Journal of Machine Learning Research 16 (2015) 961-992 Submitted 10/13; Revised 11/14; Published 5/15

Combined ℓ_1 and Greedy ℓ_0 Penalized Least Squares for Linear Model Selection[*]

Piotr Pokarowski POKAR@MIMUW.EDU.PL
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Banacha 2, 02-097 Warsaw, Poland

Jan Mielniczuk MIEL@IPIPAN.WAW.PL
Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75, 00-662 Warsaw, Poland

Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warsaw, Poland

Editor: Tong Zhang

Abstract

We introduce a computationally effective algorithm for a linear model selection consisting of three steps: screening–ordering–selection (SOS). Screening of predictors is based on the thresholded Lasso that is ℓ_1 penalized least squares. The screened predictors are then fitted using least squares (LS) and ordered with respect to their $|t|$ statistics. Finally, a model is selected using greedy generalized information criterion (GIC) that is ℓ_0 penalized LS in a nested family induced by the ordering. We give non-asymptotic upper bounds on error probability of each step of the SOS algorithm in terms of both penalties. Then we obtain selection consistency for different (n, p) scenarios under conditions which are needed for screening consistency of the Lasso. Our error bounds and numerical experiments show that SOS is worth considering alternative for multi-stage convex relaxation, the latest quasiconvex penalized LS. For the traditional setting ($n > p$) we give Sanov-type bounds on the error probabilities of the ordering–selection algorithm. It is surprising consequence of our bounds that the selection error of greedy GIC is asymptotically not larger than of exhaustive GIC.

Keywords: linear model selection, penalized least squares, Lasso, generalized information criterion, greedy search, multi-stage convex relaxation

1. Introduction

Literature concerning linear model selection has been lately dominated by analysis of the *least absolute shrinkage and selection operator* (Lasso) that is ℓ_1 penalized least squares for the 'large p - small n scenario', where n is number of observations and p is number of all predictors. For a broad overview of the subject we refer to Bühlmann and van de Geer (2011). It is known that consistency of selection based on the Lasso requires strong regularity

[*] This paper is dedicated to Jacek Koronacki on the occasion of his 70th birthday.

of an experimental matrix named *irrepresentable conditions* which are rather unlikely to hold in practice (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). However, consistency of the Lasso predictors or consistency of the Lasso estimators of the linear model parameters is proved under weaker assumptions such as the *restricted isometry property* (RIP). The last condition means that singular values of normalized experimental submatrices corresponding to small sets of predictors are uniformly bounded away from zero and infinity. Under those more realistic conditions and provided that a certain lower bound on the absolute values of model parameters called *beta-min condition* holds, the Lasso leads to consistent screening, that is the set of nonzero Lasso coefficients S contains with large predetermined probability the uniquely defined true model T. This property explains Bühlmann's suggestion that one should interpret the second 's' in 'Lasso' as 'screening' rather than 'selection' (see discussion of Tibshirani, 2011) and the task is now to remove the spurious selected predictors. To this aim two-stage procedures as the adaptive or the thresholded Lasso have been proposed (cf. Zou, 2006; Huang et al., 2008; Meinshausen and Yu, 2009; Zhou, 2009, 2010; van de Geer et al., 2011). They yield selection consistency under strong version of the beta-min condition and without such strengthening tend to diminish the number of selected spurious predictors, but, similarly to the Lasso they yield screening consistency only. Alternative approaches require minimization of *least squares* (LS) penalized by quasiconvex functions that are closer to the ℓ_0 penalty then ℓ_1 (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2010a,b; Zhang and Zhang, 2012; Huang and Zhang, 2012; Zhang, 2013; Wang et al., 2014). These methods lead to consistent selection under RIP and considerably weaker version of the beta-min condition, nevertheless are more computationally demanding.

Regularization is required when a model matrix is not a full rank or when $n < p$, but for the traditional regression when an experimental plan is of full rank and $n > p$ it is possible to construct a computationally effective and selection consistent two-stage *ordering–selection* (OS) procedure, as follows. First, a full model F using LS is fitted, predictors are ordered with respect to their $|t|$ statistics from the fit and finally, a submodel of F in a nested family pertaining to the ordering is selected using thresholding as in Rao and Wu (1989), Bunea et al. (2006) or *generalized information criterion* (GIC) as in Zheng and Loh (1995). The OS algorithm can be treated as *greedy* ℓ_0 penalized LS because it requires computing a criterion function for $2p$ models only instead of all 2^p models. Frequently, sufficient conditions on an experimental plan and a vector of true coefficients for consistency of such procedures are stated in terms of the *Kullback-Leibler divergence* (KL) of the true model from models which lack at least one true predictor (Zheng and Loh, 1995; Shao, 1998; Chen and Chen, 2008; Casella et al., 2009; Pötscher and Schneider, 2011; Luo and Chen, 2013). In particular, a bound on the probability of selection error in Shao (1998) closely resembles the Sanov theorem in information theory on bounds of probability of a non-typical event using the KL divergence.

