entropy

MDPI

Article

Selection Consistency of Lasso-Based Procedures for Misspecified High-Dimensional Binary Model and Random Regressors

Mariusz Kubkowski [1,2,†] and Jan Mielniczuk [1,2,*,†]

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland; m.kubkowski@ipipan.waw.pl

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

* Correspondence: j.mielniczuk@ipipan.waw.pl

† These authors contributed equally to this work.

Received: 13 November 2019; Accepted: 24 January 2020; Published: 28 January 2020

Abstract: We consider selection of random predictors for a high-dimensional regression problem with a binary response for a general loss function. An important special case is when the binary model is semi-parametric and the response function is misspecified under a parametric model fit. When the true response coincides with a postulated parametric response for a certain value of parameter, we obtain a common framework for parametric inference. Both cases of correct specification and misspecification are covered in this contribution. Variable selection for such a scenario aims at recovering the support of the minimizer of the associated risk with large probability. We propose a two-step selection Screening-Selection (SS) procedure which consists of screening and ordering predictors by Lasso method and then selecting the subset of predictors which minimizes the Generalized Information Criterion for the corresponding nested family of models. We prove consistency of the proposed selection method under conditions that allow for a much larger number of predictors than the number of observations. For the semi-parametric case when distribution of random predictors satisfies linear regressions condition, the true and the estimated parameters are collinear and their common support can be consistently identified. This partly explains robustness of selection procedures to the response function misspecification.

Keywords: high-dimensional regression; loss function; random predictors; misspecification; consistent selection; subgaussianity; generalized information criterion; robustness

1. Introduction

Consider a random variable $(X, Y) \in R^p \times \{0, 1\}$ and a corresponding response function defined as a posteriori probability $q(x) = P(Y = 1|X = x)$. Estimation of the a posteriori probability is of paramount importance in machine learning and statistics since many frequently applied methods, e.g. logistic or tree-based classifiers, rely on it. One of the main estimation methods of q is a parametric approach for which the response function is assumed to have parametric form

$$q(x) = q_0(\beta^T x) \tag{1}$$

for some fixed β and known $q_0(x)$. If Equation (1) holds, that is the underlying structure is correctly specified, then it is known that

$$\beta = \text{argmin}_{b \in R^p} - \{E_{X,Y}(Y \log q_0(b^T X) + (1 - Y) \log(1 - q_0(b^T X))\}, \tag{2}$$

or, equivalently (cf., e.g., [1])

$$\beta = \text{argmin}_b E_X KL(q(X), q_0(X^T b)), \tag{3}$$

where $E_X f(X)$ is the expected value of a random variable $f(X)$ and $KL(q(X), q_0(X^T b))$ is Kullback–Leibler distance between the binary distributions with success probabilities $q(X)$ and $q_0(X^T b)$:

$$KL(q(X), q_0(X^T b)) = q(X) \log \frac{q(X)}{q_0(X^T b)} + (1 - q(X)) \log \frac{1 - q(X)}{1 - q_0(X^T b)}.$$

The equalities in Equations (2) and (3) form the theoretical underpinning of (conditional) maximum likelihood (ML) method as the expression under the expected value in Equation (2) is the conditional log-likelihood of Y given X in the parametric model. Moreover, it is a crucial property needed to show that ML estimates of β under appropriate conditions approximate β.

However, more frequently than not, the model in Equation (1) does not hold, i.e. response q is misspecified and ML estimators do not approximate β, but the quantity defined by the right-hand side of Equation (3), namely

$$\beta^* = \text{argmin}_b E_X KL(q(X), q_0(X^T b)), \tag{4}$$

Thus, parametric fit using conditional ML method, which is the most popular approach to modeling binary response, also has very intuitive geometric and information-theoretic flavor. Indeed, fitting a parametric model, we try to approximate the β^* which yields averaged KL projection of unknown q on set of parametric models $\{q_0(b^T x)\}_{b \in R^p}$. A typical situation is a semi-parametric framework the true response function satisfies when

$$q(x) = \tilde{q}(\beta^T x) \tag{5}$$

for some unknown $\tilde{q}(x)$ and the model in Equation (1) is fitted where $\tilde{q} \neq q_0$. An important problem is then how β^* in Equation (4) relates to β in Equation (5). In particular, a frequently asked question is what can be said about a support of $\beta = (\beta_1, \ldots, \beta_p)^T$, i.e. the set $\{i : \beta_i \neq 0\}$, which consists of indices of predictors which truly influence Y. More specifically, an interplay between supports of β and analogously defined support of β^* is of importance as the latter is consistently estimated and the support of ML estimator is frequently considered as an approximation of the set of true predictors. Variable selection, or equivalently the support recovery of β in high-dimensional setting, is one of the most intensively studied subjects in contemporary statistics and machine learning. This is related to many applications in bioinformatics, biology, image processing, spatiotemporal analysis, and other research areas (see [2–4]). It is usually studied under a correct model specification, i.e. under theassumption that data are generated following a given parametric model (e.g., logistic or, in the case of quantitative Y, linear model).

Consider the following example: let $\tilde{q}(x) = q_L(x^3)$, where $q_L(x) = e^x/(1 + e^x)$ is the logistic function. Define regression model by $P(Y = 1|X) = \tilde{q}(\beta^T X) = q_L((X_1 + X_2)^3)$, where $X = (X_1, \ldots, X_p)$ is $N(0, I_{p \times p})$-distributed vector of predictors, $p > 2$ and $\beta = (1, 1, 0, \ldots, 0) \in R^p$. Then, the considered model will obviously be misspecified when the family of logistic models is fitted. However , it turns out in this case that, as X is elliptically contoured, $\beta^* = \eta\beta = \eta(1, 1, 0, \ldots, 0)$ and $\eta \neq 0$ (see [5]) and thus supports of β and β^* coincide. Thus, in this case, despite misspecification variable selection , i.e. finding out that X_1 and X_2 are the only active predictors, it can be solved using the methods described below.

For recent contributions to the study of Kullback–Leibler projections on logistic model (which coincide with Equation (4) for a logistic loss, see below) and references, we refer to the works of Kubkowski and Mielniczuk [6], Kubkowski and Mielniczuk [7] and Kubkowski [8]. We also refer to the work of Lu *et al.* [9], where the asymptotic distribution of adaptive Lasso is studied under misspecification in the case of fixed number of deterministic predictors. Questions of robustness

analysis evolve around an interplay between β and β^*, in particular under what conditions the directions of β and β^* coincide (cf. the important contribution by Brillinger [10] and Ruud [11]).

In the present paper, we discuss this problem in a more general non-parametric setting. Namely, the minus conditional log-likelihood $-(y \log q_0(b^T x + (1-y)\log(1-q_0(b^T x))$ is replaced by a general loss function of the form

$$l(b, x, y) = \rho(b^T x, y), \tag{6}$$

where $\rho : R \times \{0,1\} \to R$ is some function, $b, x \in R^p$, $y \in \{0,1\}$, and

$$R(b) = E_{X,Y} l(b, X, Y)$$

is the associated risk function for $b \in R^p$. Our aim is to determine a support of β^*, where

$$\beta^* = \text{argmin}_{b \in R^{p_n}} R(b). \tag{7}$$

Coordinates of β^* corresponding to non-zero coefficients are called active predictors and vector β^* the pseudo-true vector.

The most popular loss functions are related to minus log-likelihood of specific parametric models such as logistic loss

$$l_{logist}(b, x, y) = -yb^T x + \log(1 + \exp(b^T x))$$

related to $q_0(b^T x) = \exp(b^T x)/(1 + \exp(b^T x)$, probit loss

$$l_{probit}(b, x, y) = -y \log \Phi(b^T x) + (1-y)\log(1 - \Phi(b^T x))$$

related to $q_0(b^T x) = \Phi(b^T x)$, or quadratic loss $l_{lin}(b, x, y) = (y - b^T x)^2/2$ related to linear regression and quantitative response. Other losses which do not correspond to any parametric model such as Huber loss (see [12]) are constructed with a specific aim to induce certain desired properties of corresponding estimators such as robustness to outliers. We show in the following that variable selection problem can be studied for a general loss function imposing certain analytic properties such as its convexity and Lipschitz property.

For fixed number p of predictors smaller than sample size n, the statistical consequences of misspecification of a semi-parametric regression model were intensively studied by H. White and his collaborators in the 1980s. The concept of a projection on the fitted parametric model is central to these investigations which show how the distribution of maximum likelihood estimator of β^* centered by β^* changes under misspecification (cf. e.g., [13,14]). However , for the case when $p > n$, the maximum likelihood estimator, which is a natural tool for fixed $p \leq n$ case, is ill-defined and a natural question arises: What can be estimated and by what methods?

The aim of the present paper is to study the above problem in high-dimensional setting. To this end, we introduce two-stage approach in which the first stage is based on Lasso estimation (cf., e.g., [2])

$$\hat{\beta}_L = \text{argmin}_{b \in R^{p_n}} \{R_n(b) + \lambda_L \sum_{i=1}^{p_n} |b_i|\} \tag{8}$$

where $b = (b_1, \ldots, b_{p_n})^T$ and the empirical risk $R_n(b)$ corresponding to $R(b)$ is

$$R_n(b) = n^{-1} \sum_{i=1}^{n} \rho(b^T X_i, Y_i).$$

Parameter $\lambda_L > 0$ is Lasso penalty, which penalizes large l_1-norms of potential candidates for a solution. Note that the criterion function in Equation (8) for $\rho(s, y) = \log(1 + \exp(-s(2y-1))$ can be viewed as penalized empirical risk for the logistic loss. Lasso estimator is thoroughly studied in the case of the linear model when considered loss is square loss (see, e.g., [2] and [4] for references

and overview of the subject) and some of the papers treat the case when such model is fitted to Y, which is not necessarily linearly dependent on regressors (cf. [15]). In this case, regression model is misspecified with respect to linear fit. However, similar results are scarce for other scenarios such as logistic fit under misspecification in particular. One of the notable exceptions is Negahban *et al.* [16], who studied the behavior of Lasso estimate i for a general loss function and possibly misspecified models.

The output of the first stage is Lasso estimate $\hat{\beta}_L$. The second stage consists in ordering of predictors according to the absolute values of corresponding non-zero coordinates of Lasso estimator and then minimization of Generalized Information Criterion (GIC) on the resulting nested family. This is a variant of SOS (Screening-Ordering-Selection) procedure introduced in [17]. Let $\hat{s}^*$ be the model chosen by GIC procedure.

Our main contributions are as follows:

- We prove that under misspecification when the sample size grows support $\hat{s}^*$ coincides with support of β^* with probability tending to 1. In the general framework allowing for misspecification this means that selection rule $\hat{s}^*$ is consistent, i.e. $P(\hat{s}^* = s^*) \to 1$ when $n \to \infty$. In particular, when the model in Equation (1) is correctly specified this means that we recover the support of the true vector β with probability tending to 1.
- We also prove approximation result for Lasso estimator when predictors are random and ρ is a convex Lipschitz function (cf. Theorem 1).
- A useful corollary of the last result derived in the paper is determination of sufficient conditions under which active predictors can be separated from spurious ones based on the absolute values of corresponding coordinates of Lasso estimator. This makes construction of nested family containing s^* with a large probability possible.
- Significant insight has been gained for fitting of parametric model when predictors are elliptically contoured (e.g., multivariate normal). Namely, it is known that in such situation $\beta^* = \eta\beta$, i.e. these two vectors are collinear ([5]). Thus, in the case when $\eta \neq 0$ we have that support s^* of β^* coincides with support s of β and the selection consistency of two-step procedure proved in the paper entails direction and support recovery of β. This may be considered as a partial justification of a frequent observation that classification methods are robust to misspecification of the model for which they are derived (see, e.g., [5,18]).

We now discuss how our results relate to previous results. Most of the variable selection methods in high-dimensional case are studied for deterministic regressors; here, our results concern random regressors with subgaussian distributions. Note that random regressors scenario is much more realistic for experimental data than deterministic one. The stated results to the best of our knowledge are not available for random predictors even when the model is correctly specified. As to novelty of SS procedure, for its second stage we assume that the number of active predictors is bounded by a deterministic sequence k_n tending to infinity and we minimize GIC on family $\mathcal{M}$ of models with sizes satisfying also this condition. Such exhaustive search has been proposed in [19] for linear models and extended to GLMs in [20] (cf. [21]). In these papers, GIC has been optimized on all possible subsets of regressors with cardinality not exceeding certain constant k_n. Such method is feasible for practical purposes only when p_n is small. Here, we consider a similar set-up but with important differences: $\mathcal{M}$ is a data-dependent small nested family of models and optimization of GIC is considered in the case when the original model is misspecified. The regressors are supposed random and assumptions are carefully tailored to this case. We also stress the fact that the presented results also cover the case when the regression model is correctly specified and Equation (5) is satisfied.

In numerical experiments, we study the performance of grid version of logistic and linear SOS and compare it to its several Lasso-based competitors.

The paper is organized as follows. Section 2 contains auxiliaries, including new useful probability inequalities for empirical risk in the case of subgaussian random variables (Lemma 2). In Section 3, we prove a bound on approximation error for Lasso when the loss function is convex and Lipschitz

and regressors are random (Theorem 1). This yields separation property of Lasso. In Theorems 2 and 3 of Section 4, we prove GIC consistency on nested family, which in particular can be built according to the order in which the Lasso coordinates are included in the fitted model. In Section 5.1, we discuss consequences of the proved results for semi-parametric binary model when distribution of predictors satisfies linear regressions condition. In Section 6, we numerically compare the performance of two-stage selection method for two closely related models, one of which is a logistic model and the second one is misspecified.

