iSpy the Correct Features: Best Subset Selection with LASSO and Elastic-Net Regularization

Ameya Shere
New York University
as12366@nyu.edu

Neil Menghani
New York University
nlm326@nyu.edu

1 Introduction

1.1 Feature Selection

One common task in the field of data science is selecting the best subset from a given set of features. Let us consider a dataset of n samples represented as M-dimensional random vectors, where each component of these vectors is called a "feature." Each of these samples is associated with some output value that has been calculated using some unknown function of the features. It would be reasonable to assume, especially if M is large, that not all of these features are equally weighted in their contribution to the output. If one feature is highly correlated with another feature, then, intuitively, this second feature does not contribute much more to the calculation of the output beyond the contribution of the first feature. It may be useful to separate the "important" features from the "unimportant" ones: this is the task of feature selection. In other words, feature selection aims to find a subset of features that together can fully or most closely specify the output value without the rest of the features.

More formally, we can define the feature selection problem as follows. Consider i.i.d. pairs $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_i, Y_i), ..., (\mathbf{X}_n, Y_n)$, where $(\mathbf{X}_i, Y_i) \in \mathbb{R}^M \times \mathbb{R}$ or $(\mathbf{X}_i, Y_i) \in \mathbb{R}^M \times \{0, 1\}$, and these pairs are distributed as $(\mathbf{X}, Y)$ according to some unknown distribution with probability measure $\mathbb{P}$. The goal of feature selection is to find unknown "true subset" $I^* \subseteq \{1, ..., M\}$ such that the regression function $r(x) := \mathbb{E}[Y|\mathbf{X} = x] = g(\sum_{j \in I^*} \beta_j^* x_j)$, where g is a known link function and the β_j^* are the coefficients of the "true model" which calculates $r(x)$ based only on the features whose indices are in I^*. One can consider different types of models for calculating $r(x)$ by varying the link function g. For the choice $g(z) = z$, the regression function defines a linear regression model: $r(x) = \sum_{j \in I^*} \beta_j^* x_j$. For the choice $g(z) = \sigma(z) = \frac{1}{1+e^{-z}}$ (i.e., the logistic function), the regression function defines a logistic regression model: $r(x) = \sigma(\sum_{j \in I^*} \beta_j^* x_j)$ (Bunea, 2008).

1.2 Regularization

Let β^* denote the vector that contains β_j^* for indices $j \in I^*$, and 0 for all other indices, and let $\hat{\beta}$ be an estimator of β^*. In regression, the most accurate model is defined by the $\hat{\beta}$ that minimizes the risk $\mathcal{R}_{\beta^*}(\hat{\beta}) = \mathbb{E}[l(\beta^*, \hat{\beta})]$, where l denotes the loss function. Choosing l to be the squared loss (i.e., $l = (\beta^* - \hat{\beta})^2$), the risk becomes the mean-squared error, which can be decomposed according to the Bias-Variance decomposition: $\mathcal{R}_{\beta^*} = \mathbb{E}[(\beta^* - \hat{\beta})^2] = \text{Var}_{\beta^*}(\hat{\beta}) + (\mathbb{E}_{\beta^*}[\beta^* - \hat{\beta}])^2$.

It is very uncommon for any estimator to have low bias and low variance. Rather, the relationship between these two quantities is more of a tradeoff in that a decrease in one means an increase in the other. For this reason, when estimators have very low bias, they often will have high variance, which is symptomatic of a common pitfall of regression models called overfitting. A model is overfitted if it fits the given data too exactly and thus will not generalize well to other samples.

One way to alleviate this problem is through regularization. The idea behind regularization is to strategically introduce bias into a model in order to reduce the amount of variance. This bias is added via a penalty term in the estimator $\hat{\beta}$. For example, if we take $\hat{\beta}$ to be the least-squares estimator, regularization in the l_2 norm, also known as ridge regression, would yield estimator $\hat{\beta}_\lambda = \text{argmin}_{\beta\in\mathbb{R}^M}\frac{1}{n}\sum_{i=1}^{n}(\beta^\top X_i - Y_i)^2 + \lambda||\beta||_2$ for $\lambda > 0$, where $\lambda||\beta||_2$ is the penalty term. The penalty term varies depending on the type of regularization. LASSO regularization, which is regularization in the l_1 norm, has penalty term $\lambda||\beta||_1$ for $\lambda > 0$. Elastic-net regularization, which is regularization in both the l_1 and l_2 norms, has penalty term $\lambda_1||\beta||_1 + \lambda_2||\beta||_2$ for $\lambda_1, \lambda_2 > 0$. These two regularization techniques can be applied to the linear and logistic regression models to improve the accuracy of feature selection (Bunea, 2008).

2 Theoretical Foundation

2.1 Summary of Bunea's paper

In the paper, "Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization," Florentina Bunea investigates feature selection using the l_1- and $l_1 + l_2$-penalized linear and logistic regression models. The paper first establishes identifiability and stability conditions on the design matrix (i.e., the matrix of dimension $n \times M$ containing the samples $\mathbf{X}_i$ as rows) for use in the results of Sections 2 and 3. Using these conditions, the paper examines the four models. For the l_1-penalized models, the penalty coefficient is $\lambda = 2r(\delta, n, M)$, where δ is a user-specified parameter. For $l_1 + l_2$-penalized models, the penalty coefficients are $\lambda_1 = 2r(\delta, n, M)$ and $\lambda_2 = c(r, B)$, where $\max_{j\in I^*}|\beta_j^*| \leq B$.

In Section 2, for each of these models, the paper proves upper bounds in probability on the l_1 difference between the model's estimator and β^*. These bounds are of the form Crk^*(for constant $C > 0$) on the sparse l_1 balls $|\hat{\beta} - \beta^*|_1$, where $\hat{\beta}$ is the linear regression estimator for β^*, and $|\tilde{\beta} - \beta^*|_1$, where $\tilde{\beta}$ is the logistic regression estimator for β^*. Note that the balls are sparse because k^* will often be of much lower dimension than M (i.e., many of the components of β_i^* for $1 \leq i \leq M$ will be 0). The idea behind these upper bounds is to show the order of the error between the estimator and the vector of true coefficients in the l_1 norm (Bunea, 2008).