In our contribution we introduce a computationally effective three-step algorithm for linear model selection based on a *screening–ordering–selection* (SOS) scheme. Screening of predictors is based on a version of the thresholded Lasso proposed by Zhou (2009, 2010) and yields the screening set S such that $|S| \leq n$. Next, an implementation of the OS algorithm described above proposed by Zheng and Loh (1995) is applied. We give non-asymptotic upper bounds on error probability of each step of the SOS algorithm in terms of the Lasso and GIC penalties (Theorem 1). As a consequence of proved bounds we obtain selection consistency for different (n, p) scenarios under weak conditions which are sufficient for screening consistency of the Lasso. Our assumptions allow for strong correlation between predictors, in particular replication of spurious predictors is possible.

The SOS algorithm is an improvement of the new version of the thresholded Lasso and turns out to be a promising competitor to *multi-stage convex relaxation* (MCR), the latest quasiconvex penalized LS (Zhang, 2010b, 2013). The condition on correlation of predictors assumed there seems to be stronger than ours, whereas the beta-min condition may be weaker (Section 5). In our simulations for $|T| \ll n \ll p$ scenario, SOS was faster and more accurate than MCR (Section 8).

For case $n > p$ we also give a bound on probability of selection error of the OS algorithm. Our bound in this case is more general than in Shao (1998) as we allow ordering of predictors, $p = p_n \to \infty$, $|T| = |T_n| \to \infty$ or the GIC penalty may be of order n (Theorem 2). It is surprising consequence of Theorems 1-2 that the probability of selection error of greedy GIC is asymptotically not larger than of exhaustive GIC. Thus employment of greedy search dramatically decreases computational cost of l_0 penalized LS minimization without increasing selection error probability.

As a by-product we obtained a strengthened version of the nonparametric sparse oracle inequality for the Lasso proved by Bickel et al. (2009) and, as its consequence, more tight bounds on prediction and estimation error (Theorem 4). We simplified and strengthened an analogous bound for the thresholded Lasso given by Zhou (2009, 2010) (Theorem 1 part T1). It is worth noticing that all results are proved simultaneously for two versions of the algorithm: for the Lasso used in practice when a response is centered and predictors are standardized as well as for its formal version for which an intercept corresponds to a dummy predictor.

The paper is organized as follows. In Section 2 the SOS algorithm is introduced and in Section 3 we study properties of geometric characteristics pertaining to an experimental matrix and a vector of coefficients which are related to identifiability of a true model. Section 4 contains our main results that is bounds on selection error probabilities for the SOS and OS algorithm. In Section 5 we briefly discuss the MCR algorithm and compare error bounds for SOS and MCR. Section 6 treats properties of post-model selection estimators pertaining to SOS and MCR. Section 7 contains improved bounds on the Lasso estimation and prediction. Section 8 presents a simulational study. Concluding remarks are given in Section 9. Appendix contains detailed proofs of the stated results.

2. Selection Algorithm

The aim of this section is to describe the proposed selection algorithm. As in the first step of the algorithm we use the Lasso estimator to screen predictors and since in the literature there exist two versions of the Lasso for the linear model which differ in the treatment of the intercept, we start this section by defining two parametrizations of the linear model related to these versions of the Lasso. Next we state a general definition encompassing both cases, present our implementation of the SOS scheme and finally we discuss its computational complexity.

2.1 Linear Regression Model Parametrizations

We consider a general regression model of real-valued responses having the following structure

$$y_i = \mu(x_{i.}) + \varepsilon_i, \qquad i = 1, 2, \dots, n,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are iid $N(0, \sigma^2)$, $x_{i.} \in \mathbf{R}^p$, and $p = p_n$ may depend on n. In a vector form we have

$$y = \mu + \varepsilon, \tag{1}$$

where $\mu = (\mu(x_{1.}), \dots, \mu(x_{n.}))^T, \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ and $y = (y_1, \dots, y_n)^T$.

