Some proposal of the high dimensional PU learning classification procedure

Konrad Furmańczyk[0000−0002−7683−4787], Marcin Dudziński[0000−0003−4242−8411], and Diana Dziewa-Dawidczyk[0000−0001−9486−1685]

Institute of Information Technology, Warsaw University of Life Sciences, Warsaw, Poland
{konrad_furmanczyk,marcin_dudzinski,diana_dziewa_dawidczyk}@sggw.edu.pl

Abstract. In our work, we propose a new classification method for positive and unlabeled (PU) data, called the LassoJoint classification procedure, which combines the thresholded Lasso approach in the first two steps with the joint method based on logistic regression, introduced by Teisseyre et. al. [12], in the last step. We prove that, under some regularity conditions, our procedure satisfies the screening property. We also conduct some simulation study in order to compare the proposed classification procedure with the oracle method. Prediction accuracy of the proposed method has been verified for some selected real datasets.

Keywords: positive unlabeled learning · logistic regression · empirical risk minimization · thresholded Lasso.

1 Introduction

Learning from positive and unlabeled (PU in short) data is an approach, where training data contains only positive and unlabeled examples, which means that the true labels $Y \in \{0,1\}$ are not observed directly, since only surrogate variable $S \in \{0,1\}$ is observable. This surrogate variable equals 1 - if an example is labeled, or 0 - if otherwise. The PU datasets appear in a large number of applications. For example, they often appear while dealing with the so-called under-reporting data from medical surveys, fraud detection and ecological modeling. Some other interesting examples of the under-reporting survey data may be found in Bekker and Davis [1] and Teisseyre et al. [12].

Suppose that X is a feature vector and, as mentioned earlier, $Y \in \{0,1\}$ denotes a true class label and $S \in \{0,1\}$ is a variable indicating, whether an example is labeled or not (then, $S = 1$ or $S = 0$, respectively). We apply a commonly used assumption, called the Selected Completely At Random (SCAR) condition, which states that the labeled examples are randomly selected from a set of positives examples, independently from X, i.e. $P(S = 1|Y = 1, X) = P(S = 1|Y = 1)$. Let $c = P(S = 1|Y = 1)$. The parameter c is called the label frequency and plays a key role in the PU learning problem. The primary objective of our note is to introduce a new PU learning classification procedure

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

leading to the estimation of the posterior probability $f(x) = P(Y = 1|X = x)$. The three basic methods of this estimation have been proposed so far. They consist in minimizing the empirical risk of logistic loss function and are known as: the naive method, the weighted method, and the joint method (the last one has been recently introduced in the paper of Teisseyre et. al. [12]). All of these approaches have been thoroughly described in [12]. As the joint method will be applied in our procedure's construction, some details regarding this method will be presented in the next section of our article. We have named our proposed classification method as the LassoJoint procedure, since it is a three-step approach combining the thresholded Lasso procedure with the joint method from Teisseyre et. al. [12]. Namely, in its two first steps we perform - for some prespecified level - the thresholded Lasso procedure, in order to obtain the support for coefficients of a feature vector X, while in the third step we apply - on the previously determined support - the joint method. Apart from the works, where different learning methods applying logistic regression for PU data have been proposed, there are also some other interesting articles, where various machine learning tools in the PU learning problems have been used. In this context, it is worthwile to mention: the papers of Hou [7] and Guo [5], where the generative adversial networks (GAN) for the PU problem have been employed, the work of Mordelet and Vert [10], where the bagging Support Vector Machine (SVM) approach for the PU data have been applied, and an article of Song and Raskutti [10], where the multidimentional PU problem with regard to the features selection has been investigated, and where the so-called PUlasso design has been established. It turns out that the LassoJoint procedure, which we propose in our work, is computationally simple and efficient in comparison to the other existing methods where the PU problem is considered. The simplicity and efficiency of our approach have been confirmed by the conducted simulation study.
The remainder of the paper is structured as follows. Namely, in Section 2 we describe our classification procedure in detail, in particular we also prove that the introduced method is the so-called screening procedure (i.e., it selects with a high probability the most significant predictors and the number of selected features is not greater that the sample size), as the screening property is necessary to apply the joint method in the final step of the procedure. In turn, in Section 3 we carry out some numerical study, in order to check the efficiency of the proposed approach, while in Section 4 we summarize and conclude our research. The results of numerical experiments on real data are given in Supplement [1].