2. Definitions and Auxiliary Results

In the following, we allow random vector (X, Y), $q(x)$, and p to depend on sample size n, i.e., $(X, Y) = (X^{(n)}, Y^{(n)}) \in R^{p_n} \times \{0, 1\}$ and $q_n(x) = P(Y^{(n)} = 1 | X^{(n)} = x)$. We assume that n copies $X_1^{(n)}, \ldots, X_n^{(n)}$ of a random vector $X^{(n)}$ in R^{p_n} are observed together with corresponding binary responses $Y_1^{(n)}, \ldots, Y_n^{(n)}$. Moreover, we assume that observations $(X_i^{(n)}, Y_i^{(n)})$, $i = 1, \ldots, n$ are independent and identically distributed (iid). If this condition is satisfied for each n, but not necessarily for different n and m, i.e. distributions of $(X_i^{(n)}, Y_i^{(n)})$ may be different from that of $(X_j^{(m)}, Y_j^{(m)})$ or they may be dependent for $m \neq n$, then such framework is called a triangular scenario. A frequently considered scenario is the sequential one. In this case, when sample size n increases, we observe values of new predictors additionally to the ones observed earlier. This is a special case of the above scheme as then $X_i^{(n+1)} = (X_i^{(n)T}, X_{i,p_n+1}, \ldots, X_{i,p_{n+1}})^T$. In the following, we skip the upper index n if no ambiguity arises. Moreover, we write $q(x) = q_n(x)$. We impose a condition on distributions of random predictors assume that coordinates X_{ij} of X_i are subgaussian $Subg(\sigma_{jn}^2)$ with subgaussianity parameter σ_{jn}^2, i.e. it holds that (see [22])

$$E \exp(tX_{ij}) \leq \exp(t^2 \sigma_{jn}^2 / 2) \tag{9}$$

for all $t \in R$. This condition basically says that the tails of of X_{ij} do not decrease more slowly than tails of normal distribution $N(0, \sigma_{jn}^2)$. For future reference, let

$$s_n^2 = \max_{j=1,\ldots,p_n} \sigma_{jn}^2$$

and assume in the following that

$$\gamma^2 := \limsup_n s_n^2 < \infty. \tag{10}$$

We assume moreover that $X_{i1}, \ldots, X_{ip_n}$ are linearly independent in the sense that their arbitrary linear combination is not constant almost everywhere. We consider a general form of response function $q(x) = P(Y = 1 | X = x)$ and assume that for the given loss function β^*, as defined in Equation (7), exists and is unique. For $s \subseteq \{1, \ldots, p_n\}$, let $\beta^*(s)$ be defined as in Equation (7) when minimum is taken over b with support in s. We let

$$s^* = \text{supp}(\beta^*(\{1, \ldots, p_n\}) = \{i \leq p_n : \beta_i^* \neq 0\},$$

denote the support of $\beta^*(\{1, \ldots, p_n\})$ with $\beta^*(\{1, \ldots, p_n\}) = (\beta_1^*, \ldots, \beta_{p_n}^*)^T$.

Let $v_\pi = (v_{j_1}, \ldots, v_{j_k})^T \in R^{|\pi|}$ for $v \in R^{p_n}$ and $\pi = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p_n\}$. Let $\beta_{s^*}^* \in R^{|s^*|}$ be $\beta^* = \beta^*(\{1, \ldots, p_n\})$ restricted to its support s^*. Note that if $s^* \subseteq s$, then provided projections are unique (see Section 2) we have

$$\beta_{s^*}^* = \beta^*(s^*) = \beta^*(s)_{s^*}.$$

Note that this implies that for every superset $s \supseteq s^*$ of s the projection $\beta^*(s)$ on the model pertaining to s is obtained by appending projection $\beta^*(s^*)$ with appropriate number of zeros. Moreover, let

$$\beta^*_{min} = \min_{i \in s^*} |\beta^*_i|.$$

We remark that β^*, s^* and β^*_{min} may depend on n. We stress that β^*_{min} is an important quantity in the development here as it turns out that it may not decrease too quickly in order to obtain approximation results for $\hat{\beta}^*_L$ (see Theorem 1). Note that, when the parametric model is correctly specified, i.e. $q(x) = q_0(\beta^T x)$ for some β with l being an associated log-likelihood loss , if s is the support of β, we have $s = s^*$.

First, we discuss quantities and assumptions needed for the first step of SS procedure.

We consider cones of the form:

$$\mathcal{C}_\varepsilon = \{\Delta \in R^{p_n} \colon \ ||\Delta_{s^{*c}}||_1 \leq (3+\varepsilon)||\Delta_{s^*}||_1\}, \tag{11}$$

where $\varepsilon > 0$, $s^{*c} = \{1, \ldots, p_n\} \setminus s^*$ and $\Delta_{s^*} = (\Delta_{s^*_1}, \ldots, \Delta_{s^*_{|s^*|}})$ for $s^* = \{s^*_1, \ldots, s^*_{|s^*|}\}$. Cones $\mathcal{C}_\varepsilon$ are of special importance because we prove that $\hat{\beta}_L - \beta^* \in \mathcal{C}_\varepsilon$ (see Lemma 3). In addition, we note that since l^1-norm is decomposable in the sense that $||v_A||_1 + ||v_{A^c}||_1 = ||v||_1$ the definition of the cone above can be stated as

$$\mathcal{C}_\varepsilon = \{\Delta \in R^{p_n} \colon \ ||\Delta||_1 \leq (4+\varepsilon)||\Delta_{s^*}||_1\}.$$

Thus, $\mathcal{C}_\varepsilon$ consists of vectors which do not put too much mass on the complement of s^*. Let $H \in R^{p_n \times p_n}$ be a fixed non-negative definite matrix. For cone $\mathcal{C}_\varepsilon$, we define a quantity $\kappa_H(\varepsilon)$ which can be regarded as a restricted minimal eigenvalue of a matrix in high-dimensional set-up:

$$\kappa_H(\varepsilon) = \inf_{\Delta \in \mathcal{C}_\varepsilon \setminus \{0\}} \frac{\Delta^T H \Delta}{\Delta^T \Delta}. \tag{12}$$

In the considered context, H is usually taken as hessian $D^2R(\beta^*)$ and, e.g., for quadratic loss, it equals EX^TX. When H is non-negative definite and not strictly positive definite its smallest eigenvalue $\lambda_1 = 0$ and thus $\inf_{\Delta \in R^p \setminus \{0\}} \frac{\Delta^T H \Delta}{\Delta^T \Delta} = \lambda_1 = 0$. That is why we have to restrict minimization in Equation (12) in order to have $\kappa_H(\varepsilon) > 0$ in the high-dimensional case. As we prove that $\Delta_0 = \hat{\beta}_L - \beta^* \in \mathcal{C}_\varepsilon$ and would use $0 < \kappa_H(\varepsilon) \leq \Delta_0^T H \Delta_0 / \Delta_0^T \Delta_0$ it is useful to restrict minimization in Equation (12) to $\mathcal{C}_\varepsilon \setminus \{0\}$. Let R and R_n be the risk and the empirical risk defined above. Moreover, we introduce the following notation:

$$W(b) = R(b) - R(\beta^*), \tag{13}$$

$$W_n(b) = R_n(b) - R_n(\beta^*), \tag{14}$$

$$B_p(r) = \{\Delta \in R^{p_n} \colon \ ||\Delta||_p \leq r\}, \text{ for } p = 1, 2, \tag{15}$$

$$S(r) = \sup_{b \in R^{p_n} : b - \beta^* \in B_1(r)} |W(b) - W_n(b)|. \tag{16}$$

Note that $ER_n(b) = R(b)$. Thus, $S(r)$ corresponds to oscillation of centred empirical risk over ball $B_1(r)$. We need the following Margin Condition (MC) in Lemma 3 and Theorem 1:

(MC) There exist $\vartheta, \varepsilon, \delta > 0$ and non-negative definite matrix $H \in R^{p_n \times p_n}$ such that for all b with $b - \beta^* \in \mathcal{C}_\varepsilon \cap B_1(\delta)$ we have

$$R(b) - R(\beta^*) \geq \frac{\vartheta}{2}(b - \beta^*)^T H (b - \beta^*).$$

The above condition can be viewed as a weaker version of strong convexity of function R (when the right-hand side is replaced by $\vartheta||b-\beta^*||^2$) in the restricted neighbourhood of β^* (namely, in the intersection of ball $B_1(\delta)$ and cone $\mathcal{C}_\varepsilon$). We stress the fact that H is not required to be positive definite, as in Section 3 we use Condition (MC) together with stronger conditions than $\kappa_H(\varepsilon) > 0$ which imply that right hand side of inequality in (MC) is positive. We also do not require here twice differentiability of R. We note in particular that Condition (MC) is satisfied in the case of logistic loss, X being bounded random variable and $H = D^2R(\beta^*)$ (see [23–25]). It is also easily seen that that (MC) is satisfied for quadratic loss, X such that $E||X||_2^2 < \infty$ and $H = D^2R(\beta^*)$. Similar condition to (MC) (called Restricted Strict Convexity) was considered in [16] for empirical risk R_n:

$$R_n(\beta^*+\Delta) - R_n(\beta^*) \geq DR_n(\beta^*)^T\Delta + \kappa_L||\Delta||^2 - \tau^2(\beta^*)$$

for all $\Delta \in C(3, s^*)$, some $\kappa_L > 0$, and tolerance function τ. Note however that MC is a deterministic condition, whereas Restricted Strict Convexity has to be satisfied for random empirical risk function.

Another important assumption, used in Theorem 1 and Lemma 2, is the Lipschitz property of ρ :

(LL) $\exists L > 0\ \forall b_1, b_2 \in R, y \in \{0,1\}\colon\ |\rho(b_1,y) - \rho(b_2,y)| \leq L|b_1 - b_2|$.

Now, we discuss preliminaries needed for the development of the second step of SS procedure. Let $|w|$ stand for dimension of w. For the second step of the procedure we consider an arbitrary family $\mathcal{M} \subseteq 2^{\{1,\ldots,p_n\}}$ of models (which are identified with subsets of $\{1,\ldots,p_n\}$ and may be data-dependent) such that $s^* \in \mathcal{M}, \forall w \in \mathcal{M}\colon\ |w| \leq k_n$ a.e. and $k_n \in N_+$ is some deterministic sequence. We define Generalized Information Criterion (GIC) as:

$$GIC(w) = nR_n(\hat{\beta}(w)) + a_n|w|, \tag{17}$$

where

$$\hat{\beta}(w) = \operatorname*{arg\,min}_{b \in R^{p_n}\colon\ b_{w^c}=0_{|w^c|}} R_n(b)$$

is ML estimator for model w as minimization above is taken over all vectors b with support in w. Parameter $a_n > 0$ is some penalty factor depending on the sample size n which weighs how important is the complexity of the model described by the number of its variables $|w|$. Typical examples of a_n include:

- AIC (Akaike Information Criterion): $a_n = 2$;
- BIC (Bayesian Information Criterion): $a_n = \log n$; and
- EBIC(d) (Extended BIC): $a_n = \log n + 2d \log p_n$, where $d > 0$.

AIC, BIC and EBIC were introduced by Akaike [26], Schwarz [27], and Chen and Chen [19], respectively. Note that for $n \geq 8$ BIC penalty is larger than AIC penalty and in its turn EBIC penalty is larger than BIC penalty.

We study properties of $S_k(r)$ for $k = 1, 2$, where:

$$S_k(r) = \sup_{b \in D_k : b-\beta^* \in B_2(r)} |(W_n(b) - W(b)| \tag{18}$$

and is the maximal absolute value of the centred empirical risk $W_n(\cdot)$ and sets D_k for $k = 1, 2$ are defined as follows:

$$D_1 = \{b \in R^{p_n}\colon\ \exists w \in \mathcal{M}\colon\ |w| \leq k_n \wedge s^* \subset w \wedge \operatorname{supp} b \subseteq w\}, \tag{19}$$

$$D_2 = \{b \in R^{p_n}\colon\ \operatorname{supp} b \subset s^*\}. \tag{20}$$

The idea here is simply to consider sets D_i consisting of vectors having no more that k_n non-zero coordinates. However, for $s^* \leq k_n$, we need that for $b \in D_i$, we have $|\operatorname{supp}(b-\beta^*)| \leq k_n$, what we

exploit in Lemma 2. This entails additional condition in the definition of D_1. Moreover, in Section 4, we consider the following condition $C_\epsilon(w)$ for $\epsilon > 0$, $w \subseteq \{1, \ldots, p_n\}$ and some $\theta > 0$:

$C_\epsilon(w)$: $R(b) - R(\beta^*) \geq \theta ||b - \beta^*||_2^2$ for all $b \in R^{p_n}$ such that supp $b \subseteq w$ and $b - \beta^* \in B_2(\epsilon)$.