Let $X = [x_{1.}, \dots, x_{n.}]^T = [x_1, \dots, x_p]$ be the $n \times p$ matrix of experiment. We consider two linear parametrizations of (1). The first parametrization is:

$$\mu = \alpha^* + X\beta^*, \tag{2}$$

where $\alpha^* \in \mathbf{R}$ is an intercept and $\beta^* \in \mathbf{R}^p$ is a vector of coefficients corresponding to predictors. The second parametrization is

$$\mu = X\beta^*, \tag{3}$$

where the intercept is either set to 0 or is incorporated into vector β^* and treated in the same way as all other coefficients in the linear model. In order to treat both parametrizations in the same way we write $\mu = \tilde{X}\tilde{\beta}^*$ where, with $\mathbb{1}_n$ denoting a column of ones, $\tilde{X} = [\mathbb{1}_n, X]$ and $\tilde{\beta}^* = (\alpha^*, \beta^{*T})^T$ in the case of (2) and $\tilde{X} = X$ and $\tilde{\beta}^* = \beta^*$ in the case of (3). We note that (3) is convenient for theoretical considerations and simulations on synthetic data, but (2) is natural for real data applications and occurs as a default option in popular statistical software.

Let $J \subseteq \{1, 2, \dots, p\} = F$ be an arbitrary subset of the full model F and $|J|$ the number of its elements, X_J is a submatrix of X with columns having indices in J, β_J is a subvector of β with columns having indices in J. Moreover, let $\tilde{X}_J = [\mathbb{1}_n, X_J]$ and $\tilde{\beta}_J = (\alpha, \beta_J^T)^T$ in the case of (2) or $\tilde{X}_J = X_J$ and $\tilde{\beta}_J = \beta_J$ in the case of (3). $\tilde{H}_J$ will stand for a projection matrix onto the subspace spanned by columns of $\tilde{X}_J$. Linear model pertaining to predictors being columns of X_J will be frequently identified as J. We will also denote by $T = T_n$ a true model that is a model such that $T = \mathrm{supp}(\beta^*) = \{j \in F : \beta_j^* \neq 0\}$ for some β^* such that $\mu = \tilde{X}\tilde{\beta}^*$. The uniqueness of T and β^* for a given n will be discussed in Section 3.

2.2 Practical and Formal Lasso

The Lasso introduced in Tibshirani (1996) is a popular method of estimating β^* in the linear model. For discussion of properties of the Lasso see for example Tibshirani (2011) and Bühlmann and van de Geer (2011). When using the Lasso for data analytic purposes parametrization (2) is considered, vector of responses y is centered and columns of X are standardized. The standardization step is usually omitted in formal analysis in which parametrization (3) is assumed, α is taken to be 0 and X consists of meaningful predictors only, without column of ones corresponding to intercept. Alternatively, columns of X are normalized by their norms (see for example formula 2.1 in Bickel et al., 2009). Here, in order to accommodate considered approaches in one definition we introduce a general form of the Lasso. Let H_0 be an $n \times n$ projection matrix, where H_0 is specified as a vector centering matrix $\mathbb{I}_n - \mathbb{1}_n\mathbb{1}_n^T/n$ in the case of the applied version of the Lasso pertaining to parametrization (2) and the identity matrix $\mathbb{I}_n$ for the formal Lasso corresponding to (3). Moreover, let

$$D = \text{diag}(||H_0x_j||)_{j=1}^p, \quad X_0 = H_0XD^{-1}, \quad X_0 = [x_{01}, \ldots, x_{0p}], \quad y_0 = H_0y \tag{4}$$

and $\theta^* = D\beta^*$, $\mu_0 = H_0\mu$. For estimation of β^*, data (X_0, y_0) will be used. Note that for the first choice of orthogonal projection in the definition of X_0 columns in X are normalized by their norms whereas for the second they are standardized (centered and divided by their standard deviations). Consider the case of (2) and denote by H_{0J} projection onto $\text{sp}\{(H_0x_j)_{j\in J}\}$. Observe that as $\text{sp}\{\mathbb{1}_n, (x_j)_{j\in J}\} = \text{sp}\{\mathbb{1}_n\} \oplus \text{sp}\{(H_0x_j)_{j\in J}\}$ and consequently $\tilde{H}_J = H_{0J} + \mathbb{1}_n\mathbb{1}_n^T/n$, we have that

$$\mathbb{I}_n - \tilde{H}_J = (\mathbb{I}_n - H_{0J})H_0. \tag{5}$$

The above equality trivially holds also in the case of (3).

