Received: 25 September 2019 | Revised: 22 January 2021 | Accepted: 24 May 2021

DOI: 10.1111/sjos.12546

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Improving Lasso for model selection and prediction

Piotr Pokarowski[1] | **Wojciech Rejchel**[2] | **Agnieszka Sołtys**[1] | **Michał Frej**[1] | **Jan Mielniczuk**[3,4]

[1]Institute of Applied Mathematics and Mechanics, University of Warsaw, Warsaw, Poland

[2]Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland

[3]Institute of Computer Sciences, Polish Academy of Sciences, Warsaw, Poland

[4]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Correspondence
Wojciech Rejchel, Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Poland.
Email: wrejchel@gmail.com

Abstract

It is known that the Thresholded Lasso (TL), SCAD or MCP correct intrinsic estimation bias of the Lasso. In this paper we propose an alternative method of improving the Lasso for predictive models with general convex loss functions which encompass normal linear models, logistic regression, quantile regression, or support vector machines. For a given penalty we order the absolute values of the Lasso nonzero coefficients and then select the final model from a small nested family by the Generalized Information Criterion. We derive exponential upper bounds on the selection error of the method. These results confirm that, at least for normal linear models, our algorithm seems to be the benchmark for the theory of model selection as it is constructive, computationally efficient and leads to consistent model selection under weak assumptions. Constructivity of the algorithm means that, in contrast to the TL, SCAD or MCP, consistent selection does not rely on the unknown parameters as the cone invertibility factor. Instead, our algorithm only needs the sample size, the number of predictors and an upper bound on the noise parameter. We show in numerical experiments on synthetic and real-world datasets that an implementation of our algorithm is more accurate than implementations of studied concave regularizations. Our procedure

is included in the R package `DMRnet` and available in the CRAN repository.

KEYWORDS

convex loss function, empirical process, generalized information criterion, high-dimensional regression, penalized estimation, selection consistency

1 | INTRODUCTION

Sparse high-dimensional predictive models, where the number of true predictors t is significantly smaller than the sample size n and the number of all predictors p greatly exceeds n, have been a focus of research in statistical machine learning in recent years. The Lasso algorithm, that is the minimum loss method regularized by sparsity inducing the ℓ_1 penalty, is the main tool of fitting such models (Bühlmann & van de Geer, 2011; Tibshirani, 2011). However, it has been shown that the model selected by the Lasso is usually too large and that for asymptotically consistent model selection it requires the *irrepresentable condition* on an experimental matrix (Meinshausen & Bühlmann, 2006; Shen et al., 2012; Zhao & Yu, 2006) which is too restrictive in general. The model's dimension can be reduced without loss of the quality using the Thresholded Lasso (TL) algorithm, which selects variables with largest absolute values of the Lasso coefficients (Ye & Zhang, 2010; Zhou, 2009) or by solving more computationally demanding minimization of a loss with a folded concave penalty (FCP) as SCAD (Fan & Li, 2001), MCP (Zhang, 2010a) or capped l_1-penalty (Shen et al., 2012; Zhang, 2010b). TL, FCP, and similar methods lead to consistent selection under weaker assumptions such as the *restricted isometry property* (Fan et al., 2014; Shen et al., 2012; Wang et al., 2013, 2014; Zhang, 2010a, 2010b; Zhang & Zhang, 2012). In Pokarowski and Mielniczuk (2015) one introduced an algorithm called *Screening–Ordering–Selection* (SOS) for linear model selection, which reduces successfully the model selected by the Lasso. SOS is based on the variant of TL proposed by Zhou (2009) and leads to consistent model selection under assumptions similar to the restricted isometry property.

In the paper we consider two algorithms that improve the Lasso, that is, they are model selection consistent under weaker assumptions than the Lasso. Moreover, the considered procedures are computationally simpler than FCP methods that use nonconvex penalties. The first algorithm is the well-known TL. Its model selection consistency in the normal linear model is proven in (Ye & Zhang, 2010, theorem 8) provided that conditions, which seem to be "minimal," are satisfied. Our first contribution is an extension of this result (given in Theorem 1) to Generalized Linear Models (GLM). However, the TL algorithm is not constructive, because it is not known how to choose the threshold. Therefore, in the current paper we propose the second improvement of the Lasso which is the *Screening–Selection* (SS) algorithm. It is a two-step procedure: in the first step (screening) one computes the Lasso estimator $\hat{\beta}$ with penalty λ and orders its nonzero coefficients according to their decreasing absolute values. In the second step (selection) one chooses the model which minimizes the Generalized Information Criterion (GIC) with penalty $\lambda^2/2$ in a nested family induced by the ordering. Thus, the SS algorithm (Algorithm 1 below) can be viewed as the Lasso with adaptive thresholding based on GIC. We prove that this procedure is model selection consistent in a wide class of models containing linear models with the subgaussian noise (Theorem 3), GLM (Theorem 2), and models with convex (possibly nondifferentiable) loss

functions (Theorem 4) as in quantile regression or support vector machines (SVM). The obtained results are exponential upper bounds on the selection error of SS in terms of λ, which parallel the known bounds for TL (Ye & Zhang, 2010, theorem 8) or FCP (Fan et al., 2014, corollaries 3 and 5). For GLM our results are obtained on the basis of exponential inequalities for subgaussian random variables. In the case of predictive models with general convex loss functions we use methods from the empirical process theory.

The SS algorithm is a simplification and a generalization of the SOS algorithm from Pokarowski and Mielniczuk (2015) that was proposed only for normal linear models. We show that the ordering step in SOS can be done using separability of the Lasso instead of using t-statistics. Separability means that, under mild assumptions, coordinates of the Lasso corresponding to relevant predictors are larger in absolute values than those corresponding to irrelevant predictors. Moreover, we establish that the SS algorithm is model selection consistent beyond normal linear models. The new procedure can be applied to the various statistical predictive models including models with quantitative or qualitative response variables. We can use "classical" models (the normal linear model, the logistic model) as well as modern problems involving piecewise-linear loss functions (as in quantile regression or SVMs).

Our results state that, in contrast to TL and FCP methods, SS is constructive in the linear model with the subgaussian noise, because it does not rely on the unknown parameters as the true vector $\mathring{\beta}$ or the cone invertibility factors. Indeed, we will establish in Section 2 that the choice of the tuning parameter λ as

$$\lambda = \sqrt{\frac{2\sigma^2 \log p}{n}}(1 + o(1)), \tag{1}$$

leads to model selection consistency of the SS algorithm. Therefore, λ only depends on n, p and σ^2 that is an upper bound on the noise parameter. The assumption that σ^2 is known is common in the literature investigating theoretical properties of variable selection procedures (Bühlmann & van de Geer, 2011; Fan et al., 2014; Ye & Zhang, 2010).

Moreover, we will also prove that model selection consistency of the SS algorithm holds under weaker conditions than for competitive procedures. For instance, the algorithm of Wang et al. (2013) requires that

$$\frac{\log p}{\delta} = o(1) \quad \text{and} \quad \frac{t}{\mathring{\beta}_{\min}}\sqrt{\frac{\log p}{n}} = o(1), \tag{2}$$

where δ is a scaled Kullback–Leibler (K-L) distance between the true set and its submodels (defined in (15)) and $\mathring{\beta}_{\min}$ is the minimal signal strength. However, for the SS procedure we need that

$$\frac{\log p}{\delta} = O(1) \quad \text{and} \quad \frac{1}{\mathring{\beta}_{\min}}\sqrt{\frac{\log p}{n}} = O(1), \tag{3}$$

which, among others, weakens significantly the *beta-min* condition. More detailed description of the SS algorithm and comparisons to other procedures is given in Section 2. This analysis enables us to claim that for subgaussian linear models the SS algorithm seems to be the benchmark for the theory of model selection as it is constructive, computationally efficient and leads

to consistent model selection under weak assumptions. Similarly to TL or FCP, the SS algorithm becomes non-constructive, if we go beyond the linear model with the subgaussian noise. Notice that in (1), (2), and (3) we assume, similar to Wang et al. (2013), that predictors are scaled to have the l_2-norm equal to $\sqrt{n}$. However, in the rest of the paper we will use more convenient notation that the l_2-norm of predictors is one. Therefore, the expression $\sqrt{n}$ will not appear in analogs of (1), (2), and (3) in Section 2.

Although TL, FCP, or SS algorithms use the Lasso estimators only for one value of the penalty, which is convenient for theoretical analysis, the practical Lasso implementations return coefficient estimators for all possible penalty values as in the R package `LARS` described in Efron et al. (2004) or for a given net of them as in the R package `glmnet` described in Friedman et al. (2010). Similarly, using a net of penalty values, the FCP algorithm has been implemented for linear models in the R package `SparseNet` (Mazumder et al., 2011) and for logistic models in the R package `ncvreg` (Breheny & Huang, 2011). Our contribution is also the SSnet algorithm (Algorithm 2 below), which is a *practical* version of the SS algorithm. SSnet uses `glmnet` to calculate the Lasso for a net of penalty values, then again it selects the final model from a small family by minimizing GIC. In numerical experiments we investigate properties of SSnet in model selection as well as prediction and compare them to competitive procedures. Using synthetic and real-world data sets we show that SSnet is more accurate than implementations of FCP. The variant of SSnet is contained in the R package `DMRnet` and available in the CRAN repository.

GIC is a popular procedure in choosing the final model in variable selection. In the literature there are many papers investigating its properties, for instance (Kim et al., 2012; Wang et al., 2013) in linear models, (Fan & Tang, 2013; Hui et al., 2015) in GLM, Katayama and Imori (2014) in multivariate linear regression, Zhang et al. (2016) for SVM and Kim and Jeon (2016) for general convex loss functions. GIC is often applied to pathwise algorithms under a common assumption that the true model is on this path. This condition excludes Lasso-based pathwise algorithms as it is only fulfilled under restrictive assumptions. One can overcome this problem using a three-step procedures, for instance the Lasso and nonconvex penalized regression (Wang et al., 2013) or the Lasso and thresholding (Kim & Jeon, 2016). However, it makes the algorithm computationally more complex or one has to find a threshold that recovers the true model on the path, respectively. In contrast, the first step of the proposed procedure is related only to the Lasso. Indeed, we need only that the model with correctly separated predictors is on the path. This is guaranteed for the Lasso under mild assumptions. Therefore, it makes our procedure simpler and computationally more efficient.

The paper is organized as follows: in the next section we describe subgaussian GLM and algorithms that we work with. Moreover, we establish bounds for the selection error of the proposed procedures. These results are extended to models with general convex contrasts in Section 3. In Section 4 we investigate properties of estimators on simulated and real datasets. The paper is concluded in Section 5. All proofs and auxiliary results are relegated to Appendix.

2 | SUBGAUSSIAN GLMS

In this section we start with definitions of considered models and estimation criteria. Next, we present model selection algorithms and state exponential upper bounds on their selection errors.

2.1 | Models

The way we model data will encompass normal linear and logistic models as premier examples. Our assumptions are stated in their most general form which allows proving exponential bounds for probability of the selection error without obscuring their essentiality. In the paper we consider independent data $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$, where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ for $i = 1, 2, \dots, n$. We assume that for some *true* $\mathring{\beta} \in \mathbb{R}^p$ and a known differentiable *cumulant* function $\gamma : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}y_i = \mathring{\gamma}(x_i^T \mathring{\beta}) \quad \text{for} \quad i = 1, 2, \dots, n, \tag{4}$$

where $\mathring{\gamma}$ denotes the derivative of γ. Note that (4) is satisfied in particular by the GLMs with a canonical link function and a nonlinear regression with an additive error. Let $\eta_i = x_i^T \mathring{\beta}$. For $\eta = (\eta_1, \dots, \eta_n)^T$ we define $\vec{\gamma}(\eta) = (\gamma(\eta_1), \dots, \gamma(\eta_n))^T$ and similarly $\vec{\mathring{\gamma}}(\eta) = (\mathring{\gamma}(\eta_1), \dots, \mathring{\gamma}(\eta_n))^T$.

Let $X = [x_{.1}, \dots, x_{.p}] = [x_1, \dots, x_n]^T$ be a $n \times p$ matrix of experiment and $J \subseteq \{1, 2, \dots, p\} = F$ be an arbitrary subset of the *full model* F, $\bar{J} = F \setminus J$. As J may be viewed as a sequence of zeros and ones on F, $|J| = |J|_1$ denotes cardinality of J. Let β_J be a subvector of β with elements having indices in J, X_J be a submatrix of X with columns having indices in J and $r(X_J)$ denotes the rank of X_J. Moreover, let $H_J = X_J(X_J^T X^J)^{-1} X_J^T$ be an orthogonal projection matrix onto the subspace spanned by columns of X_J. Linear model pertaining to predictors being columns of X_J will be frequently identified as J. In particular, let T denotes a *true model* that is $T = \mathrm{supp}(\mathring{\beta}) = \{j \in F : \mathring{\beta}_j \neq 0\}$ and $t = |T|$. Next, for $\beta \in \mathbb{R}^p$ and $q \geq 1$ let $|\beta|_q = (\sum_{j=1}^p |\beta_j|^q)^{1/q}$ be the ℓ_q norm. The only exception is the ℓ_2 norm, for which we will use the special notation $\|\beta\|$.

In the further argumentation important roles are played by

$$\delta_k = \min_{J \subset T,\ |T \setminus J| = k} \|(I - H_J) X_T \mathring{\beta}_T\|^2, \tag{5}$$

for $k = 1, \dots, t-1$. They describe how much the true value $X\mathring{\beta} = X_T \mathring{\beta}_T$ differs from its projections on submodels of true set T. For the normal linear model they are related to the K-L divergence between two normal densities (Pokarowski & Mielniczuk, 2015, section 3). Finally, we define the sum of balls

$$\mathbb{B} \equiv \mathbb{B}(X, \mathring{\beta}, \bar{t}) = \bigcup_{J \supset T, r(X_J) = |J| \leq \bar{t}} \{\beta_J : \|X_J(\mathring{\beta}_J - \beta_J)\|^2 \leq \delta_{t-1}\}. \tag{6}$$

We assume that $t \leq \bar{t} < n \wedge p$, which implies that (6) consists of *sparse* vectors.