In Section 3, the paper first shows that for each of the four models, the empirical subset $\hat{I}$ will contain all the indices in I^* with high probability. Specifically, this will occur for varying signal strengths. The signal of the data is constrained by lower bounds on $\min_{j\in I^*}|\beta_j^*|$; in other words, the larger the value of a true coefficient is, the stronger its signal. The paper first shows that correct inclusion of the true features occurs for moderate signals where $\min_{j\in I^*}|\beta_j^*| \geq Crk^*$ for constant $C > 0$. Under stronger conditions, the correct inclusion occurs for even weaker signals, where $\min_{j\in I^*}|\beta_j^*| \geq 2r$ in the linear models, and $\min_{j\in I^*}|\beta_j^*| \geq 3.5r + 3(1+\frac{1}{r})\epsilon \vee 3.5r + (1+\frac{1}{r})\epsilon$ in the logistic models, for r and ϵ as given in Propositions 3.3 and 3.4, respectively. The data will always contain unavoidable fluctuations of order $\frac{1}{\sqrt{n}}$ under which any signal will be obscured, but weak signals above this level will be detectable (Bunea, 2008).

Finally, the paper gives confidences in terms of δ and M for the probability of correct subset selection in each of the four models. Bunea also shows that the estimator $\hat{I}$ in each model is an asymptotically consistent estimator of I^*. The results of the paper indicate that the subsets selected by these four models can recover coefficients of order $O(\sqrt{\frac{\log n}{n}})$, for M polynomial in n and for $\delta = \frac{1}{n}$, and that the subsets selected are the correct subsets with high probability (Bunea, 2008).

Though the paper argues this similar result for all four models, it does discuss conditions where the elastic-net model is more beneficial than the LASSO model. Theorem 2.2 states that for the LASSO linear model, $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \leq \frac{4}{b}rk^*) \geq 1 - \delta$ for $0 < b \leq 1$ and the given r. If the *Condition Stabil* defined in the paper is only satisfied for small values of b, the error is asymptotically unbounded as $b \to 0$ (Bunea, 2008). Intuitively, this means that the LASSO linear model is not reliable for feature selection if the correlation matrix is "unstable." This notion of "instability" refers to the case when a small decrease in the diagonal elements of the correlation matrix that correspond to the true features results in a non-semi-positive-definite matrix (Bunea, 2008). To better understand this, consider a highly correlated matrix. By definition, a correlation matrix is symmetric and diagonally-dominant, and thus symmetric positive semi-definite. In a highly correlated matrix, if we lower an element on the diagonal (which is initially at 1), then the matrix may no longer be diagonally-dominant, because high correlation values elsewhere in that row may become greater than the diagonal value. If *Condition Stabil* only holds for values of b close to 0 (note that b is the amount by which we decrease diagonal elements), then the matrix is highly correlated. Thus, the result in Theorem 2.2 intuitively means that the LASSO linear model does not calculate accurate coefficients for highly correlated matrices.

However, for the elastic-net linear model, Theorem 2.3 says that $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \leq \frac{4.25}{b+c} rk^*) \geq 1 - \delta$, for $0 < b \leq 1$, $c = \frac{r}{2B}$, and the given r. Since $0 < \delta < 1$ and M is a natural number, $r > 0$ as defined in Theorem 2.3. This means that $c > 0$, so even if *Condition Stabil* only holds for small values of b, the error of the estimator still has a finite upper bound, though, depending on the value of c, this bound may be loose (Bunea, 2008). Similarly, for the LASSO logistic model, Theorem 2.4 states that $\mathbb{P}(|\tilde{\beta} - \beta^*|_1 \leq \frac{4}{sb} rk^* + (1 + \frac{1}{r})\epsilon) \geq 1 - \delta$ for $0 < b \leq 1$, $s = (1 + e^{6LD})^{-4}$ and the given $0 < |\beta^*|_1 \leq D$, L, r, and ϵ. Since $(1 + \frac{1}{r})\epsilon \approx \frac{n}{2^{M \vee n}}$ is asymptotically negligible, the error of the estimator is again unbounded in limit as $b, s \to 0$. Intuitively, this means that when the correlation matrix is "unstable" and $|\beta^*|_1$ and the $X_{i,j}$'s are large, the LASSO logistic model is not reliable for feature selection (Bunea, 2008). However, for the elastic-net logistic model, Theorem 2.7 shows that again, since $c > 0$, there will be a finite bound on the error even if s and b are close to 0 (Bunea, 2008). Thus, the results of the paper indicate that feature selection with the $l_1 + l_2$-penalized model performs better than feature selection with the l_1-penalized model for highly correlated matrices, but in general, these models will recover coefficients of the same order and select the correct subset with high probability.

3 Experiment

3.1 Generating data

In this experiment, we generate data and implement linear and logistic regression with l_1 and $l_1 + l_2$ regularization to determine how well these models are able to detect the best subset of features for the hyperparameter bounds suggested in Bunea's paper (Menghani and Shere, 2021). We examine how feature selection for all four models is impacted by different sample sizes and different levels of correlation in the extraneous features. In each iteration of the experiment, we generate an $n \times M$ design matrix $\mathbf{X}$, an M-vector of true coefficients β^*, and an n-vector y representing the response. Then, we implement the four models to determine how well each of them can select the correct features as well as how closely the coefficients match the true coefficients of the data.

To build the simulated dataset $\mathbf{X}$, we first fix n (number of samples), M (total number of features), k^* (number of correct features), B (upper bound in absolute value on the maximum component of the true coefficients β^*), δ ($1-$ the confidence level), γ (parameter representing the proportion of the data in extraneous columns consisting of noise). The n and γ values will vary by iteration, but we fix $M = 10$, $B = 50$, $k^* = 5$, and $\delta = 0.05$ for all iterations. We then generate k^* correct features from a Gaussian distribution with mean 0 and variance 1. From these features we generate the remaining $(M - k^*)$ incorrect features as a random linear combination of the correct features with the proportion γ of the data being replaced by Gaussian noise. Note that the I^* can be inferred directly from k^*, as we operate under the assumption that the first k^* features in $\mathbf{X}$ are the correct features and all other features after that are incorrect features. Next, we center and normalize the data, an essential assumption of the paper (Bunea, 2008).

Once we have generated $\mathbf{X}$, we generate some true coefficients β^*, from which we can generate

y. To generate β^*, we use k^* coefficients from a uniform distribution over $[-B, B]$ and pad the coefficient vector with $(M - k^*)$ zeros to represent no contribution from the remaining features. Then we can generate y as a known function of $\mathbf{X}$ using $y = \mathbf{X}\beta^* + \epsilon$, where ϵ represents some Gaussian noise on the order of $\frac{1}{\sqrt{n}}$. For the logistic regression, we apply the sigmoid function to y and apply a threshold of 0.5 to make y a binary vector.