For $a = (a_j) \in \mathbf{R}^k$, let $|a| = \sum_{j=1}^k |a_j|$ and $||a|| = (\sum_{j=1}^k a_j^2)^{1/2}$ be ℓ_1 and ℓ_2 norms, respectively. As J may be viewed as sequence of zeros and ones on F, $|J|$ denotes cardinality of J.

General form of the Lasso estimator of β is defined as follows

$$\hat{\beta} = \text{argmin}_\beta\{||H_0(y - X\beta)||^2 + 2r_L|D\beta|\} = D^{-1}(\text{argmin}_\theta\{||y_0 - X_0\theta||^2 + 2r_L|\theta|\}), \tag{6}$$

where a parameter $r_L = r_{nL}$ is a penalty on l_1 norm of a potential estimator of β. Thus in the case of parametrization (2) the Lasso estimator of β may be defined without using extended matrix $\tilde{X}$ by applying H_0 to $y - X\beta$ that is by centering it. In the case of parametrization (3) $H_0 = \mathbb{I}_n$ and the usual definition of the Lasso used in formal analysis is obtained. We remark that the approaches used in theoretical considerations for which columns of X are not normalized as in Bühlmann and van de Geer (2011) or Zhang (2013) formally correspond to (6) with $H_0 = \mathbb{I}_n$ and $D = dI_p$, where $d = \max_{1\leq j\leq p} ||x_j||$.

Note that in the case of parametrization (2) $\hat{\beta}$ is subvector corresponding to β of the following minimizer

$$\text{argmin}_{\tilde{\beta}}\{||y - \tilde{X}\tilde{\beta}||^2 + 2r_L|D\beta|\} = \text{argmin}_{\alpha,\beta}\{||y - \alpha\mathbb{1}_n - X\beta||^2 + 2r_L|D\beta|\}, \tag{7}$$

where the equality of minimal values of expressions appearing in (6) and (7) is obtained when the expression $||y - \alpha\mathbb{1}_n - X\beta||^2$ is minimized with respect to α for fixed β. However, omitting centering projection H_0 in (6) when the first column of X consists of ones and corresponds to intercept, leads to lack of invariance of $\hat{\beta}$ when the data are shifted by a constant and yields different estimates that those used in practice. This is a difference between the Lasso and the LS estimator: LS estimator has the same form regardless of which of the two parametrizations (2) or (3) is applied. Using (5) we have for the LS estimator $\hat{\beta}_J^{LS}$ in model J that the sum of squared residuals for the projection $\tilde{H}y$ equals

$$R_J = ||(\mathbb{I}_n - \tilde{H}_J)y||^2 = ||(\mathbb{I}_n - H_{0J})y_0||^2 = ||y_0 - X_{0J}\hat{\theta}_J^{LS}||^2 \tag{8}$$

and

$$\hat{\beta}_J^{LS} = D^{-1}\hat{\theta}_J^{LS}, \quad \hat{\theta}_J^{LS} = \text{argmin}_{\theta_J}||y_0 - X_{0J}\theta_J||^2.$$

2.3 Implementation of the Screening–Ordering–Selection Scheme

The SOS algorithm which is the main subject of the paper is the following implementation of the SOS scheme.

Algorithm *(SOS)*
Input: y, X and r_L, b, r.
Screening. Compute the Lasso estimator $\hat{\beta} = D^{-1}\hat{\theta}$, $\hat{\theta} = (\hat{\theta}_1, \dots \hat{\theta}_p)^T$ with a penalty parameter r_L and set $S_0 = \{j : |\hat{\theta}_j| > b\}$, $B = b(|S_0| \vee 1)^{1/2}$, $S_1 = \{j : |\hat{\theta}_j| > B\}$.
Ordering. Fit the model S_1 by ordinary LS and order predictors $\hat{O} = (j_1, j_2, \dots, j_{|S_1|})$ using values of corresponding squared t statistics $t^2_{j_1} \geq t^2_{j_2} \geq \dots \geq t^2_{j_{|S_1|}}$.
Selection. In the nested family $\mathcal{G} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, S_1\}$ choose a model $\hat{T} \equiv \hat{T}_{S_1, \hat{O}}$ according to the *generalized information criterion* (GIC) $\hat{T} = \text{argmin}_{J \in \mathcal{G}}\{R_J + |J|r\}$, where $r = r_n$ is a penalty pertaining to GIC.
Output: $\hat{T}_{SOS} = \hat{T}$, $\hat{\beta}^{SOS} = \hat{\beta}^{LS}_{\hat{T}}$.

