Manuscript submitted to *The Econometrics Journal*, pp. 1–18.

# Binary Classification with Covariate Selection through L0-Penalized Empirical Risk Minimization

Le-Yu Chen[†] and Sokbae Lee[‡§]

[†] *Institute of Economics, Academia Sinica,*
*No. 128, Section 2, Academia Road, Nankang, Taipei, 115, Taiwan.*
E-mail: lychen@econ.sinica.edu.tw
[‡] *Department of Economics, Columbia University,*
*420 West 118th Street, NewYork, NY 10027, USA.*
E-mail: sl3841@columbia.edu
[§] *Centre for Microdata Methods and Practice, Institute for Fiscal Studies,*
*7 Ridgmount St, Fitzrovia, London WC1E 7AE, UK.*

**Summary** We consider the problem of binary classification with covariate selection. We construct a classification procedure by minimizing the empirical misclassification risk with a penalty on the number of selected covariates. This optimization problem is equivalent to obtaining an $\ell_0$-penalized maximum score estimator. We derive probability bounds on the estimated sparsity as well as on the excess misclassification risk. These theoretical results are non-asymptotic and established in a high-dimensional setting. In particular, we show that our method yields a sparse solution whose $\ell_0$-norm can be arbitrarily close to true sparsity with high probability and obtain the rates of convergence for the excess misclassification risk. We implement the proposed procedure via the method of mixed integer linear programming. Its numerical performance is illustrated in Monte Carlo experiments and a real data application of the work-trip transportation mode choice.

**Keywords**: *Classification, covariate selection, penalized estimation, maximum score estimation, mixed integer optimization, finite sample property.*

## 1. INTRODUCTION

Binary classification is concerned with learning a binary classifier that can be used to categorize objects into one of two predefined statuses. It is useful for the prediction of binary outcomes and arises in a wide range of applications. Binary classification has been extensively studied in the statistics and machine learning literature. For comprehensive surveys and discussions on the binary classification methods, see e.g. Devroye, Györfi, and Lugosi (1996), Vapnik (2000), Lugosi (2002), Boucheron, Bousquet, and Lugosi (2005) and Hastie, Tibshirani, and Friedman (2009). Solving for the optimal binary classifier by minimizing the empirical misclassification risk is known as an empirical risk minimization (ERM) problem. There has been massive research interest in high dimensional classification problems where the dimension of the covariate vector can be comparable with or even larger than the available training sample size. It is known (see e.g., Bickel and Levina (2004) and Fan and Fan (2008)) that working directly with a high dimensional covariate space can result in poor classification performance. To overcome this, it is often assumed that only a small subset of covariates are important for classification and covariate selection is performed to mitigate the high dimensionality problem. See Fan, Fan, and Wu (2011) for an overview on the issues and methods for high dimensional classification.

In this paper, we study the ERM based binary classification in the setting with a high dimensional vector of covariates. We propose an $\ell_0$-penalized ERM procedure for classification by minimizing the empirical misclassification risk (over a class of linear classifiers) with a penalty on the number of selected covariates. Here, the $\ell_0$-norm of a real vector refers to the number of non-zero components of the vector. Computationally, our procedure is equivalent to obtaining an $\ell_0$-penalized version of Manski (1985)'s maximum score estimator.

Although an $\ell_1$- or $\ell_2$-based approach has been much more dominant in the literature thanks to its computational advantages, there is emerging interest in adopting an $\ell_0$-based approach since the latter is regarded as a more direct solution to the covariate selection problem. For instance, Bertsimas, King, and Mazumder (2016) took an $\ell_0$-constrained approach and developed a mixed integer optimization approach for solving the best subset selection problem in linear regression. Huang, Jiao, Liu, and Lu (2018) proposed a scalable computational algorithm for $\ell_0$-penalized least squares solutions.

It is more challenging to undertake an $\ell_0$-based approach for the binary classification problem since the empirical misclassification risk is non-convex, unlike the empirical risk in linear regression. Previously, Greenshtein (2006), Jiang and Tanner (2010) and Chen and Lee (2018) studied the best subset variable selection approach where the ERM classification problem is solved subject to a constraint on a pre-specified maximal number of selected covariates. In other words, they took the $\ell_0$-constrained approaches. It is worth pointing out that sparsity is not estimated in these approaches but rather imposed as a constraint. If the true sparsity is unknown, the $\ell_0$-constrained approach will result in selecting many more (or far less) covariates if the imposed sparsity is too large (or too small). Therefore, the $\ell_0$-constrained approach has to be assisted by cross validation or some sort of model selection procedure to optimize its performance. It is not required to know the level of true sparsity in our procedure.

We establish that our approach can yield a sparse solution for covariate selection with high probability when the Bayes classifier is of a linear classifier form and respects a sparsity condition. Specifically, we present sufficient conditions under which there is a high probability that the resulting number of selected covariates under our approach can be capped by an upper bound which can be made arbitrarily close to the unknown smallest number of covariates that are relevant for classification. Moreover, we derive a probability bound on the excess misclassification risk, which in turns gives the convergence rate as a corollary. All of our theoretical results are non-asymptotic. To the best of our knowledge, Jiang and Tanner (2010) is the only existing paper that (in addition to the $\ell_0$-constrained approach) considers the $\ell_0$-penalized ERM problem in binary classification among other things. They obtained a convergence rate for the excess misclassification risk; however, they focused on time series data and their theoretical results are asymptotic. Furthermore, they did not establish theoretical results that characterize the size of the subset of covariates selected under the $\ell_0$-penalized ERM approach. Neither did they provide numerical algorithms for solving the $\ell_0$-penalized estimation problem.

Our penalized ERM approach is also closely related to the method of structural risk minimization (see, e.g., Devroye, Györfi, and Lugosi (1996, Chapter 18)), where the best classifier is selected by solving a sequence of penalized ERM problems with the penalty depending on the increasing Vapnik-Chervonenkis (VC) dimension of the space of classifiers. As will be discussed in Section 2, our approach can also be interpreted in a similar fashion yet with a different type of complexity penalty.

For implementation, we show that the $\ell_0$-penalized ERM problem of this paper can be equivalently reformulated as a mixed integer linear programming (MILP) problem. This reformulation enables us to employ modern efficient mixed integer optimization (MIO) solvers to solve our penalized ERM problem. Well-known numerical solvers such as CPLEX and Gurobi can be used to effectively solve large-scale MILP problems. See Nemhauser and Wolsey (1999) and Bertsimas and Weismantel (2005) for classic texts on the MIO theory and applications. See also Jünger, Liebling, Naddef, Nemhauser, Pulleyblank, Reinelt, Rinaldi, and Wolsey (2009), Achterberg and Wunderling (2013) and Bertsimas, King, and Mazumder (2016, Section 2.1) for discussions on computational advances in solving the MIO problems.

The remainder of the paper is organized as follows. In Section 2, we describe the binary classification problem and set forth the $\ell_0$-penalized ERM approach. In Section 3, we establish theoretical properties of the proposed classification approach. In Section 4, we provide a computational method using the MIO approach. In Section 5, we conduct a simulation study on the finite-sample performance of our proposed approach. In Section 6, we illustrate usefulness of our approach in a real data application concerning prediction of the work-trip transportation mode choice. We then conclude the paper in Section 7. Proofs of all theoretical results of the paper are collated in Section A.