2 The proposed LassoJoint algorithm

In our considerations, we assume that we have a random vector (Y, X), where $Y \in \{0, 1\}$ and $X \in \mathbb{R}^p$ is a feature vector, and that a random sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ is distributed as (Y, X) and independent of it. In addition, we suppose that the coordinates X_{ji} of X_i, $i = 1, ..., n$, $j = 1, ..., p$, are subgaussian with a parameter σ_{jn}^2, i.e. $E \exp(uX_{ji}) \leq \exp\left(u^2\sigma_{jn}^2/2\right)$ for all $u \in \mathbb{R}$.

[1] https://github.com/kfurmanczyk/ICCS21

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

Let: $s_n^2 = \max_{1\leq j\leq p}\sigma_{jn}^2$, $\limsup_n s_n^2 < \infty$, and $P\left(Y=1 \mid X=x\right) = q(x^T\beta)$ for some function $0 < q(x) < 1$ and all $x \in \mathbb{R}^p$, where p may depend on n and $p > n$. Put: $I_0 = \{j : \beta_j \neq 0\}$, $I_1 = \{1,\ldots,p\}\backslash I_0$ and $|I_0| = p_0$. We shall assume - as in Kubkowski and Mielniczuk [9] - that the distribution of X satisfies the linear regression condition (LRC), which means that

$$E(X \mid X^T\beta) = u_0 + u_1 X^T\beta \quad \text{for some} \quad u_0, u_1 \in \mathbb{R}^p.$$

This condition is fulfilled (for all β) by the class of elliptical distributions (such that, e.g., the normal distribution or the multivariate t-Student distribution). Reasoning as in Kubkowski and Mielniczuk [9], we obtain that under (LRC), there exists η satisfying $\beta^* = \eta\beta$, where $\eta \neq 0$ if $\mathrm{cov}(Y, X^T\beta) \neq 0$, and where $\beta^* = \arg\min_\beta R(\beta)$, with R standing for the risk function given by $R\left(\beta\right) = -E_{(X,Y)}l\left(\beta, X, Y\right)$, where in turn, $l\left(\beta, X, Y\right) = Y\log\sigma(X^T\beta) + (1-Y)\log\left(1-\sigma(X^T\beta)\right)$, with σ denoting logistic funcion of the form $\sigma(X^T\beta) = \exp(X^T\beta)/\left[1+\exp(X^T\beta)\right]$. Put: $I_0^* = \left\{j : \beta_j^* \neq 0\right\}$, $I_1^* = \{1,\ldots,p\}\backslash I_0^*$. It may be observed that under (LRC), we have $I_0 = I_0^*$ and consequently that $\mathrm{Supp}(I_0) = \mathrm{Supp}(I_0^*)$. In addition, put $H(b) = E(X^TX\sigma'(X^Tb))$ and define a cone $C(d,w) = \{\Delta \in \mathbb{R}^p : \|\Delta_{w^c}\|_1 \leq d\|\Delta_w\|_1\}$, where: $w \subseteq \{1,\ldots,p\}$, $w^c = \{1,\ldots,p\}\backslash w$, $\Delta_w = (\Delta_{w_1},\ldots,\Delta_{w_k})$, for $w = (w_1,\ldots,w_k)$. Furthermore, let κ be a generalized minimal eigenvalue of the matrix $H(\beta^*)$ given by $\kappa = \inf_{\Delta\in C(3,s_0^*)}\frac{\Delta^TH(\beta^*)\Delta}{\Delta^T\Delta}$. Moreover, we also define $\beta_{\min}^*$ and $\beta_{\min}$ as $\beta_{\min}^* := \min_{j\in I_0^*}\left|\beta_j^*\right|$ and $\beta_{\min} := \min_{j\in I_0}|\beta_j|$, respectively.

