Improving Lasso for Model Selection and Prediction

Piotr Pokarowski POKAR@MIMUW.EDU.PL
Agnieszka Prochenka A.PROCHENKA@PHD.IPIPAN.WAW.PL
Micha Frej M.FREJ@MIMUW.EDU.PL
Wojciech Rejchel WREJCHEL@GMAIL.COM
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
Jan Mielniczuk MIEL@IPIPAN.WAW.PL
Institute of Computer Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland

Abstract

The Lasso, that is l_1-penalized loss estimator is a popular tool for fitting sparse models to high-dimensional data. The concave regularizations SCAD or MCP approximate more closely l_0-penalized loss, that is the Generalized Information Criterion (GIC), and correct intrinsic estimation bias of the Lasso. In this paper we propose an alternative method of improving the Lasso for predictive models which encompass linear and logistic models as premier examples. The approach, for a given penalty, orders the absolute values of the Lasso non-zero coefficients and then selects the model from a small nested family by GIC. We derive an upper bound on the methods selection error and show in numerical experiments on synthetic and real-world data sets that an implementation of our algorithm is more accurate than implementations of studied concave regularizations.

Keywords: List of keywords

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 ℓ_1 penalty, is the main tool of fitting such models (Tibshirani, 2011; Bühlmann and van de Geer, 2011). However, a few years ago 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 and Bühlmann, 2006; Zhao and Yu, 2006; Shen et al., 2012) which is too restrictive in general. Model's dimension can be reduced without loss of quality using the Thresholded Lasso (TL) algorithm, which selects variables with largest absolute values of the Lasso coefficients (Ye and Zhang, 2010; Zhou, 2009) or by solving a more computationally demanding minimization of a loss with a folded concave penalty (FCP) as SCAD (Fan and Li, 2001), MCP (Zhang, 2010a) or capped l_1-penalty (Zhang, 2010b; Shen et al., 2012). TL, FCP and similar methods lead to consistent selection under weaker assumptions such as the *restricted isometry property* (Zhang, 2010a,b; Zhang and Zhang, 2012; Shen et al., 2012; Wang et al., 2013; Fan et al., 2014; Wang et al., 2014). Recently in Pokarowski and Mielniczuk (2015) an algorithm called *Screening–Ordering–Selection*

© P. Pokarowski, A. Prochenka, M. Frej, W. Rejchel & J. Mielniczuk.

(SOS) was introduced for linear model selection, which reduces the model selected by the Lasso. SOS is based on the variant of TL proposed by Zhou (2009).

The first main contribution of the paper is that we simplify and generalize SOS in the two step *Screening–Selection* (SS) algorithm for general predictive models e.g. linear normal, logistic or Cox proportional hazard models. In the first screening step, one computes the Lasso estimator $\hat{\beta}$ with penalty λ and orders its nonzero coefficients according to their decreasing absolute values. In the second, selection step, one chooses the model which minimizes GIC with penalty $\lambda^2/2$ in a nested family induced by the ordering. Post-model selection estimator of β is the minimum loss estimator for the chosen model. Thus the SS algorithm (Algorithm 1 below) is the Lasso with adaptive thresholding based on GIC. We derive an exponential upper bound on selection error of SS in terms of λ (Theorem 2), which parallels the known bounds for TL, see Theorem 8 for linear models in Ye and Zhang (2010) or FCP, see Corollary 3 and 5 in Fan et al. (2014). However, in contrast to these methods, SS is constructive for linear models in that it relies neither on the unknown parameters as the true vector $\mathring{\beta}$ or the cone invertibility factors. Instead, λ only depends on the sample size, the number of predictors and an upper bound on the noise parameter.

Although TL, FCP, SOS 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 (algorithm and R package `LARS` described in (Efron et al., 2004) or for a given net of them (R package `glmnet` described in Friedman et al. (2010). Similarly, using a net of penalty values, the FCP algorithm has been implemented for linear (in R package `SparseNet` Mazumder et al. (2011)) and logistic models (in R package `cvplogistic` Jiang and Huang (2014)).

Our second main contribution is that we propose the SOSnet algorithm (Algorithm 2 below), which is a generalization of the SOS algorithm for general predictive models. SOSnet uses `glmnet` for a net of penalty values, then for each of them it orders the chosen predictors according to Wald statistics and finally, it selects the model from a small family by minimizing GIC. We show in numerical experiments for linear and logistic models on synthetic and real-world data sets that SOSnet is more accurate than implementations of FCP.

2. Models and Fitting Algorithms

In this section we start with definitions of considered models and estimation criteria, next we present model selection algorithms.

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 selection error without obscuring their essentiality. We consider independent data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, where $y_i \in \mathbb{R},\ x_i \in \mathbb{R}^p$ for $i = 1, 2, \ldots, n$, a known differentiable *cumulant* function $\gamma : \mathbb{R} \to \mathbb{R}$ and we assume that for some *true* $\mathring{\beta} \in \mathbb{R}^p$

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

Note that (1) is satisfied in particular by the Generalized Linear Models (GLM) and a nonlinear regression with an additive error. We will use also vectorized version of the cumulant and its

derivative. For $\eta = (\eta_1, \ldots, \eta_n)^T$ we define $\boldsymbol{\gamma}(\eta) = (\gamma(\eta_1), \ldots, \gamma(\eta_n))^T$ and similarly $\dot{\boldsymbol{\gamma}}(\eta) = (\dot{\gamma}(\eta_1), \ldots, \dot{\gamma}(\eta_n))^T$.