## 2. AN $\ell_0$-PENALIZED ERM APPROACH

Let $Y \in \{0,1\}$ be the binary outcome of an object and $X$ a $(p+1)$ dimensional covariate vector of that object. Write $X = (X_1, \widetilde{X})$, where $X_1$ is a scalar random variable that is always included and has a positive effect and $\widetilde{X}$ is the $p$ dimensional subvector of $X$ subject to covariate selection. For $x \in \mathcal{X}$, let

$$b_\theta(x) \equiv 1\left\{x_1 + \widetilde{x}'\theta \geq 0\right\}, \tag{2.1}$$

where $\mathcal{X}$ is the support of $X$, $\theta$ is a vector of parameters, and $1\left\{\cdot\right\}$ is an indicator function that takes value 1 if its argument is true and 0 otherwise.

We consider binary classification using linear classifiers of the form (2.1). Since the condition $X_1 + \widetilde{X}'\theta \geq 0$ is invariant with respect to any positive scalar that multiplies both sides of this inequality, working with the classifier (2.1) amounts to normalizing the scale by setting the coefficient of $X_1$ to be unity. For any $p$ dimensional real vector $\theta$, let $\|\theta\|_0 \equiv \sum_{j=1}^p 1\{\theta_j \neq 0\}$ be the $\ell_0$-norm of $\theta$. Assume that the researcher has a training sample of $n$ independent identically distributed (i.i.d.) observations $(Y_i, X_i)_{i=1}^n$ of $(Y, X)$. We allow the dimension $p$ to be potentially much larger than the sample size $n$. We estimate the coefficient vector $\theta$ by solving the following $\ell_0$-penalized minimization problem

$$\min_{\theta\in\Theta} \; S_n(b_\theta) + \lambda \|\theta\|_0 , \tag{2.2}$$

where $\Theta \subset \mathbb{R}^p$ denotes the parameter space, and, for any indicator function $b : \mathcal{X} \mapsto \{0,1\}$,

$$S_n(b) \equiv \frac{1}{n}\sum_{i=1}^n 1\{Y_i \neq b(X_i)\}, \tag{2.3}$$

and $\lambda$ is a given non-negative tuning parameter of the penalized minimization problem.

The function $S_n(b)$ is known as the empirical misclassification risk for the binary classifier $b$. Minimization of $S_n(b)$ over the class of binary classifiers given by (2.1) is known as an empirical risk minimization (ERM) problem. The penalized ERM approach (2.2) enforces dimension reduction by attaching a higher penalty to a classifier $b_\theta$ which uses more object covariates for classification. Let $\widehat{\theta}$ be a solution to the minimization problem (2.2). We shall refer to the resulting classifier (2.1) evaluated at $\widehat{\theta}$ as an $\ell_0$-penalized ERM classifier.

For any $m \geq 0$, let

$$\mathcal{B}_m \equiv \{b_\theta : \theta \in \Theta_m\} \tag{2.4}$$

where

$$\Theta_m \equiv \{\theta \in \Theta : \|\theta\|_0 \leq m\}. \tag{2.5}$$

That is, $\mathcal{B}_m$ is the class of all linear classifiers in (2.1) whose $\theta$ vector has no more than $m$ non-zero components. For $m \in \{0, 1, ..., p\}$, let

$$S_n^C(m) \equiv \min_{b \in \mathcal{B}_m} S_n(b). \tag{2.6}$$

Then it is straightforward to see that the minimized objective value of the penalized ERM problem (2.2) is equivalent to that of the problem

$$\min_{m \in \{0,1,...,p\}} S_n^C(m) + \lambda m. \tag{2.7}$$

In other words, our approach is akin to the method of structural risk minimization as it amounts to solving ERM problems over an increasing sequence of classifier spaces $\mathcal{B}_m$ which carries a complexity penalty $\lambda m$. In the next section, we will set forth regularity conditions on the penalty tuning parameter $\lambda$ and establish theoretical properties for our classification approach.

## 3. THEORETICAL PROPERTIES

In this section, we study theoretical properties of the $\ell_0$-penalized ERM classification approach. Let $F$ denote the joint distribution of $(Y, X)$. For any indicator function $b : \mathcal{X} \mapsto \{0, 1\}$, let

$$S(b) \equiv P(Y \neq b(X)). \tag{3.8}$$

For $x \in \mathcal{X}$, let

$$\eta(x) \equiv P(Y = 1|X = x), \tag{3.9}$$

$$b^*(x) \equiv 1\{\eta(x) \geq 0.5\}. \tag{3.10}$$

For any measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, let $\|f\|_1 = E[|f(X)|]$ denote the $L_1$-norm of $f$. The functions $\eta$ and $b^*$ as well as the $L_1$-norm $\|\cdot\|_1$ depend on the data generating distribution $F$. It is straightforward to see that, for any binary classifier $b$,

$$S(b) - S(b^*) = E[|2\eta(X) - 1| \, |b^*(X) - b(X)|] \tag{3.11}$$

so that $S(b)$ is minimized at $b = b^*$. The optimal classifier $b^*$ is known as the Bayes classifier in the classification literature.

We assess the predictive performance of the $\ell_0$-penalized ERM approach by bounding the excess risk

$$U_n \equiv S(b_{\widehat{\theta}}) - S(b^*). \tag{3.12}$$

The difference $U_n$ is non-negative by (3.11). Hence, a good classifier will result in a small value of $U_n$ with a high probability and also on average.

We impose the following assumption.

Condition 3.1. *For every data generating distribution $F$, there is a non-negative integer $d$, which may depend on $F$, such that $d \leq p$ and $b^* \in \mathcal{B}_d$.*

Let $q$ denote the smallest value of non-negative integers $d$ satisfying $b^* \in \mathcal{B}_d$. By Condition 3.1, such $q$ value is finite and always exists. Condition 3.1 implies that the Bayes classifier $b^*$ admits a linear threshold crossing structure in the sense that the equivalence

$$\eta(X) \geq 0.5 \Longleftrightarrow X_1 + \widetilde{X}'\theta \geq 0 \tag{3.13}$$

holds almost surely for some $\theta \in \Theta_q$, where $q$ can be interpreted as the sparsity level associated with $b^*$, which is unknown in this binary classification problem. Moreover, the assumption that $q \leq p$ implies that the covariate vector $(X_1, \widetilde{X})$ is rich enough to embody those relevant ones for constructing the Bayes classifier.

We now remark on the sign-matching condition (3.13) in the context of the binary response model specified below. Suppose that the outcome $Y$ is generated from a latent variable threshold crossing model (see, e.g., Manski, 1985, 1988):

$$Y = 1\{\alpha X_1 + \widetilde{X}'\beta \geq \varepsilon\}, \tag{3.14}$$

where $(\alpha, \beta)$ is the underlying data generating parameter vector with $\alpha > 0$ and $\varepsilon$ is an unobserved latent variable whose distribution satisfies that

$$Med(\varepsilon | X = x) = 0 \text{ for } x \in \mathcal{X}. \tag{3.15}$$

The model (3.14) - (3.15) encompasses the logistic regression model as a special case yet allows for a more general and unknown form of the distribution of $\varepsilon$ conditional on the covariate vector $X$. Under this model specification, we can immediately see that condition (3.13) holds with $\theta = \beta/\alpha$. It is worth pointing out that the condition for the Bayes classifier taking a linear form can also hold in a more general scenario than that specified by (3.14). See e.g. Elliott and Lieli (2013, p. 18) for further discussions.

For any two real numbers $x$ and $y$, let $x \vee y \equiv \max\{x, y\}$ and $x \wedge y \equiv \min\{x, y\}$. For any $x \geq 0$, let $\lceil x \rceil$ and $\lfloor x \rfloor$ respectively denote the integer ceiling and floor of $x$. We impose the following condition on the growing rates of $\lambda$ relative to the sample size.

