected features.

	interc.	bee2	bee4	gdd	dd rain	bioc02	bioc04	bioc08	bioc09	bioc18	bioc19	slope
Lasso												
MIP												
enetLTS												
MIProb												

	sol rad	pcurv	tcurv	TWI	EW	ITL	col nov	exp 1–2	exp 2–5	exp < 1	exp > 10
Lasso											
MIP											
enetLTS											
MIProb											

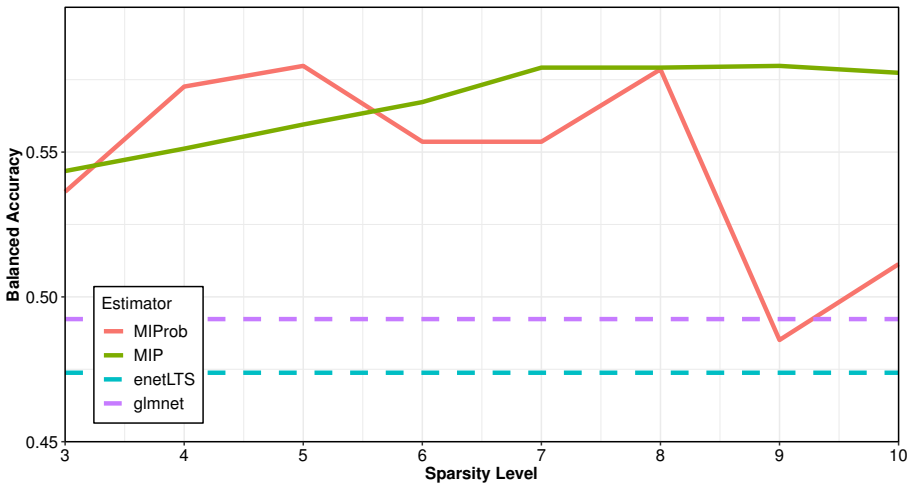

Figure 2. Balanced accuracy computed on a test set encompassing 126 points, as a function of the sparsity level k_n for MIP and MIProb (using a 10% trimming for the latter). The average balanced accuracy over 8 repetitions is shown also for Lasso and enetLTS.

Considering the other procedures, enetLTS appears to produce denser solutions (this was also observed in the simulations in Section 4), excluding only three features from the fit, and the non-robust Lasso appears to produce sparser solutions, selecting only three features—which is indeed due to the presence of outliers. This is supported by the fact that a Lasso fit after the exclusion of the outliers detected by MIProb provides

richer solutions, corresponding to clearer minima of the cross-validation error, where approximately 10 features are selected and several of these are shared with MIProb (data not shown). MIP uses the same sparsity level of MIProb but selects a different set of features, which is again due to the presence of outliers (e.g., it selects "bioc02" and the dummies related to beekeepers' experience).

Figure 3 compares Pearson residuals for MIProb and MIP estimators. The outlying cases detected by MIProb, which are highlighted in red, deviate substantially from the remaining observations and are undetected by the non-robust MIP algorithm. Moreover, focusing on the set of features selected by MIProb, Figure 4 compares the boxplots of outliers selected by MIProb against the remaining non-outlying cases. We notice that the two distributions are indeed quite different for variables such as "bee2", "gdd", "EW" and "ITL". This provides further evidence that the data set contains some outlying cases which significantly differ from the rest of the points.