Let $X = [x_{.1}, \ldots, x_{.p}] = [x_1, \ldots, x_n]^T$ be a $n \times p$ matrix of experiment and $J \subseteq \{1, 2, \ldots, p\} = F$ be an arbitrary subset of the *full model* F, $\bar{J} = F \setminus J$. As J may be viewed as 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 a rank 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 $T = \operatorname{supp}(\mathring{\beta}) = \{j \in F : \mathring{\beta}_j \neq 0\}$ and $t = |T|$.

We assume also that a *total cumulant* function $g(\beta) = \sum_{i=1}^n \gamma(x_i^T \beta)$ is convex and, additionally, *strongly convex* at $\mathring{\beta}$ in a sense that exists $c \in (0, 1]$ such that for all *sparse* $\beta \in \mathbb{B} \equiv \mathbb{B}(X, \mathring{\beta}, \bar{t})$ we have

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

where $t \leq \bar{t} < n \wedge p$,

$$\mathbb{B} = \bigcup_{J \supset T, r(X_J) = |J| \leq \bar{t}} \{\beta_J : ||X_T(\mathring{\beta}_J - \beta_J)||^2 \leq \delta_{t-1}\}, \tag{3}$$

$$\delta_{t-1} = \min_{j \in T} ||X_T \mathring{\beta}_T - x_{.j} \mathring{\beta}_j||^2. \tag{4}$$

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 $X^T X$. Let us remark that $\dot{g}(\beta) = X^T \dot{\boldsymbol{\gamma}}(X\beta)$.

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

$$\mathbb{E} \exp(u\varepsilon_i) \leq \exp(\sigma^2 u^2 / 2). \tag{5}$$

Examples. For a normal linear model $y_i \sim N(x_i^T \mathring{\beta}, \mathring{\sigma}^2)$, $\gamma(\eta_i) = \eta_i^2/2$ and (5) is fulfilled with any $\sigma \geq \mathring{\sigma}$. For a logistic model $y_i \sim \text{binom}(1, [1 + \exp(-x_i^T \mathring{\beta})]^{-1})$, $\gamma(\eta_i) = \log(1 + \exp(\eta_i))$ and as (ε_i) are bounded random variables, then (5) is satisfied with any $\sigma \geq 1/2$. It is easy to note that for a linear model the strong convexity assumption (2) is fulfilled with $c = 1$ and for a logistic model with

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

2.2. Fitting Algorithms

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

$$\ell(\beta) = \sum_{i=1}^n (\gamma(x_i^T \beta) - y_i x_i^T \beta) = g(\beta) - \beta^T X^T y, \tag{6}$$

where $y = (y_1, \ldots, y_n)^T$. It is easy to see that $\dot{\ell}(\beta) = X^T(\dot{\boldsymbol{\gamma}}(X\beta) - y)$, and consequently $\dot{\ell}(\mathring{\beta}) = -X^T \varepsilon$ for $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Moreover, observe that $\mathring{\beta} = \operatorname{argmin}_\beta \mathbb{E}\ell(\beta)$ and (2) is equivalent to strong convexity of ℓ in $\mathring{\beta}$ for all $\beta \in \mathbb{B}$

$$\ell(\beta) \geq \ell(\mathring{\beta}) + (\mathring{\beta} - \beta)^T X^T \varepsilon + \frac{c}{2}(\mathring{\beta} - \beta)^T X^T X (\mathring{\beta} - \beta). \tag{7}$$

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

Let $\widehat{\beta}_J^{ML} = \text{argmin}_{\beta_J} \ell(\beta_J)$ denotes a *minimum loss estimator* based on y and $\{x_{\cdot j}, j \in J\}$ and $\ell_J = \ell(\widehat{\beta}_J^{ML})$. Note that for GLM $\widehat{\beta}_J^{ML}$ coincides with a maximum likelihood estimator for a model pertaining to J. Finally, for $\beta \in \mathbb{R}^p$ and $q \geq 1$ let $|\beta|_q = (\sum_{j=1}^p |\beta_j|^q)^{1/q}$ be ℓ_q norm.

In Algorithm 1 we present two selection procedures: the first one is the Thresholded Lasso (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, named Screening–Selection (SS) procedure finds minimal value of Generalized Information Criterion (GIC) for the nested family which is constructed using ordering of the nonzero Lasso estimates.

3. Selection Error Bounds for TL and SS

In this section we present upper exponential bounds on the selection error of the TL and SS algorithms. 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 $\mathring{\beta}_{min} = \min_{j\in T} |\mathring{\beta}_j|$.

3.1. A Bound for TL

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

$$\mathcal{C}_a = \left\{\nu \in \mathbb{R}^p : |\nu_{\bar{T}}|_1 \leq \frac{1+a}{1-a}|\nu_T|_1,\, \nu_j x_{\cdot j}^T\left[\dot{\boldsymbol{\gamma}}(X(\mathring{\beta}+\nu)) - \dot{\boldsymbol{\gamma}}(X\mathring{\beta})\right] \leq 0,\, j \in \bar{T}\right\}. \quad (8)$$

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

$$\zeta_{a,q} = \inf_{\nu\in\mathcal{C}_a} \frac{\left|X^T\left[\dot{\boldsymbol{\gamma}}(X(\mathring{\beta}+\nu)) - \dot{\boldsymbol{\gamma}}(X\mathring{\beta})\right]\right|_\infty}{|\nu|_q} \quad (9)$$

We let $\zeta_a = \zeta_{a,\infty}$. In comparison to more popular restricted eigenvalues or compatibility constants, variants of SCIF enable sharper ℓ_q estimation error bounds of the Lasso for $q > 2$ (cf Corollary 4 in Ye and Zhang (2010) and Huang and Zhang (2012); Zhang and Zhang (2012).