After these preliminaries, we are now in a position to depict the proposed method. Namely, our procedure, called the LassoJoint approach, is a three-step method, which is described as follows:

(1) For available PU dataset (s_i, x_i), $i = 1,\ldots,n$, we perform the ordinary Lasso procedure (see Tibshirani [14]) for some tuning parameter $\lambda > 0$, i.e. we compute the following Lasso estimator of β^*: $\hat{\beta}^{(L)} = \arg\min_{\beta\in R^{p+1}}\hat{R}(\beta) + \lambda\sum_{j=1}^p|\beta_j|$, where $\hat{R}(\beta) = -\frac{1}{n}\sum_{i=1}^n\left[s_i\log\left(\sigma(x_i^T\beta)\right) + (1-s_i)\log\left(1-\sigma(x_i^T\beta)\right)\right]$ and subsequently, we obtain the corresponding support $\mathrm{Supp}^{(L)} = \{1 \leq j \leq p : \hat{\beta}_j^{(L)} \neq 0\}$;

(2) We perform the thresholded Lasso for some prespecified level δ and obtain the support $\mathrm{Supp}^{(TL)} = \{1 \leq j \leq p : \left|\hat{\beta}_j^{(L)}\right| \geq \delta\}$;

(3) We apply the joint method from Teisseyre et al. [12] for the predictors from $\mathrm{Supp}^{(TL)}$.

Remark. The PU problem is related to incorrect specification of the logistic model. Under the SCAR assumption, we obtain that $P(S=1|X=x) = cq(x^T\beta)$ and consequently, if $cq() \neq \sigma()$, then in step (1) we are fitting misspecified logistic model to (S, X). Generally speaking, the joint method from [12] consists in fitting the PU data to logistic function and in the minimization, with respect to β and $c = P(S=1 \mid Y=1)$, of the following empirical risk $\hat{R}\left(\beta, c\right) = -\frac{1}{n}\sum_{i=1}^n\left[s_i\log\left(c\sigma(x_i^T\beta)\right) + (1-s_i)\log\left(1-c\sigma(x_i^T\beta)\right)\right]$, where σ stands for logistic function and $\{(s_i, x_i)\}$ is the sample of observations from the distribution of a random vector (S, X).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

The newly proposed LassoJoint procedure is similar to the LassoSD approach introduced in Furmańczyk and Rejchel [3]. The only difference between these two methods is that, we apply the joint method from [12] in the last step of our procedure - contrary to the procedure from [3], where the authors use multiple hypotheses testing in its final stage. The introduced LassoJoint procedure is determined by the two parameters λ and δ, which may depend on n. The selection of λ and δ is possible, if we impose the following conditions - denoted as the assumptions (A1)-(A4):

(A1) The generalized eigenvalue κ, of the matrix $H(\beta^*)$, is such that $m \leq \kappa \leq M$, for some $0 < m < M$;
(A2) $p_0^2 \log(p) = o(n)$, $\log(p) = o(n\lambda^2)$, $\lambda^2 p_0^2 \log(np) = o(1)$, as $n \to \infty$;
(A3) $p_0 + \frac{c_n^2}{\delta^2} \leq n$, where $c_n = 10\frac{\sqrt{p_0}}{\kappa}\lambda$;
(A4) $\beta_{\min} \geq \left(\delta + c_n/\sqrt{p_0}\right)/\eta$.

Clearly, in view of (LRC), the condition from (A4) is equivalent to the constraint that $\beta^*_{\min} \geq \delta + c_n/\sqrt{p_0}$. In addition, due to (A2), we get that $c_n \to 0$.

The main strictly theoretical result of our work is the following assertion.

Theorem 1. *(Screening property) Under the conditions (LRC) and (A1)-(A4), we have that with a probability at least* $1 - \epsilon_n$*, where* $\epsilon_n \to 0$*:*

(a) $\left|\text{Supp}^{(TL)}\right| \leq p_0 + \frac{c_n^2}{\delta^2} \leq n$,
(b) $I_0 \subset \text{Supp}^{(TL)}$.

The presented theorem states that the proposed LassoJoint procedure is the so-called screening procedure, which means that (in the first two steps) this method selects, with a high probability, the most significant predictors of the model and that the number of selected features is not greater than the sample size n. This screening property guarantees that with a high probability, we may apply the joint procedure from Teisseyre et. al. [12], based on fitting logistic regression to PU data. The proof of Theorem 1 uses the following lemma, which straightforwardly follows from Theorem 4.9 in Kubkowski [8].