3.2 Iterating over n and γ

We perform the experiment on four models: linear and logistic regression with l_1 regularization (LASSO) and $l_1 + l_2$ regularization (elastic-net). For each model, we iterate over values of n (sample size) from 100 to 10,000 and values of γ (noise proportion) from 0 to 0.9. For each iteration, we perform the process described in Subsection 3.1 to generate $\mathbf{X}$, β^*, and y. Then, we fit the corresponding model to $\mathbf{X}$ and y using regularization parameters based on the lower bound suggested in Bunea. For linear regression, we set the l_1 regularization parameter as $2r$, where $r = 8L\frac{\log\frac{4M}{\delta}}{n}$, and the l_2 regularization parameter as c, where $c = \frac{r}{2B}$ (Bunea, 2008). For logistic regression, we set the l_1 regularization parameter as $2r$, where $r = (6+4\sqrt{2})L\sqrt{\frac{2\log 2(M)}{n}} + 2L\sqrt{\frac{2\log\frac{2M}{\delta}}{n}} + \frac{1}{4n}$, and the l_2 regularization parameter as c, where $c = \frac{r}{2B}$ (Bunea, 2008). Based on the $\hat{\beta}$ returned, we determine how many of the k^* true features the model selects, how many of the $(M - k^*)$ false features the model selects, and the l_1- and l_2-norm of the difference between the fitted coefficients $\hat{\beta}$ and the true coefficients β^*.

4 Results

4.1 Linear regression models

The results indicate that for both the LASSO and elastic-net linear regression models, as the number of samples increases, the subset selection improves. In both 1 and 2, we observe that the sample size does not seem to affect the number of true features selected, which concentrates mostly around 4. However, we see that the number of wrong features selected by the model concentrates slightly more closely around 1 as the sample size gets larger.

In addition, we observe that as n increases, the l_1 and l_2 error of the coefficients, $|\beta^* - \hat{\beta}|_1$, and $|\beta^* - \hat{\beta}|_2$ respectively, both decrease. This result indicates that not only does the accuracy of the subset selection improve with sample size, but the values of the coefficients calculated by the models also approach the true coefficient values.

Unlike the sample size results, the results in Figures 3 and 4 indicate that changing the strength of the correlation in the design matrix does not have an effect on either the subset selection performance or the error in the coefficients.

4.2 Logistic regression models

In the logistic regression models, the results are quite different. For the logistic LASSO model, Figures 5 and 7 indicate that neither the sample size nor the strength of the correlation in the design matrix have any effect on the subset selection or the error in the coefficients. In fact, 0 features are selected no matter what, and the model shrinks all coefficients to 0.

For the logistic elastic-net model, these two variables do have an effect. In 6, we see that as the sample size increases, the number of true features selected tends to concentrate around higher and higher values. However, the number of features wrongly selected also increases with sample size. This result indicates that overall subset selection performance worsens as sample size increases. On the other hand, Figure 8 shows that as the correlation in the design matrix decreases, fewer wrong features are selected, which indicates an improvement in subset selection.

Interestingly, though, neither the sample size nor the level of correlation in the design matrix seem to have any effect on the error between the coefficients computed by the model and the true coefficients. This result implies that while subset selection may be close to correct, the actual values of the coefficients computed by the model are wrong.

5 Discussion

The results for the linear regression models lend some support to Corollary 3.6 from the paper that $\hat{I}$ is an asymptotically consistent estimator of I^* (Bunea, 2008). Since we observe a slight improvement in the accuracy of the subset selection as the sample size increases, it is plausible that very large sample sizes will have highly accurate feature selection results.

The results for the logistic regression models also partially support the theoretical findings from the paper. The paper states that $l_1 + l_2$-regularized models perform better for feature selection on highly correlated data matrices than do l_1-regularized models. The results from our experiment show that this is partly true, because the logistic elastic-net model selects a much more accurate subset than the logistic LASSO model (which selects 0 features) in the highly correlated $\gamma = 0$ case.

In addition, the bound given in Theorem 2.7 on $|\tilde{\beta} - \beta^*|_1$ indicates that as the value of b increases (and if s is constant), then the coefficients calculated by the model get closer to the true coefficients. Note that an increase in b means that *Condition Stabil* is satisfied for values of b further from 0, which means the design matrix is less highly correlated.

In our experiment, s, which is dependent on L and D, is close to the same value most of the time. The results for the logistic elastic-net model do show that as we decrease the level of correlation in the design matrix, the accuracy of subset selection improves. However, from the paper, one would expect that the l_1 error between the calculated and true coefficients should also decrease as the level of correlation decreases. Counterintuitively, our findings do not affirm this hypothesis, because we found that the level of correlation in the matrix has no effect on the error in the coefficients.

It is not clear why this last result occurs. One possible explanation is that the amount of data used in our experiment is not at the scale required for the theoretical results from the paper to manifest in full. In fact, simply using more data could resolve many of the discrepancies we found between our experiment and the paper. For example, though subset selection was close to accurate for a few of the models for large sample sizes, no model correctly identified all of the true feature columns and none of the non-true feature columns. For this reason, we were not able to verify the probability bounds given in the paper for feature selection in either model or the bounds given on the size of the signal that can be detected in the coefficients. If future work in this area can repeat these experiments with more samples, it would be interesting to see if the results more closely match the findings in the paper.

One final note is that although the paper discusses how its results might be useful in the high-dimensional statistical regime, we unfortunately did not have enough compute power to run experiments on design matrices with large n and, on top of that, with $M >> n$. For this reason, we had to run our experiments in the low-dimensional regime. However, we would encourage anyone doing future research on this topic to attempt these experiments in the high-dimensional regime to verify if the results discussed in the paper still hold true.

6 Acknowledgments

We would like to acknowledge Professor Jonathan Niles-Weed for his feedback and support.

References

Florentina Bunea. 2008. Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194.

Neil Menghani and Ameya Shere. 2021. Codebase for 'best subset selection' paper. https://colab.research.google.com/drive/1wWjtMBoWYNFLLVc9MqdU4oL9gH01n8UG#scrollTo=nfLolhCjyEsU.