The OS algorithm is intended for the case $p < n$ and is a special case of SOS for which S_1 is taken equal to F.

We note that empty set in the definition of $\mathcal{G}$ corresponds to $\mu = 0$ in the case of parametrization (3) and $\mu = \alpha^*$ in the case of (2). It is easy to check also that

$$\frac{t^2_j}{n - |S_1|} = \frac{R_{S_1 \setminus \{j\}} - R_{S_1}}{R_{S_1}}, \tag{9}$$

thus ordering with respect to decreasing values of (t^2_j) in the second step of the procedure is the same as ordering of $(R_{S_1 \setminus \{j\}})$ in decreasing order.

2.4 Computational Complexity of the SOS Algorithm

There are many approximate algorithms for the Lasso estimator (6) as quadratic program solvers or coordinate descent in Friedman et al. (2010). The popular LARS method proposed in Efron et al. (2004) can be used to compute exactly, in finitely many steps, the whole Lasso regularized solution path which is piecewise linear with respect to r_L. It has been shown recently in Mairal and Yu (2012) that, in the worst case, the number of linear segments of this path is exactly $(3^p + 1)/2$, so the overall computational cost of the Lasso is $O(3^p pn)$, see Rosset and Zhu (2007). Hence, by the most popular criterion of computational complexity LARS does not differ from, for example, an exhaustive search for the ℓ_0 penalized LS problem. However, experience with data suggests that the number of linear segments of the LARS regularization path is typically $O(n)$, so LARS execution requires $O(np \min(n, p))$ flops, see Rosset and Zhu (2007) and Bühlmann and van de Geer (2011), chapter 2.12. Thus taking into account the result in Mairal and Yu (2012) on uniform approximation of the Lasso regularization paths, for typical data set the Lasso may be considered computationally efficient (cf. also discussion on the page 7 in Zhang (2013)).

In Section 4 we will discuss conditions on X and β^*_T, under which S_1 includes a unique true model T and $|S_1| \leq n$ or even $|S_1| \leq 4|T|$ with high probability. In this case we can use LS to fit a linear model, thus the ordering step takes $O(n|S_1|^2)$ calculations by the

QR decomposition of the matrix X_{0S_1}. Computing $(R_J)_{J\in\mathcal{G}}$ in the selection step demands also only one QR decomposition of X_{0S_1} with columns ordered according to $\hat{O}$. Indeed, let $X_{0S_1} = QW$, where an orthogonal matrix $Q = [q_1, \ldots, q_{|S_1|}]$. The following iterative procedure can be used

$$R_\emptyset = ||y_0||^2; \;\; \textbf{for} \;\; k = 1, \ldots, |S_1| \;\; \textbf{do} \;\; R_{\{1,\ldots,k\}} = R_{\{1,\ldots,k-1\}} - (q_k^T y_0)^2 \;\; \textbf{endfor}.$$

Observe, that from (9) the ordering part demands GIC only for $|S_1|$ models that is for $S_1 \setminus \{j\}$, $j \in S_1$. Thus two last parts of the SOS algorithm or, equivalently, the OS algorithm demands GIC only for $2|S_1|$ models instead of all $2^{|S_1|}$ and we can call it *greedy* ℓ_0 penalized LS.

We conclude that the SOS algorithm is computationally efficient and the most time expensive part of it is the screening. The same conclusion follows from our simulations described in Section 8.

3. A True Model Identifiability

In this section we consider two types of linear model characteristics which will be used to quantify the difficulty of selection or, equivalently, a true model identifiability problem, and we study the interplay between them.

3.1 Kullback-Leibler Divergences

Let T be given true model that is $T \subseteq F$ such that $\mu = \tilde{X}\tilde{\beta}^* = \tilde{X}_T\tilde{\beta}^*_T$ and $T = \text{supp}(\beta^*_T) = \{j \in F : \beta^*_{j,T} \neq 0\}$. For $J \subseteq F$ define

$$\delta(T \parallel J) = ||(\mathbb{I}_n - \tilde{H}_J)\tilde{X}_T\tilde{\beta}^*_T||^2.$$