We observe also that, although Conditions (MC) and $C_\epsilon(w)$ are similar, they are not equivalent, as they hold for $v = b - \beta^*$ belonging to different sets: $B_1(r) \cap \mathcal{C}_\varepsilon$ and $B_2(\epsilon) \cap \{\Delta \in R^{p_n}:$ supp $\Delta \subseteq w\}$, respectively. If the minimal eigenvalue λ_{min} of matrix H in Condition (MC) is positive and Condition (MC) holds for $b - \beta^* \in B_1(r)$ (instead of for $b - \beta^* \in \mathcal{C}_\varepsilon \cap B_1(r)$), then we have for $b - \beta^* \in B_2(r/\sqrt{p_n}) \subseteq B_1(r)$:

$$R(b) - R(\beta^*) \geq \frac{\vartheta}{2}(b - \beta^*)^T H(b - \beta^*) \geq \frac{\vartheta \lambda_{min}}{2}||b - \beta^*||_2^2.$$

Furthermore, if λ_{max} is the maximal eigenvalue of H and Condition $C_\epsilon(w)$ holds for all $v = b - \beta^* \in B_2(r)$ without restriction on supp b, then we have for $b - \beta^* \in B_1(r) \subseteq B_2(r)$:

$$R(b) - R(\beta^*) \geq \theta ||b - \beta^*||_2^2 \geq \frac{\theta}{\lambda_{max}}(b - \beta^*)^T H(b - \beta^*).$$

Thus, Condition (MC) holds in this case. A similar condition to Condition $C_\epsilon(w)$ for empirical risk R_n was considered by Kim and Jeon [28] (formula (2.1)) in the context of GIC minimization. It turns out that Condition $C_\epsilon(w)$ together with $\rho(\cdot, y)$ being convex for all y and satisfying Lipschitz Condition (LL) are sufficient to establish bounds which ensure GIC consistency for $k_n \ln p_n = o(n)$ and $k_n \ln p_n = o(a_n)$ (see Corollaries 2 and 3). First, we state the following basic inequality. $W(v)$ and $S(r)$ are defined above the definition of Margin Condition.

Lemma 1. *(Basic inequality). Let $\rho(\cdot, y)$ be convex function for all y. If for some $r > 0$ we have*

$$u = \frac{r}{r + ||\hat{\beta}_L - \beta||_1}, \quad v = u\hat{\beta}_L + (1 - u)\beta^*,$$

then

$$W(v) + \lambda ||v - \beta^*||_1 \leq S(r) + 2\lambda ||v_{s^*} - \beta^*_{s^*}||_1.$$

The proof of the lemma is moved to the Appendix A. It follows from the lemma that, as in view of decomposability of l^1-distance we have $||v - \beta^*||_1 = ||(v - \beta^*)^*_s||_1 + ||(v - \beta^*)_{s^{*c}}||_1$, when $S(r)$ is small we have $||(v - \beta^*)_{s^{*c}}||_1$ is not large in comparison with $||(v - \beta^*)^*_s||_1$.

Quantities $S_k(r)$ are defined in Equation (18). Recall that $S_2(r)$ is an oscillation taken over ball $B_2(r)$, whereas $S_i, i = 1, 2$ are oscillations taken over $B_1(r)$ ball with restriction on the number of nonzero coordinates.

Lemma 2. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz Condition (LL). Assume that X_{ij} for $j \geq 1$ are subgaussian $Subg(\sigma_{jn}^2)$, where $\sigma_{jn} \leq s_n$. Then, for $r, t > 0$:*

1. $P(S(r) > t) \leq \frac{8Lrs_n\sqrt{\log(p_n \vee 2)}}{t\sqrt{n}}$,
2. $P(S_1(r) \geq t) \leq \frac{8Lrs_n\sqrt{k_n \ln(p_n \vee 2)}}{t\sqrt{n}}$,
3. $P(S_2(r) \geq t) \leq \frac{4Lrs_n\sqrt{|s^*|}}{t\sqrt{n}}$.

The proof of the Lemma above, which relies on Chebyshev inequality , symmetrization inequality (see Lemma 2.3.1 of [29]), and Talagrand–Ledoux inequality ([30], Theorem 4.12), is moved to the Appendix. In the case when β^* does not depend on n and thus its support does not change, Part 3 implies in particular that $S_2(r)$ is of the order $n^{-1/2}$ in probability.

3. Properties of Lasso for a General Loss Function and Random Predictors

The main result in this section is Theorem 1. The idea for the proof is based on fact that, if $S(r)$ defined in Equation (16) is sufficiently small (condition $S(r) \leq \bar{C}\lambda r$ is satisfied), then $\hat{\beta}_L$ lies in a ball $\{\Delta \in R^{p_n} : \ ||\Delta - \beta^*||_1 \leq r\}$ (see Lemma 3). Using a tail inequality for $S(r)$ proved in Lemma 2, we obtain Theorem 1. Note that $\kappa_H(\varepsilon)$ has to be bounded away from 0 (condition $2|s^*|\lambda \leq \kappa_H(\varepsilon)\vartheta\tilde{C}r$). Convexity of $\rho(\cdot, y)$ below is understood as convexity for both $y = 0, 1$.

Lemma 3. *Let $\rho(\cdot, y)$ be convex function and assume that $\lambda > 0$. Moreover, assume margin Condition (MC) with constants $\vartheta, \epsilon, \delta > 0$ and some non-negative definite matrix $H \in R^{p_n \times p_n}$. If for some $r \in (0, \delta]$ we have $S(r) \leq \bar{C}\lambda r$ and $2|s^*|\lambda \leq \kappa_H(\varepsilon)\vartheta\tilde{C}r$, where $\bar{C} = \varepsilon/(8 + 2\varepsilon)$ and $\tilde{C} = 2/(4 + \varepsilon)$, then*

$$||\hat{\beta}_L - \beta^*||_1 \leq r.$$

The proof of the lemma is moved to the Appendix.

The first main result provides an exponential inequality for $P\left(||\hat{\beta}_L - \beta^*||_1 \leq \beta^*_{min}/2\right)$. The threshold $\beta^*_{min}/2$ is crucial there as it ensures separation: $\max_{i \in s^{*c}} |\hat{\beta}_{L,i}| \leq \min_{i \in s^*} |\hat{\beta}_{L,i}|$ (see proof of Corollary 1).

Theorem 1. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz Condition (LL). Assume that $X_{ij} \sim Subg(\sigma^2_{jn})$, β^* exists and is unique, margin Condition (MC) is satisfied for $\varepsilon, \delta, \vartheta > 0$, non-negative definite matrix $H \in R^{p_n \times p_n}$ and let*

$$\frac{2|s^*|\lambda}{\vartheta\kappa_H(\varepsilon)} \leq \tilde{C} \min\left\{\frac{\beta^*_{min}}{2}, \delta\right\},$$

where $\tilde{C} = 2/(4 + \varepsilon)$. Then,

$$P\left(||\hat{\beta}_L - \beta^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \geq 1 - 2p_n e^{-\frac{n\epsilon^2\lambda^2}{A}},$$

where $A = 128L^2(4 + \varepsilon)^2 s_n^2$.

Proof. Let

$$m = \min\left\{\frac{\beta^*_{min}}{2}, \delta\right\}.$$

Lemmas 3 and 2 imply that:

$$P\left(||\hat{\beta}_L - \beta^*||_1 > \frac{\beta^*_{min}}{2}\right) \leq P\left(||\hat{\beta}_L - \beta^*||_1 > m\right) \leq P\left(S(m) > \bar{C}\lambda m\right)$$
$$\leq 2p_n e^{-\frac{n\epsilon^2\lambda^2}{128L^2(4+\varepsilon)^2 s_n^2}}.$$

□

Corollary 1. *(Separation property) If assumptions of Theorem 1 are satisfied,*

$$\lambda = \frac{8Ls_n(4 + \varepsilon)\phi}{\varepsilon}\sqrt{\frac{2\log(2p_n)}{n}}$$

for some $\phi > 1$ and $\kappa_H(\varepsilon) > d$ for some $d, \varepsilon > 0$ for large n, $|s^*|\lambda = o(\min\{\beta^*_{min}, 1\})$, then

$$P\left(||\hat{\beta}_L - \beta^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \to 1.$$

Moreover,

$$P\left(\max_{i\in s^{*c}}|\hat{\beta}_{L,i}| \leq \min_{i\in s^{*}}|\hat{\beta}_{L,i}|\right)\to 1.$$

Proof. The first part of the corollary follows directly from Theorem 1 and the observation that:

$$P\left(||\hat{\beta}_L-\beta^*||_1>\frac{\beta^*_{min}}{2}\right)\leq e^{\log(2p_n)-\frac{n\epsilon^2\lambda^2}{128L^2(4+\epsilon)^2s_n^2}}=e^{\log(2p_n)(1-\phi^2)}\to 0.$$

Now, we prove that condition $||\hat{\beta}_L-\beta^*||_1\leq\beta^*_{min}/2$ implies separation property

$$\max_{i\in s^{*c}}|\hat{\beta}_{L,i}|\leq\min_{i\in s^{*}}|\hat{\beta}_{L,i}|. \tag{21}$$

Indeed, observe that for all $j\in\{1,\ldots,p_n\}$ we have:

$$\frac{\beta^*_{min}}{2}\geq||\hat{\beta}_L-\beta^*||_1\geq|\hat{\beta}_{L,j}-\beta^*_j|. \tag{22}$$

If $j\in s^*$, then using triangle inequality yields:

$$|\hat{\beta}_{L,j}-\beta^*_j|\geq|\beta^*_j|-|\hat{\beta}_{L,j}|\geq\beta^*_{min}-|\hat{\beta}_{L,j}|.$$

Hence, from the above inequality and Equation (22), we obtain for $j\in s^*$: $|\hat{\beta}_{L,j}|\geq\beta^*_{min}/2$. If $j\in s^{*c}$, then $\beta^*_j=0$ and Equation (22) takes the form: $|\hat{\beta}_{L,j}|\leq\beta^*_{min}/2$. This ends the proof. □

We note that the separation property in Equation (21) means that when λ is chosen in an appropriate manner, recovery of s^* is feasible with a large probability if all predictors corresponding to absolute value of Lasso coefficient exceeding a certain threshold are chosen. The threshold unfortunately depends on unknown parameters of the model. However, separation property allows to restrict attention to nested family of models and thus to decrease significantly computational complexity of the problem. This is dealt with in the next section. Note moreover that if γ in Equation (10) is finite than λ defined in the Corollary is of order $(\log p_n/n)^{1/2}$, which is the optimal order of Lasso penalty in the case of deterministic regressors (see, e.g., [2]).

4. GIC Consistency for a a General Loss Function and Random Predictors

Theorems 2 and 3 state probability inequalities related to behavior of GIC on supersets and on subsets of s^*, respectively. In a nutshell, we show for supersets and subsets separately that the probability that the minimum of GIC is not attained at s^* is exponentially small. Corollaries 2 and 3 present asymptotic conditions for GIC consistency in the aforementioned situations. Corollary 4 gathers conclusions of Theorem 1 and Corollaries 1, 2, and 3 to show consistency of SS procedure (see [17]for consistency of SOS procedure for a linear model with dieterministic predictors) in case of subgaussian variables. Note that in Theorem below we want to consider minimization of GIC in Equation (23) over all supersets of s^* as in our applications $\mathcal{M}$ is data dependent. As the number of such possible subsets is at least $\binom{p_n-|s^*|}{k_n-|s^*|}$, the proof has to be more involved than using reasoning based on Bonferroni inequality.

Theorem 2. *Assume that $\rho(\cdot,y)$ is convex, Lipschitz function with constant $L>0$, $X_{ij}\sim Subg(\sigma^2_{jn})$, condition $C_\epsilon(w)$ holds for some $\epsilon,\theta>0$ and for every $w\subseteq\{1,\ldots,p_n\}$ such that $|w|\leq k_n$. Then, for any $r<\epsilon$, we have:*

$$P(\min_{w\in\mathcal{M}:s^*\subset w}GIC(w)\leq GIC(s^*))\leq 2p_ne^{-\frac{a_n^2}{k_nB}}+2p_ne^{-\frac{nD}{k_n}}, \tag{23}$$

where $B=32nL^2r^2k_ns_n^2$ and $D=\theta^2r^2/512L^2s_n^2$.