We assume also that a *total cumulant* function

$$g(\beta) = \sum_{i=1}^n \gamma(x_i^T \beta), \tag{7}$$

is convex and, additionally, *strongly convex* at $\mathring{\beta}$ in a sense that there exists $c \in (0, 1]$ such that for all $\beta \in \mathbb{B}$ we have

$$g(\beta) \geq g(\mathring{\beta}) + (\beta - \mathring{\beta})^T \dot{g}(\mathring{\beta}) + \frac{c}{2}(\mathring{\beta} - \beta)^T X^T X (\mathring{\beta} - \beta). \tag{8}$$

We note that this crucial property of the total cumulant is slightly weaker than an usual definition of strong convexity which would have a second derivative of g at $\mathring{\beta}$ in place of X^TX. Let us remark that $\dot{g}(\beta) = X^T\vec{\mathring{\gamma}}(X\beta)$.

Moreover, we assume that centred responses $\varepsilon_i = y_i - \mathbb{E}y_i$ have a *subgaussian distribution* with the same constant σ, that is for $i = 1, 2, \dots, n$ and $u \in \mathbb{R}$ we have

$$\mathbb{E}\exp(u\varepsilon_i) \le \exp(\sigma^2 u^2/2). \tag{9}$$

Examples. Two most important representatives of GLM are the normal linear model and logistic regression. In the normal linear model

$$y_i = x_i^T\mathring{\beta} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where the noise variables ε_i are independent and normally distributed $N(0, \mathring{\sigma}^2)$. Therefore, assumptions (4), (8), and (9) are satisfied with $\gamma(\eta_i) = \eta_i^2/2, c = 1$ and any $\sigma \ge \mathring{\sigma}$, respectively. In logistic regression the response variable is dichotomous $y_i \in \{0, 1\}$ and we assume that

$$P(y_i = 1) = \frac{\exp(x_i^T\mathring{\beta})}{\exp(x_i^T\mathring{\beta}) + 1}, \quad i = 1, \dots, n.$$

In this model assumptions (4) and (8) are satisfied with $\gamma(\eta_i) = \log(1 + \exp(\eta_i))$ and

$$c = \min_i \min_{\beta\in\mathbb{B}} \exp(x_i^T\beta)/(1 + \exp(x_i^T\beta))^2,$$

respectively. Finally, as (ε_i) are bounded random variables, then (9) is satisfied with any $\sigma \ge 1/2$.

2.2 | Fitting algorithms

For estimation of $\mathring{\beta}$ we consider a *loss function*

$$\ell(\beta) = \sum_{i=1}^{n} [\gamma(x_i^T\beta) - y_i x_i^T\beta] = g(\beta) - \beta^T X^T y, \tag{10}$$

where $y = (y_1, \dots, y_n)^T$. It is easy to see that $\dot{\ell}(\beta) = X^T(\vec{\mathring{\gamma}}(X\beta) - y)$, and consequently $\dot{\ell}(\mathring{\beta}) = -X^T\varepsilon$ for $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$. Moreover, observe that $\mathring{\beta} = \operatorname{argmin}_\beta \mathbb{E}\ell(\beta)$ and (8) is equivalent to strong convexity of ℓ at $\mathring{\beta}$ for all $\beta \in \mathbb{B}$

$$\ell(\beta) \ge \ell(\mathring{\beta}) + (\mathring{\beta} - \beta)^T X^T\varepsilon + \frac{c}{2}(\mathring{\beta} - \beta)^T X^T X(\mathring{\beta} - \beta). \tag{11}$$

Let $\widehat{\beta}_J^{\mathrm{ML}} = \operatorname{argmin}_{\beta_J} \ell(\beta_J)$ denote a *minimum loss estimator* based on y and $\{x_j, j \in J\}$. Denote $\ell_J = \ell(\widehat{\beta}_J^{\mathrm{ML}})$. Note that for GLM estimator $\widehat{\beta}_J^{\mathrm{ML}}$ coincides with a maximum likelihood estimator for a model pertaining to J. In Algorithm 1 we present two selection procedures: the first one is the well-known TL method which consists of retaining only these variables for which absolute values of their Lasso estimators exceed a certain threshold τ. The second one is novel and named

Algorithm 1. TL and SS

Input: y, X and λ, τ
Screening (Lasso) $\widehat{\beta} = \text{argmin}_\beta \{\ell(\beta) + \lambda|\beta|_1\}$;

Thresholded Lasso $\widehat{T}_{TL} = \{j : |\widehat{\beta}_j| > \tau\}$;

Selection (GIC)
order nonzero $|\widehat{\beta}_{j_1}| \geq \ldots \geq |\widehat{\beta}_{j_s}|$, where $s = |\text{supp}\widehat{\beta}|$;
set $\mathcal{J} = \left\{\{j_1\}, \{j_1, j_2\}, \ldots, \text{supp}\widehat{\beta}\right\}$;
$\widehat{T}_{SS} = \text{argmin}_{J\in\mathcal{J}} \left\{\ell_J + \lambda^2/2|J|\right\}$.

Output: $\widehat{T}_{\text{TL}}, \widehat{T}_{\text{SS}}$

the SS procedure. It finds the minimal value of the GIC for the nested family which is constructed using ordering of the nonzero Lasso coefficients.

In the classical (low-dimensional) case model selection is often based on the Bayesian information criterion that is similar to the second step of Algorithm 1 with $\lambda^2 = \log n$. Model selection properties of this procedure are proven in Shao (1997). In the current paper we validate that information criteria are also useful in the high-dimensional case. As it will be shown the important difference is that in the high-dimensional scenario parameter λ^2 should be proportional to $\log p$ instead of $\log n$ (at least for linear models with the subgaussian noise).

2.3 | A selection error bound for TL

In order to make the exposition simpler we assume that columns of X are normalized in such a way that $\|x_j\| = 1$ for $j = 1, \ldots, p$. Moreover, let $\mathring{\beta}_{\min} = \min_{j\in T} |\mathring{\beta}_j|$.

First we generalize a characteristic of linear models which quantifies the degree of separation between the true model T and other models, which was introduced in Ye and Zhang (2010). For $a \in (0, 1)$ consider a signed pseudo-cone

$$\mathcal{C}_a = \left\{\nu \in \mathbb{R}^p : |\nu_{\overline{T}}|_1 \leq \frac{1+a}{1-a}|\nu_T|_1,\ \nu_j x_j^T \left[\vec{\hat{\gamma}}(X(\mathring{\beta} + \nu)) - \vec{\hat{\gamma}}(X\mathring{\beta})\right] \leq 0,\ j \in \overline{T}\right\}. \tag{12}$$

For $q \geq 1$ and $a \in (0, 1)$ let a Sign-restricted pseudo-Cone Invertibility Factor (SCIF) be defined as

$$\zeta_{a,q} = \inf_{0\neq\nu\in\mathcal{C}_a} \frac{\left|X^T \left[\vec{\hat{\gamma}}(X(\mathring{\beta} + \nu)) - \vec{\hat{\gamma}}(X\mathring{\beta})\right]\right|_\infty}{|\nu|_q}. \tag{13}$$

Notice that in the linear model the numerator of (13) is simply $|X^T X\nu|_\infty$. In the case $n > p$ one usually uses the minimal eigenvalue of the matrix $X^T X$ to express the strength of correlations between predictors. Obviously, in the high-dimensional scenario this value is zero. Therefore, SCIF can be viewed as an useful analog of the minimal eigenvalue for the case $p > n$.

We let $\zeta_a = \zeta_{a,\infty}$. In comparison to more popular restricted eigenvalues (Bickel et al., 2009) or compatibility constants (van de Geer & Bühlmann, 2009), variants of SCIF enable sharper ℓ_q

estimation error bounds of the Lasso for $q > 2$ (Huang & Zhang, 2012; Ye & Zhang, 2010; Zhang & Zhang, 2012).

The following lemma is a main tool in proving model selection consistency of the TL algorithm. For the linear model it was stated in (Ye & Zhang, 2010, theorem 3). We generalize it to GLM.

Lemma 1. *If ℓ is convex and $a \in (0,1)$, then on $\{|X^T\varepsilon|_\infty \le a\lambda\}$ we have*

$$|\hat{\beta} - \mathring{\beta}|_q \le (1+a)\lambda\zeta_{a,q}^{-1}.$$

Next, we bound the selection error of the TL algorithm in GLM.

Theorem 1. *If ℓ is convex, $(\varepsilon_i)_i$ are subgaussian with σ and for numbers $a_1, a_2 \in (0,1)$ we have*

$$2a_1^{-2}a_2^{-1}\sigma^2 \log p \le \lambda^2 \le (1+a_1)^{-2}\zeta_{a_1}^2\tau^2 < (1+a_1)^{-2}\zeta_{a_1}^2\mathring{\beta}_{\min}^2/4,$$

then

$$\mathbb{P}(\hat{T}_{\mathrm{TL}} \ne T) \le 2\exp\left(-\frac{(1-a_2)a_1^2\lambda^2}{2\sigma^2}\right). \tag{14}$$

Constant a_2 is used to remove multiplicative factor p from the exponential bound at the expense of slightly diminishing the exponent in (14). Note that assumptions of Theorem 1 stipulate that the truncation level τ is contained in the interval $[(1+a_1)\lambda\zeta_{a_1}^{-1}, \mathring{\beta}_{\min}/2)$, whose both endpoints are unknown, because ζ_{a_1} and $\mathring{\beta}_{\min}$ are unknown. Therefore, in practice τ cannot be chosen that makes the TL algorithm nonconstructive. Analogous theorems for FCP for linear models and logistic regression can be found in Fan et al. (Fan et al., 2014, corollaries 3 and 5). However, they require an additional assumption on the minimal eigenvalue of $X_T^T X_T$ and the proof is more difficult. Moreover, the choice of the tuning parameters in these methods also requires unknown ζ_a or $\mathring{\beta}_{\min}$.

2.4 | A selection error bound for SS

A scaled K-L distance between T and its submodels is given in (Shen et al., 2012, 2013) as

$$\delta = \min_{J \subset T} \frac{\|(I - H_J)X_T\mathring{\beta}_T\|^2}{|T \setminus J|}, \tag{15}$$

or just $\min_{k=1,\dots,t-1}\ \delta_k/k$, if we use notation (5). Different variants of the K-L distance have been often used in the consistency analysis of selection algorithms (Pokarowski & Mielniczuk, 2015, section 3.1), but δ defined in (15) seems to lead to optimal results (Shen et al., 2013, theorem 1). Now we introduce technical constants $a_1, \dots, a_4$ that allow to avoid ad hoc coefficients in the main results and simplify asymptotic considerations. For given $1/2 < a_1 < 1$ define $a_2 = 1 - (1 - \log(1-a_1))(1-a_1)$, $a_3 = 2 - 1/a_1$ and $a_4 = \sqrt{a_1 a_2}$. Note that a_2, a_3, and a_4 are functions of a_1 and obviously if $a_1 \to 1$, then $a_2, a_3, a_4 \to 1$. Moreover, for two real numbers a, b we denote $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.

Theorem 2. *Assume (4), (8), (9) and that for $a_1 \in (1/2, 1)$ we have*

$$\frac{2\sigma^2 \log p}{a_3 a_2 a_1 c} \vee \frac{\sigma^2 t}{(1-a_1)^2 c} \leq \lambda^2 < \frac{c\delta_{t-1}}{16(\bar{t}-t)} \wedge \frac{c\delta}{(1+\sqrt{2(1-a_1)})^2} \wedge \frac{\zeta_{a_4}^2 \mathring{\beta}_{\min}^2}{4(1+a_4)^2}. \tag{16}$$

Then

$$\mathbb{P}(\hat{T}_{\mathrm{SS}} \neq T) \leq 4.5 \exp\left(-\frac{a_2(1-a_1)c\lambda^2}{2\sigma^2}\right). \tag{17}$$

Selection consistency, that is asymptotic correctness of $\hat{T}_{\mathrm{SS}}$, now easily follows.

Corollary 1. *Assume that $t = o(\log p)$ for $n \to \infty$. Moreover, set $a_1 = 1 - \sqrt{\frac{t}{2\log p}}$ and $\lambda^2 = \frac{2\sigma^2 \log p}{a_3 a_2 a_1 c}$. Then $\lambda^2 = c^{-1}(2\sigma^2 \log p)(1 + o(1))$. If additionally $\mathring{\beta}$ is asymptotically identifiable, that is,*

$$\overline{\lim_n} \frac{2\sigma^2 \log p / c}{\frac{c\delta_{t-1}}{16(\bar{t}-t)} \wedge (c\delta) \wedge (\zeta_{a_4}^2 \mathring{\beta}_{\min}^2/16)} < 1, \quad \text{then} \quad \mathbb{P}(\hat{T}_{\mathrm{SS}} \neq T) = o(1).$$

Remark 1. Theorem 2 determines conditions on GLM and the SS algorithm for which the bound (17) on the selection error of SS holds. Corollary 1 states the easy interpretable result: if the true model is asymptotically identifiable, then SS with minimal admissible λ is asymptotically consistent. Although the identifiability condition is not effectively verifiable, λ can be explicitly given for linear models as

$$\lambda = \sqrt{2\sigma^2 \log p}(1 + o(1)), \tag{18}$$

and for logistic models as

$$\lambda = \sqrt{(\log p)/(2c)}(1 + o(1)), \tag{19}$$

since $\sigma \geq 1/2$. Let us consider subgaussian linear models and assume that σ^2 is known, which is a common condition in the literature investigating theoretical properties of variable selection procedures (Bühlmann & van de Geer, 2011; Fan et al., 2014; Ye & Zhang, 2010). Then the parameter λ of SS is given constructively provided that $t = o(\log p)$. In contrast, TL or FCP are not constructive, because they require an additional parameter τ, that depends on unknown identifiability constants as SCIF.