Lemma 1. *Under the assumptions (A1)-(A2), we obtain that with a probability at least* $1 - \epsilon_n$*, with* ϵ_n *satisfying* $\epsilon_n \to 0$*, the following property holds:* $\left\|\hat{\beta}^{(L)} - \beta^*\right\|_2 \leq c_n$*, where* $c_n \to 0$ *and* $\|x\|_2 = \sqrt{\sum_{j=1}^p x_j^2}$ *for* $x \in R^p$.

Proof. First, we prove the relation stated in (a). By the definition of $\hat{\beta}_j^{(L)}$, we get $\sum_{j \in I_1 \cap \text{Supp}^{(TL)}} \left(\hat{\beta}_j^{(L)}\right)^2 \geq \delta^2 \left|I_1 \cap \text{Supp}^{(TL)}\right|$. Hence,

$$\begin{aligned}\left|I_1 \cap \text{Supp}^{(TL)}\right| &\leq \frac{1}{\delta^2} \sum_{j \in I_1 \cap \text{Supp}^{(TL)}} \left(\hat{\beta}_j^{(L)}\right)^2 \\ &= \frac{1}{\delta^2} \sum_{j \in I_1 \cap \text{Supp}^{(TL)}} \left(\hat{\beta}_j^{(L)} - \beta_j^*\right)^2 \leq \frac{1}{\delta^2} \left\|\hat{\beta}^{(L)} - \beta^*\right\|_2^2,\end{aligned}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

and

$$\left|\text{Supp}^{(TL)}\right| \leq \left|I_0 \cap \text{Supp}^{(TL)}\right| + \left|I_1 \cap \text{Supp}^{(TL)}\right| \leq p_0 + \tfrac{1}{\delta^2}\left\|\hat{\beta}^{(L)} - \beta^*\right\|_2^2.$$

It follows from the cited lemma that with a probability at least $1-\epsilon_n$, where $\epsilon_n \to 0$, we have $\left|\text{Supp}^{(TL)}\right| \leq p_0 + \frac{c_n^2}{\delta^2}$. Combining this inequality with (A3), we obtain (a). Thus, we only need to prove the property in (b).

Since $\left\{\min_{j\in I_0}\left(\hat{\beta}_j^{(L)}\right)^2 \geq \delta^2\right\} \subseteq \left\{I_0 \subset \text{Supp}^{(TL)}\right\}$, it is sufficient to show that

$$P\left(\min_{j\in I_0}\left(\hat{\beta}_j^{(L)}\right)^2 \geq \delta^2\right) \geq 1-\epsilon_n. \tag{1}$$

Let:
$\hat{\beta}_{I_0}^{(L)} := \left\{\hat{\beta}_j^{(L)} : j \in I_0\right\}$, $\beta_{I_0}^* := \left\{\beta_j^* : j \in I_0\right\}$. As $\left\|\hat{\beta}^{(L)} - \beta^*\right\|_2^2 \geq \left\|\hat{\beta}_{I_0}^{(L)} - \beta_{I_0}^*\right\|_2^2$, we have from the given lemma that with a probability at least $1-\epsilon_n$,

$$p_0 \min_{j\in I_0}\left(\hat{\beta}_j^{(L)} - \beta_j^*\right)^2 \leq \left\|\hat{\beta}_{I_0}^{(L)} - \beta_{I_0}^*\right\|_2^2 = \sum_{j\in I_0}\left(\hat{\beta}_j^{(L)} - \beta_j^*\right)^2 \leq c_n^2.$$

Hence, $\min_{j\in I_0}\left|\hat{\beta}_j^{(L)} - \beta_j^*\right| \leq c_n/\sqrt{p_0}$. In addition, by the triangle inequality, we obtain that for $j \in I_0$, $\left|\hat{\beta}_j^{(L)}\right| \geq \left|\beta_j^*\right| - \left|\hat{\beta}_j^{(L)} - \beta_j^*\right|$ and therefore, $\min_{j\in I_0}\left|\hat{\beta}_j^{(L)}\right| \geq \min_{j\in I_0}\left|\hat{\beta}_j^*\right| - c_n/\sqrt{p_0}$. This and (A4) imply that with a probability at least $1-\epsilon_n$, $\min_{j\in I_0}\left(\hat{\beta}_j^{(L)}\right) \geq \delta^2$, which yields (1) and consequently (b).