Proof. If $s^* \subset w \in \mathcal{M}$ and $\hat{\beta}(w) - \beta^* \in B_2(r)$,, then in view of inequalities $R_n(\hat{\beta}(s^*)) \leq R_n(\beta^*)$ and $R(\beta^*) \leq R(b)$ we have:

$$\begin{aligned} R_n(\hat{\beta}(s^*)) - R_n(\hat{\beta}(w)) &\leq \sup_{b \in D_1 \colon b-\beta^* \in B_2(r)} (R_n(\beta^*) - R_n(b)) \\ &\leq \sup_{b \in D_1 \colon b-\beta^* \in B_2(r)} ((R_n(\beta^*) - R(\beta^*)) - (R_n(b) - R(b))) \\ &\leq \sup_{b \in D_1 \colon b-\beta^* \in B_2(r)} |R_n(b) - R(b) - (R_n(\beta^*) - R(\beta^*))| \\ &= S_1(r). \end{aligned}$$

Note that $a_n(|w| - |s^*|) \geq a_n$. Hence, if we have for some $w \supset s^*$: $GIC(w) \leq GIC(s^*)$, then we obtain $nR_n(\hat{\beta}(s^*)) - nR_n(\hat{\beta}(w))) \geq a_n(|w| - |s^*|)$ and from the above inequality we have $S_1(r) \geq a_n/n$. Furthermore, if $\hat{\beta}(w) - \beta^* \in B_2(r)^c$ and $r < \epsilon$, then consider:

$$v = u\hat{\beta}(w) + (1-u)\beta^*,$$

where $u = r/(r + ||\hat{\beta}(w) - \beta^*||_2)$. Then

$$||v - \beta^*||_2 = u||\hat{\beta}(w) - \beta^*||_2 = r \cdot \frac{||\hat{\beta}(w) - \beta^*||_2}{r + ||\hat{\beta}(w) - \beta^*||_2} \geq \frac{r}{2},$$

as function $x/(x+r)$ is increasing with respect to x for $x > 0$. Moreover, we have $||v - \beta^*||_2 \leq r < \epsilon$. Hence, in view of $C_\epsilon(w)$ condition, we get:

$$R(v) - R(\beta^*) \geq \theta||v - \beta^*||_2^2 \geq \frac{\theta r^2}{4}.$$

From convexity of R_n, we have:

$$R_n(v) \leq u(R_n(\hat{\beta}(w)) - R_n(\beta^*)) + R_n(\beta^*) \leq R_n(\beta^*).$$

Let supp v denote the support of vector v. We observe that supp $v \subseteq$ supp $\hat{\beta}(w) \cup$ supp $\beta^* \subseteq w$, hence $v \in D_1$. Finally, we have:

$$S_1(r) \geq R_n(\beta^*) - R(\beta^*) - (R_n(v) - R(v)) \geq R(v) - R(\beta^*) \geq \frac{\theta r^2}{4}.$$

Hence, we obtain the following sequence of inequalities:

$$\begin{aligned} P(\min_{w \in \mathcal{M} : s^* \subset w} GIC(w) \leq GIC(s^*)) & \\ \leq P(S_1(r) \geq \frac{a_n}{n}, \forall w \in \mathcal{M} \colon \hat{\beta}(w) - \beta^* \in B_2(r)) & \\ + P(\exists w \in \mathcal{M} : s^* \subset w \wedge \hat{\beta}(w) - \beta^* \in B_2(r)^c) \leq P(S_1(r) \geq \frac{a_n}{n}) + P(S_1(r) \geq \frac{\theta r^2}{4}) & \\ \leq 2p_n e^{-\frac{a_n^2}{32nL^2r^2k_n s_n^2}} + 2p_n e^{-\frac{n\theta^2 r^2}{512L^2 k_n s_n^2}}. & \end{aligned}$$

□

Corollary 2. *Assume that the conditions of Theorem 2 hold and for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \dots, p_n\}$ such that $|w| \leq k_n$, $k_n \ln(p_n \vee 2) = o(n)$ and $\liminf_{n\to\infty} \frac{D_n a_n}{k_n \log(2p_n)} > 1$, where $D_n^{-1} = 128L^2 s_n^2 \phi/\theta$ for some $\phi > 1$. Then, we have*

$$P(\min_{w\in\mathcal{M}: s^*\subset w} GIC(w) \leq GIC(s^*)) \to 0.$$

Proof. We the choose allb radius r of $B_2(r)$ in a special way. Namely, we take:

$$r_n^2 = \frac{512\phi^2 L^2 s_n^2 \log(2p_n) k_n}{n\theta^2}$$

for some $\phi > 1$. In view of assumptions $r_n \to 0$. Consider n_0 such that $r_n < \epsilon$ for all $n \geq n_0$. Hence, the second term of the upper bound in Equation (23) for $r = r_n$ is equal to:

$$2p_n e^{-\frac{n\theta^2 r_n^2}{512L^2 k_n s_n^2}} = e^{\log(2p_n)(1-\phi^2)} \to 0.$$

Similarly, the first term of the upper bound in Equation (23) is equal to:

$$2p_n e^{-\frac{a_n^2}{32nL^2 r_n^2 k_n s_n^2}} = e^{\log(2p_n)\left(1-\frac{a_n^2\theta^2}{128^2 L^4 k_n^2 s_n^4 \phi^2 \log^2(2p_n)}\right)} = e^{\log(2p_n)\left(1-\frac{D_n^2 a_n^2}{k_n^2 \log^2(2p_n)}\right)} \to 0.$$

These two convergences end the proof.

□

The most restrictive condition of Corollary 2 is $\liminf_{n\to\infty} \frac{D_n a_n}{k_n \log(2p_n)} > 1$ which is slightly weaker than $k_n \ln(p_n \vee 2) = o(a_n)$. The following remark proved in the Appendix gives sufficient conditions for consistency of BIC and EBIC penalties, which do not satisfy condition $k_n \log(p_n) = o(a_n)$.

Remark 1. *If in Corollary 2 we assume $D_n \geq A$ for some $A > 0$, then condition $\liminf_{n\to\infty} \frac{D_n a_n}{k_n \log(2p_n)} > 1$ holds when:*

1) $a_n = \log n$ and $p_n < \frac{n^{\frac{A}{k_n(1+u)}}}{2}$ for some $u > 0$.
2) $a_n = \log n + 2\gamma \log p_n$, $k_n \leq C$ and $2A\gamma - (1+u)C \geq 0$, where $C, u > 0$.
3) $a_n = \log n + 2\gamma \log p_n$, $k_n \leq C$, $2A\gamma - (1+u)C < 0$, $p_n < Bn^\delta$, where $\delta = \frac{A}{(1+u)C - 2A\gamma}$ and $B = 2^{-(1+u)C}$.

Theorem 3 is an analog of Theorem 2 for subsets of s^*.

Theorem 3. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$, condition $C_\epsilon(s^*)$ holds for some $\epsilon, \theta > 0$, and $8a_n|s^*| \leq \theta n \min\{\epsilon^2, \beta_{min}^{*2}\}$. Then, we have:*

$$P(\min_{w\in\mathcal{M}: w\subset s^*} GIC(w) \leq GIC(s^*)) \leq \sqrt{2} e^{-n \min\{\epsilon, \beta_{min}^*\}^2 E},$$

where $E = \theta^2/2^{12} L^2 s_n^2 |s^*|$

Proof. Suppose that for some $w \subset s^*$ we have $GIC(w) \leq GIC(s^*)$. This is equivalent to:

$$nR_n(\hat{\beta}(s^*)) - nR_n(\hat{\beta}(w)) \geq a_n(|w| - |s^*|).$$

In view of inequalities $R_n(\hat{\beta}(s^*)) \leq R_n(\beta^*)$ and $a_n(|w| - |s^*|) \geq -a_n|s^*|$, we obtain:

$$nR_n(\beta^*) - nR_n(\hat{\beta}(w)) \geq -a_n|s^*|.$$

Let $v = u\hat{\beta}(w) + (1-u)\beta^*$ for some $u \in [0,1]$ to be specified later. From convexity of ρ, we consider:

$$nR_n(\beta^*) - nR_n(v) \geq nu(R_n(\beta^*) - R_n(\hat{\beta}(w))) \geq -ua_n|s^*| \geq -a_n|s^*|. \tag{24}$$

We consider two cases separately:

(1) $\beta^*_{min} > \epsilon$.
First, observe that

$$8a_n|s^*| \leq \theta\epsilon^2 n, \tag{25}$$

which follows from our assumption. Let $u = \epsilon/(\epsilon + ||\hat{\beta}(w) - \beta^*||_2)$ and

$$v = u\hat{\beta}(w) + (1-u)\beta^*. \tag{26}$$

Note that $||\hat{\beta}(w) - \beta^*||_2 \geq ||\beta^*_{s^*\setminus w}||_2 \geq \beta^*_{min}$. Then, as function $d(x) = x/(x+c)$ is increasing and bounded from above by 1 for $x, c > 0$, we obtain:

$$\epsilon \geq ||v - \beta^*||_2 = \frac{\epsilon||\hat{\beta}(w) - \beta^*||_2}{\epsilon + ||\hat{\beta}(w) - \beta^*||_2} \geq \frac{\epsilon\beta^*_{min}}{\epsilon + \beta^*_{min}} > \frac{\epsilon^2}{2\epsilon} = \frac{\epsilon}{2}. \tag{27}$$

Hence, in view of $C_\epsilon(s^*)$ condition, we have:

$$R(v) - R(\beta^*) > \theta\frac{\epsilon^2}{4}.$$

Using Equations (24)–(26) and the above inequality yields:

$$S_2(\epsilon) \geq R_n(\beta^*) - R(\beta^*) - (R_n(v) - R(v)) > \theta\frac{\epsilon^2}{4} - \frac{a_n}{n}|s^*| \geq \frac{\theta\epsilon^2}{8}.$$

Thus, in view of Lemma 2, we obtain:

$$P(\min_{w\in\mathcal{M}: w\subset s^*} GIC(w) \leq GIC(s^*)) \leq P\left(S_2(\epsilon) > \frac{\theta\epsilon^2}{8}\right) \leq \sqrt{2}e^{-\frac{n\theta^2\epsilon^2}{4096L^2s_n^2|s^*|}}. \tag{28}$$

(2) $\beta^*_{min} \leq \epsilon$.
In this case, we take $u = \beta^*_{min}/(\beta^*_{min} + ||\hat{\beta}(w) - \beta^*||_2)$ and define v as in Equation (26). Analogously, as in Equation (27), we have:

$$\frac{\beta^*_{min}}{2} \leq ||v - \beta^*||_2 \leq \beta^*_{min}.$$

Hence, in view of $C_\epsilon(s^*)$ condition, we have:

$$R(v) - R(\beta^*) \geq \theta\frac{\beta^{*2}_{min}}{4}.$$

Using Equation (24) and the above inequality yields:

$$S_2(\beta^*_{min}) \geq R_n(\beta^*) - R(\beta^*) - (R_n(v) - R(v)) \geq \theta\frac{\beta^{*2}_{min}}{4} - \frac{a_n}{n}|s^*| \geq \frac{\theta}{8}\beta^{*2}_{min}.$$

Thus, in view of Lemma 2, we obtain:

$$P(\min_{w\in\mathcal{M}: w\subset s^*} GIC(w) \leq GIC(s^*)) \leq P\left(S_2(\beta^*_{min}) \geq \frac{\theta}{8}\beta^{*2}_{min}\right) \leq \sqrt{2}e^{-\frac{n\theta^2\beta^{*2}_{min}}{2^{12}L^2s_n^2|s^*|}}. \tag{29}$$

By combining Equations (28) and (29), the theorem follows. □

Corollary 3. *Assume that loss $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$, condition $C_\epsilon(s^*)$ holds for some $\epsilon, \theta > 0$ and $a_n|s^*| = o(n \min\{1, \beta_{min}^*\}^2)$, then*

$$P(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \to 0.$$

Proof. First, observe that as $a_n \to \infty$

$$a_n|s^*| = o(n \min\{1, \beta_{min}^*\}^2)$$

implies

$$|s^*| = o(n \min\{1, \beta_{min}^*\}^2),$$

and thus in view of Theorem 3 we have

$$P(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \to 0.$$

□

5. Selection Consistency of SS Procedure

In this section, we combine the results of the two previous sections to establish consistency of a two-step SS procedure. It consists in construction of a nested family of models $\mathcal{M}$ using magnitude of Lasso coefficients and then finding the minimizer of GIC over this family. As $\mathcal{M}$ is data dependent to establish consistency of the procedure we use Corollaries 2 and 3 in which the minimizer of GIC is considered over *all* subsets and supersets of s^*.

SS (Screening and Selection) procedure is defined as follows:

1. Choose some $\lambda > 0$.
2. Find $\hat{\beta}_L = \arg\min_{b \in R^{p_n}} R_n(b) + \lambda||b||_1$.
3. Find $\hat{s}_L = \operatorname{supp} \hat{\beta}_L = \{j_1, \ldots, j_k\}$ such that $|\hat{\beta}_{L,j_1}| \geq \ldots \geq |\hat{\beta}_{L,j_k}| > 0$ and $j_1, \ldots, j_k \in \{1, \ldots, p_n\}$.
4. Define $\mathcal{M}_{SS} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_k\}\}$.
5. Find $\hat{s}^* = \arg\min_{w \in \mathcal{M}_{SS}} GIC(w)$.

The SS procedure is a modification of SOS procedure in [17] designed for linear models. Since ordering step considered in [17] is omitted in the proposed modification, we abbreviate the name to SS.

Corollary 4 and Remark 2 describe the situations when SS procedure is selection consistent. In it, we use the assumptions imposed in Sections 2 and 3 together with an assumption that support of s^* contains no more than k_n elements, where k_n is some deterministic sequence of integers. Let $\mathcal{M}_{SS}$ is nested family constructed in the step 4 of SS procedure.