The following lemma is a generalization of Theorem 3 in Ye and Zhang (2010).

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

Theorem 2 *If ℓ is convex, $(\varepsilon_i)_i$ are subgaussian with σ and for $a_1, a_2 \in (0,1)$ 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

$$\mathbb{P}(\hat T_{TL} \ne T) \le 2\exp\Big(-\frac{(1-a_2)a_1^2\lambda^2}{2\sigma^2}\Big). \tag{10}$$

Constant a_2 is used to remove multiplicative factor p from the exponential bound at the expense of slightly diminishing the exponent in (10). Note that assumptions of Theorem2 stipulate that truncation level τ is contained in interval $[(1+a_1)\lambda\zeta_{a_1}^{-1}, \mathring\beta_{min}/2)$. Analogous theorem for FCP, see Corollary 3 for linear models and Corollary 5 for logistic models in Fan et al. (2014), requires an additional assumption on on the minimal eigenvalue of $X_T^TX_T$ and the proof is more difficult, but for both methods a condition on τ requires unknown ζ_a or $\mathring\beta_{min}$.

3.2. A Bound for SS

Let H_J be an orthogonal projection matrix onto the subspace spanned by columns of X_J. A scaled K-L distance between T and its submodels, see Shen et al. (2012, 2013) is

$$\delta = \min_{J\subset T} \frac{||(I-H_J)X_T\mathring\beta_T||^2}{|T\setminus J|}. \tag{11}$$

Different variants of the K–L distance have been often used in the consistency analysis of selection algorithms, cf Section 3.1 in Pokarowski and Mielniczuk (2015), but δ defined above seems to lead to optimal results, cf Theorem 1 in Shen et al. (2013). Technical constants $a_1, \dots, a_4$ allow to avoid *ad-hoc* coefficients in the bounds 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 $a_4 = \sqrt{a_1a_2}$. Note that a_2, a_3 and a_4 are functions of a_1 and obviously if $a_1 \to 1$, then $a_2, a_3, a_4 \to 1$.

Theorem 3 *Assume (1)–(5) and that for $a_1 \in (1/2, 1)$*

$$\frac{2\sigma^2\log p}{a_3a_2a_1c} \vee \frac{\sigma^2 t}{(1-a_1)^2c} \le \lambda^2 < \frac{\delta c}{(1+\sqrt{2(1-a_1)})^2} \wedge \frac{\zeta_{a_4}^2\mathring\beta_{min}^2}{4(1+a_4)^2}. \tag{12}$$

Then

$$\mathbb{P}(\hat T_{SS} \ne T) \le 4.5\exp\Big(-\frac{a_2(1-a_1)c\lambda^2}{2\sigma^2}\Big). \tag{13}$$

Selection consistency that is asymptotic correctness of $\hat T_{SS}$ now easily follows.

Corollary 4 *Assume that $t = o(\log p)$ for $n \to \infty$ and set $a_1 = 1 - \sqrt{\frac{t}{2\log p}}$, $\lambda^2 = \frac{2\sigma^2\log p}{a_3a_2a_1c}$. Then $\lambda^2 = c^{-1}(2\sigma^2\log p)(1+o(1))$. If additionally $\mathring\beta$ is asymptotically identifiable:*

$$\overline{\lim_n} \frac{2\sigma^2\log p}{(c^2\delta)\wedge(\zeta_{a_4}^2\mathring\beta_{min}^2/16)} < 1, \quad \textit{then} \quad \mathbb{P}(\hat T_{SS} \ne T) = o(1).$$

Remarks. 1. Theorem 3 for linear models may be obtained directly, analogously as in Pokarowski and Mielniczuk (2015). The resulting lower bound on λ^2 is $\lambda^2 \geq 2\sigma^2 \log p/(a_1 a_2 a_3)$ without the additional condition $\lambda^2 \geq \sigma^2 t/(1-a_1)^2$ assumed in Theorem 3 (recall that $c = 1$).

2. Theorem 2 determines the interval of admissible λ, a parameter of SS, for which a bound in (13) holds. Corollary 1 states more easily interpretable result: for λ equal to lower endpoint of the above interval SS is asymptotically correct provided that the true model is asymptotically identifiable. Although identifiability condition is not effectively verifiable, λ can be explicitly given for linear models as

$$\lambda = \sqrt{2\sigma^2 \log p}(1 + o(1)) \tag{14}$$

and for logistic models as

$$\lambda = \sqrt{(\log p)/(2c)}(1 + o(1)), \tag{15}$$

since $\sigma \geq 1/2$. Thus for linear models a parameter of SS is given constructively in contrast to TL or FCP which require an additional parameter τ, depending on identifiability constants as SCIF. In literature concerning Lasso and its modifications the smallest possible λ is taken as the default value as then the algorithm is 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* in the Conclusions.