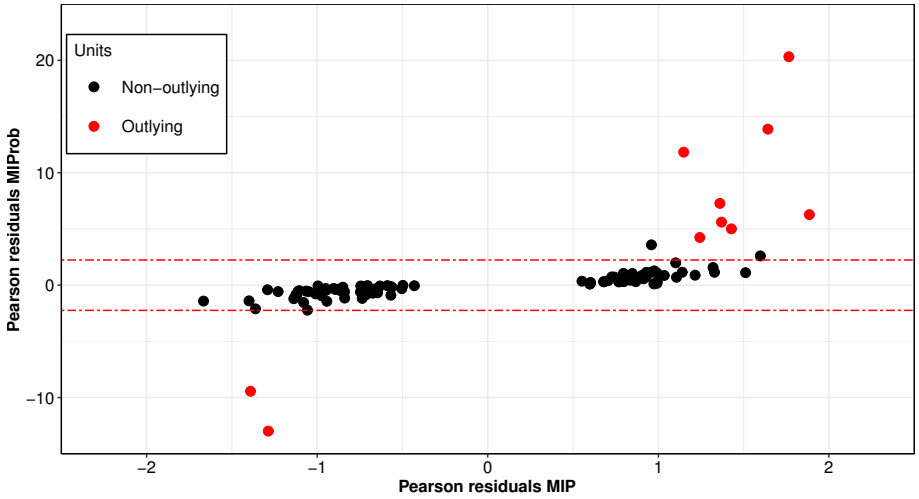

Figure 3. Pearson residuals for MIProb and MIP. Outlying cases detected by MIProb are highlighted in red. Horizontal red lines represent the 0.0125 and 0.9875 quantiles of the standard normal distribution.

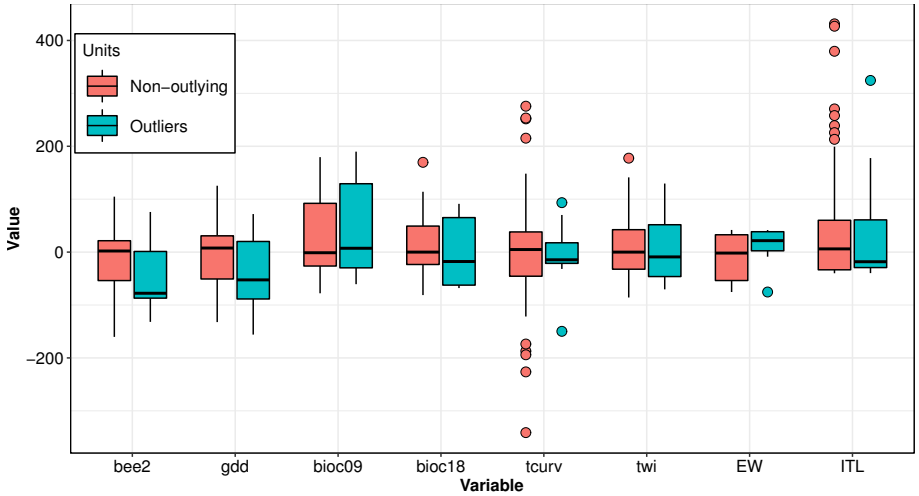

Figure 4. Box plots comparing the values assumed by the features selected by MIProb contrasting outlying and non-outlying case. The values of each feature are scaled to have zero median and MAD equal to the average MAD across columns.

6. Discussion

We propose a discrete approach based on L_0-constraints to simultaneously perform feature selection and multiple outlier detection for logistic regression models. This is important since modern (binary) classification studies often encompass a large number of

features, which tends to increase the probability of data contamination. Outliers need to be detected and treated appropriately, since they can hinder classical estimation methods. Specifically, we focus on the logistic slippage model, which leads to the exclusion (or trimming) of the most influential cases from the fit, and a "strong" sparsity assumption on the coefficients. To solve such a double combinatorial problem, we rely on state-of-the-art solvers for mixed-integer conic programming which, unlike existing heuristic methods for robust and sparse logistic regression, provide guarantees of optimality even if the algorithm is stopped before convergence. Our proposal, MIProb, provides robust and sparse estimates with an optional ridge-like penalization term.

MIProb, outperforms existing methods in our simulation study. It provides sparser solutions with lower false positive and negative rates for both feature selection and outlier detection while maintaining stronger predictive power under most settings. Moreover, MIProb performs very well in our honey bee overwintering survival application. Based on three years of publicly available data from Pennsylvania beekeepers, it outperforms existing heuristic methods in terms of predictive power, robustness and sparsity of the estimates, and it produces results consistent with previous studies [18]. In particular, we found that weather variables appear to be strong contributors. Winter total precipitation and growing degree days are positively associated with honey bee survival, while the mean temperature of the driest quarter and the precipitation of the warmest quarter show a negative association. Moreover, our results indicate that the lower the exposure to pesticide (i.e., as their distance increases) the higher honey bee survival is. These findings are important in order to understand the main drivers of honey bee loss and highlight the importance of multi-source data to study and predict honey bee overwintering survival.