In view of (5) we obtain

$$\delta(T \parallel J) = ||(\mathbb{I}_n - H_{0J})H_0\tilde{X}_T\tilde{\beta}^*_T||^2 = ||(\mathbb{I}_n - H_{0J})H_0X_T\beta^*_T||^2 = ||(\mathbb{I}_n - H_{0J})X_{0T}\theta^*_T||^2. \quad (10)$$

Let $KL(\tilde{\beta}^*_T \parallel \tilde{\beta}_J) = \mathbf{E}_{\tilde{\beta}^*_T}\log(f_{\tilde{\beta}^*_T}/f_{\tilde{\beta}_J})$ be the *Kullback-Leibler divergence* of the normal density $f_{\tilde{\beta}^*_T}$ of $N(\tilde{X}_T\tilde{\beta}^*_T, \sigma^2\mathbb{I}_n)$ from the normal density $f_{\tilde{\beta}_J}$ of $N(\tilde{X}_J\tilde{\beta}_J, \sigma^2\mathbb{I}_n)$. Let $\Sigma = X_0^TX_0$ be a *coherence matrix* if H_0 is the identity matrix and *a correlation matrix* if $H_0 = \mathbb{I}_n - \mathbb{1}_n\mathbb{1}_n^T/n$. Let Σ_J stands for a submatrix of Σ with columns having indices in J and let $\lambda_{min}(\Sigma_J)$, $\lambda_{max}(\Sigma_J)$ denote extremal eigenvalues of Σ_J. The following proposition lists the basic properties of the parameter δ. Observe also that $\delta(T \parallel J)$ is a parameter of non-centrality of χ^2 distribution of R_J that is $R_J \sim \chi^2_{n-|J|}(\delta(T \parallel J))$.

Proposition 1

$$(i) \quad \delta(T \parallel J) = 2\sigma^2 \min_{\tilde{\beta}_J} KL(\tilde{\beta}^*_T \parallel \tilde{\beta}_J) = 2\sigma^2 \min_{\tilde{\beta}_J} KL(\tilde{\beta}_J \parallel \tilde{\beta}^*_T).$$

$$(ii) \quad \delta(T \parallel J) = \min_{\theta_J}\left|\left|[X_{0,T\setminus J}, X_{0,J}]\begin{pmatrix}\theta^*_{T\setminus J}\\ \theta_J\end{pmatrix}\right|\right|^2 \geq \lambda_{min}(\Sigma_{J\cup T})||\theta^*_{T\setminus J}||^2 \quad (11)$$

The following scaled Kullback-Leibler divergence will be employed in our main results in Section 4.

$$\delta(T,s) = \min_{j\in T, J\supseteq T, |J|\leq s} \delta(T \parallel J \setminus \{j\}).$$

This coefficient was previously used to prove selection consistency in Zheng and Loh (1995); Chen and Chen (2008); Luo and Chen (2013) and to establish asymptotic law of post-selection estimators in Pötscher and Schneider (2011). Similar coefficients appear in proofs of selection consistency in Shao (1998) and Casella et al. (2009). Obviously, $\delta(T,s)$ is a nonincreasing function of s.

Identifiability of a true model is stated in the proposition below in terms of

$$\delta(T) = \min_{J\not\supseteq T, |J|\leq |T|} \delta(T \parallel J).$$

Proposition 2 *There exists at most one true model T such that $\delta(T) > 0$.*

Assume by contradiction that T' is a different true model, that is we have $T' = \mathrm{supp}(\tilde{\beta})$ for some $\tilde{\beta}$ such that $\mu = \tilde{X}\tilde{\beta}$. Then by symmetry we can assume $|T| \leq |T'|$. Hence $|T' \setminus T| > 0$ and $\delta(T') \leq \delta(T' \parallel T) = 0$.
It is easy to see that if $\delta(T) > 0$ then columns of X_T are linearly independent and, consequently, there exists at most one $\tilde{\beta}^*_T$ such that $\mu = \tilde{X}_T\tilde{\beta}^*_T$.

In Section 4.2 we infer identifiability of a true model T from Proposition 2 and the following inequality

$$\delta(T,p) \leq \delta(T). \tag{12}$$

Indeed, for any J such that $J \not\supseteq T$ and $|J| \leq |T|$ there exists $j \in T$ such that $J \subseteq F \setminus \{j\}$. Thus we obtain $\delta(T \parallel F \setminus \{j\}) \leq \delta(T \parallel J)$ and minimizing both sides yields (12).