In the literature concerning the Lasso and its modifications the smallest possible λ is taken as the default value, because it makes the algorithm asymptotically consistent for the largest class of models (the same approach is adopted for prediction and estimation). Such λ will be called the *safest choice*. It is interesting that for linear models GIC with λ given by (18) was originally derived from the minimax perspective by Foster and George (1994). They called such selection the risk inflation criterion (RIC), because it asymptotically minimises the maximum predictive risk with respect to the oracle for the orthogonal matrix of experiment X.

Remark 2. A generic combination of the penalized log-likelihood (as TL or FCP) with GIC is considered in Fan and Tang (2013). In the first step the method computes a path of models indexed by λ and next GIC is used to choose the final model. They assume that the true model has to be on this path and use GIC with the penalty asymptotically larger than $\log p$. Thus, their results are

weaker and need more restrictive assumptions, which are given in (Fan & Tang, 2013, section 6). For instance, if the cumulant function γ, defined in (4), has uniformly bounded second derivative, then we do not require its third derivative in contrast to Fan and Tang (2013). Moreover, using Corollary 1 and assuming that the number c is constant as in Fan and Tang (2013), the last step of our algorithm uses GIC with the safest choice λ instead of $K \log p$ as in Fan and Tang (2013) for some $K \to \infty$ with $n \to \infty$. It is worth to note that their results are obtained using the empirical processes theory, while the proof of Theorem 2 is based on elementary exponential inequalities for subgaussian variables given in Section 2.6.

2.5 | A selection error bound for SS in subgaussian linear models

In this part of the paper we show that SS is constructive for the linear model with the subgaussian noise. The main difference between Theorem 2 and the following Theorem 3 is that the lower bound on λ^2 in Theorem 3 does not depend on the dimension of the true model t and the parameter c (because $c = 1$ for subgaussian linear models).

Theorem 3. *Consider the linear model with the subgaussian noise. Assume that there exists positive number a such that $a \leq \sqrt{1 - (1 + \log(2 \log p))/(2 \log p)}$ and*

$$\frac{2\sigma^2 \log p}{a^4} \leq \lambda^2 < \frac{\zeta_a^2 \mathring{\beta}_{\min}^2 \wedge \delta}{4(1+a)^2}.$$

Then

$$P(\hat{T}_{\mathrm{SS}} \neq T) \leq 5 \exp\left(-\frac{a^2(1-a^2)\lambda^2}{2\sigma^2}\right). \tag{20}$$

The safest choice of λ in Theorem 3 does not depend on unknown expressions, which justifies the claim that our algorithm is constructive in the linear model with the subgaussian noise. Next, we compare the above result to Wang et al. (2013) and Fan et al. (2014).

Remark 3. The algorithm in Wang et al. (2013) has three steps: the Lasso, nonconvex penalized linear regression and GIC with the parameter $C \log p$, where $C \to \infty$ with $n \to \infty$. Obviously, the SS algorithm is computationally faster, because it does not need the most time-consuming second step. Moreover, the first two steps of their algorithm form a variant of FCP, so their parameters are not given constructively as we explain in the discussion after Theorem 1. Finally, their assumptions are stronger than ours. Indeed, conditions in (Wang et al., 2013, theorems 3.2 and 3.5) lead to $n/(\delta t) = O(1)$ and $t \log p / n = o(1)$, where δ is defined in (15). From these two facts and (Wang et al., 2013, expression (3.4)) we obtain

$$\frac{\log p}{\delta} = o(1) \quad \text{and} \quad \frac{t\sqrt{\log p}}{\mathring{\beta}_{\min}} = o(1). \tag{21}$$

If we fix σ^2 and ζ_a^2 as in (Wang et al., 2013, expression (3.4)), then in Theorem 3 we require only that

$$\frac{\log p}{\delta} = O(1) \quad \text{and} \quad \frac{\sqrt{\log p}}{\mathring{\beta}_{\min}} = O(1).$$

In (Wang et al., 2013, theorem 3.6) the second condition in (21) is weakened at the price of stronger conditions on the the design matrix. However, their assumptions are still more restrictive assumptions than ours.

Remark 4. (Fan et al., 2014, subsection 3.1) consider an algorithm based on nonconvex penalization in linear models. Their procedure is model selection consistent, but is also not constructive. Indeed, in (Fan et al., 2014, corollary 3) the parameter λ has to be in the interval, whose both endpoints depends on unknown $\|\beta^*_{\mathcal{A}}\|_{\min}, \sqrt{s}, \kappa_{\text{linear}}$ that are analogs of $\mathring{\beta}_{min}, t, \xi_{a_1}$ in the current paper. In (Fan et al., 2014, remark 5) it is shown that $\sqrt{s}$ can be avoided, but their algorithm is still not constructive in contrast to the SS algorithm.

2.6 | Exponential bounds for subgaussian vectors

This part of the paper is devoted to exponential inequalities for subgaussian random vectors. They are interesting by themselves and can be used in different problems than we consider. In the current paper they are main probabilistic tools that are needed to prove Theorems 1–3. Specifically, in Lemma 2 (iii) we generalize the Wallace inequality for χ^2 distribution Wallace (1959) to the subgaussian case using the inequality for the moment generating function in Lemma 2 (ii). The last inequality is proved by the decoupling technique as in the proof of theorem 2.1 in Hsu et al. (2012).

Lemma 2. *Let $\varepsilon \in \mathbb{R}^n$ be a vector of zero-mean independent errors having subgaussian distribution with a constant σ, $v \in \mathbb{R}^n$, $0 < a < 1$ and H be an orthogonal projection such that $tr(H) = m$. Then*

(i) for $\tau > 0$

$$\mathbb{P}(\varepsilon^T v/\|v\| \geq \tau) \leq \exp\left(-\frac{\tau^2}{2\sigma^2}\right), \tag{22}$$

(ii)

$$\mathbb{E}\exp\left(\frac{a}{2\sigma^2}\varepsilon^T H\varepsilon\right) \leq \exp\left(-\frac{m}{2}\log(1-a)\right),$$

(iii) for $\tau > 1$

$$\mathbb{P}(\varepsilon^T H\varepsilon \geq m\sigma^2\tau) \leq \exp\left(-\frac{m}{2}\left(\tau - 1 - \log\tau\right)\right). \tag{23}$$

3 | EXTENSION TO GENERAL CONVEX CONTRASTS

In this part of the paper we investigate properties of the SS algorithm beyond GLM as well. The main assumption, that will be required, is convexity of the contrast function. We show that the SS algorithm is very flexible procedure that can be applied successfully to the various spectrum of practical problems.

First, for $\beta \in \mathbb{R}^p$ and a contrast function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we define a loss function

$$\ell(\beta) = \sum_{i=1}^{n} \phi(\beta^T x_i, y_i).$$

Considering the normal linear model one usually uses the quadratic contrast

$$\phi(\beta^T x_i, y_i) = (y_i - \beta^T x_i)^2,$$

as we have done in Section 2. However, it is well known that the quadratic contrast is very sensitive to the distribution of errors ε_i and does not work well, if this distribution is heavy-tailed and outliers appear. To overcome this difficulty we can use the absolute contrast

$$\phi(\beta^T x_i, y_i) = |y_i - \beta^T x_i|.$$

Next, working with dichotomous y_i we can apply logistic regression that belongs to GLM and has been considered in Section 2. In this case we have

$$\phi(\beta^T x_i, y_i) = -y_i \beta^T x_i + \log[1 + \exp(\beta^T x_i)].$$

But there are also very popular and efficient algorithms called SVM that use, for instance, the following quadratic hinge contrast

$$\phi(\beta^T x_i, y_i) = [\max(0, 1 - y_i \beta^T x_i)]^2.$$

Our main assumption is that the contrast function ϕ is convex with respect to β. All examples given above satisfy this property. Notice that they need not be differentiable nor decomposable, as in (10) for GLM, into the sum of the nonrandom cumulant γ and the random linear term $y_i \beta^T x_i$. The SS algorithm for convex contrasts is the same as in Algorithm 1.

We add few definitions and notations to those in the previous parts of the paper. We start with defining two balls: the first one is the l_1-ball $B_1(r) = \{\beta : |\beta - \mathring{\beta}|_1 \le r\}$ with radius $r > 0$. The second one is the l_2-ball $B_{2,J}(r) = \{\beta_J : \|X_J(\mathring{\beta} - \beta_J)\|^2 \le r^2\}$ with radius $r > 0$, where J is a (sparse) subset of $\{1, \ldots, p\}$ such that $T \subset J, r(X_J) = |J| \le \bar{t}$. Recall that $\mathring{\beta}$ is, as previously, a minimizer of $\mathbb{E}\ell(\beta)$. Besides, let $B_J = B_{2,J}\left(\sqrt{\delta_{t-1}}\right)$, where δ_{t-1} is defined in (5). In further argumentation key roles are played by:

$$Z(r) = \sup_{\beta \in B_1(r)} \left| \ell(\beta) - \mathbb{E}\ell(\beta) - [\ell(\mathring{\beta}) - \mathbb{E}\ell(\mathring{\beta})] \right|,$$

and

$$U_J(r) = \sup_{\beta \in B_{2,J}(r)} \left| \ell(\beta) - \mathbb{E}\ell(\beta) - [\ell(\mathring{\beta}) - \mathbb{E}\ell(\mathring{\beta})] \right|,$$

which are empirical processes over l_1 and l_2-balls, correspondingly. We need also the compatibility factor that is borrowed from Bühlmann and van de Geer (2011) and is an analog of SCIF defined in (13). Namely, for arbitrary $a \in (0, 1)$ a compatibility factor is

$$\kappa_a = \inf_{0 \ne \beta \in C_a} \frac{\beta^T X^T X \beta}{|\beta_T|_1^2}, \tag{24}$$

where C_a is a simplified version of (12), namely

$$C_a = \left\{ \nu \in \mathbb{R}^p : |\nu_{\overline{T}}|_1 \le \frac{1+a}{1-a} |\nu_T|_1 \right\}.$$

Convexity of the contrast function is the main assumption in this section. However, similarly to the previous section we need also the following *strong convexity* of $\mathbb{E}\ell(\beta)$ at $\mathring{\beta}$: there exists $c_1 \in (0,1]$ ($c_2 \in (0,1]$, respectively) such that for each $\beta_1 \in B_1(\mathring{\beta}_{\min})$ ($\beta_2 \in \mathbb{B}$ defined in (6), respectively) we have for $i = 1, 2$

$$\mathbb{E}\ell(\beta_i) - \mathbb{E}\ell(\mathring{\beta}) \geq \frac{c_i}{2}(\beta_i - \mathring{\beta})^T X^T X (\beta_i - \mathring{\beta}). \tag{25}$$

Notice that in (25) we require the expected loss $\mathbb{E}\ell(\beta)$, not the loss $\ell(\beta)$, to be strongly convex. Therefore, the condition (25) can be satisfied easily even if the contrast function ϕ is not differentiable, for instance for absolute or quadratic hinge contrasts (see Remark 6). For GLM in section 2 the condition (11) is equivalent to (25) for $i = 2$, that will be explained in Remark 7.

To prove exponential bounds for GLM in subsection 2.4 we use subgaussianity that allows us to obtain probabilistic inequalities in Lemma 2. In this section we need the analog of (22) of the form: there exists $L > 0$ and constants $K_1, K_2 > 0$ such that for each $0 < r \leq \mathring{\beta}_{\min}$ and $z \geq 1$ we have

$$P\left(\frac{Z(r)}{r} > K_1 L z\sqrt{\log(2p)}\right) \leq \exp\left(-K_2 \log(2p)\, z^2\right). \tag{26}$$

Besides, the inequality (23) is replaced by the following: there exists $L > 0$ and constants $K_3, K_4 > 0$ such that for each $0 < r \leq \sqrt{\delta_{t-1}}$, $z \geq 1$ and J such that $T \subset J, r(X_J) = |J| \leq \bar{t}$ we have

$$P\left(\frac{U_J(r)}{r} > K_3 L z\sqrt{|J|}\right) \leq \exp(-K_4|J|z^2). \tag{27}$$

The detailed comparison between assumptions and results for models in this section and those for GLM in Theorem 2 is given in Remarks 7 and 8 after the main result of this section, which is now stated.

Theorem 4. *Fix $a_1, a_2 \in (0,1)$ and let K_i be universal constants. Assume that (25), (26), (27), and*

$$K_1 \max\left(\frac{\log p}{a_1^2}, \frac{\log p}{c_2}, \frac{t}{a_2 c_2}\right) L^2 \leq \lambda^2 \leq \tag{28}$$

$$\leq K_2 \min\left[\frac{c_2\delta}{(1+\sqrt{2a_2})^2}, \frac{c_2\delta_{t-1}}{\bar{t}-t}, (1-a_1)^2 c_1^2 \kappa_{a_1}^2 \mathring{\beta}_{\min}^2\right]. \tag{29}$$

Then

$$P\left(\hat{T}_{\mathrm{SS}} \neq T\right) \leq K_3 \exp\left[-\frac{K_4\lambda^2}{L^2}\min\left(a_1^2, a_2 c_2\right)\right]. \tag{30}$$

Theorem 4 bounds exponentially the selection error of the SS algorithm. It extends Theorem 2 to the wide class of convex contrast functions. In particular, these contrasts can be nondifferentiable as in quantile regression or SVM. In Remarks 5 and 6 we discuss assumptions (26), (27), and (25) of Theorem 4, respectively. The detail comparison to Theorem 2 is given in Remarks 7 and 8. There we argue that Theorem 4 applied to GLM is only slightly worse than Theorem 2, which is devoted to GLM.