4. Extension to general convex contrasts

In this part of the paper we investigate properties of the SS algorithm beyond GLM. In fact, 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 (succesfully) applied 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 standard 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 before. 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, for instance, 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 the logistic regression that belongs to GLM and has been considered ealier. 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 support vector machines that use, for instance, the following 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 in the above paragraph satisfy this property. Notice that they need not be differentiable nor decompose, as in (6) for GLM, into the sum of the nonrandom cumulant γ and the random linear term $y_i \beta^T x_i$. In this section we prove the exponential upper bound of the error of the SS algorithm based on these contrast functions that is almost the same as in Theorem 3. To do it we strongly exploit the empirical process theory. Obviously, the SS algorithm 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 $B_1(r) = \{\beta : |\beta - \mathring{\beta}|_1 \leq r\}$ with radius $r > 0$. The

second one is the l_2-ball $B_{2,J}(r) = \{\beta_J : ||X_J(\mathring{\beta} - \beta_J)||^2 \leq r^2\}$ with radius $r > 0$, where J is a (sparse) subset of $\{1, \ldots, p\}$ and $T \subset J$. Recall that $\mathring{\beta}$ is, as previously, a minimizer $\mathbb{E}\ell(\beta)$. Besides, let $B_J = B_J\left(\sqrt{\delta_{t-1}}\right)$, where δ_{t-1} is defined in (4). 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 an analog of SCIF defined in (9). Namely, for arbitrary $a \in (0, 1)$ a compatibility factor is

$$\kappa_a = \inf_{0 \neq \beta \in \mathcal{C}_a} \frac{\beta^T X^T X \beta}{|\beta_T|_1^2}, \tag{16}$$

where $\mathcal{C}_a$ is a simplified version of (8), namely

$$\mathcal{C}_a = \left\{\nu \in \mathbb{R}^p : |\nu_{\bar{T}}|_1 \leq \frac{1+a}{1-a}|\nu_T|_1\right\}.$$

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 $\beta_1 \in B_1(\mathring{\beta}_{min})$ ($\beta_2 \in B$, respectively) we have for $i = 1, 2$

$$\mathbb{E}\ell(\beta_i) - \mathbb{E}\ell(\mathring{\beta}) \geq \frac{c_i}{2}(\beta_i - \mathring{\beta})^T X^T X (\beta_i - \mathring{\beta}). \tag{17}$$

Notice that we require the expected loss $\mathbb{E}\ell(\beta)$, not the loss $\ell(\beta)$, to be strongly convex. Besides, while considering GLM the condition (7) is equivalent to (17) for $i = 2$. Finally, to prove exponential bounds for GLM we use subgaussianity that allows as to obtain probabilistic inequalities in Lemma 10. In this section we need the analog of (23) 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

$$P\left(\frac{Z(r)}{r} > K_1 L z\sqrt{\log(2p)}\right) \leq \exp\left(-K_2 \log(2p)\, z^2\right). \tag{18}$$

Besides, the inequality (24) is replaced by the following: there exists $L > 0$ and constants $K_1, K_2 > 0$ such that for each $0 < r \leq \sqrt{\delta_{t-1}}$, $z \geq 1$ and $J : T \subset J, r(X_J) = |J| \leq \bar{t}$ we have

$$P\left(\frac{U_J(r)}{r} > K_1 L z\sqrt{|J|}\right) \leq \exp(-K_2|J|z^2). \tag{19}$$

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

Theorem 5 *Fix $a \in (0, 1)$. Assume that (17), (18), (19) and*

$$K_1 \max(\log p, t/a)\, \frac{L^2}{c_2} \leq \lambda^2 \leq K_2 \min\left[\frac{c_2\delta}{(1+\sqrt{2a})^2}, \frac{c_2\delta_{t-1}}{\bar{t}-t}, \left(c_1\kappa_a\mathring{\beta}_{min}\right)^2\right]. \tag{20}$$

Then

$$P\left(\hat{T}_{SS} \neq T\right) \leq K_1 \exp\left(-K_2 \frac{ac_2\lambda^2}{L^2}\right). \tag{21}$$

Theorem 5 bounds exponentially the error of the SS algorithm in the case that the contrast function is quite general convex function. It is similar (in fact, slightly worse) than Theorem 3 that considers GLM. Below we discuss its assumptions and compare it in detail with Theorem 3.

Remark 6 . *The main assumptions of Theorem 5 are convexity of the contrast function and conditions (18) and (19). They can be proved using tools from the empirical process theory such that concentration inequalities (Massart, 2000), the Symmetrization (van der Vaart and Wellner, 1996, Lemma 2.3.1) and Contraction Lemma (Ledoux and Talagrand, 1991, Theorem 4.12). It is quite remarkable that to get (18) or (19) 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 $x, y, 0 < r \leq \mathring{\beta}_{min}$ and $\beta, \tilde{\beta} \in B_1(r)$*

$$|\phi(\beta^T x, y) - \phi(\tilde{\beta}^T x, y)| \leq L|\beta^T x - \tilde{\beta}^T x|. \tag{22}$$

This fact follows from Massart (2000, Theorem 9) and Bühlmann and van de Geer (2011, Lemma 14.20). We show in the appendix that to get (19) we need similar argumentation and (22) to be satisfied for all $x, y, r \in \left(0, \sqrt{\delta_{t-1}}\right), J : T \subset J, r(X_J) = |J| \leq \bar{t}$ and $\beta, \tilde{\beta} \in B_{2,J}(r)$. Notice that logistic and absolute contrast functions satisfy (22) with $L = 2$ and $L = 1$, respectively. Moreover, every convex function is locally Lipschitz, so (22) is also satisfied for remaining contrasts (but in these cases L depends on n).