Condition 3.2. *$\lambda = c\sqrt{n^{-1}\ln(p \vee n)}$ for some constant $c > 0$.*

Let

$$m_0 \equiv q \vee (p \wedge \lfloor \lambda^{-1} \rfloor), \tag{3.16}$$
$$r_n \equiv q \ln(p \vee n). \tag{3.17}$$

The estimate $\left\|\widehat{\theta}\right\|_0$ corresponds to the number of covariates selected under the $\ell_0$-penalized ERM approach. We now provide a result on the statistical behavior of $\left\|\widehat{\theta}\right\|_0$, which sheds lights on the dimension reduction performance of our penalized estimation approach.

Theorem 3.1. *Assume $q \geq 1$. Given Conditions 3.1 and 3.2, for all given $\sigma > 0$ and $\epsilon \in (0, 1)$, there is a universal constant $M_\sigma$, which depends only on $\sigma$, such that*

$$P\left(\left\|\widehat{\theta}\right\|_0 > s\right) \leq j_0 e^{-\sigma r_n} \tag{3.18}$$

*where*

$$s \equiv (1+\epsilon)q + \epsilon, \tag{3.19}$$

$$j_0 \equiv \left\lceil \frac{\ln(m_0) - \ln(\epsilon)}{\left|\ln(2\sqrt{M_\sigma}) - \ln(c)\right|} \right\rceil, \tag{3.20}$$

*provided that the constant* $c$ *in Condition 3.2 is sufficiently large such that*

$$c \geq 2\sqrt{M_\sigma}\,(1+\epsilon)\,\epsilon^{-1}, \tag{3.21}$$

*and the inequality*

$$4\,(k+1)\ln(M_\sigma k \ln(p \vee n)) \leq k\ln(p \vee n) + 6\,(k+1)\ln 2 \tag{3.22}$$

*holds for any integer* $k$ *that satisfies*

$$q \leq k \leq [m_0 \vee \lfloor s \rfloor \vee ((j_0 - 1)\,q + \lfloor \sqrt{m_0} \rfloor)] \wedge p.$$

By (3.16), (3.20) and Condition 3.2, we see that

$$j_0 = O(\ln(q \vee \sqrt{n})) \tag{3.23}$$

so that, for any $\sigma > 0$, the term $e^{\sigma r_n}$ grows much faster than $j_0$ as $p \vee n$ grows. Thus, for any fixed $\epsilon \in (0,1)$, we can deduce from Theorem 3.1 that $P\left(\left\|\widehat{\theta}\right\|_0 > s\right) \longrightarrow 0$ as $p \vee n \longrightarrow \infty$. Moreover, this theorem implies that our approach is effective in reducing the covariate dimension in the sense that, with high probability in large samples, the number of selected covariates is capped above by the quantity (3.19), which can be made arbitrarily close to the true sparsity $q$ in the classification problem. Specifically, if $\epsilon$ turns out to be smaller than $1/(q+1)$, the result (3.18) implies that $P\left(\left\|\widehat{\theta}\right\|_0 > q+1\right)$ tends to zero exponentially in $r_n$.

The next theorem characterizes the predictive performance of the $\ell_0$-penalized ERM approach.

Theorem 3.2. *Under the setup and assumptions stated in Theorem 3.1, the following result holds:*

$$P(U_n > 3\lambda s) \leq (1 + j_0)\,e^{-\sigma r_n}. \tag{3.24}$$

Theorem 3.2 implies that the tail probability of $U_n$ decays to zero exponentially in $r_n$. Moreover, inequality (3.24) together with the fact that $U_n \leq 1$ immediately implies that

$$E[U_n] \leq (1 + j_0)\,e^{-\sigma r_n} + 3\lambda s. \tag{3.25}$$

Taking $\sigma = 2$ and using (3.23) and Condition 3.2, we can therefore deduce that

$$E[U_n] = O\left(q\sqrt{n^{-1}\ln(p \vee n)}\right), \tag{3.26}$$

which converges to zero whenever

$$q^2 \ln(p \vee n) = o(n). \tag{3.27}$$

The rate condition (3.27) allows the case that

$$\ln p = O(n^\alpha) \text{ and } q = o(n^{1/2 - \alpha/2}) \text{ for } 0 < \alpha < 1. \tag{3.28}$$

In other words, the $\ell_0$-penalized ERM classification approach is risk-consistent even when the number of potentially relevant covariates ($p$) grows exponentially in sample size, provided that the number of truly effective covariates ($q$) can only grow at a polynomial rate.

We note that both probability bounds (3.18) and (3.24) depend on the data distribution $F$ only through $q$ which is the smallest value of non-negative integers $d$ satisfying Condition 3.1. If the value of $q$ is uniform over $F$, our results hold uniformly over data generating processes.

We provide some further remarks on the convergence rate result (3.26). Condition 3.1 implies that the space $\mathcal{B}_q$ contains the Bayes classifier $b^*$. Thus, if the value of $q$ were known, one could performed classification via the $\ell_0$-constrained ERM approach where the empirical risk $S_n(b)$ is minimized with respect to $b \in \mathcal{B}_q$. The lower bound on the VC dimension of the classifier space $\mathcal{B}_q$ grows at rate $O(q \ln(p/q))$ (Abramovich and Grinshtein (2019, Lemma 1)). Hence, the rate $O(\sqrt{n^{-1} q \ln(p/q)})$ is the minimax optimal rate at which the excess risk converges to zero under this constrained estimation approach (Devroye, Györfi, and Lugosi (1996, Theorem 14.5)). In view of this, suppose $p$ grows at a polynomial or exponential rate in $n$. Then our rate result (3.26) is nearly oracle in the sense that, when $q$ grows at rate $O(\ln n)$, the rate (3.26) remains close within some $\ln n$ factor to the optimal rate attained under the case of known $q$. Moreover, both rates coincide and reduce to $O(\sqrt{n^{-1} \ln p})$ when the value of $q$ does not increase with the sample size.

We conclude this section by commenting on an alternative choice of the penalization tuning parameter $\lambda$. In view of the link to structural risk minization in (2.7), we may consider a sequence of tuning parameters $\{\lambda_m : 0, 1, ..., p\}$ in

$$\min\nolimits_{m \in \{0,1,\ldots,p\}} S_n^C(m) + \lambda_m m. \tag{3.29}$$

For instance, one may set $\lambda_m = c\sqrt{\{n(m \vee 1)\}^{-1} \ln(p \vee n)}$ for some constant $c > 0$. However, it would be computationally demanding to solve (3.29) since it requires solving $p$ subproblems of $S_n^C(m)$ in (2.6). We leave this possible extension for a topic for future research.

## 4. COMPUTATIONAL ALGORITHMS

While the ERM approach to binary classification is theoretically sound, its implementation is computationally challenging and is known to be an *NP* (Non-deterministic Polynomial time) hard problem (Johnson and Preparata, 1978). Florios and Skouras (2008) developed a mixed integer optimization (MIO) based computational method and provided numerical evidence demonstrating effectiveness of the MIO approach to solving the ERM type optimization problems. Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2018) adopted the MIO solution approach to solving the optimal treatment assignment problem which is closely related to the ERM based classification problem. The MIO approach is also useful for solving problems of variable selection through $\ell_0$-norm constraints. See Bertsimas, King, and Mazumder (2016) and Chen and Lee (2018) who proposed MIO based computational algorithms to solving the $\ell_0$-constrained regression and classification problems respectively.