Remark 5. The important assumptions of Theorem 4 are conditions (26) and (27). They can be proved using tools from the empirical process theory such that concentration inequalities (Massart, 2000), the Symmetrization Lemma (van der Vaart & Wellner, 1996, lemma 2.3.1) and the Contraction Lemma (Ledoux & Talagrand, 1991, theorem 4.12). It is quite remarkable that to get (26) or (27) we need only one new condition. Namely, we need that the contrast function is Lipschitz in the following sense: there exists $L > 0$ such that for all $x_i, y, 0 < r \leq \mathring{\beta}_{\min}$ and $\beta, \tilde{\beta} \in B_1(r)$

$$|\phi(\beta^T x_i, y) - \phi(\tilde{\beta}^T x_i, y)| \leq L|\beta^T x_i - \tilde{\beta}^T x_i|. \tag{31}$$

Indeed, (26) with $K_1 = 8\sqrt{2}, K_2 = 4$ follows from the above-mentioned tools and can be established as in (Bühlmann & van de Geer, 2011, lemma 14.20) combined with (Massart, 2000, theorem 9). On the other hand, to get (27) with $K_3 = 8, K_4 = 2$ we need (31) to be satisfied for all $x_i, y, 0 < r \leq \sqrt{\delta_{t-1}}, J : T \subset J, r(X_J) = |J| \leq \bar{t}$ and $\beta, \tilde{\beta} \in B_J$. This fact can be obtained as in (Bühlmann & van de Geer, 2011, lemma 14.19) combined again with (Massart, 2000, theorem 9). Notice that the logistic and the absolute contrast function satisfy (31) with $L = 2$ and $L = 1$, respectively. The property (31) is also satisfied for the quadratic hinge contrast, but in this case L depends on n.

Remark 6. The condition (25) is often called the margin condition in the literature. For quadratic and logistic contrasts we have considered it in the previous section. To prove it for SVM with the quadratic hinge contrast one can use methods based on the modulus of convexity (Bartlett et al., 2006, lemma 7). For linear models with the absolute contrast it can be established analogously to (Kim & Jeon, 2016, Lemma 3), if densities of noise variables are lower-bounded in a neighborhood of the origin.

Remark 7 (Similarities to Theorem 2). We compare Theorem 2 to Theorem 4 applied to GLM. We can calculate that for quadratic and logistic contrasts we have

$$\ell(\beta) - \ell(\mathring{\beta}) = -(\beta - \mathring{\beta})^T X^T y + g(\beta) - g(\mathring{\beta}),$$

and

$$\mathbb{E}\ell(\beta) - \mathbb{E}\ell(\mathring{\beta}) = -(\beta - \mathring{\beta})^T X^T \mathbb{E}y + g(\beta) - g(\mathring{\beta}),$$

where g is a total cumulant function (7). Therefore, the condition (25) for l_2-balls is the same as (11). Besides, we have for $\varepsilon = y - \mathbb{E}y$ that

$$\ell(\beta) - \mathbb{E}\ell(\beta) - [\ell(\mathring{\beta}) - \mathbb{E}\ell(\mathring{\beta})] = (\beta - \mathring{\beta})^T X^T \varepsilon. \tag{32}$$

In Theorem 2 we suppose that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and subgaussian. These assumptions are used to establish (22) and (23), that are crucial in the proof of Theorem 2. Notice that (32) implies that for GLM we have $Z(r)/r \leq \left|X^T\varepsilon\right|_\infty$ and $U_J(r)/r \leq \sqrt{\varepsilon^T H_J \varepsilon}$. Therefore, assumptions (26) and (27) are analogs of (22) and (23), respectively. Moreover, the condition (28) and the result in (30) differ only in constants from their counterparts in Theorem 2.

Remark 8 (Differences from Theorem 2). The SS algorithm consists of two steps. In the last paragraph we have clarified that the theoretical analysis of the second step (selection) in Theorem 2 is not significantly simpler than for general models with convex contrasts. However, we can find

differences while investigating the first step (screening based on the lasso). Working with GLM we exploit differentiability of contrasts and the useful decomposition (10). Due to that the right-hand side of (16) is usually better than (29), because in Theorem 4 we have to assume (25) also with respect to l_1-balls and c_1^2 appears in (29).

4 | EXPERIMENTS

While convenient for theoretical analysis TL, FCP, or SS algorithms use the Lasso estimators only for one value of the penalty, the practical Lasso implementations return coefficient estimators for a given net of them as in the R package `glmnet` described in Friedman et al. (2010). Similarly, using a net of penalty values, the Minimax Concave Penalty (MCP) algorithm, a popular realization of FCP, has been implemented for linear and logistic models in the R packages `SparseNet`, Mazumder et al. (2011), and `ncvreg`, Breheny and Huang (2011), respectively. Thus, we propose a net modification of the SS algorithm, which we call SSnet and state in Algorithm 2. In the first step this procedure calculates the Lasso for a net of m values of *an input grid*: $\lambda_1^L, \dots, \lambda_m^L$. Then the final model is chosen using GIC in a similar way to the SS algorithm.

We remind that for linear models λ depends on the noise variance, which is unknown in practice. Estimation of σ^2 is quite difficult, especially in the high-dimensional scenario. Computing the Lasso for the whole net of tuning parameters is a simple way to overcome this problem. Obviously, in the second step of our procedure we use GIC, so we require an estimator of σ^2. But in this step we work with models, whose dimensionality is significantly reduced by the Lasso. Therefore, the classical estimator of σ^2, which uses residual sum of squares, can be applied.

We performed numerical experiments fitting sparse linear and logistic models to high-dimensional benchmark simulations and real data sets. We investigated properties in model selection and prediction of the SSnet algorithm and its competitor, which was a net version of MCP. The fair comparison of these two net algorithms is difficult, because their input grids depend on these algorithm themselves and their runs. Therefore, we decide to introduce *an output grid*, which is the same for both algorithms $\lambda_1 < \dots < \lambda_r$ in Algorithm 2. We compare the algorithms

Algorithm 2. SSnet

Input: y, X and (input grid $: \lambda_1^L < \dots < \lambda_m^L$, output grid $: \lambda_1 < \dots < \lambda_r$)
Screening (Lasso)
for $k = 1$ **to** m **do**
  $\widehat{\beta}^{(k)} = \operatorname{argmin}_\beta \left\{ \ell(\beta) + \lambda_k^L |\beta|_1 \right\}$;
  order nonzero $|\widehat{\beta}_{j_1}^{(k)}| \geq \dots \geq |\widehat{\beta}_{j_{s_k}}^{(k)}|$, where $s_k = |\operatorname{supp} \widehat{\beta}^{(k)}|$;
  set $\mathcal{J}_k = \left\{ \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, j_2, \dots, j_{s_k}\} \right\}$
end for;
Selection (GIC)
$\mathcal{J} = \bigcup_{k=1}^m \mathcal{J}_k$;
for $k = 1$ **to** r **do**
  $\widehat{T}_k = \operatorname{argmin}_{J \in \mathcal{J}} \left\{ \ell_J + \lambda_k^2 / 2|J| \right\}$;
end for;
Output: $\widehat{T}_1, \dots, \widehat{T}_r$.

TABLE 1 Plan of experiments for linear models

	n	p	$\mathring{\beta}$	ρ	σ^2	SNR
N.1.5	100	3000	$\mathring{\beta}^{(1)}$	.5	4	2.3
N.1.7	100	3000	$\mathring{\beta}^{(1)}$	.7	4	2.6
N.1.9	100	3000	$\mathring{\beta}^{(1)}$	.9	4	3
N.2.5	200	2000	$\mathring{\beta}^{(2)}$	.5	7	2.4
N.2.7	200	2000	$\mathring{\beta}^{(2)}$	.7	7	2.3
N.2.9	200	2000	$\mathring{\beta}^{(2)}$	.9	7	2.2

Abbreviation: SNR, Signal to Noise Ratio.

using plots, which describe the interplay between a prediction error (PE) and a model dimension (MD). In these plots we show the averaged PE and MD over simulation runs (datasets) for distinct values from the output grid. In particular, we want to find the minimum PE and the optimal penalty. Basing only on input grids, it would be more difficult to interpret such averaging as well as to locate the PE minimum. Finally, using such plots makes our procedure more *user-friendly*. Namely, observing the relation between dimensionality of a model and its prediction accuracy a user can decide, which estimator is best suited for both: describing the phenomena and his/her expectations.

4.1 | Simulated data

For linear models we studied the performance of two algorithms: SSnet and MCP computed using the R package `SparseNet` Mazumder et al. (2011) for the default nine values of γ and 50 values of λ. Our algorithm used the R package `glmnet` Friedman et al. (2010) to compute the Lasso estimators for 50 lambdas on a log scale.

We generated samples (y_i, x_i), $i = 1, \ldots, n$ from the normal linear model. Two vectors of parameters were considered: $\mathring{\beta}^{(1)} = (3, 1.5, 0, 0, 2, 0^T_{p-5})^T$, as in Wang et al. (2013) as well as $\mathring{\beta}^{(2)} = (0^T_{p-10}, s_1 \cdot 2, s_2 \cdot 2, \ldots, s_{10} \cdot 2)^T$, where s_l equals 1 or -1 with equal probability, $l = 1, \ldots, 10$ chosen separately for every run as in experiment 2 in Wang et al. (2014). The rows of X were iid p-dimensional vectors $x_i \sim N(0_p, \Xi)$. We considered auto-regressive structure of covariance matrix that is $\Xi = \left(\rho^{|i-j|}\right)_{i,j=1}^{p}$ for $\rho = 0.5, 0.7, 0.9$. The columns of X were centered and normalized so that $\|x_j\|^2 = n$ and $\varepsilon \sim N(0_n, \sigma^2 I_n)$. The plan of experiments is presented in Table 1 with SNR meaning a *Signal to Noise Ratio*. For every experiment the results were based on $N = 1000$ simulation runs.

The output grid was chosen as $\lambda_k^2 = c_k \cdot \log(p) \cdot \sigma^2, k = 1, \ldots, r$, where $(c_k)_{k=1}^r = (.25, .5, \ldots, 7.5)$. We reported the mean MD that is $|\widehat{T_k}|, k = 1, \ldots, r$ and the mean squared PE on new dataset with 1000 observations equalling $\|X\mathring{\beta} - X\widehat{\beta}_{\widehat{T_k}}\|^2/(n\sigma^2)$, where $\widehat{\beta}_{\widehat{T_k}}$ is the postselection OLS estimator. Values of $(\mathrm{MD}(k), \mathrm{PE}(k))$ for the models chosen by GIC $=$ GIC(λ_k) were calculated and averaged over simulations.

The results are presented in two first columns of Figure 1. The two vertical lines indicate models chosen using GIC with $c_k = 2.5$: the black one for SSnet and the red one for `SparseNet`. The blue vertical line denotes the true MD.

For logistic models we compared the performance of two algorithms: SSnet and MCP implemented in the R package `ncvreg` for the default value of $\gamma = 3$ and 100 values of λ. As for

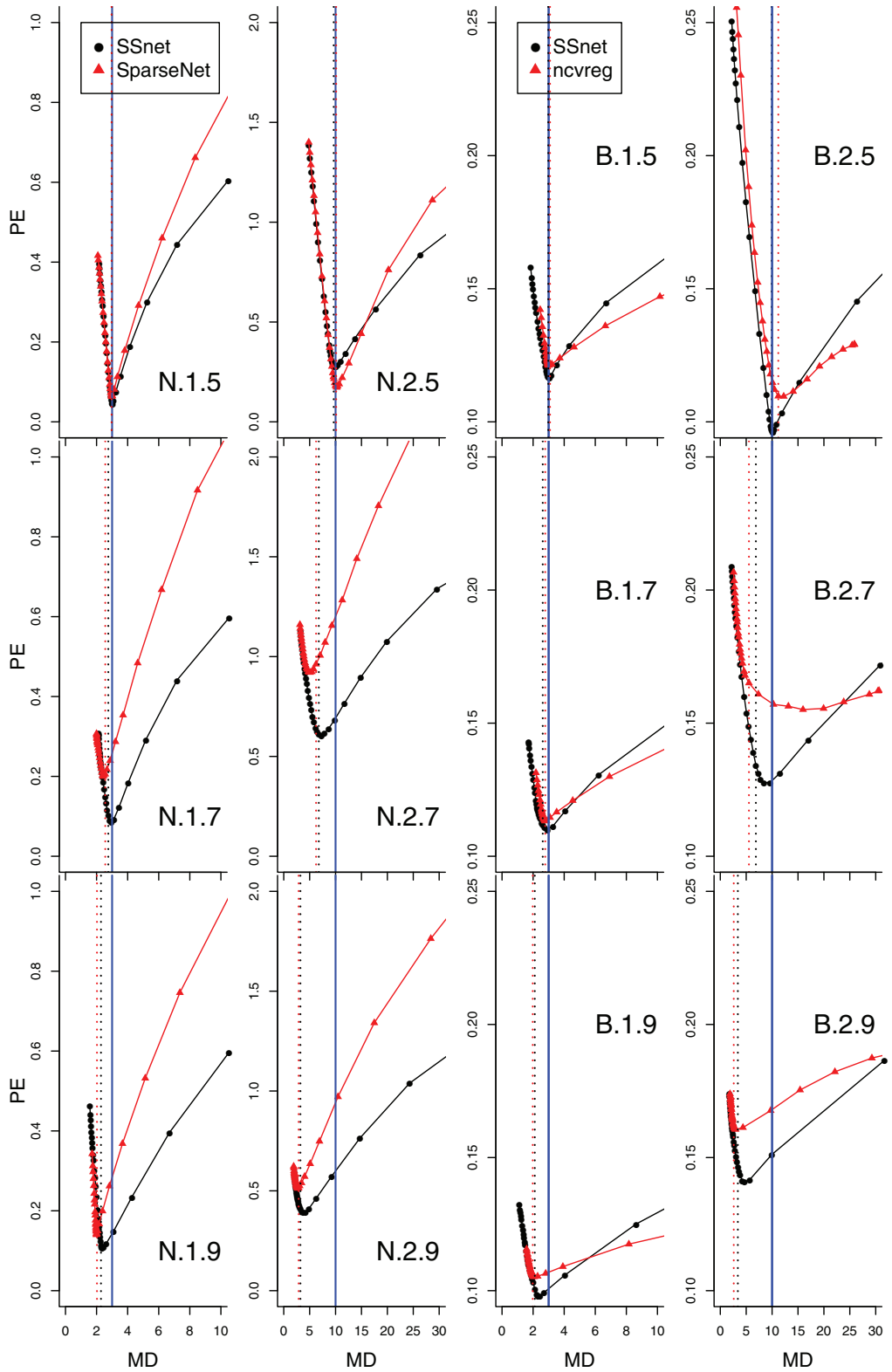

FIGURE 1 Results for simulated data [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Plan of experiments for logistic models