Remark 7 *The condition (17) is often called the "margin condition" in the literature. For quadratic and logistic contrasts it has been considered ealier. To prove it for SVM contrasts one can use methods based on the modulus of convexity (Bartlett et al., 2006, Lemma 7).*

Remark 8 (Comparison to Theorem 3 - Similarities) *As we have already mentioned Theorem 5 can be also 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. Therefore, the condition (17) for l_2-balls is the same as (7). Besides, we have for $\varepsilon = y - \mathbb{E}y$ that

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

Therefore, we can calculate that for GLM we have $Z(r)/r \leq \left|X^T\varepsilon\right|_\infty$ and $U_J(r)/r \leq \sqrt{\varepsilon^T H_J \varepsilon}$, that simplifies bounds in (18) and (19). It makes them analogous to those in Lemma 10 (i) and (iii). Previously they have been obtained using properties of errors $\varepsilon_1, \ldots, \varepsilon_n$. Now they are proved using the empirical process theory in more general framework. Lemma 10 (i) and (iii) are intended to GLM, so they should give better results than the general methods. And they do, but only in relation to constants K_i. Finally, notice that the left-hand side in condition (20) and the result in (21) are the same (again with respect to the constants) to those in Theorem 3.

Remark 9 (Comparison to Theorem 3 - Differences) *The SS algorithm consists of two steps. In the last paragraph we have mentioned that the theoretical analysis of the second step (selection) is similar in GLM and models with convex contrasts. However, we can find differences while investigating the first step (screening based on the lasso). In general, it is related to the fact that the properties of the lasso in GLM (so we work with differentiable contrasts that can be nicely decomposed) are better and/or better studied. In Theorem 5 we assume (17) with also respect to l_1-balls, which makes the right-hand side of (20) usually worse than in (12), because c_1^2 appears in (20).*

Summarizing, using the empirical process theory we are able to prove that the SS algorithm works in a satisfactory way even in quite general models with convex contrasts.

5. Proofs

In the following subsections we present auxiliary exponential inequalities for subgaussian random variables, upper-bound selection error of TL, and two parts of selection error of SS.

5.1. Exponential Bounds for Subgaussian Vectors

In the following lemma we develop auxiliary probabilistic tools. Specifically, in lemma 10 (iii) we generalize Wallace inequality for χ^2 distribution Wallace (1959) to the subgaussian case using the inequality for the moment generating function in lemma 10 (ii). The last inequality is proved by the decoupling technique as in the proof of Theorem 2.1 in Hsu et al. (2012).

Lemma 10 *Let $\varepsilon \in \mathbb{R}^n$ be a vector of zero-mean independent errors having subgaussian distribution with a constant σ, $\nu \in \mathbb{R}^n$, $0 < a < 1$ and H be a orthogonal projection such that $tr(H) = m$. Then*

(i) for $\tau > 0$

$$\mathbb{P}(\varepsilon^T \nu / ||\nu|| \geq \tau) \leq \exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{23}$$

(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 $\tau > 1$

$$\mathbb{P}(\varepsilon^T H \varepsilon \geq m\sigma^2\tau) \leq \exp\left(-\frac{m}{2}(\tau - 1 - \log\tau)\right) \tag{24}$$

Proof [Proof of Lemma 10] Let $Z = \varepsilon^T \nu / ||\nu||$ and $a > 0$. From Markov 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

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

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

From Markov inequality and part (ii) of the Lemma we have

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

Thus after minimization the last expression w.r.t a we obtain part (iii). ■

5.2. Proof of Lemma 1

Let $\mathcal{A}_a = \{|X^T\varepsilon|_\infty \leq a\lambda\}$ and $\hat{\varepsilon} = y - \dot{\boldsymbol{\gamma}}(X\hat{\beta})$. We have $\dot{\ell}(\hat{\beta}) = -X^T\hat{\varepsilon}$ and from the Karush-Kuhn-Tucker (KKT) theorem we obtain equations

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

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

$$\begin{aligned} D(\nu) &= \nu^T X^T\big[\dot{\boldsymbol{\gamma}}(X\hat{\beta}) - \dot{\boldsymbol{\gamma}}(X\mathring{\beta})\big] = \nu_T^T X_T^T(\varepsilon - \hat{\varepsilon}) + \nu_{\bar{T}}^T X_{\bar{T}}^T(\varepsilon - \hat{\varepsilon}) \\ &\leq |\nu_T|_1(|X_T^T\varepsilon|_\infty + |X_T^T\hat{\varepsilon}|_\infty) + |\nu_{\bar{T}}|_1(|X_{\bar{T}}^T\varepsilon|_\infty - \lambda) \leq |\nu_T|_1(1+a)\lambda + |\nu_{\bar{T}}|_1(a-1)\lambda. \quad (25) \end{aligned}$$

Then letting $\nu = \Delta_J$ for $J \subseteq \bar{T}$ 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 and Zhang, 2012). Hence $(1-a)|\nu_{\bar{T}}|_1 \leq (1+a)|\nu_T|_1$. Thus, on $\mathcal{A}_a$, $\Delta \in \mathcal{C}_a$ and from the definition of ζ_a we obtain using KKT again

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

5.3. Proof of Theorem 2

First we will prove that $\mathcal{A}_a \subseteq \{\hat{T}_{TL} = T\}$. From Lemma 1 and assumptions we have on $\mathcal{A}_a$

$$|\Delta|_\infty \leq (1+a)\lambda\zeta_a^{-1} \leq \tau < \mathring{\beta}_{min}/2. \quad (26)$$

Thus using (26) twice we have for $j \in T$ and $k \notin T$

$$|\hat{\beta}_j| \geq |\mathring{\beta}_j| - |\hat{\beta}_j - \mathring{\beta}_j| > \mathring{\beta}_{min} - \mathring{\beta}_{min}/2 > \tau \geq |\hat{\beta}_k - \mathring{\beta}_k| = |\hat{\beta}_k| \quad (27)$$