Motivated by these previous works, we now present an MIO based computational method for solving the $\ell_0$-penalized ERM problem. Given that $Y_i \in \{0,1\}$, solving the problem (2.2) amounts to solving

$$\min_{\theta\in\Theta} \frac{1}{n}\sum\nolimits_{i=1}^{n}\left[Y_i - (2Y_i-1)\,1\{X_{1i}+\widetilde{X}_i'\theta \geq 0\}\right] + \lambda\,\|\theta\|_0\,. \tag{4.30}$$

We assume that the parameter space $\Theta$ takes the form $\Theta = \prod_{j=1}^{p}\left[\underline{c}_j, \overline{c}_j\right]$, where $\underline{c}_j$ and $\overline{c}_j$ are lower and upper parameter bounds such that $-\infty < \underline{c}_j \leq \theta_j \leq \overline{c}_j < \infty$ for $j \in \{1,...,p\}$. We specify sufficiently large magnitudes of $\underline{c}_j$ and $\overline{c}_j$ so as to include parameter values $\theta$ for which condition (3.13) holds.

Our implementation builds on the method of mixed integer optimization. Specifically, we note that the minimization problem (4.30) can be equivalently reformulated as the following mixed integer linear programming (MILP) problem:

$$\min_{\theta\in\boldsymbol{\Theta},d_1,...,d_n,e_1,...,e_p} \frac{1}{n}\sum\nolimits_{i=1}^{n}\left[Y_i - (2Y_i-1)\,d_i\right] + \lambda\sum\nolimits_{j=1}^{p} e_j \tag{4.31}$$

subject to

$$(d_i - 1)\,M_i \leq X_{1i} + \widetilde{X}_i'\theta < d_i(M_i+\delta),\ i \in \{1,...n\}, \tag{4.32}$$

$$e_j\underline{c}_j \leq \theta_j \leq e_j\overline{c}_j,\ j\in\{1,...,p\}, \tag{4.33}$$

$$d_i \in \{0,1\},\ i\in\{1,...n\}, \tag{4.34}$$

$$e_j \in \{0,1\},\ j\in\{1,...,p\}, \tag{4.35}$$

where $\delta$ is a given small and positive real scalar (e.g. $\delta = 10^{-6}$ as in our numerical study), and

$$M_i \equiv \max_{\theta\in\Theta}\left|X_{1i}+\widetilde{X}_i'\theta\right| \text{ for } i \in \{1,...,n\}. \tag{4.36}$$

We now explain the equivalence between (4.30) and (4.31). Given $\theta$, the inequality constraints (4.32) and the dichotomization constraints (4.34) enforce that $d_i = 1\{X_{1i} + \widetilde{X}_i'\theta \geq 0\}$ for $i \in \{1,...n\}$. Moreover, the on-off constraints (4.33) and (4.35) ensure that, whenever $e_j = 0$, the value $\theta_j$ must also be zero so that the $j$th component of the covariate vector $\widetilde{X}$ is excluded in the resulting $\ell_0$-penalized ERM classifier. The sum $\sum_{j=1}^{p} e_j$ thus captures the number of non-zero components of the vector $\theta$. As a result, both minimization problems (4.30) and (4.31) are equivalent. This equivalence enables us to employ modern MIO solvers to solve for $\ell_0$-penalized ERM classifiers. For implementation, note that the values $(M_i)_{i=1}^{n}$ in the inequality constraints (4.32) can be computed by formulating the maximization problem in (4.36) as linear programming problems, which can be efficiently solved by modern numerical solvers. Hence these values can be easily computed and stored as inputs to the MILP problem (4.31).

## 5. SIMULATION STUDY

In this section, we conduct simulation experiments to study the performance of our approach. We consider a simulation setup similar to that of Chen and Lee (2018, Section 5) and use the following data generating design. Let $V = (V_1, ..., V_p)$ be a multivariate normal random vector with mean zero and covariance matrix $\Sigma$ with its element $\Sigma_{i,j} = (0.25)^{|i-j|}$. The binary outcome is generated according to the following specification:

$$Y = 1\{X_1 + \widetilde{X}'\theta^* \geq \sigma(X)\xi\},$$

where $\theta^*$ denotes the true data generating parameter value, $X = (X_1, \widetilde{X})$ is a $(p+1)$ dimensional covariate vector with $X_1 = V_1$ and $\widetilde{X} = (1, V_2, ..., V_p)$, and $\xi$ is a random variate that is independent of $V$ and follows the standard logistic distribution. The constant term in $\widetilde{X}$ is included to capture the regression intercept. We set $\theta_1^* = 0$ and $\theta_j^* = 0$ for $j \in \{3, ..., p\}$. The coefficient $\theta_2^*$ is chosen to be non-zero such that, among all the covariates in $\widetilde{X}$, only the variable $\widetilde{X}_2 = V_2$ is relevant in the data generating processes (DGP). We consider the following two specifications for $\theta_2^*$ and $\sigma(X)$:

$$\text{DGP(i)} : \theta_2^* = -0.55 \text{ and } \sigma(X) = 0.2.$$

$$\text{DGP(ii)} : \theta_2^* = -1.85 \text{ and } \sigma(X) = 0.2\left(1 + 2\left(V_1 + V_2\right)^2 + \left(V_1 + V_2\right)^4\right).$$

We used 100 simulation repetitions in each Monte Carlo experiment. For each simulation repetition, we generated a training sample of $n = 100$ observations for estimating the coefficients $\theta$ and a validation sample of 5000 observations for evaluating the out-of-sample classification performance. We considered simulation configurations with $p \in \{10, 200\}$ to assess the classifier's performance in both the low and high dimensional binary classification problems.

We specified the parameter space $\Theta$ to be $[-10, 10]^p$ for the MIO computation of the $\ell_0$-penalized ERM classifiers. Throughout this paper, we used the MATLAB implementation of the Gurobi Optimizer to solve the MIO problems (4.31). Moreover, all numerical computations were done on a desktop PC (Windows 7) equipped with 32 GB RAM and a CPU processor (Intel i7-5930K) of 3.5 GHz. To reduce computation cost of solving the $\ell_0$-penalized ERM problems, we set the MIO solver time limit to be one hour beyond which we forced the solver to stop early and used the best discovered feasible solution to construct the resulting $\ell_0$-penalized ERM classifier.

The main computational cost arises from solving the MIO problem (4.31) for given values of $(M_i)_{i=1}^n$. The problem (4.36) for computing the $(M_i)_{i=1}^n$ values are nonetheless computationally very light. For every simulated dataset in our study, it took less than 0.3 CPU seconds to complete the computation for obtaining all the $(M_i)_{i=1}^n$ values.

For implementation, it remains to specify an exact form of the penalty parameter $\lambda$ in (2.2). We set

$$\lambda = v\left[\ln\ln\left(p \vee n\right)\right]\sqrt{\ln\left(p \vee n\right)/n}, \tag{5.37}$$

where $v$ is a tuning constant which remains to be calibrated. The form (5.37) implies that the value $c$ in Condition 3.2 is taken to be $v \ln\ln(p \vee n)$, which will satisfy inequality (3.21) and hence validate the probability bound (3.18) when $p \vee n$ is sufficiently large. Moreover, by the risk upper bound (3.25), the convergence rate result (3.26) continues to hold up to a factor of $\ln\ln\left(p \vee n\right)$.