	n	p	$\mathring{\beta}$	ρ
B.1.5	300	3000	$\mathring{\beta}^{(1)}$	.5
B.1.7	300	3000	$\mathring{\beta}^{(1)}$	.7
B.1.9	300	3000	$\mathring{\beta}^{(1)}$	.9
B.2.5	500	2000	$\mathring{\beta}^{(2)}$	.5
B.2.7	500	2000	$\mathring{\beta}^{(2)}$	.7
B.2.9	500	2000	$\mathring{\beta}^{(2)}$	.9

linear models, SSnet called the R package `glmnet` Friedman et al. (2010) to compute the Lasso estimators for 20 lambdas on a default log scale. We performed experiments very similar to those for linear models, changing only n and the number of simulation runs to $N = 500$. The plan of experiments is shown in the Table 2. Random samples were generated according to the binomial distribution. We reported PE defined as misclassification frequency on new data set with 1000 observations. The results organized in a similar way as for the linear models are shown in columns 3 and 4 of Figure 1. The two vertical lines indicate models chosen using GIC with $c_k = 2$, the black one for SSnet and the red one for `ncvreg`.

Summarizing the results of the simulation study, one can observe that SSnet for linear models turned out to have equal or lower PE in almost all of the experimental setups. The differences are most visible in setups with autocorrelation structure with $\rho = 0.7$. The value $c_k = 2.5$ in GIC usually gave satisfactory results. The mean execution time of SSnet was approximately three times faster than for `SparseNet`. SSnet for logistic regression gave higher accuracy as `ncvreg`, but execution time of SSnet was approximately 10 times longer than `ncvreg`. The value $c_k = 2$ in GIC usually gave satisfactory results.

4.2 | Real datasets

The methylation data set was described in Hannum et al. (2013). It consists of the age of 656 human individuals together with values of phenotypic features such as gender and body mass index and of genetic features, which are methylation states of 485 577 CpG markers. Methylation was recorded as a fraction representing the frequency of methylation of a given CpG marker across the population of blood cells taken from a single individual. In our comparison we used only genetic features from which we extracted 193,870 most relevant CpGs according to onefold t-tests with Benjamini–Hochberg adjustment, FDR = 0.05. We compared the root mean squared errors (PE) and model dimensions (MD) for SSnet and `SparseNet`. To calculate PE we used 10-fold cross-validation. For each value of hyperparameter $c_k = .25, .5, \ldots, 7.5$ values of (MD(k), PE(k)) for the models chosen by GIC = GIC(λ_k) were calculated and averaged over 10-folds. The results are presented in Figure 2. `SparseNet` yields a path of models for each value of parameter $\gamma = g1, \ldots, g9$. We present results for $g1$, corresponding to the Lasso, for $g9$, close to the best subset, and for an intermediate value $g8$ in Figure 2. Remarkably, SSnet gives uniformly smaller PE than `SparseNet` for all $MD \geq 3$. The two vertical lines indicate models chosen using GIC with $c_k = 2.5$: the black one for SSnet and the red one for `SparseNet`.

A logistic model was fitted to the breast cancer data described in Gravier et al. (2010) which concerns small, invasive carcinomas without auxiliary lymph node involvement to predict

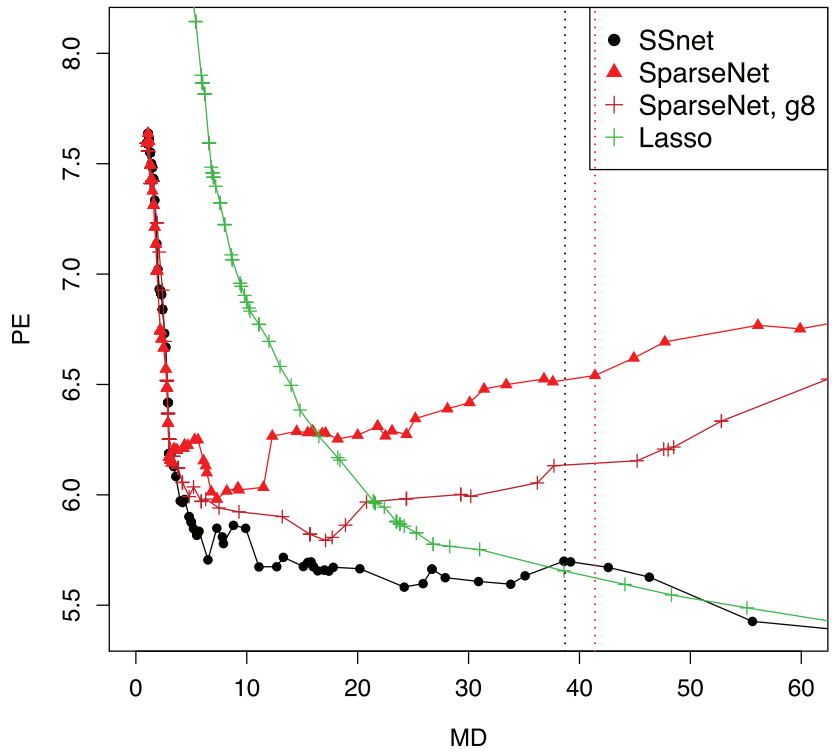

FIGURE 2 Results for the methylation dataset [Colour figure can be viewed at wileyonlinelibrary.com]

metastasis of small node-negative breast carcinoma. There were 168 patients: 111 with no event after diagnosis labeled as good, and 57 with early metastasis labeled as poor. The number of predictors in this data was 2905. We compared the mean errors of binary prediction (PE) and MDs for SSnet and `ncvreg`. For each of 80 hyperparameters $c_k = 0.001; 0.005; 0.01; 0.02, \ldots, 0.1, 0.15, \ldots, 1, \ldots$ 50 values of (MD(k); PE(k)) for the models chosen by GIC = GIC(λ_k) were calculated and averaged over 10 folds. Again to calculate PE we used 10-fold cross-validation. The results are presented in Figure 3. It is hard to find the minimal of PE for `ncvreg`. If we increased the net of c_k, maybe we would obtain a smaller PE for ncvreg, but the model would be significantly larger than for SSnet. The algorithms work comparably, but again SSnet was three times longer. The two vertical lines indicate models chosen using GIC with $c_k = 2$: the black one for SSnet and the red one for `ncvreg`.

5 | CONCLUSIONS

In the paper we propose the SS algorithm which is an alternative method to TL and FCP of improving the Lasso. Our approach encompasses fundamental models for prediction of continuous as well as of binary responses and the main results are stated jointly for both of them. Its assumptions are stated in the most general form which allows proving exponential bound without obscuring the essence of the results and comparing the bounds for both models. By simplifying SOS to SS we were able to simplify reasoning used for SOS and then extend them from normal linear models to general predictive models.

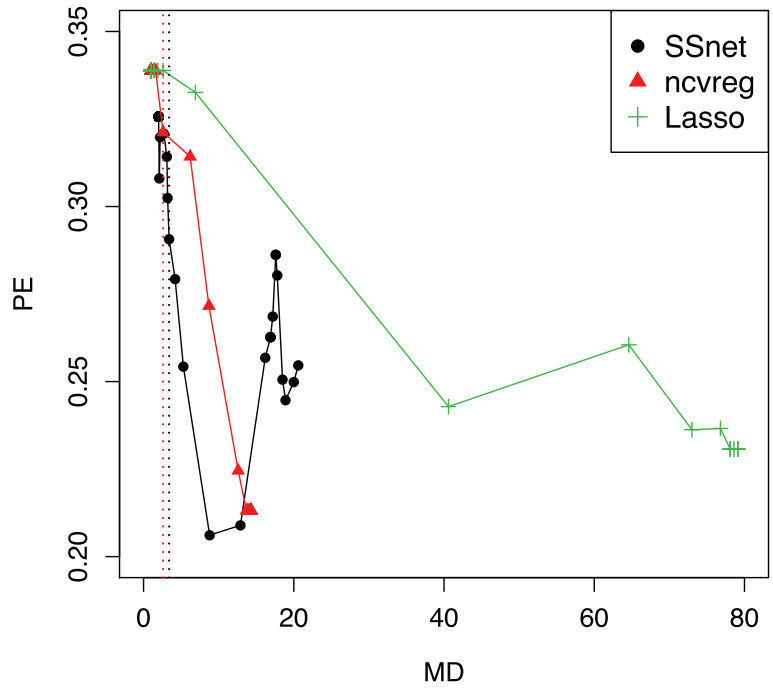

FIGURE 3 Results for the breast cancer dataset [Colour figure can be viewed at wileyonlinelibrary.com]

We propose the algorithm SSnet, which is a generalization of the SS algorithm for general predictive models. Using a net of parameters, SSnet avoids the problem of choosing one specific λ. The gap between theoretical results for SS and the SSnet algorithm is similar to the difference between theory for FCP and it implementations `SparseNet` or `ncvreg`. Numerical experiments reveal that for normal linear models SSnet is more accurate than `SparseNet` and three times faster, whereas for logistic models performance of SSnet is also better than the performance of `ncvreg` with computing times 3–10 times longer.

We have shown in simulations (dotted vertical lines in Figure 1) that predictively optimal λ for normal linear models equals approximately $\sqrt{2.5\sigma^2 \log p}$, which is close to (18) and for logistic models is $\sqrt{2 \log p}$, which together with (19) suggests that $c \approx 1/4$. The relations between the safest choice λ discussed in Remark 1 and predictively optimal λ are important applications of our theory.

In the package `DMRnet` one can find the SOSnet algorithm, which is a net version of SOS from Pokarowski and Mielniczuk (2015). SOSnet can be viewed as SSnet with additional step ("Ordering") based on refitting estimators and calculating Wald statistics. Results of our numerical experiments, which are not presented here, suggest that SOSnet can improve the quality of SSnet, especially in the linear model.

ORCID

Wojciech Rejchel https://orcid.org/0000-0003-1148-1439

REFERENCES

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association*, *101*, 138–156.

Bickel, P., Ritov, Y., & Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, *37*, 1705–1732.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, *5*, 232–253.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, *42*, 819–849.

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *75*, 531–552.

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, *22*, 1947–1975.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., de Rycke, Y., Pierga, J. Y., Lucchesi, C., Reyal, F., Fourquet, A., Roman-Roman, S., Radvanyi, F., Sastre-Garau, X., Asselain, B., & Delattre, O. (2010). Prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*, *49*(12), 1125–1125, 1134.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, *49*(2), 359–367.

Hsu, D., Kakade, S. M., & Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, *17*, 1–6.

Huang, J., & Zhang, C. H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, *13*, 1839–1864.

Hui, F. K. C., Warton, D. I., & Foster, S. D. (2015). Tuning parameter selection for the adaptive Lasso using ERIC. *Journal of the American Statistical Association*, *110*, 262–269.

Katayama, S., & Imori, S. (2014). Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *Journal of Multivariate Analysis*, *132*, 138–150.

Kim, Y., & Jeon, J. J. (2016). Consistent model selection criteria for quadratically supported risks. *The Annals of Statistics*, *44*, 2467–2496.

Kim, Y., Kwon, S., & Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, *13*, 1037–1057.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces: Isoperimetry and processes*. Springer.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, *28*, 863–884.

Mazumder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*, 1125–1138.

Meinshausen, N., & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*, 1436–1462.

Pokarowski, P., & Mielniczuk, J. (2015). Combined ℓ_0 and ℓ_1 penalized least squares for linear model selection. *Journal of Machine Learning Research*, *16*, 961–992.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*, 221–242.

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, *107*, 223–232.

Shen, X., Pan, W., Zhu, Y., & Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, *65*, 807–832.

Tibshirani, R. (2011). Regression shrinkage and selection via the Lasso: A retrospective. *Journal of the Royal Statistical Society Series B*, *73*, 273–282.

van de Geer, S. (2002). M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, *108*, 55–69.

van de Geer, S., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, *3*, 1360–1392.
van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer Verlag.
Wallace, D. L. (1959). Bounds on normal approximations to Student's and the chi-square distributions. *Annals of Mathematical Statistics*, *30*, 1121–1130.
Wang, L., Kim, Y., & Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics*, *41*, 2505–2536.
Wang, Z., Liu, H., & Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, *42*, 2164–2201.
Ye, F., & Zhang, C. H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the l_q loss in l_r balls. *Journal of Machine Learning Research*, *11*, 3519–3540.
Zhang, C. H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*, 894–942.
Zhang, C. H., & Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*, 576–593.
Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, *11*, 1081–1107.
Zhang, X., Wu, Y., Wang, L., & Li, R. (2016). A consistent information criterion for support vector machines in diverging model spaces. *Journal of Machine Learning Research*, *17*, 466–491.
Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.
Zhou, S. (2009). *Thresholding procedures for high dimensional variable selection and statistical estimation* (pp. 2304–2312). NIPS.