Corollary 4. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$ and β^* exists and is unique. If $k_n \in N_+$ is some sequence, margin Condition (MC) is satisfied for some $\vartheta, \delta, \varepsilon > 0$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$ and the following conditions are fulfilled:*

- $|s^*| \leq k_n$,
- $P(\forall w \in \mathcal{M}_{SS} : |w| \leq k_n) \to 1$,
- $\liminf_n \kappa_H(\varepsilon) > 0$ *for some* $\varepsilon > 0$, *where H is non-negative definite matrix and $\kappa_H(\varepsilon)$ is defined in Equation (12)*,
- $\log(p_n) = o(n\lambda^2)$,
- $k_n\lambda = o(\min\{\beta_{min}^*, 1\})$,

- $k_n \log p_n = o(n)$,
- $k_n \log p_n = o(a_n)$,
- $a_n k_n = o(n \min\{\beta^*_{min}, 1\}^2)$,

then for SS procedure we have

$$P(\hat{s}^* = s^*) \to 1.$$

Proof. In view of Corollary 1, following from the separation property in Equation (22) we obtain $P(s^* \in \mathcal{M}_{SS}) \to 1$. Let:

$$A_1 = \{\min_{w \in \mathcal{M}_{SS}: w \supset s^*, |w| \leq k_n} GIC(w) \leq GIC(s^*)\},$$
$$A_2 = \{\min_{w \in \mathcal{M}_{SS}: w \supset s^*, |w| > k_n} GIC(w) \leq GIC(s^*)\},$$
$$B = \{\forall w \in \mathcal{M}_{SS} : |w| \leq k_n\}.$$

Then, we have again from the fact that $A_2 \cap B = \emptyset$, union inequality and Corollary 2:

$$P(\min_{w \in \mathcal{M}_{SS}: w \supset s^*} GIC(w) \leq GIC(s^*)) = P(A_1 \cup A_2) = P(A_1 \cup (A_2 \cap B^c))$$
$$\leq P(A_1) + P(B^c) \to 0. \tag{30}$$

In an analogous way, using $|s^*| \leq k_n$ and Corollary 3 yields:

$$P(\min_{w \in \mathcal{M}_{SS}: w \subset s^*} GIC(w) \leq GIC(s^*)) \to 0. \tag{31}$$

Now, observe that in view of definition of $\hat{s}^*$ and union inequality:

$$P(\hat{s}^* = s^*) = P(\min_{w \in \mathcal{M}_{SS}: w \neq s^*} GIC(w) > GIC(s^*))$$
$$\geq 1 - P(\min_{w \in \mathcal{M}_{SS}: w \subset s^*} GIC(w) \leq GIC(s^*))$$
$$- P(\min_{w \in \mathcal{M}_{SS}: w \supset s^*} GIC(w) \leq GIC(s^*)).$$

Thus, $P(\hat{s}^* = s^*) \to 1$ in view of the above inequality and Equations (30) and (31). □

5.1. Case of Misspecified Semi-Parametric Model

Consider now the important case of the misspecified semi-parametric model defined in Equation (5) for which function $\tilde{q}$ is unknown and may be arbitrary. An interesting question is whether information about β can be recovered when misspecification occurs. The answer is positive under some additional assumptions on distribution of random predictors. Assume additionally that X satisfies

$$E(X|\beta^T X) = u_0 + u\beta^T X, \tag{32}$$

where β is the true parameter. Thus, regressions of X given $\beta^T X$ have to be linear. We stress that conditioning $\beta^T X$ involves only the true β in Equation (5). Then, it is known (cf. [10], [11] and [5]) that $\beta^* = \eta\beta$ and $\eta \neq 0$ if $\mathrm{Cov}(Y, X) \neq 0$. Note that because β and β^* are collinear and $\eta \neq 0$ it follows that $s = s^*$. This is important in practical applications as it shows that a position of the optimal separating direction given by β can be consistently recovered. It is also worth mentioning that if Equation (32) is satisfied the direction of β coincides with the direction of the first canonical vector. We refer to the work of Kubkowski and Mielniczuk [7] for the proof and to the work o Kubkowski and Mielniczuk [6] for discussion and up-to date references to this problem. The linear regressions condition in Equation (32) is satisfied, e.g., by elliptically contoured distribution, in particular by multivariate normal. We note

that it is proved in [18] that Equation (32) approximately holds for the majority of β. When Equation (32) holds exactly, proportionality constant η can be calculated numerically for known $\tilde{q}$ and β. We can state thus the following result provided Equation (32) is satisfied.

Corollary 5. *Assume that Equation (32) and the assumptions of Corollary 4 are satisfied. Moreover,* $\mathrm{Cov}(Y, X) \neq 0$. *Then,* $P(\hat{s}^* = s) \to 1$.

Remark 2. *If* $p_n = O(e^{cn^\gamma})$ *for some* $c > 0$, $\gamma \in (0, 1/2)$, $\xi \in (0, 0.5 - \gamma)$, $u \in (0, 0.5 - \gamma - \xi)$, $k_n = O(n^\xi)$, $\lambda = C_n\sqrt{\log(p_n)/n}$, $C_n = O(n^u)$, $C_n \to +\infty$, $n^{-\frac{\gamma}{2}} = O(\beta^*_{min})$, $a_n = dn^{\frac{1}{2}-u}$, *then assumptions imposed on asymptotic behavior of parameters in Corollary 4 are satisfied.*

Note that p_n is allowed to grow exponentially: $\log p_n = O(n^\gamma)$, however β^*_{min} may not decrease to 0 too quickly with regard to growth of p_n: $n^{-\frac{\gamma}{2}} = O(\beta^*_{min})$.

Remark 3. *We note that, to apply Corollary 4 to the two-step procedure based on Lasso, it is required that* $|s^*| \leq k_n$ *and that the support of Lasso estimator with probability tending to* 1 *contains no more than* k_n *elements. Some results bounding* $|\operatorname{supp} \hat{\beta}_L|$ *are available for deterministic X (see [31]) and for random X (see [32]), but they are too weak to be useful for EBIC penalties. The other possibility to prove consistency of two-step procedure is to modify it in the first step by using thresholded Lasso (see [33]) corresponding to* k'_n *largest Lasso coefficients where* $k'_n \in N$ *is such that* $k_n = o(k'_n)$. *This is a subject of ongoing research.*

6. Numerical Experiments

6.1. Selection Procedures

We note that the original procedure is defined for a single λ only. In the simulations discussed below, we implemented modifications of SS procedure introduced in Section 5. In practice, it is generally more convenient to consider in the first step some sequence of penalty parameters $\lambda_1 > \ldots > \lambda_m > 0$ instead of only one λ in order to avoid choosing the "best" λ. For the fixed sequence $\lambda_1, \ldots, \lambda_m$, we construct corresponding families $\mathcal{M}_1, \ldots, \mathcal{M}_m$ analogously to $\mathcal{M}$ in Step 4 of the SS procedure. Thus, we arrive at the following SSnet procedure, which is the modification of SOSnet procedure in [17]. Below, $\tilde{b}$ is a vector b with first coordinate corresponding to intercept omitted, $b = (b_0, \tilde{b}^T)^T$:

1. Choose some $\lambda_1 > \ldots > \lambda_m > 0$.
2. Find $\hat{\beta}^{(i)}_L = \underset{b \in R^{p_n+1}}{\arg\min}\, R_n(b) + \lambda_i ||\tilde{b}||_1$ for $i = 1, \ldots, m$.
3. Find $\hat{s}^{(i)}_L = \operatorname{supp} \hat{\tilde{\beta}}^{(i)}_L = \{j^{(i)}_1, \ldots, j^{(i)}_{k_i}\}$ where $j^{(i)}_1, \ldots, j^{(i)}_{k_i}$ are such that $|\hat{\beta}^{(i)}_{L,j^{(i)}_1}| \geq \ldots \geq |\hat{\beta}^{(i)}_{L,j^{(i)}_{k_i}}| > 0$ for $i = 1, \ldots, m$.
4. Define $\mathcal{M}_i = \{\{j^{(i)}_1\}, \{j^{(i)}_1, j^{(i)}_2\}, \ldots, \{j^{(i)}_1, j^{(i)}_2, \ldots, j^{(i)}_{k_i}\}\}$ for $i = 1, \ldots, m$.
5. Define $\mathcal{M} = \{\emptyset\} \cup \bigcup_{i=1}^{m} \mathcal{M}_i$.
6. Find $\hat{s}^* = \underset{w \in \mathcal{M}}{\arg\min}\, GIC(w)$, where

$$GIC(w) = \min_{b \in R^{p_n+1} : \operatorname{supp} \tilde{b} \subseteq w} nR_n(b) + a_n(|w| + 1).$$

Instead of constructing families $\mathcal{M}_i$ for each λ_i in SSnet procedure, λ can be chosen by cross-validation using 1SE rule (see [34]) and then SS procedure is applied for such λ. We call this procedure SSCV. The last procedure considered was introduced by Fan and Tang [35] and is Lasso procedure with penalty parameter $\hat{\lambda}$ chosen in a data-dependent way analogously to SSCV. Namely,

it is the minimizer of GIC criterion with $a_n = \log(\log n) \cdot \log p_n$ for which ML estimator has been replaced by Lasso estimator with penalty λ. Once $\hat{\beta}_L(\hat{\lambda}_L)$ is calculated, then $\hat{s}^*$ is defined as its support. The procedure is called LFT in the sequel.

We list below versions of the above procedures along with R packages that were used to choose sequence $\lambda_1, \ldots, \lambda_m$ and computation of Lasso estimator. The following packages were chosen based on selection performance after initial tests for each loss and procedure:

- SSnet with logistic or quadratic loss: `ncvreg`;
- SSCV or LFT with logistic or quadratic loss: `glmnet`; and
- SSnet, SSCV or LFT with Huber loss (cf. [12]): `hqreg`.

The following functions were used to optimize R_n in GIC minimization step for each loss:

- logistic loss: `glm.fit` (package `stats`);
- quadratic loss: `.lm.fit` (package `stats`); and
- Huber loss: `rlm` (package `rlm`).

Before applying the investigated procedures, each column of matrix $\mathbb{X} = (X_1, \ldots, X_n)^T$ was standardized as Lasso estimator $\hat{\beta}_L$ depends on scaling of predictors. We set length of λ_i sequence to $m = 20$. Moreover, in all procedures we considered only λ_i for which $|\hat{s}_L^{(i)}| \leq n$ because, when $|\hat{s}_L^{(i)}| > n$, Lasso and ML solutions are not unique (see [32,36]). For Huber loss, we set parameter $\delta = 1/10$ (see [12]). The number of folds in SSCV was set to $K = 10$.

Each simulation run consisted of L repetitions, during which samples $\mathbb{X}_k = (X_1^{(k)}, \ldots, X_n^{(k)})^T$ and $\mathbf{Y}_k = (Y_1^{(k)}, \ldots, Y_n^{(k)})^T$ were generated for $k = 1, \ldots, L$. For kth sample $(\mathbb{X}_k, \mathbf{Y}_k)$ estimator $\hat{s}_k^*$ of set of active predictors was obtained by a given procedure as the support of $\hat{\tilde{\beta}}(\hat{s}_k^*)$, where

$$\hat{\beta}(\hat{s}_k^*) = (\hat{\beta}_0(\hat{s}_k^*), \hat{\tilde{\beta}}(\hat{s}_k^*)^T)^T = \underset{b \in R^{p_n+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho(b^T X_i^{(k)}, Y_i^{(k)})$$

is ML estimator for kth sample. We denote by $\mathcal{M}^{(k)}$ the family $\mathcal{M}$ obtained by a given procedure for kth sample.

In our numerical experiments we have computed the following measures of selection performance which gauge co-direction of true parameter β and $\hat{\beta}$ and the interplay between s^* and $\hat{s}^*$:

- $ANGLE = \frac{1}{L} \sum_{k=1}^{L} \arccos |\cos \angle(\tilde{\beta}_0, \hat{\tilde{\beta}}(\hat{s}_k^*))|$, where

$$\cos \angle(\tilde{\beta}, \hat{\tilde{\beta}}(\hat{s}_k^*)) = \frac{\sum_{j=1}^{p_n} \beta_j \hat{\beta}_j(\hat{s}_k^*)}{||\tilde{\beta}||_2 ||\hat{\tilde{\beta}}(\hat{s}_k^*)||_2}$$

  and we let $\cos \angle(\tilde{\beta}, \hat{\tilde{\beta}}(\hat{s}_k^*)) = 0$, if $||\tilde{\beta}||_2 ||\hat{\tilde{\beta}}(\hat{s}_k^*)||_2 = 0$,
- $P_{inc} = \frac{1}{L} \sum_{k=1}^{L} I(s^* \in \mathcal{M}^{(k)})$,
- $P_{equal} = \frac{1}{L} \sum_{k=1}^{L} I(\hat{s}_k^* = s^*)$.
- $P_{supset} = \frac{1}{L} \sum_{k=1}^{L} I(\hat{s}_k^* \supseteq s^*)$.

Thus, $ANGLE$ is equal an of angle between true parameter (with intercept omitted) and its post model-selection estimator averaged over simulations, P_{inc} is a fraction of simulations for which family $\mathcal{M}^{(k)}$ contains true model s^*, and P_{equal} and P_{supset} are the fractions of time when SSnet chooses true model or its superset, respectively.

6.2. Regression Models Considered

To investigate behavior of two-step procedure under misspecification we considered two similar models with different sets of predictors. As sets of predictors differ, this results in correct specification of the first model (Model M1) and misspecification of the second (Model M2).

Namely, in Model M1, we generated n observations $(X_i, Y_i) \in R^{p+1} \times \{0,1\}$ for $i = 1, \ldots, n$ such that:

$$X_{i0} = 1, X_{i1} = Z_{i1}, X_{i2} = Z_{i2}, X_{ij} = Z_{i,j-7} \text{ for } j = 10, \ldots, p,$$
$$X_{i3} = X_{i1}^2, X_{i4} = X_{i2}^2, X_{i5} = X_{i1}X_{i2},$$
$$X_{i6} = X_{i1}^2 X_{i2}, X_{i7} = X_{i1}X_{i2}^2, X_{i8} = X_{i1}^3, X_{i9} = X_{i2}^3,$$

where $Z_i = (Z_{i1}, \ldots, Z_{ip})^T \sim \mathcal{N}_p(0_p, \Sigma)$, $\Sigma = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1,1)$. We consider response function $q(x) = q_L(x^3)$ for $x \in R$, $s = \{1,2\}$ and $\beta_s = (1,1)^T$. Thus,

$$\begin{aligned} P(Y_i = 1 | X_i = x_i) &= q(\beta_s^T x_{i,s}) = q(x_{i1} + x_{i2}) = q_L((x_{i1} + x_{i2})^3) \\ &= q_L(x_{i1}^3 + x_{i2}^3 + 3x_{i1}^2 x_{i2} + 3x_{i1}x_{i2}^2) \\ &= q_L(3x_{i6} + 3x_{i7} + x_{i8} + x_{i9}). \end{aligned}$$

We observe that the last equality implies that the above binary model is correctly specified with respect to family of fitted logistic models and X_6, X_7, X_8 and X_9 are four active predictors, whereas the remaining ones play no role in prediction of Y. Hence, $s^* = \{6,7,8,9\}$ and $\beta_{s^*}^* = (3,3,1,1)^T$ are, respectively, sets of indices of active predictors and non-zero coefficients of projection onto family of logistic models.