For practical applications, we recommend calibrating the tuning scalar $v$ via the method of cross validation. Yet, for this simulation study, we used a simple heuristic rule and set

$$v \equiv h\left(1 - h\right), \tag{5.38}$$

$$h \equiv \min_{t \in [-10, 10]} \frac{1}{n}\sum\nolimits_{i=1}^{n} 1\left\{Y_i \neq 1\left\{X_{1i} + t \geq 0\right\}\right\}. \tag{5.39}$$

The value $h$ in (5.39) can also be computed via the MIO approach by simply removing from the MIO problem (4.31) the constraints (4.33) and (4.35) as well as the binary controls $(e_1, .., e_p)$ and the penalty part in the objective function. This computation is much faster as it is concerned with one-dimensional optimization. The rationale behind the choice (5.38) is as follows. Note that (5.39) corresponds to an ERM classification using classifier (2.1) where $\widetilde{X}$ only consists of the intercept term, and (5.38) corresponds to an estimate of the variance of the misclassification loss under such a simple classification rule. Intuitively speaking, the value $v$ captures the variability of the empirical risk under a parsimonious covariate space specification. From the bias and variance tradeoff perspective, when this variability is small, we may as well increase the classifier flexibility by attaching a small penalty in the penalized ERM procedure so as to induce a richer set of selected covariates for classification.

Let logit_lasso denote the $\ell_1$-penalized logistic regression approach (see, e.g., Friedman, Hastie, and Tibshirani, 2010). The logit_lasso estimation approach is a computationally attractive approach that can be used to estimate high dimensional binary response models. We compared in simulations the performance of our method to that of the logit_lasso approach. We used the MATLAB implementation of the well known **glmnet** computational algorithms (Qian, Hastie, Friedman, Tibshirani, and Simon, 2013) for solving the logit_lasso estimation problems. We did not penalize the coefficient of the covariate $X_1$ so that, as in the simulations of the $\ell_0$-penalized ERM approach, this variable would always be included in the resulting classifier constructed under the logit_lasso approach. We calibrated the lasso penalty parameter value over a sequence of 100 values via the 10-fold cross validation procedure. We used the default setup of **glmnet** for constructing this tuning sequence among which we reported results based on the following two choices, $\{\lambda_{opt}^{lasso}, \lambda_{1se}^{lasso}\}$, of the penalty parameter value. The value $\lambda_{opt}^{lasso}$ refers to the lasso penalty parameter value that minimized the cross validated misclassification risk, whereas $\lambda_{1se}^{lasso}$ denotes the largest penalty parameter value whose corresponding cross validated misclassification risk still falls within the one standard error of the cross validated misclassification risk evaluated at $\lambda_{opt}^{lasso}$. The choice $\lambda_{1se}^{lasso}$ induces a more parsimonious estimating model and is known as the "one-standard-error" rule, which is also commonly employed in the statistical learning literature (e.g., Hastie, Tibshirani, and Friedman, 2009).

We considered the following performance measures. Let $\widehat{\theta}$ denote the estimated coefficients under a given classification approach. For the logit_lasso approach, we derived $\widehat{\theta}$ by dividing the lasso-penalized logistic regression coefficients of the variables $\widetilde{X}$ by the magnitude of that of the variable $X_1$. We can easily deduce that $b_{\theta^*}(X) = 1\{X_1 + \widetilde{X}'\theta^* \geq 0\}$ is the Bayes classifier in this simulation design. To assess the classification performance, we report the relative risk, which is the ratio of the misclassification risk evaluated at the classifier $b_{\widehat{\theta}}$ over that evaluated at the Bayes classifier. In each simulation repetition, we approximated the out-of-sample misclassification risk using the generated validation sample. Let *in_RR* and *out_RR* respectively denote the average of in-sample and that of out-of-sample relative risks over all the simulation repetitions.

We also examine the covariate selection performance of the classification method. We say that a covariate $\widetilde{X}_j$ is effectively selected if and only if the magnitude of $\widehat{\theta}_j$ is larger than a small tolerance level (e.g., $10^{-6}$ as used in our numerical study) which is distinct from zero in numerical computation. Let *Corr_sel* be the proportion of the variable $\widetilde{X}_2$ being effectively selected. Let *Orac_sel* be the proportion of obtaining an oracle covariate selection outcome where the variable $\widetilde{X}_2$ was the only one that was effectively selected among all the variables in $\widetilde{X}$. Let *Num_irrel* denote the average number of effectively selected covariates whose true DGP coefficients are zero.

### *5.1. Simulation Results*

We now present in Tables 1 and 2 the simulation results under the setups of DGP(i) and DGP(ii) respectively. From these two tables, we find that, regarding the in-sample classification performance, our method outperformed the two logit_lasso based approaches across almost all the DGP configurations in the simulation. For the out-of-sample classification performance, we see that the $\ell_0$-penalized ERM classifier dominated the logit_lasso classifiers across all simulation scenarios and this dominance was more evident in the high dimensional setup with $p = 200$.

**Table 1.** Comparison of classification methods under DGP(i)

| | $p = 10$ | | | $p = 200$ | | |
|---|---|---|---|---|---|---|
| method | $\ell_0$-ERM | logit_lasso | | $\ell_0$-ERM | logit_lasso | |
| | | $\lambda_{opt}^{lasso}$ | $\lambda_{1se}^{lasso}$ | | $\lambda_{opt}^{lasso}$ | $\lambda_{1se}^{lasso}$ |
| *Corr_sel* | 0.98 | 1 | 0.99 | 0.94 | 0.99 | 0.87 |
| *Orac_sel* | 0.95 | 0 | 0 | 0.83 | 0 | 0 |
| *Num_irrel* | 0.03 | 3.05 | 1.45 | 0.15 | 7.26 | 2.62 |
| *in_RR* | 0.828 | 0.870 | 1.134 | 0.778 | 0.843 | 1.237 |
| *out_RR* | 1.094 | 1.168 | 1.304 | 1.139 | 1.313 | 1.471 |

**Table 2.** Comparison of classification methods under DGP(ii)

| | $p = 10$ | | | $p = 200$ | | |
|---|---|---|---|---|---|---|
| method | $\ell_0$-ERM | logit_lasso | | $\ell_0$-ERM | logit_lasso | |
| | | $\lambda_{opt}^{lasso}$ | $\lambda_{1se}^{lasso}$ | | $\lambda_{opt}^{lasso}$ | $\lambda_{1se}^{lasso}$ |
| *Corr_sel* | 0.91 | 1 | 0.91 | 0.83 | 0.95 | 0.84 |
| *Orac_sel* | 0.91 | 0 | 0 | 0.82 | 0 | 0 |
| *Num_irrel* | 0.01 | 2.99 | 1.48 | 0.05 | 9.86 | 2.72 |
| *in_RR* | 0.893 | 0.969 | 1.095 | 0.884 | 0.804 | 1.069 |
| *out_RR* | 1.071 | 1.160 | 1.248 | 1.103 | 1.271 | 1.289 |

Concerning the covariate selection results, both Tables 1 and 2 indicate that all the three classification approaches had high *Corr_sel* rates and hence were effective for selecting the relevant variable $\widetilde{X}_2$. However, the good performance in the *Corr_sel* criterion might just be a consequence of overfitting, which may result in excessive selection of irrelevant variables and thus adversely impact on the out-of-sample classification performance. From the results on the *Num_irrel* performance measure, we note that the numbers of irrelevant variables selected under the two logit_lasso based approaches remained quite large relatively to those under the $\ell_0$-penalized ERM approach even though all these approaches exhibited the effect of shrinking the covariate space dimension. In fact, we observe non-zero and high values of *Orac_Sel* for the $\ell_0$-classifier across all the simulation setups whereas the two logit_lasso classifiers could not induce any oracle variable selection outcome in the simulation. These covariate selection performance results help to explain that the risk performance dominance of the $\ell_0$-penalized ERM approach could be observed even in the DGP(i) simulations where the logistic regression model was correctly specified.

## 6. EMPIRICAL APPLICATION

We illustrate usefulness of the $\ell_0$-penalized ERM approach in an empirical application of work-trip transportation mode choice. We used the transportation mode dataset analyzed by Horowitz (1993). This dataset has also been extensively used as an illustrating empirical example in the discrete choice econometric literature (e.g., see Florios and Skouras (2008), Benoit and Van den Poel (2012, Section 4.3), Chen and Lee (2018) and the references therein).