How to cite this article: Pokarowski, P., Rejchel, W., Sołtys, A., Frej, M., & Mielniczuk, J. (2021). Improving Lasso for model selection and prediction. *Scandinavian Journal of Statistics*, 1–33. https://doi.org/10.1111/sjos.12546

APPENDIX. PROOFS AND AUXILIARY RESULTS

In the following subsections we present proofs of the results stated in the paper.

A.1 Proof of Lemma 6

Let $Z = \varepsilon^T v/\|v\|$ and $a > 0$. From Markov's inequality we obtain

$$\mathbb{P}(Z \geq \tau) \leq e^{-a}\mathbb{E}e^{aZ} \leq e^{-a\tau + a^2\sigma^2/2}.$$

Minimizing the last expression w.r.t. a gives part (i).

Let $\xi \sim N(0, I_n)$ be a vector of iid standard normal errors independent of ε. We have

$$\mathbb{E}\exp\left(\frac{a}{2\sigma^2}\varepsilon^T H\varepsilon\right) = \mathbb{E}\,\mathbb{E}\left[\exp\left(\frac{\sqrt{a}}{\sigma}\xi^T H\varepsilon\right)\Big|H\varepsilon\right] = \mathbb{E}\exp\left(\frac{\sqrt{a}}{\sigma}\xi^T H\varepsilon\right)$$
$$= \mathbb{E}\,\mathbb{E}\left[\exp\left(\frac{\sqrt{a}}{\sigma}\xi^T H\varepsilon\right)\Big|\xi^T H\right] \leq \mathbb{E}\exp\left(\frac{a}{2}\xi^T H\xi\right).$$

Thus part (ii) follows from a known formula for the moment generating function of the χ^2 distribution.

From Markov's inequality and the part (ii) of this lemma we have

$$\begin{aligned}\mathbb{P}(\varepsilon^T H\varepsilon \geq m\sigma^2\tau) &\leq \exp\left(-\frac{am\tau}{2}\right)\mathbb{E}\exp\left(\frac{a}{2\sigma^2}\varepsilon^T H\varepsilon\right)\\ &\leq \exp\left(-\frac{m}{2}\left(a\tau+\log(1-a)\right)\right).\end{aligned}$$

Thus, minimizing the last expression w.r.t a we obtain part (iii). ∎

A.2 Proof of Lemma 1

Let

$$\mathcal{A}_a = \{|X^T\varepsilon|_\infty \leq a\lambda\}, \tag{A1}$$

and $\hat{\varepsilon} = y - \vec{\mathring{\gamma}}(X\hat{\beta})$. We have $\mathring{\ell}(\hat{\beta}) = -X^T\hat{\varepsilon}$ and from the Karush-Kuhn-Tucker (KKT) theorem we obtain equations

$$X^T\hat{\varepsilon} = \lambda\left[\mathbb{I}(\hat{\beta}>0) - \mathbb{I}(\hat{\beta}<0) + u\mathbb{I}(\hat{\beta}=0)\right] \quad \text{for} \quad u\in[-1,1].$$

Let $\Delta = \hat{\beta} - \mathring{\beta}$ and $\nu \in \mathbb{R}^p$ be such that $\operatorname{sgn}(\nu_{\overline{T}}) = \operatorname{sgn}(\Delta_{\overline{T}})$. We have $\nu_J^T X_J^T\hat{\varepsilon} = \lambda|\nu_J|_1$ for $J \subseteq \overline{T}$ and consequently

$$\begin{aligned}D(\nu) &= \nu^T X^T\left[\vec{\mathring{\gamma}}(X\hat{\beta}) - \vec{\mathring{\gamma}}(X\mathring{\beta})\right] = \nu_T^T X_T^T(\varepsilon-\hat{\varepsilon}) + \nu_{\overline{T}}^T X_{\overline{T}}^T(\varepsilon-\hat{\varepsilon})\\ &\leq |\nu_T|_1(|X_T^T\varepsilon|_\infty + |X_T^T\hat{\varepsilon}|_\infty) + |\nu_{\overline{T}}|_1(|X_{\overline{T}}^T\varepsilon|_\infty - \lambda)\\ &\leq |\nu_T|_1(1+a)\lambda + |\nu_{\overline{T}}|_1(a-1)\lambda.\end{aligned} \tag{A2}$$

Then letting $\nu = \Delta_J$ for $J \subseteq \overline{T}$ we have $D(\nu) \leq 0$. Moreover, for $\nu = \Delta$ we have from convexity of g that

$$D(\nu) = (\hat{\beta} - \mathring{\beta})^T[\dot{g}(\hat{\beta}) - \dot{g}(\mathring{\beta})] \geq 0.$$

Indeed, $D_0(\beta_1,\beta_2) = (\beta_1-\beta_2)^T[\dot{g}(\beta_1) - \dot{g}(\beta_2)]$ is the symmetrized Bregman divergence (Huang & Zhang, 2012). Hence $(1-a)|\nu_{\overline{T}}|_1 \leq (1+a)|\nu_T|_1$. Thus, on $\mathcal{A}_a$, $\Delta \in \mathcal{C}_a$ and from the definition of ζ_a we obtain using KKT again

$$\zeta_{a,q}|\Delta|_q \leq \left|X^T\left[\vec{\mathring{\gamma}}(X\hat{\beta})) - \vec{\mathring{\gamma}}(X\mathring{\beta})\right]\right|_\infty \leq |X^T\hat{\varepsilon}|_\infty + |X^T\varepsilon|_\infty \leq (1+a)\lambda.$$ ∎

A.3 Proof of Theorem 1

First, we will prove that $\mathcal{A}_a \subseteq \{\hat{T}_{TL} = T\}$ for $\mathcal{A}_a$ defined in (A1). From Lemma 1 and assumptions we have on $\mathcal{A}_a$

$$|\Delta|_\infty \leq (1+a)\lambda\zeta_a^{-1} \leq \tau < \mathring{\beta}_{\min}/2, \tag{A3}$$

where we recall that $\Delta = \hat{\beta} - \mathring{\beta}$. Thus using (A3) twice we have for $j \in T$ and $k \notin T$

$$|\hat{\beta}_j| \geq |\mathring{\beta}_j| - |\hat{\beta}_j - \mathring{\beta}_j| > \mathring{\beta}_{\min} - \mathring{\beta}_{\min}/2 > \tau \geq |\hat{\beta}_k - \mathring{\beta}_k| = |\hat{\beta}_k|, \tag{A4}$$

and it follows that $\mathcal{A}_a \subseteq \{\hat{T}_{TL} = T\}$. Moreover, the assumptions of this theorem imply

$$-a_1^2\lambda^2 + 2\sigma^2 \log p \leq -(1-a_2)a_1^2\lambda^2.$$

Hence, using Lemma 2 (i) we easily obtain

$$\mathbb{P}(\hat{T}_{TL} \neq T) \leq \mathbb{P}(\mathcal{A}_{a_1}^c) \leq 2p \exp\left(-\frac{a_1^2\lambda^2}{2\sigma^2}\right) \leq 2\exp\left(-\frac{(1-a_2)a_1^2\lambda^2}{2\sigma^2}\right). \quad \blacksquare$$

A.4 Proof of Theorem 2

Let us observe that the consecutive steps of the SS algorithm constitute a decomposition of the selection error into two parts:

$$\{\hat{T} \neq T\} = \{T \notin \mathcal{J}\} \cup \{T \in \mathcal{J}, \hat{T} \neq T\}. \tag{A5}$$

Therefore, Theorem 2 follows easily from (A6) and (A10) below.

Having in mind that for given $a_1 \in (1/2, 1)$ we let $a_2 = 1 - (1 - \log(1-a_1))(1-a_1)$, $a_3 = 2 - 1/a_1$ and $a_4 = \sqrt{a_1 a_2}$, by arguments similar to those in the proof of Theorem 1 we obtain $\mathcal{A}_{a_4} \subseteq \{T \in \mathcal{J}\}$. Moreover, assumptions $0 < c \leq 1$ and $\frac{2\sigma^2 \log p}{a_3 a_2 a_1 c} \leq \lambda^2$ imply

$$-a_4^2\lambda^2 + 2\sigma^2 \log p \leq -a_2 a_1 c\lambda^2 + 2\sigma^2 \log p \leq -(1-a_3)a_2 a_1 c\lambda^2 = -a_2(1-a_1)c\lambda^2.$$

As a result

$$\mathbb{P}(T \notin \mathcal{J}) \leq \mathbb{P}(\mathcal{A}_{a_4}^c) \leq 2p \exp\left(-\frac{a_4^2\lambda^2}{2\sigma^2}\right) \leq 2\exp\left(-\frac{a_2(1-a_1)c\lambda^2}{2\sigma^2}\right). \tag{A6}$$

Now we bound $\mathbb{P}\left(T \in \mathcal{J}, \hat{T} \neq T\right)$. In Lemmas 4 and 5, given below, we bound probability that in the second step of the SS algorithm GIC chooses a subset of the true set, that is,

$$\{T \in \mathcal{J}, \hat{T} \subset T\}$$

or a superset of T, that is,

$$\{T \in \mathcal{J}, \hat{T} \supset T\}.$$

These lemmas state that both components of the selection error set are included in the critical sets of the following form

$$C_J(\tau) = \{\varepsilon^T H_J \varepsilon \geq \tau\} \tag{A7}$$

or

$$C_{J \ominus T}(\tau) = \{\varepsilon^T (H_J - H_T)\varepsilon \geq \tau\}, \tag{A8}$$

where J is such that $T \subset J$, $r(X_J) = |J| \leq \bar{t}$. We consider only supersets J that $r(X_J) = |J|$, because GIC corresponding to the superset J such that $r(X_J) < |J|$ is larger than GIC corresponding to a superset J_1 such that $J_1 \subsetneq J$ and $r(X_{J_1}) = |J_1|$.

Let us define $\tau_0 = \frac{1}{1-a_1}$, $\tau_1 = \frac{(1-a_1)c\lambda^2}{t\sigma^2}$ and $\tau_2 = \frac{a_1 c\lambda^2}{\sigma^2}$. Under our assumptions we have $2 < \tau_0 < \tau_1 < \tau_2$. Let $f_2(\tau) = 1 - (1 + \log\tau)/\tau$ for $\tau > 1$. Of course f_2 is increasing, $f_2(1) = 0$ and $f_2(\tau) \to 1$ for $\tau \to \infty$. Consequently $a_2 = f_2(\tau_0) < f_2(\tau_1) < f_2(\tau_2)$, which means that

$$a_2\tau_r < \tau_r - 1 - \log\tau_r \text{ for } r = 1, 2. \tag{A9}$$

From Lemma 4, Lemma 5, Lemma 2 (iii) and (A9) we get

$$\begin{aligned}
\mathbb{P}\left(T \in \mathcal{J}, \hat{T} \neq T\right) &\leq \mathbb{P}\left(C_T(t\sigma^2\tau_1)\right) + \sum_{J \supset T : r(X_J) = |J| \leq \tilde{t}} \mathbb{P}\left(C_{J \ominus T}(|J \setminus T|\sigma^2\tau_2)\right) \\
&\leq \exp\left(-ta_2\tau_1/2\right) + \sum_{m=1}^{p-t} \binom{p-t}{m} \exp\left(-ma_2\tau_2/2\right) \\
&\leq \exp\left(-ta_2\tau_1/2\right) + \sum_{m=1}^{p-t} \frac{1}{m!} \exp\left(\frac{-m}{2}(a_2\tau_2 - 2\log p)\right).
\end{aligned}$$

Using $\exp(d) - 1 \leq \log(2)^{-1} d$ for $0 \leq d \leq \log(2)$ and the fact that probability is not greater than 1 we obtain

$$\mathbb{P}\left(T \in \mathcal{J}, \hat{T} \neq T\right) \leq \exp\left(-ta_2\tau_1/2\right) + (\log 2)^{-1} \exp\left(-(a_2\tau_2 - 2\log p)/2\right).$$

For $a_3 \in (0, 1)$, assumption $\frac{2\sigma^2 \log p}{a_3 a_2 a_1 c} \leq \lambda^2$ implies $-a_2\tau_2 + 2\log p \leq -(1 - a_3)a_2\tau_2$, therefore

$$\mathbb{P}\left(T \in \mathcal{J}, \hat{T} \neq T\right) \leq \exp\left(-ta_2\tau_1/2\right) + (\log 2)^{-1} \exp\left(-(1 - a_3)a_2\tau_2/2\right), \tag{A10}$$

that finishes the proof of Theorem 2. ∎

Before we state Lemma 4 and 5 we introduce few notations. For $k = 1, \ldots, t-1$ and δ_k in (5) we define

$$\begin{aligned}
\mathcal{E}_k(\tau) &= \{\exists J \subset T,\ |T \setminus J| = k :\ \ell_J - \ell_T \leq \tau\}, \\
B_k &= \{\beta_T :\ \|X_T\beta_T - X_T\mathring{\beta}_T\|^2 \leq \delta_k\}.
\end{aligned} \tag{A11}$$

First, we prove the following lemma which will be used in the proof of Lemma 4.