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

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

Hence, using Lemma 10 (i) we easily obtain

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

5.4. Proof of Theorem 3

Let us observe that the consecutive steps of SS constitute decomposition of the selection error into 2 parts: $\{\hat{T} \neq T\} = \{T \notin \mathcal{J}\} \cup \{T \in \mathcal{J}, \hat{T} \neq T\}$. Therefore Theorem 3 follows easily from (28) and (30) 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 Theorem 2 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

$$\mathbb{P}\big(T \notin \mathcal{J}\big) \leq \mathbb{P}(\mathcal{A}_{a_4}^c) \leq 2p \exp\big(-\frac{a_4^2\lambda^2}{2\sigma^2}\big) \leq 2\exp\big(-\frac{a_2(1 - a_1)c\lambda^2}{2\sigma^2}\big). \tag{28}$$

Now we bound probability $\mathbb{P}\big(T \in \mathcal{J}, \hat{T} \neq T\big)$. Components of the selection error set are included in the critical sets of the following form $\mathcal{C}_J(\tau) = \{\varepsilon^T H_J \varepsilon \geq \tau\}$.

Proofs of the lemmas stated below are relegated to the Supplemental Materials.

Lemma 11 *If for $a \in (0, 1)$ we have $\lambda^2 < c\delta/\big(1 + \sqrt{2a}\big)^2$, then $\{T \in \mathcal{J}, \hat{T} \subset T\} \subseteq \mathcal{C}_T(ac\lambda^2)$.*

Lemma 12 *For $a \in (0, 1)$ we have*

$$\{T \in \mathcal{J}, \hat{T} \supset T\} \subseteq \mathcal{C}_T((1 - a)c\lambda^2) \cup \bigcup_{J \supset T} \mathcal{C}_{J \setminus T}(|J \setminus T|ac\lambda^2).$$

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

$$a_2\tau_r < \tau_r - 1 - \log\tau_r \text{ for } r = 1, 2. \tag{29}$$

From Lemma 11, Lemma 12, Lemma 10 (iii) and (29) we get

$$\mathbb{P}\big(T \in \mathcal{J}, \hat{T} \neq T\big) \leq \mathbb{P}\big(\mathcal{C}_T(t\sigma^2\tau_1)\big) + \sum_{J \supset T} \mathbb{P}\big(\mathcal{C}_{J \setminus T}(|J \setminus T|\sigma^2\tau_2)\big) \leq \exp\big(-ta_2\tau_1/2\big)$$
$$+ \sum_{m=1}^{p-t} \binom{p-t}{m} \exp\big(-ma_2\tau_2/2\big) \leq \exp\big(-ta_2\tau_1/2\big) + \sum_{m=1}^{p-t} \frac{1}{m!} \exp\big(\frac{-m}{2}(a_2\tau_2 - 2\log p)\big).$$

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

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

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

$$\mathbb{P}\big(T \in \mathcal{J}, \hat{T} \neq T\big) \leq \exp\big(-ta_2\tau_1/2\big) + (\log 2)^{-1} \exp\big(-(1 - a_3)a_2\tau_2/2\big). \tag{30}$$

6. Experiments

While convenient for theoretical analysis TL, FCP, SOS 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 it (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 (in R package `SparseNet` Mazumder et al. (2011)) and logistic models (in R package `cvplogistic` Jiang and Huang (2014)).

In order to improve SOS and SS performance, we propose a net modification of SOS called SOSnet (Algorithm 2 below), which employs estimates for m values: $\lambda_1, \ldots, \lambda_m$. This alteration results in higher accuracy of model selection and sparse prediction as shown in our experiments. The most intensive step of the algorithm, namely computations of Lasso estimators, can be performed for all λ values during one run of the `glmnet` algorithm. As in SOS the ordering step allows to choose T when the screening step is correct i.e. $\text{supp}\hat{\beta}_\lambda \supseteq T$ but ordering given by absolute values of coordinates of $\hat{\beta}_\lambda$ is wrong. An additional loop (for $l = 1$ series to o) is introduced in order to find a correct screening set having possibly small cardinality.

Algorithm 2 SOSnet

Input: y, X and $(o, \lambda \leq \lambda_1 < \ldots < \lambda_m)$
Screening (Lasso)
for $k = 1$ **to** m **do**
  $\widehat{\beta}^{(k)} = \text{argmin}_\beta \{\ell(\beta) + \lambda_k|\beta|_1\}$; order nonzero $|\widehat{\beta}^{(k)}_{j_1}| \geq \ldots \geq |\widehat{\beta}^{(k)}_{j_{s_k}}|$, where $s_k = |\text{supp}\widehat{\beta}^{(k)}|$;
  Ordering (squared Wald tests)
  for $l = 1$ **to** o **do**
    set $J = \{j_1, j_2, \ldots, j_{s_{kl}}\}$, where $s_{kl} = \lfloor \frac{s_k \cdot l}{o} \rfloor$;
    compute $\widehat{\beta}^{ML}_J$;
    set predictors in J according to squared Wald tests: $w^2_{i_1} \geq w^2_{i_2} \geq \ldots \geq w^2_{i_{s_{kl}}}$;
    set $\mathcal{J}_{kl} = \{\{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, i_2, \ldots, i_{s_{kl}}\}\}$
  end for;
end for;
Selection (GIC)
$\mathcal{J} = \bigcup_{k=1}^m \bigcup_{l=1}^o \mathcal{J}_{kl}; \widehat{T} = \text{argmin}_{J \in \mathcal{J}} \left\{\ell_J + \lambda^2/2|J|\right\}$
Output: $\widehat{T}, \widehat{\beta}^{GSSG} = \widehat{\beta}^{ML}_{\widehat{T}}$

We performed numerical experiments fitting sparse linear and logistic models to high-dimensional benchmark simulations and real data sets.