3 Numerical Study

Suppose that: $X_1, \ldots X_p$ are generated independently from $N(0,1)$, and Y_i, $i = 1, \ldots, n$, are generated from the binom$(1, p_i)$ distribution, where: $p_i = \sigma(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}), \beta_0 = 1$. The following high-dimensional models were simulated:

(M1) $p_0 = 5, p = 1.2 \cdot 10^3, n = 10^3, \beta_1 = \ldots = \beta_{p_0} = 1, \beta_{p_0+1} = \ldots = \beta_p = 0$;
(M2) $p_0 = 5, p = 1.2 \cdot 10^3, n = 10^3, \beta_1 = \ldots = \beta_{p_0} = 2, \beta_{p_0+1} = \ldots = \beta_p = 0$;
(M3) $p_0 = 5, p = 10^3, n = 10^3, \beta_1 = \ldots = \beta_{p_0} = 2, \beta_{p_0+1} = \ldots = \beta_p = 0$;
(M4) $p_0 = 20, p = 10^3, n = 10^3, \beta_1 = \ldots = \beta_{p_0} = 2, \beta_{p_0+1} = \ldots = \beta_p = 0$;
(M5) $p_0 = 5, p = 2 \cdot 10^3, n = 2 \cdot 10^3, \beta_1 = \ldots = \beta_{p_0} = 2, \beta_{p_0+1} = \ldots = \beta_p = 0$;
(M6) $p_0 = 5, p = 2 \cdot 10^3, n = 2 \cdot 10^3, \beta_1 = \ldots = \beta_{p_0} = 3, \beta_{p_0+1} = \ldots = \beta_p = 0$.

For all of the specified models, the LassoJoint method was implemented. In its first step, the Lasso method was used with some tuning parameters λ that were chosen either on the basis of 10-fold cross-validation scheme in the first scenario or by putting $\lambda = ((\log p)/n)^{1/3}$ in the second scenario. In the second step, we applied the thresholded Lasso design for $\delta = 0.5 \cdot ((\log p)/n)^{1/3}$.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

In the third - and simultaneously - the last step of our procedure, the variables selected by the thresholded Lasso method were employed to the joint method from [12] for the problem of the PU data classification. From the listed models, we randomly selected $c \cdot 100\%$ of the labeled observations of S, for $c = 0.1; 0.3; 0.5; 0.7; 0.9$. Next, we generated a test sample of size 1000 from our models and determined their accuracy percentage based on 100 MC replications of our experiments. The idea of our procedure's accuracy assesment is similar to the idea from Furmańczyk and Rejchel [4]. We applied the 'glmnet' package [2] from the R software and [13] in our computations. The results of our simulation study are collected in Table 1 (the column 'oracle' shows the accuracy of classifier that uses only the significant predictors and the true parameters of logistic models).

Table 1. Results for M1-M6

c	**model**	**scen 1**	**scen 2**	**oracle**	**model**	**scen 1**	**scen 2**	**oracle**
0.1	M1	0.526	0.597	0.808	M4	0.501	0.531	0.939
0.3	M1	0.492	0.456	0.808	M4	0.530	0.529	0.939
0.5	M1	0.598	0.530	0.808	M4	0.671	0.596	0.939
0.7	M1	0.688	0.591	0.808	M4	0.699	0.656	0.939
0.9	M1	0.743	0.623	0.808	M4	0.792	0.733	0.939
0.1	M2	0.505	0.504	0.887	M5	0.410	0.473	0.885
0.3	M2	0.586	0.514	0.887	M5	0.667	0.514	0.885
0.5	M2	0.705	0.565	0.887	M5	0.770	0.588	0.885
0.7	M2	0.770	0.636	0.887	M5	0.803	0.648	0.885
0.9	M2	0.820	0.698	0.887	M5	0.680	0.694	0.885
0.1	M3	0.516	0.568	0.885	M6	0.537	0.548	0.921
0.3	M3	0.608	0.505	0.885	M6	0.742	0.532	0.921
0.5	M3	0.708	0.567	0.885	M6	0.812	0.594	0.921
0.7	M3	0.778	0.640	0.885	M6	0.853	0.668	0.921
0.9	M3	0.820	0.706	0.885	M6	0.710	0.724	0.921

Real data experiments and all codes in R are presented in Supplement, available on https://github.com/kfurmanczyk/ICCS21.