Lemma 3. *For $b \in (0, 1)$ we have*

$$\mathcal{E}_k(bc\delta_k/2) \subseteq C_T((1 - b)^2 c^2 \delta_k/4).$$

Proof. For $\beta_T \in \partial B_k$ from assumption (11), Schwartz inequality and properties of the orthogonal projection H_T we get

$$\begin{aligned}
\mathring{\ell}(\beta_T) := \ell(\beta_T) - \ell(\mathring{\beta}_T) &\geq (\mathring{\beta}_T - \beta_T)^T X_T^T H_T \varepsilon + \frac{c}{2}(\mathring{\beta}_T - \beta_T)^T X_T^T X_T (\mathring{\beta}_T - \beta_T) \\
&\geq -\sqrt{\delta_k \varepsilon^T H_T \varepsilon} + \frac{c}{2}\delta_k.
\end{aligned}$$

Since the last expression does not depend on β_T, we have for $b \in (0,1)$

$$\mathcal{L}_k(b) := \left\{ \min_{\beta_T \in \partial B_k} \mathring{\ell}(\beta_T) \le \frac{bc\delta_k}{2} \right\} \subseteq \left\{ -\sqrt{\delta_k \varepsilon^T H_T \varepsilon} + \frac{c}{2}\delta_k \le \frac{bc\delta_k}{2} \right\}$$
$$= C_T\left(\frac{(1-b)^2 c^2 \delta_k}{4} \right).$$

Let us notice that for $J \subset T$ such that $|T \setminus J| = k$ we have $\|X_T(\hat{\beta}_J^{\mathrm{ML}} - \mathring{\beta}_T)\|^2 \ge \delta_k$, so $\hat{\beta}_J^{\mathrm{ML}} \notin \mathrm{int}(B_k)$. Since $\mathring{\ell}$ is convex and $\mathring{\ell}(\mathring{\beta}_T) = 0$ we obtain $\mathcal{L}_k(b) \supseteq \mathcal{E}_k(bc\delta_k/2)$. ∎

Lemma 4. *If for $a \in (0,1)$ we have $\lambda^2 < c\delta \Big/ \left(1 + \sqrt{2a}\right)^2$, then $\{T \in \mathcal{J}, \hat{T} \subset T\} \subseteq C_T(ac\lambda^2)$.*

Proof. From assumption of this lemma $\lambda^2 < c\delta$, so $b_k := \lambda^2 k/(c\delta_k) < 1$. Hence for $k = 1, \ldots, t-1$

$$k\lambda^2 \le b_k c\delta_k. \tag{A12}$$

Moreover, if $a \in (0,1)$ then

$$akc\lambda^2 \le (1-b_k)^2 c^2 \delta_k/4, \tag{A13}$$

because $ab \le (1-b)^2/4$ for $b = \lambda^2/(c\delta)$. The last inequality is true, if

$$b \le 1 + 2a - \sqrt{(1+2a)^2 - 1} = f_1(a). \tag{A14}$$

Indeed, it is easy to check that (A14) follows from the assumption as

$$f_1(a) = \frac{1}{1 + 2a + \sqrt{(1+2a)^2 - 1}} \ge \frac{1}{(1+\sqrt{2a})^2}.$$

Finally from (A12), Lemma 3 and (A13) we obtain, respectively

$$\{T \in \mathcal{J}, \hat{T} \subset T\} \subseteq \bigcup_{k=1}^{t-1} \mathcal{E}_k\left(\frac{k\lambda^2}{2}\right) \subseteq \bigcup_{k=1}^{t-1} \mathcal{E}_k\left(\frac{b_k c\delta_k}{2}\right)$$
$$\subseteq \bigcup_{k=1}^{t-1} C_T\left(\frac{(1-b_k)^2 c^2 \delta_k}{4}\right) \subseteq C_T(ac\lambda^2)$$

∎

Lemma 5. *For $a \in (0,1)$ we have*

$$\{T \in \mathcal{J}, \hat{T} \supset T\} \subseteq C_T((1-a)c\lambda^2) \cup \bigcup_{J \supset T : r(X_J) = |J| \le \bar{t}} C_{J \ominus T}(|J \setminus T| ac\lambda^2). \tag{A15}$$

Proof. For J such that $J \supset T, r(X_J) = |J| \le \bar{t}$ define $W_J = X_J(\mathring{\beta}_J - \hat{\beta}_J^{\mathrm{ML}})$ and $m = |J \setminus T|$. Notice that the event $\{\ell_J - \ell_T \le -m\lambda^2/2\}$ can be decomposed as

$$\{\ell_J - \ell_T \le -m\lambda^2/2, \hat{\beta}_J^{\mathrm{ML}} \in B_J\} \cup \{\ell_J - \ell_T \le -m\lambda^2/2, \hat{\beta}_J^{\mathrm{ML}} \notin B_J\}, \tag{A16}$$

where $B_J = \{\beta_J : \|X_J(\mathring{\beta} - \beta_J)\|^2 \le \delta_{t-1}\}$. For $\hat{\beta}_J^{\mathrm{ML}} \in B_J$ we have from assumption (11) and properties of orthogonal projection H_J

$$\begin{aligned}
\ell_J - \ell_T &\ge \ell(\hat{\beta}_J^{\mathrm{ML}}) - \ell(\mathring{\beta}_J) \\
&\ge (\mathring{\beta}_J - \hat{\beta}_J^{\mathrm{ML}})^T X_J^T H_J \varepsilon + \frac{c}{2}(\mathring{\beta}_J - \hat{\beta}_J^{\mathrm{ML}})^T X_J^T X_J (\mathring{\beta}_J - \hat{\beta}_J^{\mathrm{ML}}) \\
&= W_J^T H_J \varepsilon + \frac{c}{2} W_J^T W_J = \frac{1}{2c}\|cW_J + H_J\varepsilon\|^2 - \frac{1}{2c}\varepsilon^T H_J \varepsilon \\
&\ge -\frac{1}{2c}\varepsilon^T H_J \varepsilon,
\end{aligned}$$

so $\{\ell_J - \ell_T \le -m\lambda^2/2, \hat{\beta}_J^{\mathrm{ML}} \in B_J\} \subseteq \{\varepsilon^T H_J \varepsilon \ge mc\lambda^2\}$. The second event in the sum (A16) is obviously contained in $\{\hat{\beta}_J^{\mathrm{ML}} \notin B_J\}$. Now, we show that this event is also contained in $\{\varepsilon^T H_J \varepsilon \ge mc\lambda^2\}$. To do it, we use argumentation similar to van de Geer (2002). Thus, we define

$$\begin{aligned}
d_J &= \|X_J(\hat{\beta}_J^{\mathrm{ML}} - \mathring{\beta})\|, \quad u = \frac{\sqrt{\delta_{t-1}}}{\sqrt{\delta_{t-1}} + d_J}, \\
\tilde{\beta}_J &= u\hat{\beta}_J^{\mathrm{ML}} + (1-u)\mathring{\beta}, \quad \tilde{d}_J = \|X_J(\tilde{\beta}_J - \mathring{\beta})\|.
\end{aligned}$$

The random vector $\tilde{\beta}_J$ belongs to B_J, because

$$\|X_J(\tilde{\beta}_J - \mathring{\beta})\| = u d_J \le \sqrt{\delta_{t-1}}.$$

Using convexity of the loss function ℓ we have

$$\ell(\tilde{\beta}_J) \le u\ell(\hat{\beta}_J^{\mathrm{ML}}) + (1-u)\ell(\mathring{\beta}) = u[\ell(\hat{\beta}_J^{\mathrm{ML}}) - \ell(\mathring{\beta})] + \ell(\mathring{\beta}) \le \ell(\mathring{\beta}).$$

Using this fact as well as assumption (11), Schwartz inequality and properties of the orthogonal projection H_J we get

$$\begin{aligned}
0 &\ge \ell(\tilde{\beta}_J) - \ell(\mathring{\beta}) \ge (\mathring{\beta}_J - \tilde{\beta}_J)^T X_J^T H_J \varepsilon + \frac{c}{2}(\mathring{\beta}_J - \tilde{\beta}_J)^T X_J^T X_J (\mathring{\beta}_J - \tilde{\beta}_J) \\
&\ge -\sqrt{\tilde{d}_J^2 \varepsilon^T H_J \varepsilon} + \frac{c}{2}\tilde{d}_J^2.
\end{aligned}$$

It gives us

$$\varepsilon^T H_J \varepsilon \ge \frac{c^2}{4}\tilde{d}_J^2.$$

Therefore, we obtain

$$\{\hat{\beta}_J^{\mathrm{ML}} \notin B_J\} = \{\tilde{d}_J > \sqrt{\delta_{t-1}}/2\} \subset \left\{\varepsilon^T H_J \varepsilon \ge \frac{c^2 \delta_{t-1}}{16}\right\}.$$

From assumptions of Theorem 2 we know that $mc\lambda^2 \le \frac{c^2\delta_{t-1}}{16}$, that gives us $\{\hat{\beta}_J^{\mathrm{ML}} \notin B_J\} \subset \{\varepsilon^T H_J \varepsilon \ge mc\lambda^2\}$.

On the other hand, $\varepsilon^T H_J \varepsilon = \varepsilon^T H_T \varepsilon + \varepsilon^T (H_J - H_T)\varepsilon$, hence we obtain for $\tau > 0$ and $a \in (0,1)$

$$\{\varepsilon^T H_J \varepsilon \geq \tau\} \subseteq \{\varepsilon^T H_T \varepsilon \geq (1-a)\tau\} \cup \{\varepsilon^T (H_J - H_T)\varepsilon \geq a\tau\}.$$

Finally

$$\begin{aligned}\{T \in \mathcal{J}, \hat{T} \supset T\} &\subseteq \bigcup_{J \supset T : r(X_J) = |J| \leq \bar{t}} \{\ell_J - \ell_T \leq -|J \setminus T|\lambda^2/2\} \\ &\subseteq C_T((1-a)c\lambda^2) \cup \bigcup_{J \supset T : r(X_J) = |J| \leq \bar{t}} C_{J \ominus T}(|J \setminus T| a c \lambda^2). \end{aligned} \tag{A17}$$

■

A.5 Proof of Theorem 3

The idea of the proof is similar to the proof of Theorem 2. However, here we consider the linear model with the subgaussian noise that allows us to apply simpler and better arguments than in the general case of GLM. Indeed, while working with the event $\{T \in \mathcal{J}, \hat{T} \subset T\}$ we can use the result for the scalar product (Lemma 2 (i)) instead of the one for quadratic forms (Lemma 2 (ii)). Besides, when considering $\{T \in \mathcal{J}, \hat{T} \supset T\}$ we need to upper bound probability of $\{\varepsilon^T H_J \varepsilon - \varepsilon^T H_T \varepsilon \geq |J \setminus T|\lambda^2\}$ instead of $\{\varepsilon^T H_J \varepsilon \geq |J \setminus T|\lambda^2\}$. Obviously, probability of the former event is smaller.

As previously, we work with the error decomposition (A5). Probability of the event $\{T \notin \mathcal{J}\}$ can be bounded in the same way as in the proof of Theorem 2, if we put $a_1 = \sqrt{a_2} = a$. However, working with subsets and supersets of T is different.

We start with subsets of T. For $J \subset T$ we denote

$$\mu_{T \ominus J} = (H_T - H_J) X_T \mathring{\beta}_T,$$

and

$$\delta_{T \ominus J} = \|\mu_{T \ominus J}\|^2.$$

Notice that $\ell_J + |J|\lambda^2/2 \leq \ell_T + t\lambda^2/2$ is equivalent to

$$\delta_{T \ominus J} + 2\varepsilon^T \mu_{T \ominus J} + \varepsilon^T (H_T - H_J)\varepsilon \leq (t - |J|)\lambda^2.$$

Whence

$$\begin{aligned}\left\{\ell_J + |J|\lambda^2/2 \leq \ell_T + t\lambda^2/2\right\} &\subseteq \left\{-2\varepsilon^T \mu_{T \ominus J} \geq \delta_{T \ominus J} - k\lambda^2\right\} \\ &\subseteq \left\{-\frac{2\varepsilon^T \mu_{T \ominus J}}{\|\mu_{T \ominus J}\|} \geq (k\delta)^{1/2}\left(1 - \frac{\lambda^2}{\delta}\right)\right\},\end{aligned}$$

where $k = t - |J|$ and $\delta_{T \ominus J} \geq \delta_k \geq k\delta$ by the definition of δ. Thus, from Lemma 2(i) we have

$$\begin{aligned}&P\left(\exists J \subset T : \ \ell_J + |J|\lambda^2/2 \leq \ell_T + t\lambda^2/2\right) \\ &\leq P\left(\exists J \subset T : \ -\frac{\varepsilon^T \mu_{T \ominus J}}{\|\mu_{T \ominus J}\|} \geq 2^{-1}(k\delta)^{1/2}\left(1 - \frac{\lambda^2}{\delta}\right)\right) \\ &\leq \sum_{k=1}^{t} \binom{t}{k} \exp\left(-\frac{k}{8\sigma^2}\delta\left(1 - \frac{\lambda^2}{\delta}\right)^2\right)\end{aligned}$$

$$\leq \sum_{k=1}^{t} \frac{1}{k!} \exp\left(-k\left(\frac{1}{8\sigma^2}\delta\left(1-\frac{\lambda^2}{\delta}\right)^2 - \log t\right)\right)$$
$$\leq \log(2)^{-1} \exp\left(-\frac{1}{8\sigma^2}\delta\left(1-\frac{\lambda^2}{\delta}\right)^2 + \log t\right)$$
$$\leq \log(2)^{-1} t \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right).$$

For the penultimate inequality we use again the inequality $\exp(c) - 1 \leq log(2)^{-1}c$ for $0 \leq c \leq log(2)$ and the fact that probability is not greater than 1. For the last inequality above we used

$$a^2\lambda^2 \leq \frac{\delta}{4}\left(1-\frac{\lambda^2}{\delta}\right)^2,$$

which is satisfied for

$$\lambda^2/\delta \leq f_1(a), \tag{A18}$$

where

$$f_1(a) = 1 + 2a^2 - \sqrt{(1+2a^2)^2 - 1}.$$

It is easy to check that (A18) follows from the assumption $\lambda^2 \leq \delta/(2+2a)^2$ as

$$f_1(a) = \frac{1}{1+2a^2+\sqrt{(1+2a^2)^2-1}} \geq \frac{1}{(1+\sqrt{2}a)^2} \geq \frac{1}{(2+2a)^2}.$$