3.2 Restricted Eigenvalues

For $J \subseteq F$, $\bar{J} = F \setminus J$ and $c > 0$ let

$$\kappa^2(J,c) = \min_{\nu\neq 0, |\nu_{\bar{J}}|\leq c|\nu_J|} \frac{\nu^T\Sigma\nu}{\nu_J^T\nu_J} \qquad \text{and} \qquad \kappa^2(s,c) = \min_{J:|J|\leq s} \kappa(J,c).$$

Both coefficients will be called *restricted eigenvalues* of Σ. Observe that

$$\kappa^2(J,c) = \min_{\nu\neq 0, |\nu_{\bar{J}}|\leq c|\nu_J|} \frac{||X_0\nu||^2}{||\nu_J||^2} = \min_{\nu\neq 0, |\nu_{\bar{J}}|\leq c|\nu_J|} \frac{||X_0\nu_J - X_0\nu_{\bar{J}}||^2}{||\nu_J||^2}. \tag{13}$$

The coefficient $\kappa(s,c)$ is a modified version of an index introduced in Bickel et al. (2009). Modification consists in replacing X appearing in the original definition by X_0 and omitting the term $n^{-1/2}$. Pertaining parameters for a fixed set of predictors J and their various modifications were introduced and applied to bound the Lasso errors by van de Geer and Bühlmann (2009).

In order to study relations between sparse and restricted eigenvalues we set

$$\kappa^2(J,0) = \min_{\nu\neq 0, \mathrm{supp}(\nu)\subseteq J} \frac{\nu^T\Sigma\nu}{\nu^T\nu} \qquad \text{and} \qquad \kappa^2(s,0) = \min_{J:|J|\leq s} \kappa^2(J,0).$$

Note that if X_0 is defined in (4) or in remark below (6) applies we have that $\max_{1\leq j\leq p}||x_{0j}|| \leq 1$. Thus from Rayleigh-Ritz theorem we have

$$\kappa^2(J,0) = \lambda_{min}(\Sigma_J) \leq \frac{tr(\Sigma_J)}{|J|} \leq 1 \wedge \lambda_{max}(\Sigma_J). \tag{14}$$

The upper bound above equals 1 when the columns are normalized or standardized. Note that $\kappa(J,c)$ and $\kappa(s,c)$ are nonincreasing functions of both arguments. Moreover, $\kappa^2(J,c) \leq \kappa^2(J,0)$ and $\kappa^2(s,c) \leq \kappa^2(s,0)$. This holds in view of an observation that for any fixed J and $c > 0$, any ν such that $\text{supp}(\nu) \subseteq J$ satisfies $\nu = \nu_J$ and thus $|\nu_{\bar{J}}| \leq c|\nu_J|$. It is easy to show also that $\kappa^2(J,c) \to \kappa^2(J,0)$ and $\kappa^2(s,c) \to \kappa^2(s,0)$ monotonically when $c \to 0^+$. Another less obvious bound, which is used in the following is stated below.

Proposition 3 *For any $s \in \mathbf{N}$ and $c > 0$*

$$\kappa^2(s,c) \leq (\lfloor c \rfloor + 1)\kappa^2((\lfloor c \rfloor + 1)s, 0).$$

Condition $\kappa(s,c) > 0$ imposed on matrix X is called *restricted eigenvalue condition* in Bickel et al. (2009) for their slightly different κ. Proposition 3 generalizes an observation there (p. 1720) that if the restricted eigenvalue condition holds for $c \geq 1$, then all square submatrices of Σ of size $2s$ are necessarily positive definite. Indeed, the proposition above implies that $\kappa(2s,0) > 0$ from which the observation follows. Positiveness of $\kappa(T,c)$ which due to the restriction on vectors ν over which minimization is performed can hold even for $p > n$, is a certain condition on weak correlation of columns. This condition, which will be assumed later, is much less stringent than $\kappa(|T|,c) > 0$, as it allows for example replication of columns belonging to the complement of T. Moreover $\kappa(T,c) > 0$ for $c \geq 1$ implies identifiability of a true model.

Proposition 4 *There exists at most one true model T such that $\kappa(T,1) > 0$.*

It follows that if $\kappa(T,1) > 0$, then columns of X_T are linearly independent and, consequently, there exists at most one $\tilde{\beta}^*_T$ such that $\mu = \tilde{X}_T\tilde{\beta}^*_T$.

The following $\kappa - \delta$ *inequalities* follow from the Propositions 1 (ii) and the Proposition 3. We set $\theta^*_{\min} = \min_{j\in T}|\theta^*_j|$ and $t = |T|$.