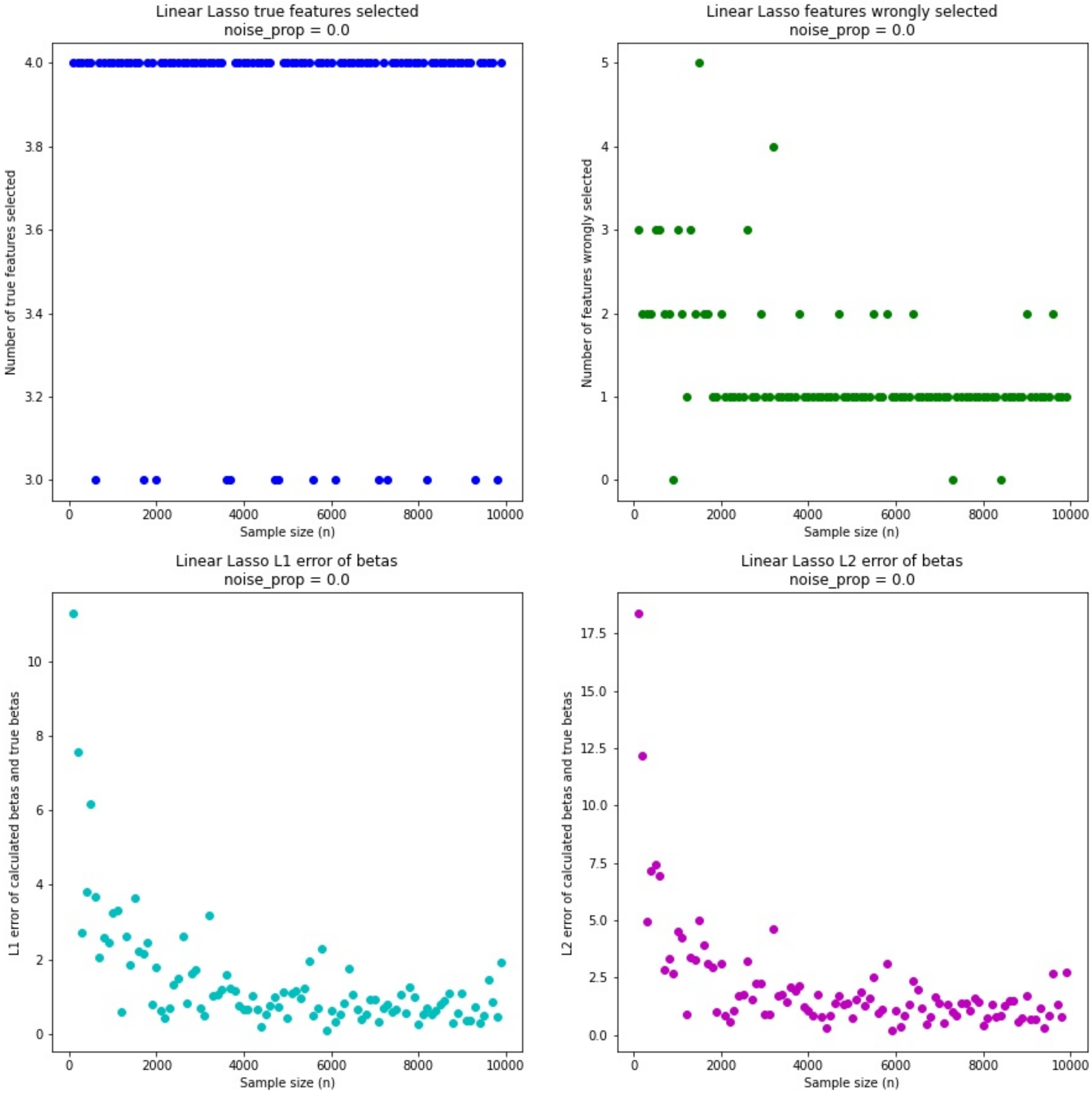

Figure 1: Plots for the linear regression with l_1 regularization shown. Top left: how many true features are identified as a function of sample size. Top right: how many incorrect features are identified as a function of sample size. Bottom left: L_1 error of the coefficients as a function of sample size. Bottom right: L_2 error of the coefficients as a function of sample size.