Our work can be extended in several directions. We are exploring additional and more complex simulation settings (e.g., higher dimensionality, collinear features, etc.). We did not experiment with the ridge-like penalty in the current paper, but this is an important tool and requires further investigation. However, computing time is currently the main bottleneck to more extensive exploration. Thus, in the future , we plan to consider additional modeling strategies that can reduce the computational burden. For instance, developing more suitable big-$\mathcal{M}$ bounds and using outer approximation techniques with dynamic constraint generation and first-order techniques as in [30]. We are also exploring more efficient tuning strategies for the sparsity and trimming levels, as well as the ridge-like parameter, if present. Utilizing approaches such as warm-starts or integrated cross-validation [56] can substantially reduce the computational burden for subsequent runs of the MIP algorithm, and allow better tuning. If the trimming level for MIProb is inflated, a re-weighting approach may also be included in order to increase the efficiency of the estimator as in [10], as well as approaches based on the forward search [45]. However, larger trimming levels might increase the computational burden, and the procedure does not take into account the feature selection process. Thus, the forward search might be combined with diagnostic methods that simultaneously study the effect of outliers and features [78]. Moreover, the theoretical properties of our procedure require further investigation, and its extension to other generalized linear models such as Poisson or multinomial regressions is of great interest.

Source code for the implementation of our procedure and to replicate our simulation and application results is openly available at https://github.com/LucaIns/SFSOD_logreg (accessed on 29 July 2021).

Author Contributions: Conceptualization, L.I.; methodology, L.I. and A.K.; project administration, L.I.; software, L.I. and A.K.; formal analysis, L.I. and A.K.; investigation, L.I. and M.C.; writing—original draft preparation, L.I.; writing—review and editing, L.I., M.C., A.K. and F.C.; supervision, F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the NIH B2D2K training grant and the Huck Institutes of the Life Sciences of Penn State.

Informed Consent Statement: Not applicable.

Data Availability Statement: Reference to honey bee survey data is reported in Section 5.

Acknowledgments: Computations were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. This content is solely the responsibility of the authors and does not necessarily represent the views of the Institute for Computational and Data Sciences.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
2. Cox, D.R.; Snell, E.J. *Analysis of Binary Data*, 2nd ed.; Chapman & Hall: London, UK, 1989.
3. Cramer, J.S. *The Origins of Logistic Regression*; Technical Report 2002-119/4; Tinbergen Institute: Amsterdam, The Netherlands, 2002.
4. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
5. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; John Wiley & Sons: New York, NY, USA, 2006.
6. She, Y.; Owen, A.B. Outlier detection using nonconvex penalized regression. *J. Am. Stat. Assoc.* **2011**, *106*, 626–639. [CrossRef]