6.1. Simulated Data

For linear models we studied the performance of two algorithms: SOSnet and MCP computed using the R package `SparseNet` Mazumder et al. (2011) for the default 9 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 and with $o = 5$.

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

We generated samples (y_i, x_i), $i = 1, \ldots, n$ from the normal linear model. Two vectors of parameters were considered: $\mathring{\beta}^{(1)} = (3, 1.5, 0, 0, 2, 0_{p-5}^T)^T$, as in Wang et al. (2013) as well as $\mathring{\beta}^{(2)} = (0_{p-10}^T, s_1 \cdot 2, s_2 \cdot 2, \ldots, s_{10} \cdot 2)^T$, 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 centred and normalized so that $||x_{\cdot j}||^2 = n$ 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. We reported mean model dimension (MD) that is $|\text{supp}(\hat{\beta})|$ and mean squared prediction error (PE) on new data set with 1000 observations equalling $\|X\mathring{\beta} - X\widehat{\beta}\|^2/(n\sigma^2)$. We chose the model using GIC with $\lambda^2 = c \cdot \log(p) \cdot \sigma^2$. For each value of hyperparameter $c = .25, .5, \ldots, 7.5$ values of $(MD(c), PE(c))$ for the models chosen by GIC=GIC(c) 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 $c = 2.5$: the black one for SOSnet and the red one for `SparseNet`. The blue vertical line denotes the true model dimension.

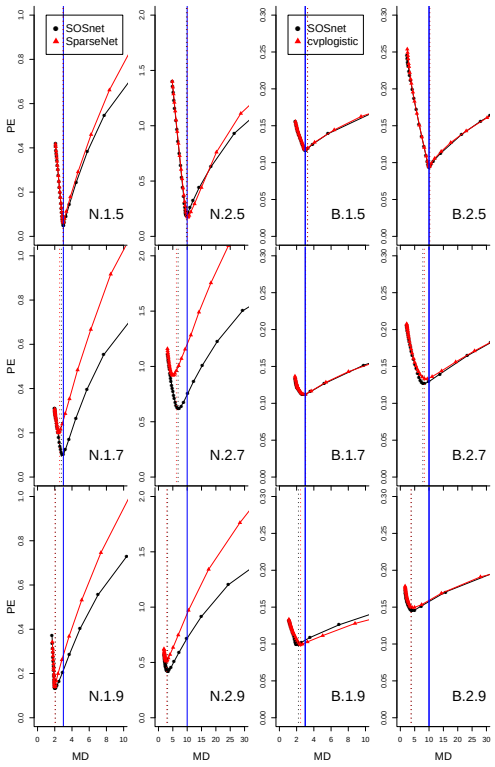
For logistic models we compared the performance of two algorithms: SOSnet and MCP implemented in the R package `cvplogistic` for the default value of $\kappa = 1/2.7$ and 100 values of λ. As for linear models, SOSnet called the R package `glmnet` Friedman et al. (2010) to compute the Lasso estimators for 20 lambdas on a default log scale and with $o = 5$. 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 prediction error 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 3–4 of Figure 1. The two vertical lines indicate models chosen using GIC with $c = 2$, the black one for SOSnet and the red one for `cvplogistic`.

Summarizing the results of the simulation study, one can observe that SOSnet 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 $c = 2.5$ in GIC usually gave satisfactory results. The mean execution time of SOSnet was approximately 1.5 times longer than for `SparseNet`. SOSnet for logistic regression gave similar accuracy as `cvplogistic` with much lower execution time: SOSnet was approximately 10 times faster. The value $c = 2$ in GIC usually gave satisfactory results.

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

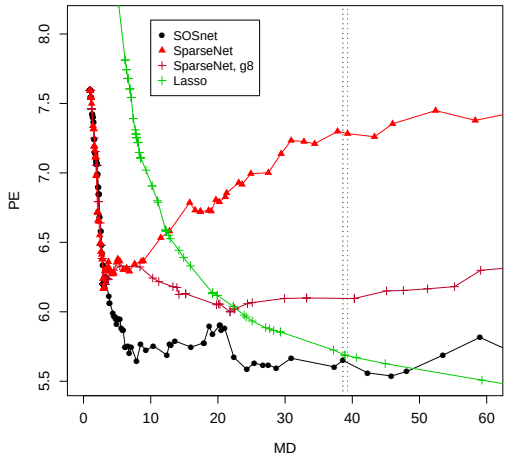
Figure 1: Results for simulated data



6.2. Real Data Sets

The methylation data set was described in Hannum et al. (2013). It consist 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=.05. We compared the root mean squared errors (PE) and model dimensions (MD) for SOSnet and `SparseNet` via 10-fold cross-validation. For each value of hyperparameter $c = .25, .5, \ldots, 7.5$ values of $(MD(c), PE(c))$ for the models chosen