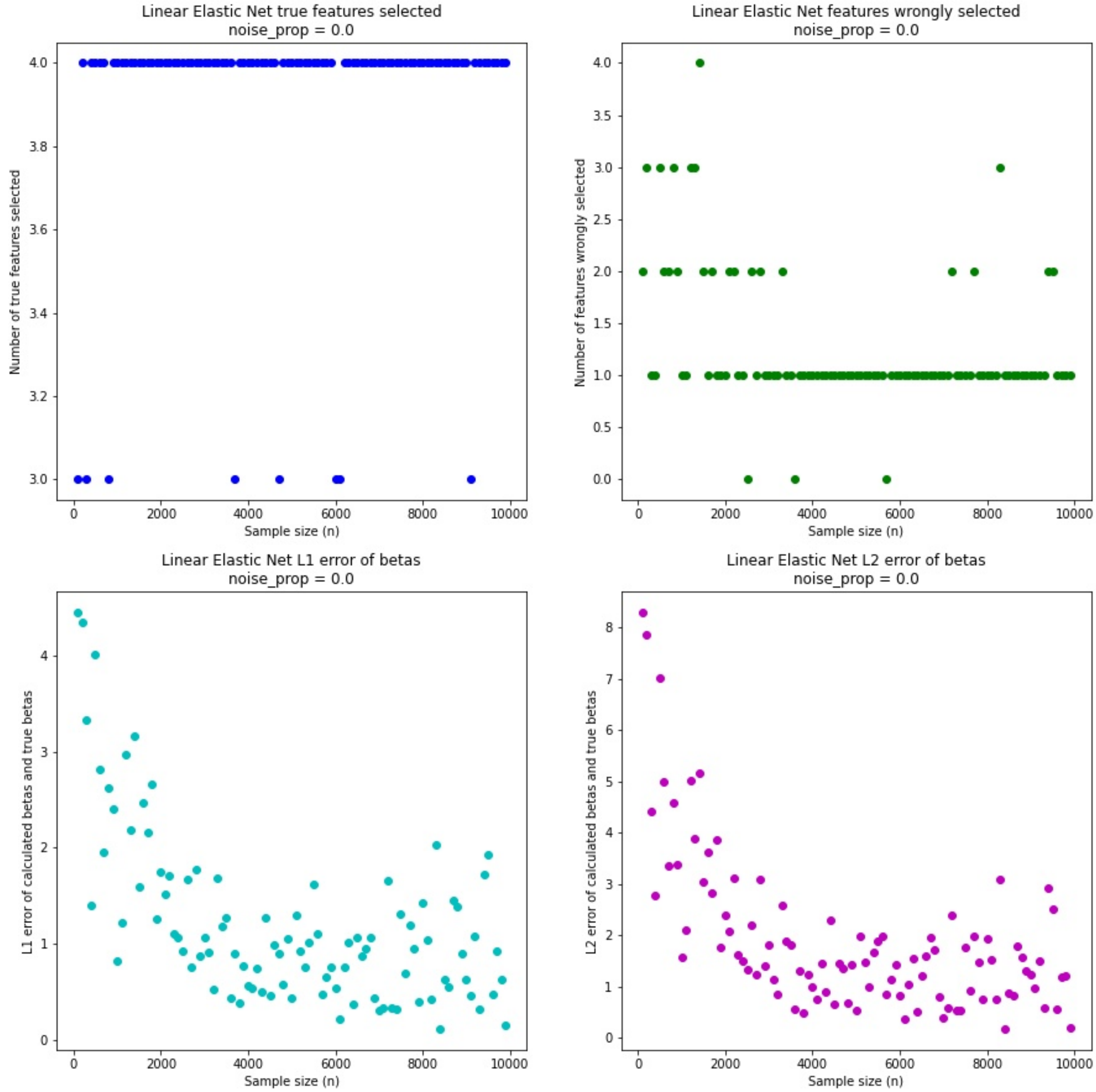

Figure 2: Plots for the linear regression with $l_1 + l_2$ regularization shown. Top left: how many true features are identified as a function of sample size. Top right: how many incorrect features are identified as a function of sample size. Bottom left: L_1 error of the coefficients as a function of sample size. Bottom right: L_2 error of the coefficients as a function of sample size.

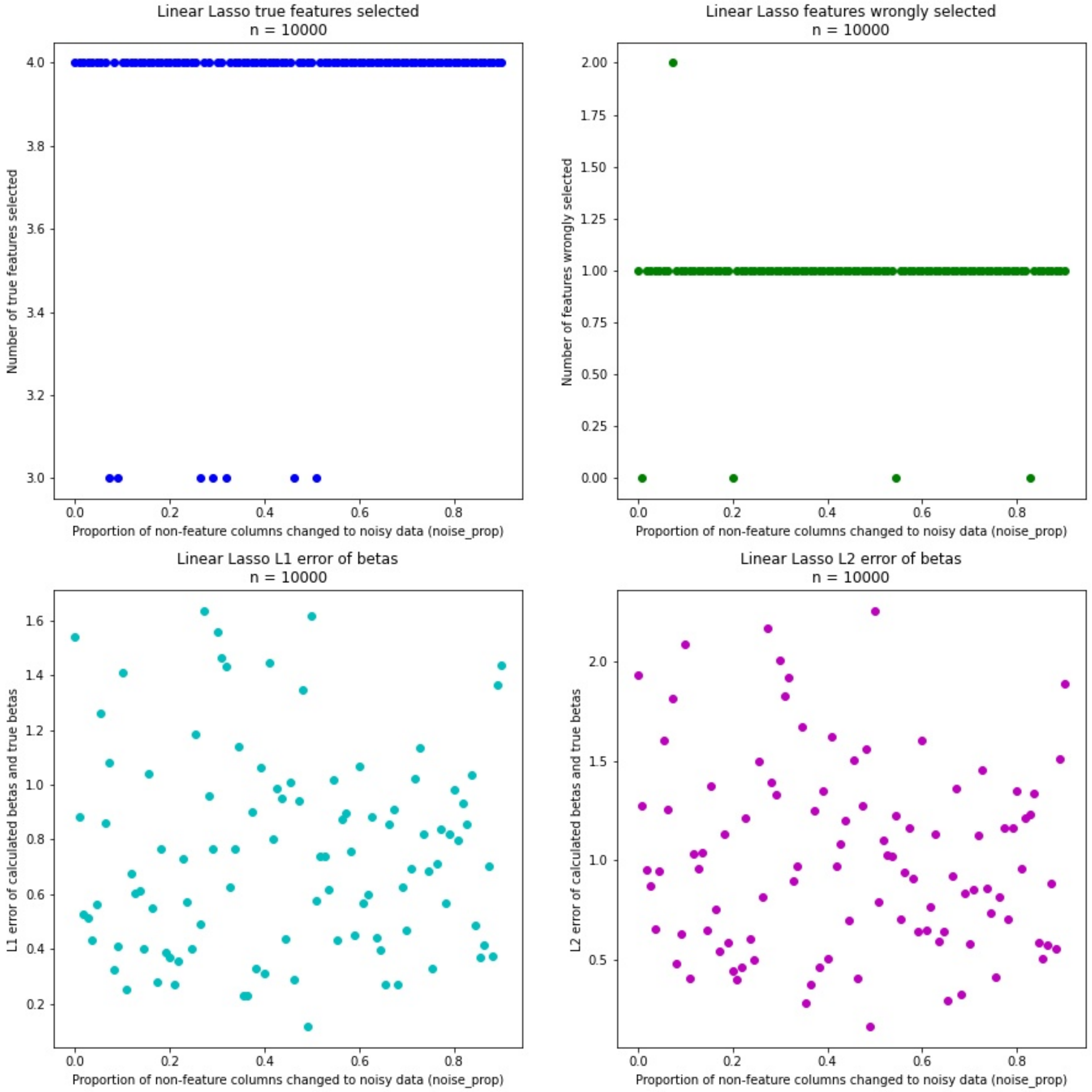

Figure 3: Plots for the linear regression with l_1 regularization shown. Top left: how many true features are identified as a function of noise proportion. Top right: how many incorrect features are identified as a function of noise proportion. Bottom left: L_1 error of the coefficients as a function of noise proportion. Bottom right: L_2 error of the coefficients as a function of noise proportion.