We considered the following parameters in numerical experiments: $n = 500, p = 150, \rho \in \{-0.9 + 0.15 \cdot k\colon\ k = 0,1,\ldots,12\}$, and $L = 500$ (the number of generated datasets for each combination of parameters). We investigated procedures SSnet, SSCV, and LFT using logistic, quadratic, and Huber (cf. [12]) loss functions. For procedures SSnet and SSCV, we used GIC penalties with:

- $a_n = \log n$ (BIC); and
- $a_n = \log n + 2 \log p_n$ (EBIC1).

In Model M2, we generated n observations $(X_i, Y_i) \in R^{p+1} \times \{0,1\}$ for $i = 1, \ldots, n$ such that $X_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^T$ and $(X_{i1}, \ldots, X_{ip})^T \sim \mathcal{N}_p(0_p, \Sigma)$, $\Sigma = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1,1)$. Response function is $q(x) = q_L(x^3)$ for $x \in R$, $s = \{1,2\}$ and $\beta_s = (1,1)^T$. This means that:

$$P(Y_i = 1 | X_i = x_i) = q(\beta_s^T x_{i,s}) = q(x_{i1} + x_{i2}) = q_L((x_{i1} + x_{i2})^3)$$

This model in comparison to Model M1 does not contain monomials of X_{i1} and X_{i2} of degree higher than 1 in its set of predictors. We observe that this binary model is misspecified with respect to fitted family of logistic models, because $q(x_{i1} + x_{i2}) \not\equiv q_L(\beta^T x_i)$ for any $\beta \in R^{p+1}$. However, in this case, the linear regressions condition in Equation (32) is satisfied for X, as it follows normal distribution (see [5,7]) . Hence, in view of Proposition 3.8 in [6], we have $s_{log}^* = \{1,2\}$ and $\beta_{log,s_{log}^*}^* = \eta(1,1)^T$ for some $\eta > 0$. Parameters n, p, ρ as well as L were chosen as for Model M1.

6.3. Results for Models M1 and M2

We first discuss the behavior of P_{inc}, P_{equal} and P_{supset} for the considered procedures. We observe that values of P_{inc} for SSCV and SSnet are close to 1 for low correlations in Model M2 for every tested loss (see Figure 1). In Model M1, P_{inc} attains the largest values for SSnet procedure and logistic loss for low correlations, which is because in most cases the corresponding family $\mathcal{M}$ is the largest among the families created by considered procedures. P_{inc} is close to 0 in Model M1 for quadratic and Huber loss, which results in low values of the remaining indices. This may be due to strong dependences between

predictors in Model M1; note that we have, e.g. $\mathrm{Cor}(X_{i1}, X_{i8}) = 3/\sqrt{15} \approx 0.77$. It is seen that in Model M1 inclusion probability P_{inc} is much lower than in Model M2 (except for negative correlations). It it also seen that P_{inc} for SSCV is larger than for LFT and LFT fails with respect to P_{inc} in M1.

In Model M1, the largest values P_{equal} are attained for SSnet with BIC penalty, the second best is SSCV with EBIC1 penalty (see Figure 2). In Model M2, P_{equal} is close to 1 for SSnet and SSCV with EBIC1 penalty and is much larger than P_{equal} for the corresponding versions using BIC penalty. We also note that choice of loss is relevant only for larger correlations. These results confirm theoretical result of Theorem 2.1 in [5], which show that collinearity holds for broad class of loss function. We observe also that, although in Model M2 remaining procedures do not select s^* with high probability, they select its superset, what is indicated by values of P_{supset} (see Figure 3). This analysis is confirmed by an analysis of $ANGLE$ measure (see Figure 4), which attains values close to 0, when P_{supset} is close to 1. Low values of $ANGLE$ measure mean that estimated vector $\hat{\tilde{\beta}}(\hat{s}_k^*)$ is approximately proportional to $\tilde{\beta}$, which is the case for Model M2, where normal predictors satisfy linear regressions condition. Note that the angles of $\hat{\tilde{\beta}}(\hat{s}_k^*)$ and $\tilde{\beta}^*$ in Model M1 significantly differ even though Model M1 is well specified. In addition, for the best performing procedures in both models and *any* loss considered, P_{equal} is much larger in Model M2 than in Model M1, even though the latter is correctly specified. This shows that choosing a simple misspecified model which retains crucial characteristics of the well specified large model instead of the latter might be beneficial.

In Model M1, procedures with BIC penalty perform better than those with EBIC1 penalty; however, the gain for P_{equal} is much smaller than the gain when using EBIC1 in Model M2. LFT procedure performs poorly in Model M1 and reasonably well in Model M2. The overall winner in both models is SSnet. SSCV performs only slightly worse than SSnet in Model M2 but performs significantly worse in Model M1.

Analysis of computing times of the first and second stages of each procedure shows that SSnet procedure creates large families $\mathcal{M}$ and GIC minimization becomes computationally intensive. We also observe that the first stage for SSCV is more time consuming than for SSnet, what is caused by multiple fitting of Lasso in cross-validation. However, SSCV is much faster than SSnet in the second stage.

We conclude that in the considered experiments SSnet with EBIC1 penalty works the best in most cases; however, even for the winning procedure, strong dependence of predictors results in deterioration of its performance. It is also clear from our experiments that a choice of GIC penalty is crucial for its performance. Modification of SS procedure which would perform satisfactorily for large correlations is still an open problem.

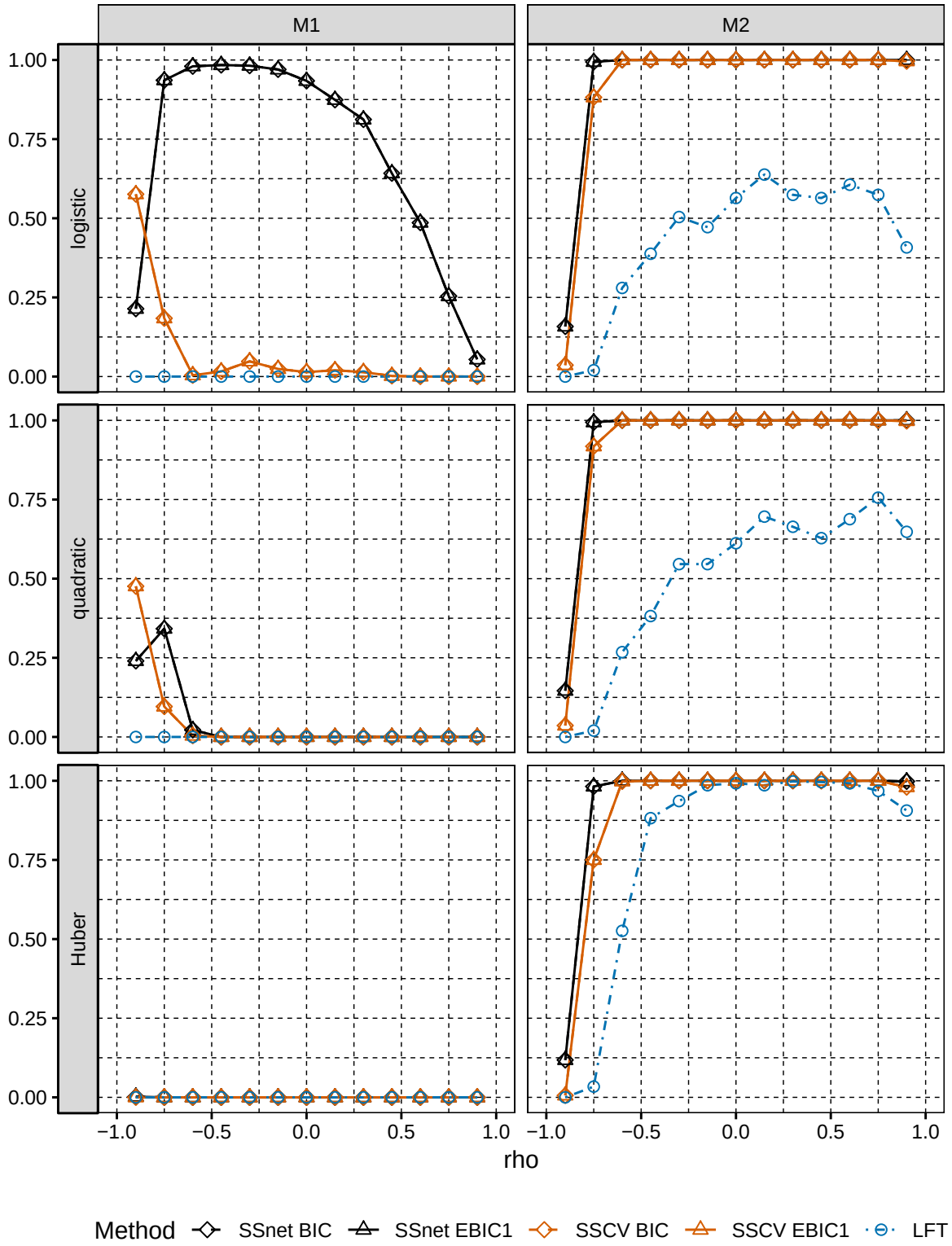

Figure 1. P_{inc} for Models M1 and M2.

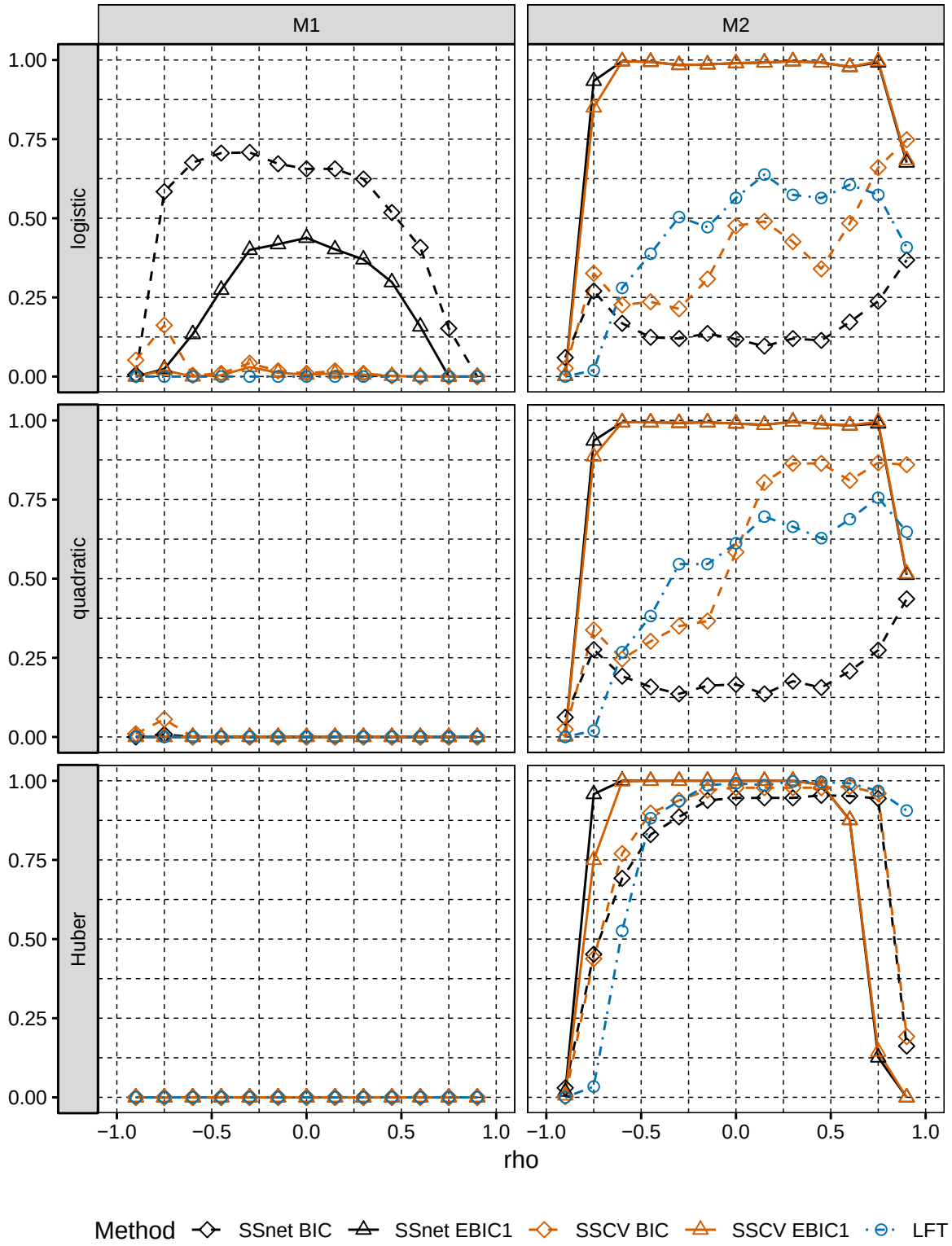

Figure 2. P_{equal} for Models M1 and M2.