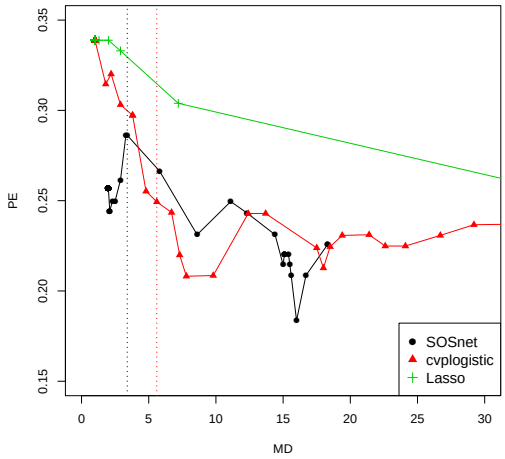
Figure 2: Results for the methylation data set



by GIC=GIC(c) 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, SOSnet gives uniformly smaller PE than `SparseNet` for all $MD \geq 3$. The two vertical lines indicate models chosen using GIC with $c = 2.5$: the black one for SOSnet 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 axillary lymph node involvement to predict 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 model dimensions (MD) for SOSnet and `cvplogistic` via 10-fold cross-validation. The results are presented in Figure 3. Minimal PE for SOSnet was smaller than for `cvplogistic`, but for a larger model. The algorithms work comparably, but again SOSnet was 63 times faster. The two vertical lines indicate models chosen using GIC with $c = 2$: the black one for SOSnet and the red one for `cvplogistic`.

Figure 3: Results for the breast cancer data set



7. Conclusions

In the paper we propose the SS algorithm which is an alternative method to TL and FCP of improving the Lasso. For linear models it 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.

Our approach encompasses fundamental model for prediction of continuous as well as of binary response and the main result is 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 linear models to general predictive models.

We propose an algorithm SOSnet, which is a generalization of the SOS algorithm for general predictive models. Using net of parameters, SOSnet avoids problem of choosing one specific λ. The gap between theoretical results for SS and the SOSnet algorithm is similar to the difference between theory for FCP and it implementations `SparseNet` or `cvpologistic`. Numerical experiments reveal that for linear models SOSnet is more accurate than `SparseNet` with comparable computing time, whereas for logistic models performance of SOSnet is on par with performance of `cvpologistic` with computing times at least 10 times shorter.

We have shown in simulations (dotted vertical lines in Figure 1) that predictively optimal λ for linear models equals approximately $\sqrt{2.5\sigma^2 \log p}$, which is close to (14) and for logistic models is $\sqrt{2 \log p}$, which together with (15) suggests that $c \approx 1/4$. The relations between the safest choice λ discussed in Remark 2 and predictively optimal λ are important applications of our theory.

Acknowledgments

We thank a bunch of people.

References

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

P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data*. Springer, New York, 2011.

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

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

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

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

E. Gravier, G. Pierron, A. Vincent-Salomon, N. Gruel, V. Raynal, A. Savignoni, Y. De Rycke, J.Y. Pierga, C. Lucchesi, F. Reyal, et al. Prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*, 49(12):1125–1125, 2010.

G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.B. Fan, Y. Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

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

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

D. Jiang and J. Huang. Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and Computing*, 24(5):871–883, 2014.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

P. Massart. About the constants in talagrands concentration inequalities for empirical processes. *The Annals of Probability*, 28:863–884, 2000.

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

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

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

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

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

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Verlag, New York, 1996.

D.L. Wallace. Bounds on normal approximations to student's and the chi-square distributions. *Annals of Mathematical Statistics*, 30:1121–1130, 1959.

L. Wang, Y. Kim, and R. Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics*, 41:2505–2536, 2013.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, 42:2164–2201, 2014.

F. Ye and C.H. Zhang. Rate minimaxity of the Lasso and Dantzig Selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942, 2010a.

C.H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27:576–593, 2012.

T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010b.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *NIPS*, pages 2304–2312, 2009.

Appendix A. Proofs and auxiliary results

Let us define for $k = 1, \ldots, t-1$

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

Obviously $\delta = \min_k \delta_k / k$. First we prove the following lemma which will be used in the proof of Lemma 3.

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

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

Proof [Proof of Lemma 5] For $\beta_T \in \partial B_k$ from assumption (5), 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 $b \in (0,1)$

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

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

Proof [Proof of Lemma 3] From assumption of the Lemma $\lambda^2 < c\delta$, so $b_k = \lambda^2 k/(c\delta_k) < 1$. Hence for $k = 1, \ldots, t-1$

$$k\lambda^2 \leq b_k c\delta_k. \tag{31}$$

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

$$akc\lambda^2 \leq (1-b_k)^2c^2\delta_k/4, \tag{32}$$

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

$$b \leq 1 + 2a - \sqrt{(1+2a)^2 - 1} = f_1(a). \tag{33}$$

Indeed, it is easy to check that (33) follows from the assumption as

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

Finally from (31), Lemma 5 and (32) we obtain, respectively

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

■

Proof [Proof of Lemma 4] For $J \supset T$ define $W_J = X_J(\mathring{\beta}_J - \hat{\beta}_J^{ML})$ and $m = |J \setminus T|$. We have from assumption (5) and properties of orthogonal projection H_J

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

so $\{\ell_J - \ell_T \leq -m\lambda^2/2\} \subseteq \{\varepsilon^T H_J \varepsilon \geq mc\lambda^2\}$.

Moreover, $\varepsilon^T H_J \varepsilon \leq \varepsilon^T H_T \varepsilon + \varepsilon^T H_{J\setminus T} \varepsilon$, hence we obtain for $\tau > 0$ and $a \in (0,1)$

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

Finally

$$\{T \in \mathcal{J}, \hat{T} \supset T\} \subseteq \bigcup_{J \supset T} \{\ell_J - \ell_T \leq -|J\setminus T|\lambda^2/2\} \subseteq \mathcal{C}_T((1-a)c\lambda^2) \cup \bigcup_{J \supset T} \mathcal{C}_{J\setminus T}(|J\setminus T|ac\lambda^2).$$

■