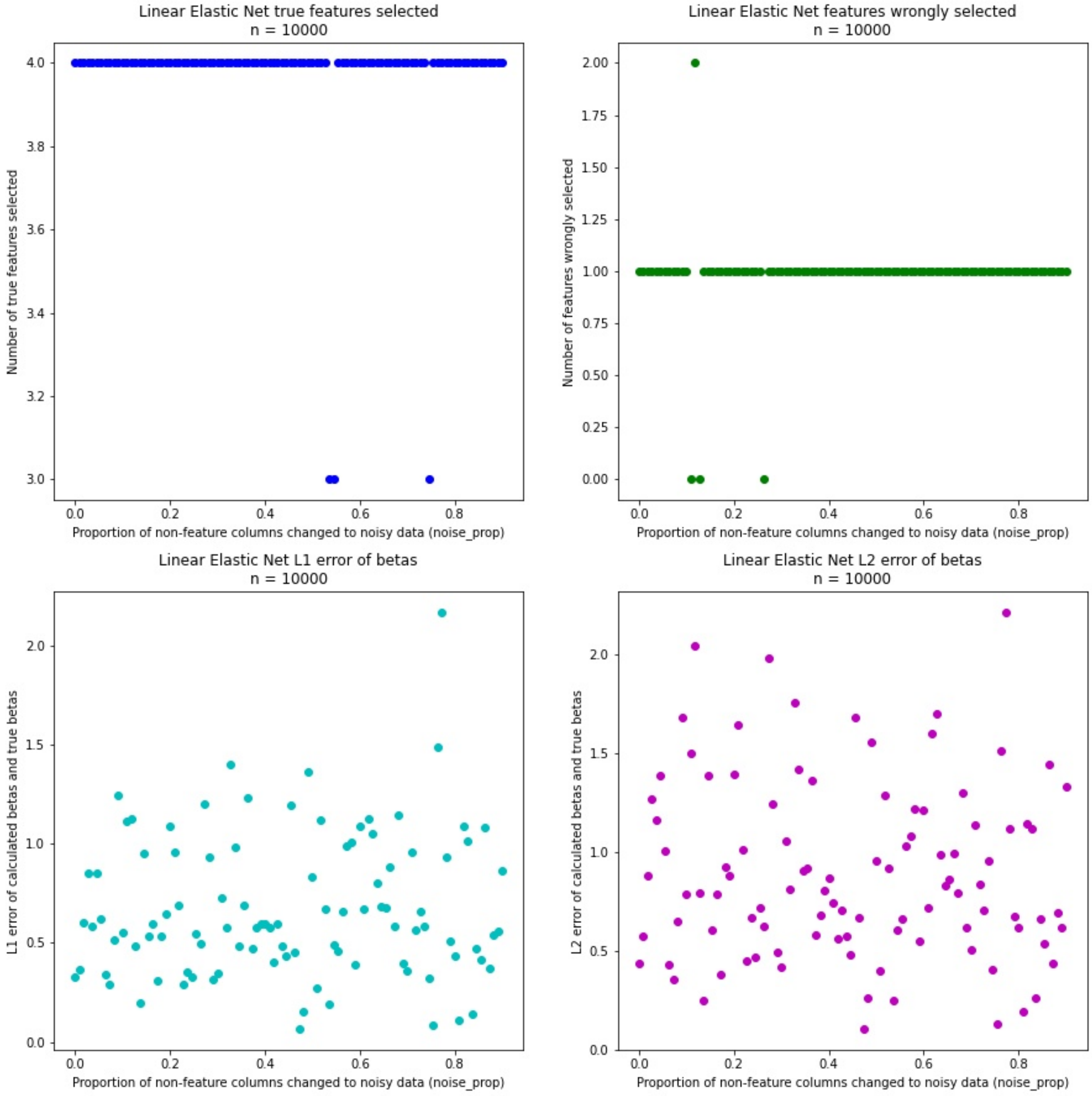

Figure 4: Plots for the linear regression with $l_1 + l_2$ regularization shown. Top left: how many true features are identified as a function of noise proportion. Top right: how many incorrect features are identified as a function of noise proportion. Bottom left: L_1 error of the coefficients as a function of noise proportion. Bottom right: L_2 error of the coefficients as a function of noise proportion.

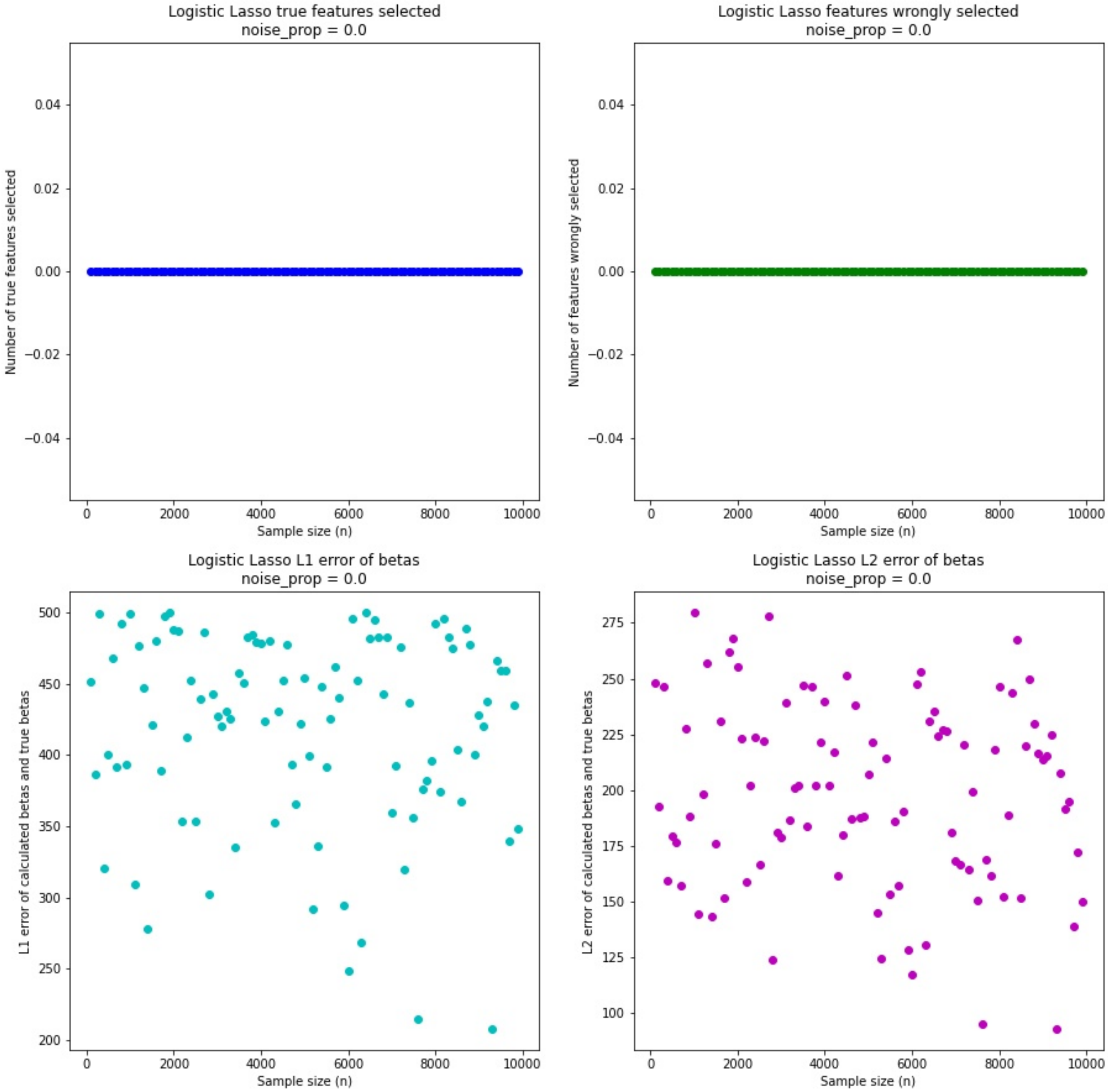

Figure 5: Plots for the logistic regression with l_1 regularization shown. Top left: how many true features are identified as a function of sample size. Top right: how many incorrect features are identified as a function of sample size. Bottom left: L_1 error of the coefficients as a function of sample size. Bottom right: L_2 error of the coefficients as a function of sample size.