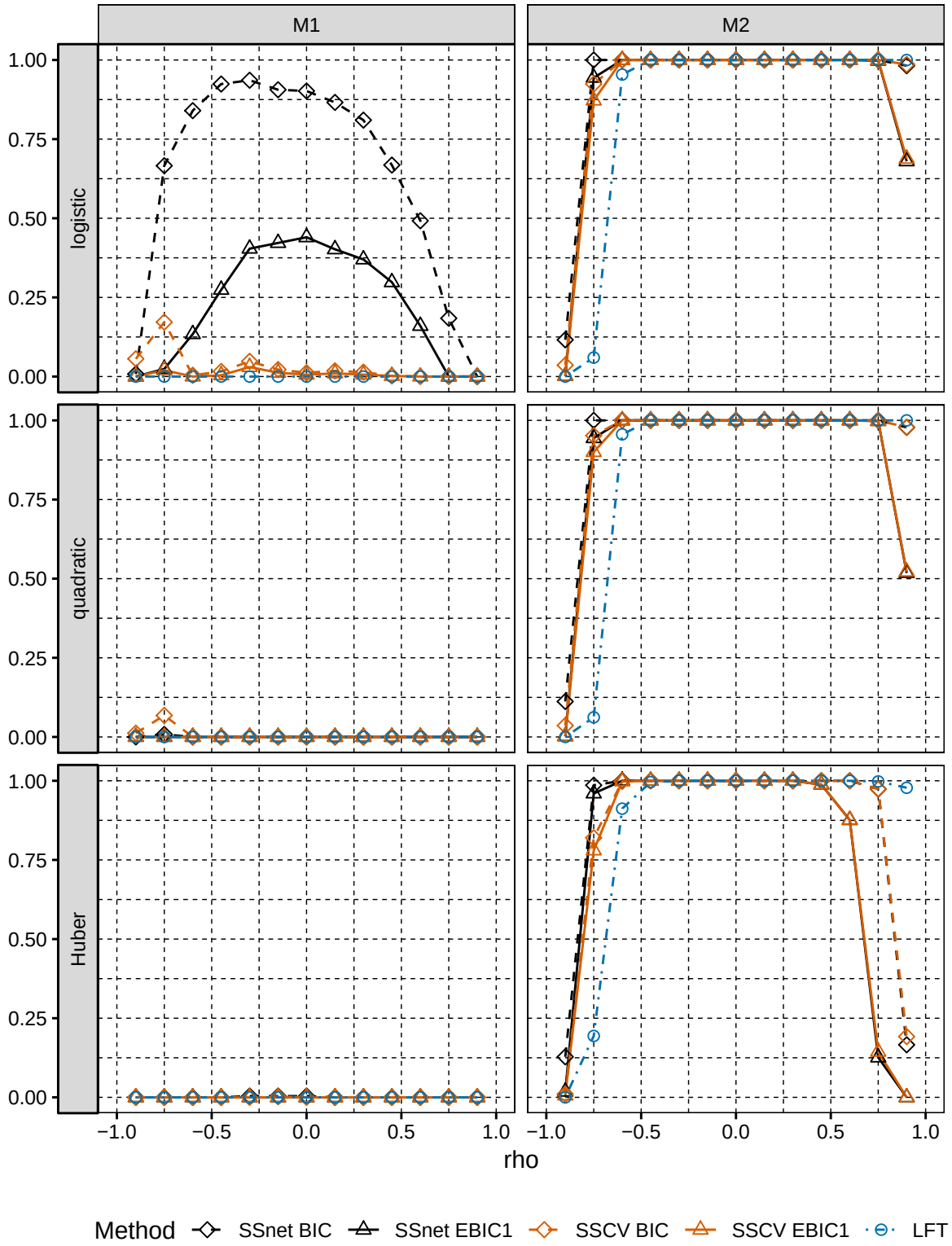

Figure 3. P_{supset} for Models M1 and M2.

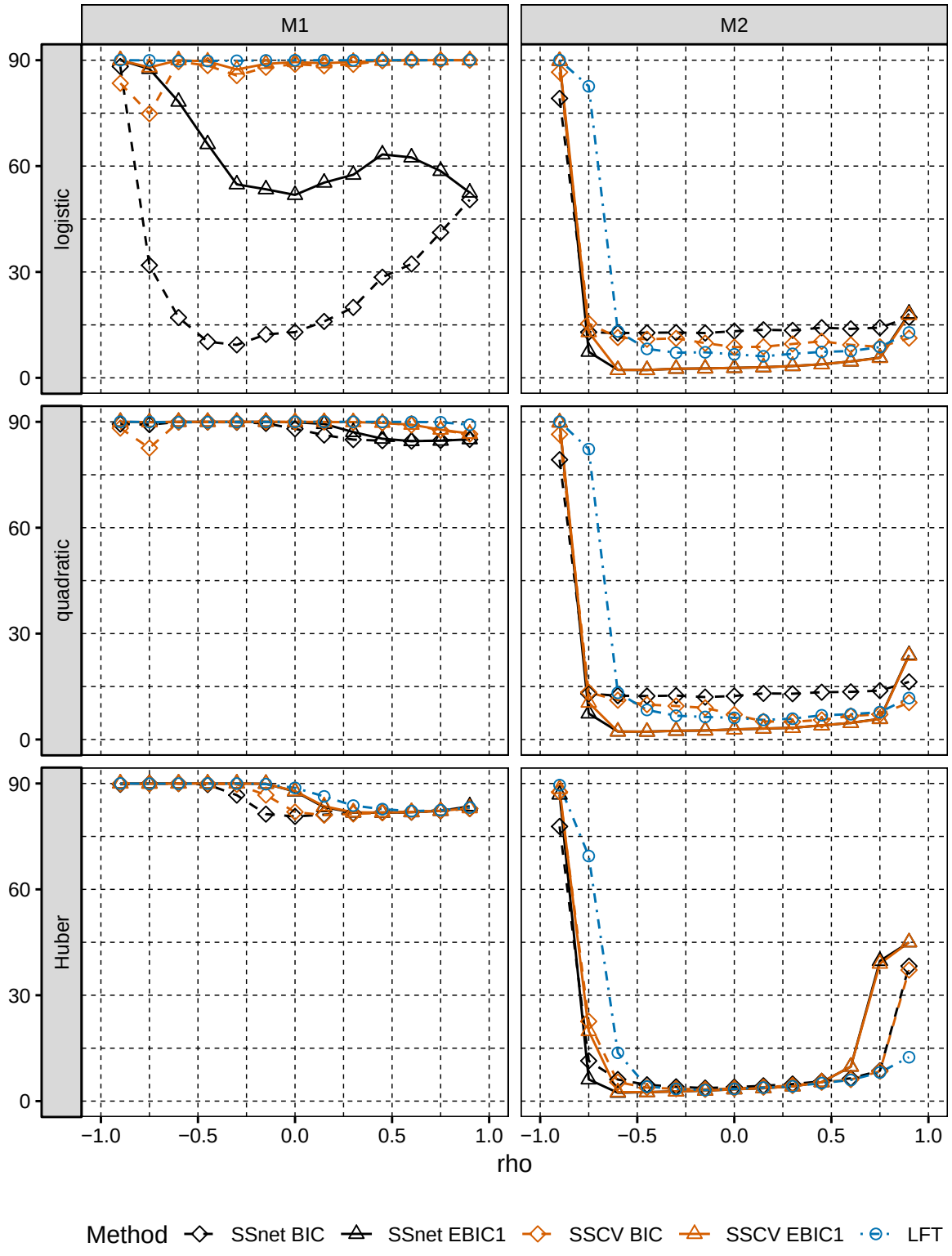

Figure 4. *ANGLE* for Models M1 and M2.

7. Discussion

In the paper, we study the problem of selecting a set of active variables in binary regression model when the number of all predictors p is much larger then number of observations n and active predictors are sparse among all predictors, i.e. their number is significantly smaller than p. We

consider a general binary model and fit based on minimization of empirical risk corresponding to a general loss function. This scenario encompasses the common case in practice when the underlying semi-parametric model is misspecified, i.e. the assumed response function is different from the true one. For random predictors, we show that in such a case the two-step procedure based on Lasso consistently estimates the support of pseudo-true vector β^*. Under linear regression conditions and semi-parametric model, this implies consistent recovery of a subset of active predictors. This partly explains why selection procedures perform satisfactorily even when the fitted model is wrong. We show that, by using the two-step procedure, we can successfully reduce the dimension of the model chosen by Lasso. Moreover, for the two-step procedure in the case of random predictors, we do not require restrictive conditions on experimental matrix needed for Lasso support consistency for deterministic predictors such as irrepresentable condition. Our experiments show satisfactory behavior of the proposed SSnet procedure with EBIC1 penalty.

Future research directions include considering the performance of SS procedure without subgaussianity assumption and for practical importance an automatic choice of a penalty for GIC criterion. Moreover, we note the existing challenge of finding a modification of SS procedure that would perform satisfactorily for large correlations is still an open problem. It would also be of interest to find conditions under which weaker than Equation (32) would lead to collinearity of β and β^* (see [18] for different angle on this problem).

Author Contributions: Both authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: The research of the second author was partially supported by Polish National Science Center grant 2015/17/*B*/*ST*6/01878.

Acknowledgments: The comments by the two referees, which helped to improve presentation of the original version of the manuscript, are gratefully acknowledged.

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

Appendix A

Proof of Lemma 1:

Proof. Observe first that function R_n is convex as ρ is convex. Moreover, from the definition of $\hat{\beta}_L$, we get the inequality:

$$W_n(\hat{\beta}_L) = R_n(\hat{\beta}_L) - R_n(\beta^*) \leq \lambda(||\beta^*||_1 - ||\hat{\beta}_L||_1). \tag{A1}$$

Note that $v - \beta^* \in B_1(r)$, as we have:

$$||v - \beta^*||_1 = \frac{||\hat{\beta}_L - \beta^*||_1}{r + ||\hat{\beta}_L - \beta^*||_1} \cdot r \leq r. \tag{A2}$$

By definition of W_n, convexity of R_n, Equation (A2) and definition of S, we have:

$$\begin{aligned} W(v) &= W(v) - W_n(v) + R_n(v) - R_n(\beta^*) \\ &\leq W(v) - W_n(v) + u(R_n(\hat{\beta}_L) - R_n(\beta^*)) \leq S(r) + uW_n(\hat{\beta}_L). \end{aligned} \tag{A3}$$

From the convexity of l_1 norm, Equations (A3) and (A1), equality $||\beta^*||_1 = ||\beta^*_{s^*}||_1$, and triangle inequality, it follows that:

$$\begin{aligned} W(v) + \lambda||v||_1 &\leq W(v) + \lambda u||\hat{\beta}_L||_1 + \lambda(1-u)||\beta^*||_1 \\ &\leq S(r) + uW_n(\hat{\beta}_L) + u\lambda(||\hat{\beta}_L||_1 - ||\beta^*||_1) + \lambda||\beta^*||_1 \\ &\leq S(r) + \lambda||\beta^*||_1 \leq S(r) + \lambda||\beta^* - v_{s^*}||_1 + \lambda||v_{s^*}||_1. \end{aligned} \tag{A4}$$

Hence,

$$W(v)+\lambda||v-\beta^*||_1=(W(v)+\lambda||v||_1)+\lambda(||v-\beta^*||_1-||v||_1)$$
$$\leq S(r)+\lambda||\beta^*-v_{s^*}||_1+\lambda||v_{s^*}||_1+\lambda(||v-\beta^*||_1-||v||_1)=S(r)+2\lambda||\beta^*-v_{s^*}||_1.$$

□

We prove now Lemma A1 needed in the proof of Lemma 2 below.

Lemma A1. *Assume that $S\sim Subg(\sigma^2)$ and T is a random variable such that $|T|\leq M$, where M is some positive constant and S and T are independent. Then, $ST\sim Subg(M^2\sigma^2)$.*

Proof. Observe that:

$$Ee^{tST}=E(E(e^{tST}|T))\leq Ee^{\frac{t^2T^2\sigma^2}{2}}\leq e^{\frac{t^2M^2\sigma^2}{2}}.$$

□

Proof of Lemma 2.