The data consist of 842 observations sampled randomly from the Washington, D.C., area transportation study. Each record in the dataset contains the following information for a single work trip of the traveler: the chosen transportation mode, the number of cars owned by the traveler's household ($CARS$), the transit out-of-vehicle travel time minus automobile out-of-vehicle travel time in minutes ($DOVTT$), the transit in-vehicle travel time minus automobile in-vehicle travel time in minutes ($DIVTT$) and the transit fare minus automobile travel cost in dollars ($DCOST$).

The binary outcome $Y$ is the traveler's chosen mode of transportation such that $Y = 1$ if the choice is automobile and 0 otherwise. We considered binary prediction of $Y$ using the linear classifier (2.1) where we specified the variable $X_1$ to be $DCOST$ and $\widetilde{X}$ to be the vector consisting of an intercept term together with those linear, quadratic, cubic polynomial and cross-interacting terms constructed from the remaining three explanatory variables $Z = (CARS, DOVTT, DIVTT)$. This specification allowed us to approximate a smooth function of $Z$ by its cubic expansion, which resulted in a covariate vector $\widetilde{X}$ of dimension $p = 20$. See Table 3 below for details of the covariate specification. All stochastic covariates were standardized to have zero mean and unit variance.

We compared the predictive performance in terms of misclassification risk of our $\ell_0$-penalized ERM approach to that of both the logit_lasso and the subset selection approaches. We randomly selected one third of the entire data observations to form a training sample for the estimation of the classifier parameters. The remaining two thirds were used as the validation sample for the estimation of the out-of-sample misclassification risks. We used the same procedure as described in the simulation study of Section 5 to implement the logit_lasso approach. For the $\ell_0$-penalized ERM approach, we multiplied the scalar $v$ computed through (5.38) and (5.39) by an additional tuning factor $g_v$, which would further shrink or enlarge the benchmark penalty parameter value given by (5.37). Specifically, we used a 5-fold cross validation procedure to choose the best tuning value $g_v$ from the multiplier set $\{1/32, 1/16, 1/8, 1/4, 1/2, 1, 2\}$. Except for this modification, the remaining implementation details for the $\ell_0$-penalized ERM approach were the same as those set forth in Section 5. In particular, we also specified the parameter space $\Theta$ to be $[-10, 10]^p$, which was conservatively large as it allowed parameters to take values of an order of magnitude much larger than standard deviations of the covariates. The same setting of parameter space was also adopted in previous studies (Florios and Skouras (2008) and Chen and Lee (2018)) on MIO based binary choice estimation with the same dataset employed here.

For the subset selection approach, we constructed the classifier by solving the $\ell_0$-constrained ERM problem (2.6) with the subset size $m$ whose value was calibrated over a set of integers ranging from 1 to $p$ via a 5-fold cross validation procedure. To solve (2.6), we applied the MIO based computational algorithm of Chen and Lee (2018, Section 4.1) using the same parameter space and computing hardware and software configurations as in the $\ell_0$-penalized ERM approach. For the MIO computation for both our $\ell_0$-penalized and the $\ell_0$-constrained ERM problems, we ran the MIO solver until convergence and thus obtained numerically exact global solutions to these optimization problems.

Table 3 shows the estimated parameter values and summarizes both the predictive performance and variable selection results in this real data exercise conducted under our $\ell_0$-penalized ERM ($\ell_0$-ERM), the logit_lasso implemented with two calibrated lasso penalty values $\lambda_{opt}^{lasso}$ and $\lambda_{1se}^{lasso}$, and the aforementioned subset selection approaches. To interpret these empirical results, we notice that the $\ell_0$-ERM and subset selection approaches had comparable predictiove performance for both the in-sample and out-of-sample settings. Moreover, the predictive performance of these two approaches slightly dominated that of the logit_lasso with the optimal penalty parameter $\lambda_{opt}^{lasso}$ but dominated to a larger extent that of the logit_lasso with the alternative penalty parameter $\lambda_{1se}^{lasso}$. For the variable selection results, we note that the estimated sparsity, defined as the number of active variables excluding the variable $DCOST$, for our $\ell_0$-ERM approach was the smallest among all these classification approaches. By contrast, the estimated sparsity for the $\lambda_{opt}^{lasso}$ based logit_lasso approach substantially outnumbered that associated with both the $\ell_0$-ERM and the cross-validated subset selection approaches. Among the stochastic variables subject to selection, we find that $CARS$ was selected across all the classification methods and its parameter estimate was also mostly of a much larger magnitude than those of other selected covariates. This dovetails with our intuition that $CARS$ could be the most important predictor for the work-trip transportation mode choice. It is also noted that nonlinear expansion terms were selected under both the subset selection and logit_lasso approaches. Such nonlinearity did not emerge under the $\ell_0$-ERM approach which, while retaining commensurate predictive power, resulted in a more parsimonious set of selected covariates that did not include any interacting or higher order expansion terms.

## 7. CONCLUSIONS

In this paper, we study the binary classification problem in a setting with a high dimensional vector of covariates. We construct a binary classification procedure by minimizing the empirical misclassification risk with a penalty on the number of selected covariates. We establish a finite-sample probability bound showing that this classification approach can yield a sparse solution for covariate selection with a high probability. We also conduct non-asymptotic analysis on the excess misclassification risk and establish its rate of convergence. For implementation, we show that the penalized empirical risk minimization problem can be solved via the method of mixed integer linear programming.

There are a couple of topics one may consider as possible extensions. First, our theoretical results in Theorems 3.1 and 3.2 do not depend on how well the non-zero coefficients of an optimal classifier are separated from zero. It would be interesting to investigate additional theoretical aspects of our approach such as variable selection consistency under further assumptions on the uniqueness of optimal classifier and minimal separation of its non-zero coefficients (see, e.g., the beta-min condition in Chapter 7.4 of Bühlmann

**Table 3.** Estimated Results for the Prediction of Transportation Mode Choice

| method | $\ell_0$-ERM | subset selection | logit_lasso | |
|---|---|---|---|---|
| | | | $\lambda_{opt}^{lasso}$ | $\lambda_{1se}^{lasso}$ |
| $DCOST$ | 1 | 1 | 1 | 1 |
| $Intercept$ | 3.3963 | 5.8741 | 4.6849 | 3.4514 |
| $CARS$ | 2.6349 | 3.3458 | 3.2005 | 1.7105 |
| $DOVTT$ | 0 | 0 | 1.2969 | 0.3891 |
| $DIVTT$ | 0.8399 | 0 | 0 | 0.0554 |
| $CARS \times DOVTT$ | 0 | 0 | 0.4190 | 0 |
| $DOVTT \times DIVTT$ | 0 | 0 | 0 | 0 |
| $CARS \times DIVTT$ | 0 | 0 | 0 | 0 |
| $CARS \times CARS$ | 0 | 0 | -1.4702 | 0 |
| $DOVTT \times DOVTT$ | 0 | 0 | 0 | 0 |
| $DIVTT \times DIVTT$ | 0 | 7.1414 | 1.6478 | 0.1845 |
| $CARS \times CARS \times CARS$ | 0 | 0 | 0 | 0 |
| $DOVTT \times DOVTT \times DOVTT$ | 0 | 0 | -0.6540 | 0 |
| $DIVTT \times DIVTT \times DIVTT$ | 0 | 0 | 0 | 0 |
| $CARS \times CARS \times DOVTT$ | 0 | 0 | 0 | 0 |
| $CARS \times CARS \times DIVTT$ | 0 | 0 | 0 | 0 |
| $DOVTT \times DOVTT \times CARS$ | 0 | 0 | 0 | 0 |
| $DOVTT \times DOVTT \times DIVTT$ | 0 | -1.6542 | -0.6859 | 0 |
| $DIVTT \times DIVTT \times CARS$ | 0 | 0 | 0 | 0 |
| $DIVTT \times DIVTT \times DOVTT$ | 0 | 0 | 0 | 0 |
| $CARS \times DOVTT \times DIVTT$ | 0 | 0 | 0 | 0 |
| estimated sparsity | 3 | 4 | 8 | 5 |
| In-sample Risk | 0.0854 | 0.0819 | 0.0996 | 0.1281 |
| Out-of-sample Risk | 0.1070 | 0.1070 | 0.1087 | 0.1212 |

and Van De Geer, 2011). Second, our proposed method is suitable for training samples with small or moderate size. It would be a natural step to develop a divide-and-conquer algorithm for a large-scale problem (see, e.g., Shi, Lu, and Song, 2018). Third, our approach might be applicable for developing sparse policy learning rules (see, e.g., Athey and Wager, 2018). These are topics for further research.