4 Conclusions

The results of our simulation study show that if c increases, then the percentage of correct classifications increases as well. They also show that the classifications obtained by applying the proposed LassoJoint method display smaller classification errors (and thus - better classification accuracy) for the models with larger signals (i.e., for the M5 and M6 models). Comparing the M3 and M5 models, we can see that with an increase of the number of significant predictors (p_0), the classification accuracy is slightly decreasing. Furthermore, in all cases - except

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2

for the situation where $c = 0.1$ - the selection of the tuning parameter λ obtained by using the cross-validation design results in better classification accuracy. In addition, we may observe that in the case when $c = 0.7$ or $c = 0.9$, our LassoJoint approach is nearly as good as the 'oracle' method. In turn, for $c = 0.1$ the classification accuracy was low - from 0.41 do 0.60, but in the 'easiest' case, i.e. when $c = 0.9$, the classification accuracy ranged from 0.7 to 0.82. Furthermore, the results of our experiments conducted on real datasets show that if c increases, then the percentage of correct classifications increases as well. In addition, these results show similar classification accuracy among all of the considered classification methods (see Supplement). The proposed new LassoJoint classification method for PU data allows for the relatively low simulation computational costs to analyze data in a high-dimensional case, i.e. when the number of predictors exceeds the size of available sample ($p > n$). We aim to devote our further research to a more detailed analysis of the introduced procedure, in particular to the examination regarding optimal selection of the model parameters.

References

1. Bekker, J., Davis, J.: Learning from positive and unlabeled data: a survey. Available from http: arxiv.org/abs/1811.04820v3 (2020)
2. Friedman, J., Hastie, T., Simon, N., Tibshirani, R.: Glmnet: Lasso and elastic-net regularized generalized linear models. R package version 2.0 (2015)
3. Furmańczyk, K., Rejchel, W.: High-dimensional linear model selection motivated by multiple testing. Statistics **54**, 152–166 (2020)
4. Furmańczyk, K., Rejchel, W.: Prediction and variable selection in high-dimensional misspecified classification. Entropy **22**(5), 543 (2020)
5. Guo, T., Xu, C., Huang, J., Wang, Y., Shi, B., Xu, C., Tao, D.: On positive-unlabeled classification in GAN. CVPR (2020)
6. Hastie, T., Fithian, W.: Inference from presence-only data; the ongoing controversy. Ecography **36**: 864–867 (2013)
7. Hou, M., Chaib-draa, B., Li, C., Zhao, Q.: Generative adversarial positive-unlabeled learning. Proceedings of the twenty-seventh International Joint Conference on Artificial Intelligence (IJCAI-18) (2018)
8. Kubkowski, M.: Misspecification of binary regression model: properties and inferential procedures. Ph.D. Thesis. Warsaw University of Technology, Warsaw (2019)
9. Kubkowski, M., Mielniczuk J.: Active set of predictors for misspecified logistic regression. Statistics **51**, 1023–1045 (2017)
10. Mordelet, F., Vert, J.P.: A bagging SVM to learn from positive and unlabeled examples. Pattern Recognition Letters (2013)
11. Song, H., Raskutti, G.: High-dimensional variable selection with presence-only data. arXiv:1711.08129v3 (2018)
12. Teisseyre, P., Mielniczuk J., Łazecka, M.: Different strategies of fitting logistic regression for positive and unlabelled data. Computational Sciences-ICCS 2020: 3–17 (2020)
13. Teisseyre, P.: Repository from https://github.com/teisseyrep/Pulogistic. Last accessed 1 Jan 2021
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B **58**, 267–288 (1996)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_2