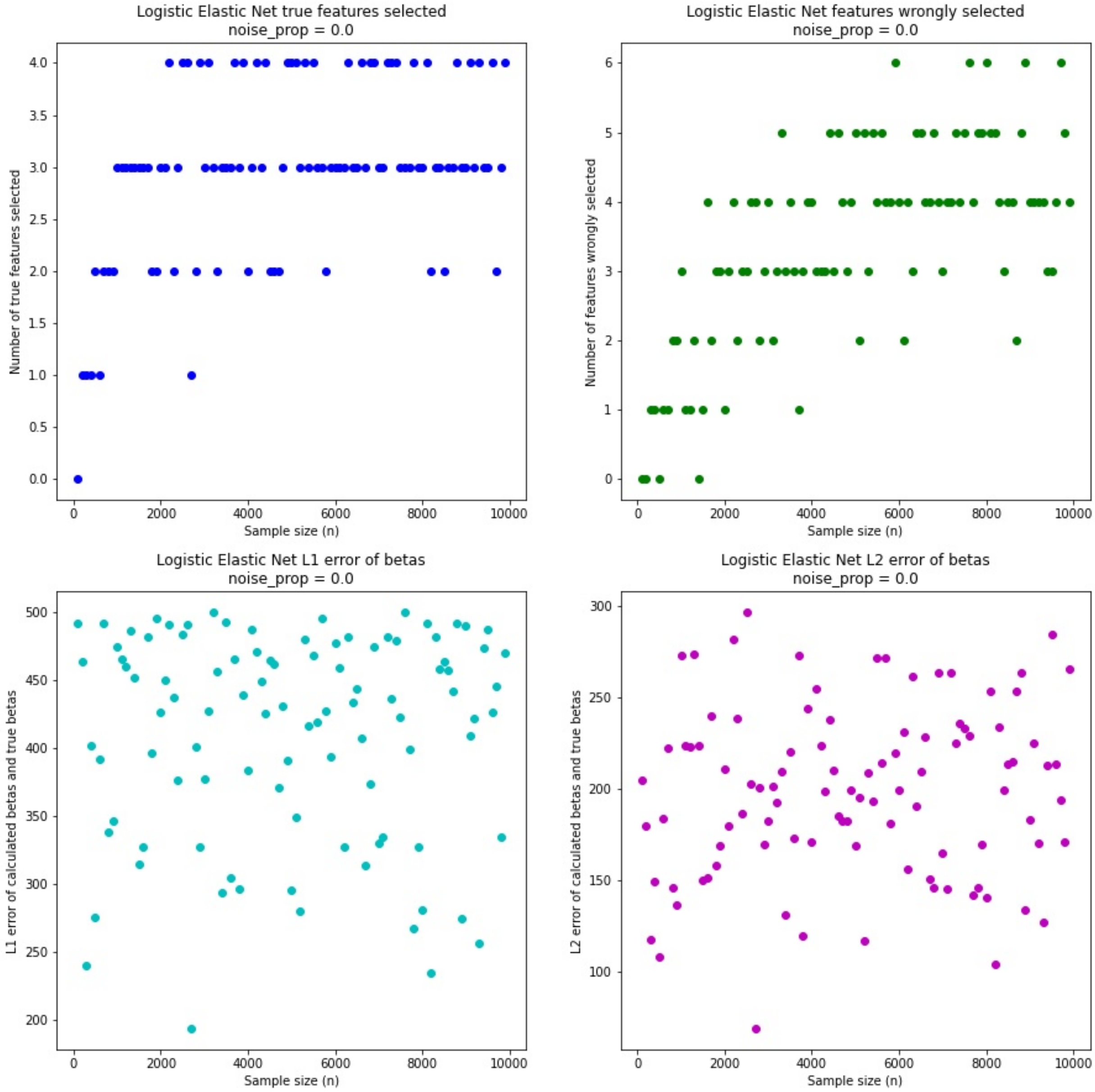

Figure 6: Plots for the logistic regression with $l_1 + l_2$ regularization shown. Top left: how many true features are identified as a function of sample size. Top right: how many incorrect features are identified as a function of sample size. Bottom left: L_1 error of the coefficients as a function of sample size. Bottom right: L_2 error of the coefficients as a function of sample size.