7. Insolia, L.; Kenney, A.; Chiaromonte, F.; Felici, G. Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics* **2021**. accepted author manuscript. [CrossRef]
8. Insolia, L.; Chiaromonte, F.; Li, R.; Riani, M. Doubly robust feature selection with mean and variance outlier detection and oracle properties. *arXiv* **2021**, arXiv:2106.11941.
9. Tibshirani, J.; Manning, C.D. Robust logistic regression using shift parameters. *arXiv* **2013**, arXiv:1305.4987.
10. Kurnaz, F.S.; Hoffmann, I.; Filzmoser, P. Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemom. Intell. Lab. Syst.* **2017**, *172*, 211–222. [CrossRef]
11. Bertsimas, D.; King, A.; Mazumder, R. Best subset selection via a modern optimization lens. *Ann. Stat.* **2016**, *44*, 813–852. [CrossRef]
12. Schrijver, A. *Theory of Linear and Integer Programming*; John Wiley & Sons: New York, NY, USA, 1986.
13. Zhang, C.H.; Zhang, T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Stat. Sci.* **2012**, *27*, 576–593. [CrossRef]
14. Bedrick, E.J.; Hill, J.R. Outlier tests for logistic regression: A conditional approach. *Biometrika* **1990**, *77*, 815–827. [CrossRef]
15. Seeley, T.D.; Visscher, P.K. Survival of honeybees in cold climates: The critical timing of colony growth and reproduction. *Ecol. Entomol.* **1985**, *10*, 81–88. [CrossRef]
16. Döke, M.A.; Frazier, M.; Grozinger, C.M. Overwintering honey bees: Biology and management. *Curr. Opin. Insect Sci.* **2015**, *10*, 185–193. [CrossRef]
17. Beyer, M.; Junk, J.; Eickermann, M.; Clermont, A.; Kraus, F.; Georges, C.; Reichart, A.; Hoffmann, L. Winter honey bee colony losses, Varroa destructor control strategies, and the role of weather conditions: Results from a survey among beekeepers. *Res. Vet. Sci.* **2018**, *118*, 52–60. [CrossRef]
18. Calovi, M.; Grozinger, C.M.; Miller, D.A.; Goslee, S.C. Summer weather conditions influence winter survival of honey bees (*Apis mellifera*) in the northeastern United States. *Sci. Rep.* **2021**, *11*, 1553. [CrossRef]
19. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10. [CrossRef]
20. Santner, T.J.; Duffy, D.E. A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1986**, *73*, 755–758. [CrossRef]
21. Fahrmeir, L.; Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Stat.* **1985**, *13*, 342–368. [CrossRef]
22. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
23. Duffy, D.E.; Santner, T.J. On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Commun. Stat.-Theory Methods* **1989**, *18*, 959–980. [CrossRef]
24. Le Cessie, S.; Van Houwelingen, J.C. Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C* **1992**, *41*, 191–201. [CrossRef]
25. Koh, K.; Kim, S.J.; Boyd, S. An interior-point method for large-scale ℓ_1-regularized logistic regression. *J. Mach. Learn. Res.* **2007**, *8*, 1519–1555.
26. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
27. Algamal, Z.Y.; Lee, M.H. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Comput. Biol. Med.* **2015**, *67*, 136–145. [CrossRef]
28. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
29. Miller, A.J. *Subset Selection in Regression*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2002.
30. Bertsimas, D.; King, A. Logistic regression: From art to science. *Stat. Sci.* **2017**, *32*, 367–384. [CrossRef]