Proof. From the Chebyshev inequality (first inequality below), symmetrization inequality (see Lemma 2.3.1 of [29]) and Talagrand–Ledoux inequality ([30], Theorem 4.12), we have for $t>0$ and $(\varepsilon_i)_{i=1,\ldots,n}$ being Rademacher variables independent of $(X_i)_{i=1,\ldots,n}$:

$$\begin{aligned}P(S(r)>t)&\leq\frac{ES(r)}{t}\\&\leq\frac{2}{tn}E\sup_{b\in R^{pn}:b-\beta^*\in B_1(r)}\left|\sum_{i=1}^n\varepsilon_i(\rho(X_i^Tb,Y_i)-\rho(X_i^T\beta^*,Y_i))\right|\\&\leq\frac{4L}{tn}E\sup_{b\in R^{pn}:b-\beta^*\in B_1(r)}\left|\sum_{i=1}^n\varepsilon_iX_i^T(b-\beta^*)\right|.\end{aligned}\tag{A5}$$

We observe that $\varepsilon_iX_{ij}\sim Subg(\sigma_{jn}^2)$ in view of Lemma A1. Hence, using independence, we obtain $\sum_{i=1}^n\varepsilon_iX_{ij}\sim Subg(n\sigma_{jn}^2)$ and thus $\sum_{i=1}^n\varepsilon_iX_{ij}\sim Subg(ns_n^2)$. Applying Hölder inequality and the following inequality (see Lemma 2.2 of [37]):

$$E\left|\left|\sum_{i=1}^n\varepsilon_iX_{ij}\right|\right|_\infty\leq\sqrt{n}s_n\sqrt{2\ln(2p_n)}\leq 2s_n\sqrt{n\ln(p_n\vee 2)}\tag{A6}$$

we have:

$$\begin{aligned}\frac{4L}{tn}E\sup_{b\in R^{pn}:b-\beta^*\in B_1(r)}\left|\sum_{i=1}^n\varepsilon_iX_i^T(b-\beta^*)\right|&\leq\frac{4Lr}{t}E\max_{j\in\{1,\ldots,p_n\}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iX_{ij}\right|\\&\leq\frac{8Lrs_n\sqrt{\log(p_n\vee 2)}}{t\sqrt{n}}.\end{aligned}$$

From this, Part 1 follows. In the proofs of Parts 2–3, the first inequalities are the same as in Equation (A5) with supremums taken on corresponding sets. Using Cauchy–Schwarz inequality, inequality $||v||_2 \leq \sqrt{|v|}||v||_\infty$, inequality $||v_\pi||_\infty \leq ||v||_\infty$ for $\pi \subseteq \{1, \ldots, p_n\}$, and Equation (A6) yields:

$$\begin{aligned} P(S_1(r) \geq t) &\leq \frac{4L}{nt} E \sup_{b \in D_1 :\, b-\beta^* \in B_2(r)} \left| \sum_{i=1}^n \varepsilon_i X_i^T (b - \beta^*) \right| \\ &\leq \frac{4Lr}{nt} E \max_{\pi \subseteq \{1,\ldots,p_n\}, |\pi| \leq k_n} \left|\left| \sum_{i=1}^n \varepsilon_i X_{i,\pi} \right|\right|_2 \\ &\leq \frac{4Lr}{nt} E \max_{\pi \subseteq \{1,\ldots,p_n\}, |\pi| \leq k_n} \sqrt{|\pi|} \left|\left| \sum_{i=1}^n \varepsilon_i X_{i,\pi} \right|\right|_\infty \\ &\leq \frac{4Lr\sqrt{k_n}}{nt} E \left|\left| \sum_{i=1}^n \varepsilon_i X_i \right|\right|_\infty \leq \frac{8Lr}{t\sqrt{n}} \sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)}. \end{aligned}$$

Similarly for $S_2(r)$, using Cauchy–Schwarz inequality, $||v_\pi||_2 \leq ||v_{s^*}||_2$, which is valid for $\pi \subseteq s^*$, definition of l_2 norm and inequality $E|Z| \leq \sqrt{EZ^2} \leq \sigma$ for $Z \sim Subg(\sigma^2)$, we obtain:

$$\begin{aligned} P(S_2(r) \geq t) &\leq \frac{4L}{nt} E \sup_{b \in D_2 :\, b-\beta^* \in B_2(r)} \left| \sum_{i=1}^n \varepsilon_i X_i^T (b - \beta^*) \right| \\ &\leq \frac{4Lr}{nt} E \max_{\pi \subseteq s^*} \left|\left| \sum_{i=1}^n \varepsilon_i X_{i,\pi} \right|\right|_2 \leq \frac{4Lr}{nt} E \left|\left| \sum_{i=1}^n \varepsilon_i X_{i,s^*} \right|\right|_2 \\ &\leq \frac{4Lr}{nt} \sqrt{E \left|\left| \sum_{i=1}^n \varepsilon_i X_{i,s^*} \right|\right|_2^2} = \frac{4Lr}{nt} \sqrt{\sum_{j \in s^*} E \left(\sum_{i=1}^n \varepsilon_i X_{ij} \right)^2} \leq \frac{4Lr}{\sqrt{nt}} \sqrt{|s^*|} s_n. \end{aligned}$$

□

Proof of Lemma 3.

Proof. Let u and v be defined as in Lemma 1. Observe that $||v - \beta^*||_1 \leq r/2$ is equivalent to $||\hat{\beta}_L - \beta^*||_1 \leq r$, as the function $f(x) = rx/(x + r)$ is increasing, $f(r) = r/2$ and $f(||\hat{\beta}_L - \beta^*||_1) = ||v - \beta^*||_1$. Let $C = 1/(4 + \varepsilon)$. We consider two cases:

(i) $||v_{s^*} - \beta^*_{s^*}||_1 \leq Cr$.

In this case, from the basic inequality (Lemma 1), we have:

$$||v - \beta^*||_1 \leq \lambda^{-1}(W(v) + \lambda||v - \beta^*||_1) \leq \lambda^{-1} S(r) + 2||v_{s^*} - \beta^*_{s^*}||_1 \leq \bar{C}r + 2Cr = \frac{r}{2}.$$

(ii) $||v_{s^*} - \beta^*_{s^*}||_1 > Cr$.

Note that $||v_{s^{*c}}||_1 < (1 - C)r$, otherwise we would have $||v - \beta^*||_1 > r$, which contradicts Equation (A2) in proof of Lemma 1. Now, we observe that $v - \beta^* \in \mathcal{C}_\varepsilon$, as we have from definition of C and assumption for this case:

$$||v_{s^{*c}}||_1 < (1 - C)r = (3 + \varepsilon)Cr < (3 + \varepsilon)||v_{s^*} - \beta^*_{s^*}||_1.$$

By inequality between l_1 and l_2 norms, the definition of $\kappa_H(\varepsilon)$, inequality $ca^2/4 + b^2/c \geq ab$, and margin Condition (MC) (which holds because $v - \beta^* \in B_1(r) \subseteq B_1(\delta)$ in view of Equation (A2)), we conclude that:

$$||v_{s^*} - \beta^*_{s^*}||_1 \leq \sqrt{|s^*|}||v_{s^*} - \beta^*_{s^*}||_2 \leq \sqrt{|s^*|}||v - \beta^*||_2 \tag{A7}$$

$$\leq \sqrt{|s^*|}\sqrt{\frac{(v-\beta^*)^T H(v-\beta^*)}{\kappa_H(\varepsilon)}}$$

$$\leq \frac{\vartheta(v-\beta^*)^T H(v-\beta^*)}{4\lambda} + \frac{|s^*|\lambda}{\vartheta\kappa_H(\varepsilon)} \leq \frac{W(v)}{2\lambda} + \frac{|s^*|\lambda}{\vartheta\kappa_H(\varepsilon)}. \tag{A8}$$

Hence, from the basic inequality (Lemma 1) and the inequality above, it follows that:

$$W(v) + \lambda||v - \beta^*||_1 \leq S(r) + 2\lambda||v_{s^*} - \beta^*_{s^*}||_1 \leq S(r) + W(v) + \frac{2|s^*|\lambda^2}{\vartheta\kappa_H(\varepsilon)}.$$

Subtracting $W(v)$ from both sides of the above inequality and using the assumption on S, the bound on $|s^*|$, and the definition of $\tilde{C}$ yields:

$$||v - \beta^*||_1 \leq \frac{S(r)}{\lambda} + \frac{2|s^*|\lambda}{\vartheta\kappa_H(\varepsilon)} \leq \bar{C}r + \frac{2|s^*|\lambda}{\vartheta\kappa_H(\varepsilon)} \leq (\bar{C} + \tilde{C})r = \frac{r}{2}.$$

□

Proof of Remark 1.

Proof. Condition $\liminf_{n\to\infty} \frac{D_n a_n}{k_n \log(2p_n)} > 1$ is equivalent to the condition that exists some $u > 0$ that for almost all n we have:

$$D_n a_n - (1+u)k_n \log(2p_n) > 0.$$

We observe that, if

$$Aa_n - (1+u)k_n \log(2p_n) > 0,$$

then the above condition is satisfied. For BIC, we have:

$$A \log n > (1+u)k_n \log(2p_n) > 0,$$

which is equivalent to the condition 1) of the Remark.
(2) We observe that using inequalities $k_n \leq C$, $2A\gamma - (1+u)C \geq 0$ and $p_n \geq 1$ yields for $n > 2^{\frac{(1+u)C}{A}}$:

$$A(\log n + 2\gamma \log p_n) - (1+u)k_n \log(2p_n) \geq A(\log n + 2\gamma \log p_n) - (1+u)C \log(2p_n)$$
$$= (2A\gamma - (1+u)C)\log p_n + A \log n - (1+u)C \log 2 \geq A \log n - (1+u)C \log 2 > 0.$$

(3) In this case, we check similarly as in (2) that

$$A(\log n + 2\gamma \log p_n) - (1+u)k_n \log(2p_n) \geq A(\log n + 2\gamma \log p_n) - (1+u)C \log(2p_n)$$
$$= (2A\gamma - (1+u)C)\log p_n + A \log n - (1+u)C \log 2 > 0$$

□

References

1. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2006.
2. Bühlmann, P.; van de Geer, S. *Statistics for High-dimensional Data*; Springer: New York, NY, USA, 2011.

3. van de Geer, S. *Estimation and Testing Under Sparsity*; Lecture Notes in Mathematics; Springer: New York, NY, USA, 2009.
4. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity*; Springer: New York, NY, USA, 2015.
5. Li, K.; Duan, N. Regression analysis under link violation. *Ann. Stat.* **1989**, *17*, 1009–1052.
6. Kubkowski, M.; Mielniczuk, J. Active set of predictors for misspecified logistic regression. *Statistics* **2017**, *51*, 1023–1045.
7. Kubkowski, M.; Mielniczuk, J. Projections of a general binary model on logistic regrssion. *Linear Algebra Appl.* **2018**, *536*, 152–173.
8. Kubkowski, M. Misspecification of binary regression model: properties and inferential procedures. Ph.D. Thesis, Warsaw University of Technology, Warsaw, Poland, 2019.
9. Lu, W.; Goldberg, Y.; Fine, J. On the robustness of the adaptive lasso to model misspecification. *Biometrika* **2012**, *99*, 717–731.
10. Brillinger, D. A Generalized linear model with 'gaussian' regressor variables. In *A Festschfrift for Erich Lehmann*; Wadsworth International Group: Belmont, CA, USA, 1982; pp. 97–113.
11. Ruud, P. Suffcient conditions for the consistency of maximum likelihood estimation despite misspecifcation of distribution in multinomial discrete choice models. *Econometrica* **1983**, *51*, 225–228.
12. Yi, C.; Huang, J. Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *J. Comput. Graph. Stat.* **2017**, *26*, 547–557, [https://doi.org/10.1080/10618600.2016.1256816]. doi:10.1080/10618600.2016.1256816.
13. White, W. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25.
14. Vuong, Q. Likelihood ratio testts for model selection and not-nested hypotheses. *Econometrica* **1989**, *57*, 307–333.
15. Bickel, P.; Ritov, Y.; Tsybakov, A. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732.
16. Negahban, S.N.; Ravikumar, P.; Wainwright, M.J.; Yu, B. A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. *Stat. Sci.* **2012**, *27*, 538–557.
17. Pokarowski, P.; Mielniczuk, J. Combined ℓ_1 and greedy ℓ_0 penalized least squares for linear model selection. *J. Mach. Learn. Res.* **2015**, *16*, 961–992.
18. Hall, P.; Li, K.C. On almost Linearity of Low Dimensional Projections from High Dimensional Data. *Ann. Stat.* **1993**, *21*, 867–889.
19. Chen, J.; Chen, Z. Extended bayesian information criterion for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771.
20. Chen, J.; Chen, Z. Extended BIC for small-n-large-p sparse GLM. *Stat. Sin.* **2012**, *22*, 555–574.
21. Mielniczuk, J.; Szymanowski, H. Selection consistency of Generalized Information Criterion for sparse logistic model. In *Stochastic Models, Statistics and their Applications*; Steland A., Rafajłowicz E., Szajowski K., Eds.; Springer: Cham, Switzerland, 2015; Volume 122, pp. 111–118.
22. Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*; Cambridge University Press: Cambridge, UK, 2012; pp. 210–268.
23. Fan, J.; Xue, L.; Zou, H. Supplement to "Strong oracle optimality of folded concave penalized estimation.". 2014. Available online: NIHMS649192-supplement-suppl.pdf (25 January 2020).
24. Fan, J.; Xue, L.; Zou, H. Strong Oracle Optimality of folded concave penalized estimation. *Ann. Stat.* **2014**, *43*, 819–849.
25. Bach, F. Self-concordant analysis for logistic regression. *Electron. J. Stat.* **2010**, *4*, 384–414.
26. Akaike, H. Statistical predictor identification. *Ann. Inst. Stat. Math.* **1970**, *22*, 203–217.
27. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464.
28. Kim, Y.; Jeon, J. Consistent model selection criteria for quadratically supported risks. *Ann. Stat.* **2016**, *44*, 2467–2496.
29. van der Vaart, A.W.; Wellner, J.A. *Weak Convergence and Empirical Processes with Applications to Statistics*; Springer: New York, NY, USA, 1996.
30. Ledoux, M.; Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*; Springe: New York, NY, USA, 1991.

31. Huang, J.; Ma, S.; Zhang, C. Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sin.* **2008**, *18*, 1603–1618.
32. Tibshirani, R. The lasso problem and uniqueness. *Electron. J. Stat.* **2013**, *7*, 1456–1490.
33. Zhou, S. Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv* **2010**, arXiv:1002.1583
34. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
35. Fan, Y.; Tang, C. Tuning parameter selection in high dimensional penalized likelihood. *J. Royal Stat. Soc. Ser. B* **2013**, *75*, 531–552.
36. Rosset, S.; Zhu, J.; Hastie, T. Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.* **2004**, *5*, 941–973.
37. Devroye, L.; Lugosi, G. *Combinatorial Methods in Density Estimation*; Springer Science & Business Media: New York, NY, USA, 2012.