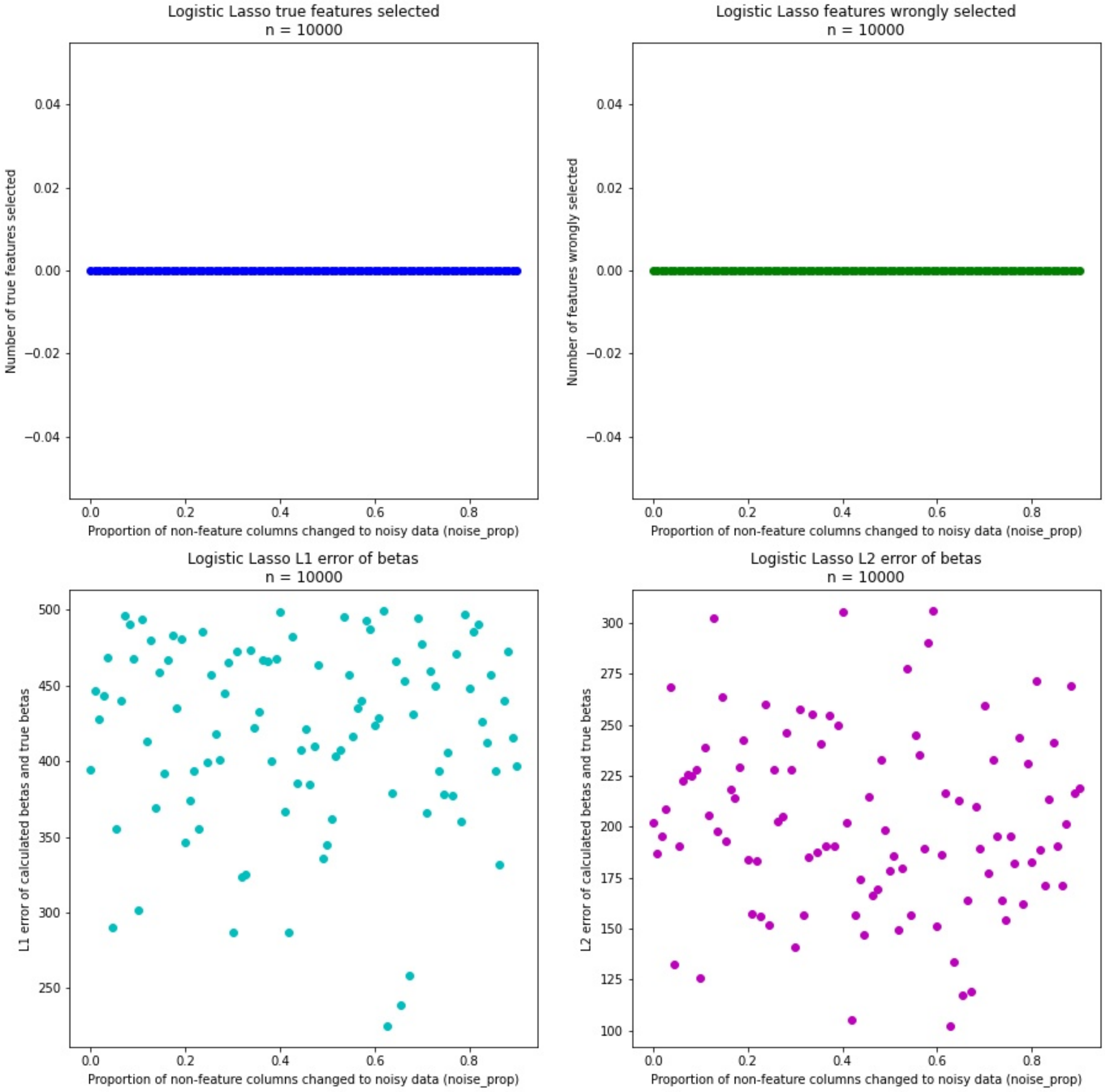

Figure 7: Plots for the logistic regression with l_1 regularization shown. Top left: how many true features are identified as a function of noise proportion. Top right: how many incorrect features are identified as a function of noise proportion. Bottom left: L_1 error of the coefficients as a function of noise proportion. Bottom right: L_2 error of the coefficients as a function of noise proportion.

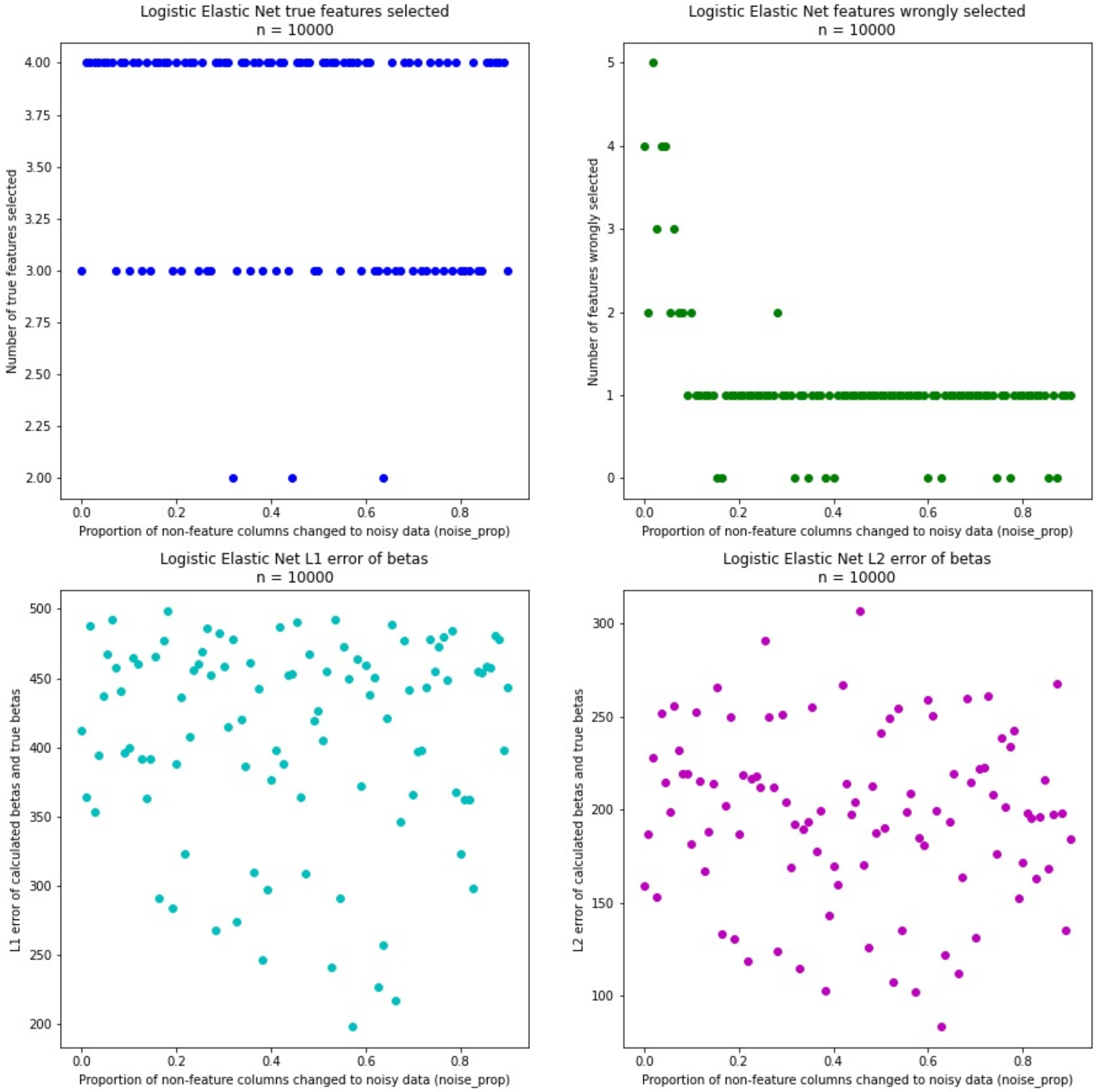

Figure 8: Plots for the logistic regression with $l_1 + l_2$ regularization shown. Top left: how many true features are identified as a function of noise proportion. Top right: how many incorrect features are identified as a function of noise proportion. Bottom left: L_1 error of the coefficients as a function of noise proportion. Bottom right: L_2 error of the coefficients as a function of noise proportion.