31. Shen, X.; Pan, W.; Zhu, Y. Likelihood-based selection and sharp parameter estimation. *J. Am. Stat. Assoc.* **2012**, *107*, 223–232. [CrossRef]
32. Copas, J.B. Binary regression models for contaminated data. *J. R. Stat. Soc. Ser. B* **1988**, *50*, 225–253. [CrossRef]
33. Imon, A.R.; Hadi, A.S. Identification of multiple outliers in logistic regression. *Commun. Stat.-Theory Methods* **2008**, *37*, 1697–1709. [CrossRef]
34. Landwehr, J.M.; Pregibon, D.; Shoemaker, A.C. Graphical methods for assessing logistic regression models. *J. Am. Stat. Assoc.* **1984**, *79*, 61–71. [CrossRef]
35. Pregibon, D. Logistic regression diagnostics. *Ann. Stat.* **1981**, *9*, 705–724. [CrossRef]
36. Carroll, R.J.; Pederson, S. On robustness in the logistic regression model. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 693–706. [CrossRef]
37. Rousseeuw, P.J.; Christmann, A. Robustness against separation and outliers in logistic regression. *Comput. Stat. Data Anal.* **2003**, *43*, 315–332. [CrossRef]
38. Bianco, A.M.; Yohai, V.J. Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*; Rieder, H., Ed.; Springer: New York, NY, USA, 1996; pp. 17–34.
39. Künsch, H.R.; Stefanski, L.A.; Carroll, R.J. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Am. Stat. Assoc.* **1989**, *84*, 460–466.
40. Croux, C.; Haesbroeck, G. Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Stat. Data Anal.* **2003**, *44*, 273–295. [CrossRef]
41. Gervini, D. Robust adaptive estimators for binary regression models. *J. Stat. Plan. Inference* **2005**, *131*, 297–311. [CrossRef]
42. Shafieezadeh-Abadeh, S.; Esfahani, P.M.; Kuhn, D. Distributionally robust logistic regression. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1576–1584.
43. Beckman, R.J.; Cook, R.D. Outlier ..... s. *Technometrics* **1983**, *25*, 119–149.
44. Atkinson, A.C.; Riani, M. *Robust Diagnostic Regression Analysis*; Springer: New York, NY, USA, 2000.
45. Atkinson, A.C.; Riani, M. Regression diagnostics for binomial data from the forward search. *J. R. Stat. Soc. Ser. D* **2001**, *50*, 63–78. [CrossRef]
46. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; John Wiley & Sons: New York, NY, USA, 1987.
47. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv* **2017**, arXiv:1707.08692.
48. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **1995**, *37*, 373–384. [CrossRef]
49. Bernholt, T. *Robust Estimators Are Hard to Compute*; Technical Report 52/2005; University of Dortmund: Dortmund, Germany, 2006.
50. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
51. Donoho, D.L.; Huber, P.J. The notion of breakdown point. In *A festschrift for Erich L. Lehmann*; Bickel, P., Doksum, K.A., Hodges, J.L., Eds.; Wadsworth: Belmont, CA, USA, 1983; pp. 157–184.
52. Christmann, A. Least median of weighted squares in logistic regression with large strata. *Biometrika* **1994**, *81*, 413–417. [CrossRef]
53. Croux, C.; Flandre, C.; Haesbroeck, G. The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Stat. Probab. Lett.* **2002**, *60*, 377–386. [CrossRef]
54. Müller, C.H.; Neykov, N. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *J. Stat. Plan. Inference* **2003**, *116*, 503–519. [CrossRef]
55. Hadi, A.S.; Luceño, A. Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Comput. Stat. Data Anal.* **1997**, *25*, 251–272. [CrossRef]
56. Kenney, A.; Chiaromonte, F.; Felici, G. MIP-BOOST: Efficient and Effective L_0 Feature Selection for Linear Regression. *J. Comput. Graph. Stat.* **2021**, 1–12. [CrossRef]
57. Kurnaz, F.S.; Hoffmann, I.; Filzmoser, P. *enetLTS: Robust and Sparse Methods for High Dimensional Linear and Logistic Regression*. R Package Version 0.1.0. 2018. Available online: https://CRAN.R-project.org/package=enetLTS (accessed on 15 July 2021).
58. Calderone, N.W. Insect Pollinated Crops, Insect Pollinators and US Agriculture: Trend Analysis of Aggregate Data for the Period 1992–2009. *PLoS ONE* **2012**, *7*, e37235. [CrossRef]
59. Chopra, S.S.; Bakshi, B.R.; Khanna, V. Economic Dependence of U.S. Industrial Sectors on Animal-Mediated Pollination Service. *Environ. Sci. Technol.* **2015**, *49*, 14441–14451. [CrossRef]
60. Morse, R.A.; Calderone, N.W. The value of honey bees as pollinators of US crops in 2000. *Bee Cult.* **2000**, *128*, 1–15.
61. Becher, M.A.; Osborne, J.L.; Thorbek, P.; Kennedy, P.J.; Grimm, V. Towards a systems approach for understanding honeybee decline: A stocktaking and synthesis of existing models. *J. Appl. Ecol.* **2013**, *50*, 868–880. [CrossRef]
62. Pettis, J.S.; Delaplane, K.S. Coordinated responses to honey bee decline in the USA. *Apidologie* **2010**, *41*, 256–263. [CrossRef]
63. Potts, S.G.; Roberts, S.P.; Dean, R.; Marris, G.; Brown, M.A.; Jones, R.; Neumann, P.; Settele, J. Declines of managed honey bees and beekeepers in Europe. *J. Apic. Res.* **2010**, *49*, 15–22. [CrossRef]
64. Oldroyd, B.P.; Nanork, P. Conservation of Asian honey bees. *Apidologie* **2009**, *40*, 296–312. [CrossRef]
65. Ellis, J.D.; Evans, J.D.; Pettis, J. Colony losses, managed colony population decline, and Colony Collapse Disorder in the United States. *J. Apic. Res.* **2010**, *49*, 134–136. [CrossRef]