## ACKNOWLEDGEMENTS

We would like to thank an anonymous referee for helpful comments. This work was supported in part by the Ministry of Science and Technology, Taiwan (MOST107-2410-H-001-029-MY2), Academia Sinica (AS-CDA-106-H01), the European Research Council (ERC-2014-CoG-646917-ROMIA), and the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the CeMMAP.

## REFERENCES

Abramovich, F., and V. Grinshtein (2019): “High-dimensional classification by sparse logistic regression,” *IEEE Transactions on Information Theory*, 65(5), 3068–3079.

Achterberg, T., and R. Wunderling (2013): “Mixed integer programming: Analyzing 12 years of progress," in *Facets of combinatorial optimization*, pp. 449–481. Springer.

Athey, S., and S. Wager (2018): "Efficient policy learning," *arXiv preprint arXiv:1702.02896.*

Benoit, D. F., and D. Van den Poel (2012): "Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution," *Journal of Applied Econometrics*, 27(7), 1174–1188.

Bertsimas, D., A. King, and R. Mazumder (2016): "Best subset selection via a modern optimization lens," *Annals of Statistics*, 44(2), 813–852.

Bertsimas, D., and R. Weismantel (2005): *Optimization over integers*, vol. 13. Dynamic Ideas Belmont.

Bickel, P. J., and E. Levina (2004): "Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, 10(6), 989–1010.

Boucheron, S., O. Bousquet, and G. Lugosi (2005): "Theory of classification: A survey of some recent advances," *ESAIM: probability and statistics*, 9, 323–375.

Bühlmann, P., and S. Van De Geer (2011): *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chen, L.-Y., and S. Lee (2018): "Best subset binary prediction," *Journal of Econometrics*, 206(1), 39–56.

Devroye, L., L. Györfi, and G. Lugosi (1996): *Probabilistic Theory of Pattern Recognition.* Springer.

Elliott, G., and R. Lieli (2013): "Predicting binary outcomes," *Journal of Econometrics*, 174(1), 15–26.

Fan, J., and Y. Fan (2008): "High dimensional classification using features annealed independence rules," *Annals of statistics*, 36(6), 2605–2637.

Fan, J., Y. Fan, and Y. Wu (2011): "High-dimensional classification," in *High-dimensional data analysis*, pp. 3–37. World Scientific.

Florios, K., and S. Skouras (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33(1), 1–22.

Greenshtein, E. (2006): "Best subset selection, persistence in high-dimensional statistical learning and optimization under $L_1$ constraint," *Annals of Statistics*, 34(5), 2367–2386.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Prediction, Inference and Data Mining*, vol. 2. Springer series in statistics New York.

Horowitz, J. L. (1993): "Semiparametric estimation of a work-trip mode choice model," *Journal of Econometrics*, 58(1), 49–70.

Huang, J., Y. Jiao, Y. Liu, and X. Lu (2018): "A Constructive Approach to $L_0$ Penalized Regression," *Journal of Machine Learning Research*, 19(10), 1–37.

Jiang, W., and M. A. Tanner (2010): "Risk Minimization for Time Series Binary Choice with Variable Selection," *Econometric Theory*, 26(5), 1437–1452.

Johnson, D., and F. Preparata (1978): "The densest hemisphere problem," *Theoretical Computer Science*, 6(1), 93–107.

Jünger, M., T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey (2009): *50 years of integer programming 1958-2008: From the early years to the state-of-the-art*. Springer Science & Business Media.

Kitagawa, T., and A. Tetenov (2018): "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 86(2), 591–616.

Lugosi, G. (2002): "Pattern Classification and Learning Theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi, pp. 1–56. Springer.

Manski, C. F. (1985): "Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83(403), 729–738.

Mbakop, E., and M. Tabord-Meehan (2018): "Model Selection for Treatment Choice: Penalized Welfare Maximization," *arXiv preprint arXiv:1609.03167*.

Nemhauser, G. L., and L. A. Wolsey (1999): *Integer and combinatorial optimization*. Wiley-Interscience.

Qian, J., T. Hastie, J. Friedman, R. Tibshirani, and N. Simon (2013): *Glmnet for Matlab* `http://www.stanford.edu/~hastie/glmnet_matlab/`.

Shi, C., W. Lu, and R. Song (2018): "A Massive Data Framework for M-Estimators with Cubic-Rate," *Journal of the American Statistical Association*, 113(524), 1698–1709.

Vapnik, V. (2000): *The nature of statistical learning theory*, vol. 2. Springer-Verlag New York.

## A. PROOFS OF THEORETICAL RESULTS

We shall use the following lemma in the proofs of Theorems 3.1 and 3.2.

Lemma 1.1. *For all $\sigma > 0$, there is a universal constant $M_\sigma$, which depends only on $\sigma$, such that*

$$P\left(\sup_{b\in\mathcal{B}_k} |S_n(b) - S(b)| > \sqrt{\frac{M_\sigma k \ln(p \vee n)}{n}}\right) \leq e^{-\sigma k \ln(p\vee n)} \tag{A.1}$$

*for any integer $k \in \{1, ..., p\}$ such that*

$$4\,(k+1)\ln\left(M_\sigma k \ln(p \vee n)\right) \leq k \ln(p \vee n) + 6\,(k+1)\ln 2. \tag{A.2}$$

The proof of Lemma 1.1 is a straightforward modification of that of Theorem 1 of Chen and Lee (2018). Hence we omit this proof here.