Now, we consider supersets of T. As previously we consider only such supersets J that $r(X_J) = |J| \leq \bar{t}$. We have $\ell_T - \ell_J = \varepsilon^T(H_J - H_T)\varepsilon/2$. Using Lemma 2 (iii), as in the proof of Theorem 2, we obtain

$$P\left(\exists J \supset T : \ \ell_J + |J|\lambda^2/2 \leq \ell_T + t\lambda^2/2\right)$$
$$= P\left(\exists J \supset T : \ \varepsilon^T(H_J - H_T)\varepsilon \geq (|J| - t)\lambda^2\right)$$
$$\leq \log(2)^{-1} p \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right).$$

Finally, we get the following inequalities

$$P(\hat{T}_{SS} \neq T) \leq \left(2 + \frac{2}{\log(2)}\right) p \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right) \leq 5\exp\left(-\frac{a^2(1-a^2)\lambda^2}{2\sigma^2}\right). \quad \blacksquare$$

A.6 Proof of Theorem 4

In section 3 we state Theorem 4 with simplified constants K_i. Its general form is given below:

Fix $a_1, a_2 \in (0,1)$. Assume that (25), (26), (27) and

$$\max\left(\frac{4K_1^2\log(2p)}{a_1^2}, \frac{32K_3^2\log p}{K_4 c_2}, \frac{K_3^2 t}{a_2 c_2}, \frac{16K_3^2(t+1)}{c_2}\right) L^2 \leq \lambda^2 \leq . \tag{A19}$$

$$\leq \min\left[\frac{c_2\delta}{(1+\sqrt{2a_2})^2}, \frac{c_2\delta_{t-1}}{4(\bar{t}-t)}, \frac{(1-a_1)^2c_1^2\kappa_{a_1}^2\mathring{\beta}_{\min}^2}{16}\right]. \tag{A20}$$

Then

$$P\left(\hat{T}_{SS} \neq T\right) \leq 4.5\exp\left[-\frac{\lambda^2}{L^2}\min\left(\frac{a_1^2K_2}{4K_1^2}, \frac{K_4c_2}{32K_3^2}, \frac{a_2K_4c_2}{K_3^2}\right)\right].$$

The main difference between proofs of Theorems 2 and 4 is that here we investigate properties of the expected loss $\mathbb{E}\ell(\cdot)$ instead of the loss $\ell(\cdot)$. It relates to the fact that the former function is more regular. To be more precise, in many parts of the proof we work with expressions of the form

$$\ell(\tilde{\beta}) - \ell(\mathring{\beta}), \tag{A21}$$

where $\tilde{\beta}$ is a random vector contained in some ball $\tilde{B}$. Clearly, we can transform (A21) into

$$\left[\mathbb{E}\ell(\tilde{\beta}) - \mathbb{E}\ell(\mathring{\beta})\right] + \left[\ell(\tilde{\beta}) - \ell(\mathring{\beta}) - \mathbb{E}\ell(\tilde{\beta}) + \mathbb{E}\ell(\mathring{\beta})\right]. \tag{A22}$$

The first term in (A22) can be handled using the regularity of the function $\mathbb{E}\ell(\cdot)$ given in assumption (25), while to work with the latter we apply assumptions (26) or (27).

As previously, we work with the error decomposition (A5) and start with bounding probability of $\{T \not\in \mathcal{J}\}$. Take $a_1 \in (0,1)$ and define $r^* = \frac{4\lambda}{(1-a_1)c_1\kappa_{a_1}}$. From (A20) we have $r^* \leq \mathring{\beta}_{\min}$. Consider the following event

$$\Omega = \left\{Z(r^*) \leq \frac{a_1\lambda r^*}{2}\right\}, \tag{A23}$$

which has probability not less than

$$1 - \exp\left(-\frac{a_1^2K_2\lambda^2}{4K_1^2L^2}\right). \tag{A24}$$

Indeed, if we take $z = \frac{a_1\lambda}{2K_1L\sqrt{\log(2p)}}$ and $r = r^*$ in (26), then using again (A20) we obtain $z \geq 1$ and (A24) follows.

Similarly to the proof of (Bühlmann & van de Geer, 2011, theorem 6.4), we can show that the event (A23) implies that

$$|\hat{\beta} - \mathring{\beta}|_1 \leq r^*/2. \tag{A25}$$

From (A25) and (A20) we can obtain separability of the Lasso. Namely, for each $j \in T, k \not\in T$ that

$$|\hat{\beta}_j| \geq |\mathring{\beta}_j| - |\hat{\beta}_j - \mathring{\beta}_j| \geq \mathring{\beta}_{\min} - |\hat{\beta} - \mathring{\beta}|_1 \geq r^*/2 \geq |\hat{\beta} - \mathring{\beta}|_1 \geq |\hat{\beta}_k - \mathring{\beta}_k| = |\hat{\beta}_k|.$$

Now we consider probability that GIC chooses a submodel of T in the second step of the SS algorithm. Using the definition (A11) we obtain

$$\{T \in \mathcal{J}, \hat{T} \subset T\} \subset \bigcup_{k=1}^{t-1} \mathcal{E}_k(k\lambda^2/2).$$

Fix k and J such that $J \subset T$, $|T \setminus J| = k$. We take arbitrary $\beta_T \in \partial B_{2,T}\left(\sqrt{\delta_k}\right)$ and recall that $\mathring{\ell}(\beta) = \ell(\beta) - \ell(\mathring{\beta})$. It is clear that $\delta_{t-1} \geq \delta_k$, so we can use (25) for β_T and obtain

$$\begin{aligned}\ell(\beta_T) - \ell(\mathring{\beta}) &= \mathbb{E}\mathring{\ell}(\beta_T) + [\mathring{\ell}(\beta_T) - \mathbb{E}\mathring{\ell}(\beta_T)] \\ &\geq \frac{c_2}{2}(\mathring{\beta} - \beta_T)^T X_T^T X_T(\mathring{\beta} - \beta_T) - U_T\left(\sqrt{\delta_k}\right) \geq \frac{c_2}{2}\delta_k - U_T\left(\sqrt{\delta_k}\right).\end{aligned}$$

Since the last expression does not depend on β_T we have

$$\min_{\beta_T \in \partial B_{2,T}\left(\sqrt{\delta_k}\right)} \mathring{\ell}(\beta_T) \geq \frac{c_2}{2}\delta_k - U_T\left(\sqrt{\delta_k}\right).$$

Proceeding as in the proof of Lemma 3 we obtain

$$\mathcal{E}_k(k\lambda^2/2) \subset \left\{k\lambda^2/2 \geq \frac{c_2}{2}\delta_k - U_T\left(\sqrt{\delta_k}\right)\right\}.$$

If we take $b_k = \frac{k\lambda^2}{c_2\delta_k}$, then

$$\mathcal{E}_k(k\lambda^2/2) \subset \left\{U_T\left(\sqrt{\delta_k}\right) \geq \frac{c_2\delta_k}{2}(1 - b_k)\right\}.$$

From (A20) we have (as in the proof of Lemma 4)

$$a_2 b_k \leq (1 - b_k)^2/4. \tag{A26}$$

Take

$$z = \frac{(1 - b_k)c_2\sqrt{\delta_k}}{2K_3 L\sqrt{t}},$$

which is not smaller than one by (A26) and (A20). Thus, using (27) with $r = \sqrt{\delta_k}$ and (A26) we have

$$P\left(U_T\left(\sqrt{\delta_k}\right) \geq \frac{c_2\delta_k}{2}(1 - b_k)\right) \leq \exp\left(-\frac{K_4 a_2 k c_2 \lambda^2}{K_3^2 L^2}\right).$$

Therefore,

$$\begin{aligned}P(T \in \mathcal{J}, \hat{T} \subset T) &\leq \sum_{k=1}^{t-1} \exp\left(-k\frac{K_4 a_2 c_2 \lambda^2}{K_3^2 L^2}\right) \\ &\leq \left[\exp\left(\frac{K_4 a_2 c_2 \lambda^2}{K_3^2 L^2}\right) - 1\right]^{-1} \leq \frac{\exp(K_4 t)}{\exp(K_4 t) - 1}\exp\left(-\frac{K_4 a_2 c_2 \lambda^2}{K_3^2 L^2}\right),\end{aligned}$$

because $\lambda^2 \geq \frac{K_3^2 t L^2}{a_2 c_2}$ by (A19).

Next, we consider choosing a supermodel of T by GIC. Fix the set $J \supset T, r(X_J) = |J| \leq \bar{t}$ and denote $m = |J \setminus T|$. Therefore, we have

$$\begin{aligned}\{-m\lambda^2/2 \geq \ell(\hat{\beta}_J) - \ell(\hat{\beta}_T)\} &\subset \{-m\lambda^2/2 \geq \mathring{\ell}(\hat{\beta}_J)\} \\ &= \{-m\lambda^2/2 \geq \mathbb{E}\mathring{\ell}(\hat{\beta}_J) + [\mathring{\ell}(\hat{\beta}_J) - \mathbb{E}\mathring{\ell}(\hat{\beta}_J)]\} \\ &\subset \{-m\lambda^2/2 \geq \mathring{\ell}(\hat{\beta}_J) - \mathbb{E}\mathring{\ell}(\hat{\beta}_J)\} =: D_J,\end{aligned}$$

because $\ell(\hat{\beta}_T) \leq \ell(\mathring{\beta})$ and $\mathbb{E}\ell(\hat{\beta}_J) \geq \mathbb{E}\ell(\mathring{\beta})$. We will prove that D_J is contained in

$$\{U_J(r) > m\lambda^2/2\}, \tag{A27}$$

for $r = 2\lambda\sqrt{\frac{m}{c_2}}$. The event D_J can be decomposed as

$$D_J = \{D_J \cap [\hat{\beta}_J \in B_{2,J}(r)]\} \cup \{D_J \cap [\hat{\beta}_J \notin B_{2,J}(r)]\}. \tag{A28}$$

It is clear that the first event on the right-hand side of (A28) is contained in (A27) and the second one is contained in $\{\hat{\beta}_J \notin B_{2,J}(r)\}$. Now we prove that $\{\hat{\beta}_J \notin B_{2,J}(r)\}$ is also contained in (A27). Our argumentation is similar to the proof of Lemma 5, but here we take $u = \frac{r}{r+d_J}$. Therefore, we obtain

$$\mathbb{E}\mathring{\ell}(\tilde{\beta}_J) = \mathring{\ell}(\tilde{\beta}_J) - \mathring{\ell}(\tilde{\beta}_J) + \mathbb{E}\mathring{\ell}(\tilde{\beta}_J) \leq -\mathring{\ell}(\tilde{\beta}_J) + \mathbb{E}\mathring{\ell}(\tilde{\beta}_J) \leq U_J(r). \tag{A29}$$

Moreover, the following bound

$$\mathbb{E}\ell(\tilde{\beta}_J) - \mathbb{E}\ell(\mathring{\beta}) \geq \frac{c_2}{2}\tilde{d}_J^2, \tag{A30}$$

is implied by (25), because $\tilde{\beta}_J \in B_{2,J}(r) \subset B_J$. This inclusion comes from (A20), because

$$\delta_{t-1} \geq \frac{4\lambda^2(\bar{t} - t)}{c_2} \geq \frac{4\lambda^2 m}{c_2}.$$

Taking (A29) and (A30) we have

$$\{\hat{\beta}_J \notin B_{2,J}(r)\} = \{d_J^2 > r^2\} = \left\{\tilde{d}_J^2 > \frac{r^2}{4}\right\} \subset \left\{U_J(r) > \frac{c_2 r^2}{8}\right\},$$

and $\frac{c_2 r^2}{8} = m\lambda^2/2$ by the choice of r. Therefore, we have just proved that D_J is contained in (A27). To finish the proof we need sharp enough upper bound of (A27). It can be obtained using (27) with $z = \frac{\lambda\sqrt{mc_2}}{4K_3 L\sqrt{|J|}}$ that gives us

$$P(D_J) \leq P(U_J(r) > m\lambda^2/2) \leq \exp\left(-\frac{K_4 m\lambda^2 c_2}{16K_3^2 L^2}\right).$$

Notice that $z \geq 1$, because using (A20)

$$\lambda^2 \geq \frac{16K_3^2 L^2(t+1)}{c_2} \geq \frac{16K_3^2 L^2|J|}{c_2 m}.$$

Thus, we obtain the following bound on probability of choosing a supermodel

$$\begin{aligned} P(T \in \mathcal{J}, \hat{T} \supset T) &\leq \sum_{m=1}^{p-t} \binom{p-t}{m} \exp\left(-m\frac{K_4\lambda^2 c_2}{16K_3^2L^2}\right) \\ &\leq \sum_{m=1}^{p-t} \frac{1}{m!} \exp\left[-m\left(\frac{K_4\lambda^2 c_2}{16K_3^2L^2} - \log(p-t)\right)\right] \\ &\leq (\log 2)^{-1} \exp\left(-\frac{K_4\lambda^2 c_2}{32K_3^2L^2}\right), \end{aligned}$$

where we use the fact that $\frac{K_4\lambda^2 c_2}{16K_3^2L^2} \geq 2\log(p-t)$ from (A20) and two inequalities $\binom{p-t}{m} \leq \frac{(p-t)^m}{m!}$ and $\exp(b) - 1 \leq (\log 2)^{-1} b$ for $b \in (0, \log 2)$. ∎