66. van Dooremalen, C.; Gerritsen, L.; Cornelissen, B.; van der Steen, J.J.M.; van Langevelde, F.; Blacquière, T. Winter survival of individual honey bees and honey bee colonies depends on level of Varroa destructor infestation. *PLoS ONE* **2012**, *7*, e36285. [CrossRef] [PubMed]
67. Morawetz, L.; Köglberger, H.; Griesbacher, A.; Derakhshifar, I.; Crailsheim, K.; Brodschneider, R.; Moosbeckhofer, R. Health status of honey bee colonies (*Apis mellifera*) and disease-related risk factors for colony losses in Austria. *PLoS ONE* **2019**, *14*, e0219293. [CrossRef]
68. Genersch, E.; Von Der Ohe, W.; Kaatz, H.; Schroeder, A.; Otten, C.; Büchler, R.; Berg, S.; Ritter, W.; Mühlen, W.; Gisder, S.; et al. The German bee monitoring project: A long term study to understand periodically high winter losses of honey bee colonies. *Apidologie* **2010**, *41*, 332–352. [CrossRef]
69. Yasrebi-de Kom, I.A.R.; Biesmeijer, J.C.; Aguirre-Gutiérrez, J. Risk of potential pesticide use to honeybee and bumblebee survival and distribution: A country-wide analysis for The Netherlands. *Divers. Distrib.* **2019**, *25*, 1709–1720. [CrossRef]
70. Switanek, M.; Crailsheim, K.; Truhetz, H.; Brodschneider, R. Modelling seasonal effects of temperature and precipitation on honey bee winter mortality in a temperate climate. *Sci. Total Environ.* **2017**, *579*, 1581–1587. [CrossRef]
71. Stout, J.C.; Morales, C.L. Ecological impacts of invasive alien species on bees. *Apidologie* **2009**, *40*, 388–409. [CrossRef]
72. vanEngelsdorp, D.; Meixner, M.D. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *J. Invertebr. Pathol.* **2010**, *103*, S80–S95. [CrossRef]
73. Steinhauer, N.; Rennich, K.; Wilson, M.; Caron, D.; Lengerich, E.; Pettis, J.; Rose, R.; Skinner, J.; Tarpy, D.; Wilkes, J.; et al. A national survey of managed honey bee 2012–2013 annual colony losses in the USA: Results from the Bee Informed Partnership. *J. Apic. Res.* **2014**, *53*, 1–18. [CrossRef]
74. Bruckner, S.; Nathalie, S.; Jonathan, E.; Anne Marie, F.; Kelly, K.; Eric, M.; Annette, M.; Meghan, M.; Elina, N.; Juliana, R.; et al. 2019–2020 Honey Bee Colony Losses in the United States: Preliminary Results. **2020**, *579*, 1581–1587. unpublished work. Available online: https://beeinformed.org/wp-content/uploads/2021/06/BIP_2020_21_Losses_Abstract_2021.06.14_FINAL_R1.pdf (accessed on 15 July 2021).
75. Steinhauer, N.; Aurell, D.; Bruckner, S.; Wilson, M.; Rennich, K.; vanEngelsdorp, D.; Williams, G. United States Honey Bee Colony Losses 2020–2021: Preliminary Results. 2021, unpublished report. Available online: https://beeinformed.org/2020/06/22/preliminary-results-of-the-2019-2020-national-honey-bee-colony-loss-survey/ (accessed on 15 July 2021).
76. Bruckner, S.; Steinhauer, N.; Engelsma, J.; Fauvel, A.M.; Kulhanek, K.; Malcom, E.; Meredith, A.; Milbrath, M.; Niño, E.; Rangel, J.; et al. 2019–2020 Honey Bee Colony Losses in the United States: Preliminary Results. Bee Informed Partnership. 2020. Available online: https://beeinformed.org/wp-content/uploads/2020/06/BIP_2019_2020_Losses_Abstract.pdf (accessed on 15 July 2021).
77. Douglas, M.R.; Sponsler, D.B.; Lonsdorf, E.V.; Grozinger, C.M. County-level analysis reveals a rapidly shifting landscape of insecticide hazard to honey bees (*Apis mellifera*) on US farmland. *Sci. Rep.* **2020**, *10*, 797. [CrossRef] [PubMed]
78. Menjoge, R.S.; Welsch, R.E. A diagnostic method for simultaneous feature selection and outlier identification in linear regression. *Comput. Stat. Data Anal.* **2010**, *54*, 3181–3193. [CrossRef]