### *A.1. Proof of Theorem 3.1*

Proof. (Proof of Theorem 3.1) We first prove the probability bound (3.18). Let $\theta^* \equiv \arg\inf_{\theta\in\Theta_q} S(b_\theta)$. Because $b^* \in \mathcal{B}_q$, it is straightforward to see that

$$\begin{aligned} U_n &= \left[S(b_{\widehat{\theta}}) - S_n(b_{\widehat{\theta}}) - \lambda \left\|\widehat{\theta}\right\|_0\right] + \left[S_n(b_{\widehat{\theta}}) + \lambda \left\|\widehat{\theta}\right\|_0 - S(b^*)\right] \\ &\leq \left[S(b_{\widehat{\theta}}) - S_n(b_{\widehat{\theta}}) - \lambda \left\|\widehat{\theta}\right\|_0\right] + \left[S_n(b^*) + \lambda \left\|\theta^*\right\|_0 - S(b^*)\right] \\ &\leq \left|S_n(b_{\widehat{\theta}}) - S(b_{\widehat{\theta}})\right| + \sup\nolimits_{b\in\mathcal{B}_q} |S_n(b) - S(b)| + \lambda q - \lambda \left\|\widehat{\theta}\right\|_0. \end{aligned} \tag{A.3}$$

Since $U_n \geq 0$, it follows from (A.3) that

$$\left\|\widehat{\theta}\right\|_0 \leq q + 2\lambda^{-1}\Delta_n(\left\|\widehat{\theta}\right\|_0 \vee q) \tag{A.4}$$

where, for any $k \geq 0$,

$$\Delta_n(k) \equiv \sup_{b \in \mathcal{B}_k} \left|S_n(b) - S(b)\right|. \tag{A.5}$$

By construction, $0 \leq S_n(b) \leq 1$ for any indicator function $b$. We thus have that

$$\lambda \left\|\widehat{\theta}\right\|_0 \leq S_n(b_{\widehat{\theta}}) + \lambda \left\|\widehat{\theta}\right\|_0 \leq 1, \tag{A.6}$$

where the second inequality above follows by evaluating the objective function in (2.2) at the $\theta$ vector whose components are all zero. By (A.6), we can deduce that

$$\left\|\widehat{\theta}\right\|_0 \leq p \wedge \left\lfloor \lambda^{-1} \right\rfloor. \tag{A.7}$$

Note that, by (A.4), (A.7) and (3.16),

$$\left\|\widehat{\theta}\right\|_0 \leq q + 2\lambda^{-1}\Delta_n(m_0). \tag{A.8}$$

Given $\sigma > 0$, let $M_\sigma$ be the universal constant stated in Lemma 1.1. Let $\delta \equiv 2c^{-1}\sqrt{M_\sigma}$ and $c$ is the constant specified in Condition 3.2. Suppose $c$ is sufficiently large such that inequality (3.21) holds. Since $\epsilon \in (0,1)$, we have that

$$\delta \leq \epsilon\,(1+\epsilon)^{-1} < 1. \tag{A.9}$$

For each positive integer $j$, let

$$m_j \equiv q + \delta\sqrt{m_{j-1}}. \tag{A.10}$$

Given that $q \geq 1$, by (A.10), we have that

$$\begin{aligned} m_j &\leq q + \delta m_{j-1} \\ &\leq q + \delta q + \delta^2 m_{j-2} \\ &\quad .... \\ &\leq q\frac{1-\delta^j}{1-\delta} + \delta^j m_0. \end{aligned}$$

Therefore, by (3.20), we have that, for all $j \geq j_0$,

$$m_j \leq q\frac{1-\delta^j}{1-\delta} + \epsilon \leq s, \tag{A.11}$$

where the second inequality follows from (A.9).

By (A.4), we have that

$$\begin{aligned} P\left(\left\|\widehat{\theta}\right\|_0 \leq m_j\right) &\leq P\left(\left\|\widehat{\theta}\right\|_0 \leq q + 2\lambda^{-1}\Delta_n(m_j)\right) \\ &\leq P\left(\left\|\widehat{\theta}\right\|_0 \leq q + 2\lambda^{-1}\Delta_n(m_j), \Delta_n(m_j) \leq \lambda c^{-1}\sqrt{M_\sigma m_j}\right) \\ &\quad + P\left(\Delta_n(m_j) > \lambda c^{-1}\sqrt{M_\sigma m_j}\right) \\ &\leq P\left(\left\|\widehat{\theta}\right\|_0 \leq m_{j+1}\right) + P\left(\Delta_n(m_j) > \lambda c^{-1}\sqrt{M_\sigma m_j}\right). \end{aligned}$$

Hence

$$P\left(\left\|\widehat{\theta}\right\|_0 > m_{j+1}\right) \leq P\left(\left\|\widehat{\theta}\right\|_0 > m_j\right) + P\left(\Delta_n(m_j) > \lambda c^{-1}\sqrt{M_\sigma m_j}\right). \tag{A.12}$$

By (A.7), $P\left(\left\|\widehat{\theta}\right\|_0 > m_0\right) = 0$. Using this fact and applying (A.12) recursively, we have that

$$P\left(\left\|\widehat{\theta}\right\|_0 > m_k\right) \leq \sum\nolimits_{j=0}^{k-1} P\left(\Delta_n(m_j) > \lambda c^{-1}\sqrt{M_\sigma m_j}\right). \tag{A.13}$$

Therefore, result (3.18) of Theorem 3.1 follows by noting that

$$\begin{aligned} & P\left(\left\|\widehat{\theta}\right\|_0 > s\right) \\ & \leq \sum\nolimits_{i=0}^{j_0-1} P\left(\Delta_n(m_i) > \lambda c^{-1}\sqrt{M_\sigma m_i}\right) && \text{(A.14)} \\ & \leq \sum\nolimits_{i=0}^{j_0-1} P\left(\Delta_n(\lfloor m_i \rfloor \wedge p) > \lambda c^{-1}\sqrt{M_\sigma\left(\lfloor m_i \rfloor \wedge p\right)}\right) && \text{(A.15)} \\ & \leq \sum\nolimits_{i=0}^{j_0-1} e^{-\sigma(\lfloor m_i \rfloor \wedge p)\ln(p\vee n)} && \text{(A.16)} \\ & \leq j_0 e^{-\sigma r_n}, && \text{(A.17)} \end{aligned}$$

where (A.14) follows from (A.11) and (A.13), (A.15) follows from the fact that $r \geq \lfloor r \rfloor \wedge p$ and $\Delta_n(r) = \Delta_n(\lfloor r \rfloor \wedge p)$ for all $r \geq 0$, and, because $q \leq m_i \leq m_0 \vee \left((j_0-1)\,q + \sqrt{m_0}\right)$ for $i \in \{0, 1, 2, ..., j_0 - 1\}$, (A.16) follows from an application of Lemma 1.1, where the value of $k$ in this lemma is taken over the range $\{q, q+1, ..., \left[m_0 \vee \left((j_0-1)\,q + \left\lfloor\sqrt{m_0}\right\rfloor\right)\right] \wedge p\}$.

### *A.2. Proof of Theorem 3.2*

Proof. We exploit the proof of Theorem 3.1 to show the probability bound (3.24). Specifically, using (A.3) and (A.5) and noting that $s \geq q$, we have that

$$U_n \leq 2\Delta_n(\left\|\widehat{\theta}\right\|_0 \vee q) + \lambda s.$$

Hence

$$\begin{aligned} P\left(U_n > 3\lambda s\right) & \leq P\left(\Delta_n(\left\|\widehat{\theta}\right\|_0 \vee q) > \lambda s, \left\|\widehat{\theta}\right\|_0 \leq s\right) + P\left(\left\|\widehat{\theta}\right\|_0 > s\right) \\ & \leq P\left(\Delta_n(\lfloor s \rfloor) > \lambda\sqrt{\lfloor s \rfloor}\right) + j_0 e^{-\sigma r_n} && \text{(A.18)} \\ & \leq P\left(\Delta_n(\lfloor s \rfloor) > c^{-1}\sqrt{M_\sigma}\lambda\sqrt{\lfloor s \rfloor}\right) + j_0 e^{-\sigma r_n} && \text{(A.19)} \\ & \leq e^{-\sigma\lfloor s \rfloor(p\vee n)} + j_0 e^{-\sigma r_n} && \text{(A.20)} \\ & \leq \left(1 + j_0\right) e^{-\sigma r_n}, && \text{(A.21)} \end{aligned}$$

where (A.18) follows from the probability bound (3.18) of Theorem 3.1, (A.19) follows from (3.21), which implies $c > \sqrt{M_\sigma}$, (A.20) follows from Lemma 1.1, and (A.21) follows from the fact that $q \leq s$.