Proposition 5 *We have*

$$\kappa^2(T,3)\theta^{*2}_{\min} \leq \delta(T,t) \tag{15}$$

and

$$\kappa^2(t,3)\theta^{*2}_{\min} \leq 4\delta(T,4t). \tag{16}$$

4. Error Bounds for the SOS and OS Algorithms

In this section we present the main result that is non-asymptotic bounds on the error probabilities for all steps of the SOS algorithm. The errors of consecutive steps of SOS constitute decomposition of the selection error into four parts. Two errors which can be possibly committed in the selection step correspond to two situations when the selected model is a proper subset or a superset of T.

4.1 Error Bounds for SOS

Let $\mathcal{S}_n$ be a family of models having no more than s predictors where s is defined below and $\mathcal{T}_n = \{S \in \mathcal{S}_n : S \supseteq T\}$ consists of all true models in $\mathcal{S}_n$. Observe that $|\mathcal{T}_n| = \sum_{k=0}^{s-t} \binom{p-t}{k}$. Moreover, let O_{S_1} denote a set of all correct orderings of S_1 that is orderings such that all true variables in S_1 precede the spurious ones. To simplify notation set $\delta_s = \delta(T, s)$, $\delta_t = \delta(T, t)$ and $\kappa = \kappa(T, 3)$. We also define two constants $c_1 = (3 + 6\sqrt{2})^{-1} \approx 0.087$ and $c_2 = (6+4\sqrt{2})^{-1} \approx 0.086$. We assume for the remaining part of the paper that $p \geq t+1 \geq 2$ as boundary cases are easy to analyze. Moreover, we assume the following condition which ensures that the size of S_1 defined in the first step of the SOS algorithm does not exceed n with large probability and consequently LS could be performed on data (y_0, X_{0S_1}). It states that

$$s = s(T) = t + \lfloor t^{1/2}\kappa^{-2} \rfloor \leq n. \tag{17}$$

Theorem 1 *(T1) If for some $a \in (0,1)$ $8a^{-1}\sigma^2 \log p \leq r_L^2 \leq b^2/36 \leq c_1^2 t^{-1}\kappa^4\theta_{min}^{*2}$, then*

$$P(S_1 \notin \mathcal{T}_n) \leq \exp\left(-\frac{(1-a)r_L^2}{8\sigma^2} \right) \left(\frac{\pi r_L^2}{8\sigma^2} \right)^{-1/2}. \tag{18}$$

(T2) If for some $a \in (0,1)$ $a^{-1}\sigma^2 \log p \leq c_2(s-t+2)^{-1}\delta_s$, then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \notin O_{S_1}) \leq \frac{3}{2} \exp\left(-\frac{(1-a)c_2\delta_s}{\sigma^2} \right) \left(\frac{\pi c_2 \delta_s}{\sigma^2} \right)^{-1/2}. \tag{19}$$

(T3) If for some $a \in (0,1)$ (a) $r < at^{-1}\delta_t$ and (b) $8a^{-1}\sigma^2 \log t \leq (1-a)^2\delta_t$, then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}_{SOS}| < t) \leq \frac{1}{2} \exp\left(-\frac{(1-a)^3\delta_t}{8\sigma^2} \right) \left(\frac{\pi(1-a)^2\delta_t}{8\sigma^2} \right)^{-1/2}. \tag{20}$$

(T4) If for some $a \in (0,1)$ $4a^{-1}\sigma^2 \log p \leq r$, then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}_{SOS}| > t) \leq \exp\left(-\frac{(1-a)r}{2\sigma^2} \right) \left(\frac{\pi r}{2\sigma^2} \right)^{-1/2}. \tag{21}$$

A regularity condition on the plan of experiment $\tilde{X}$ and the true $\tilde{\beta}^*$ induced by the assumption of Theorem 1 (T1), namely $8a^{-1}\sigma^2 \log p \leq c_1^2 t^{-1}\kappa^4\theta_{min}^{*2}$, is known as the *beta-min condition.* Its equivalent form, which is popular in the literature states that for some $a \in (0,1)$

$$\sqrt{8c_1^{-2}a^{-1}\sigma^2 t\kappa^{-4} \log p} \leq \min_{j \in T} ||H_0 x_j|| \, |\beta_j^*|. \tag{22}$$

Observe that (22) implies that $\kappa > 0$, so it guarantees identifiability of T in view of Proposition 4.

Note that bounds in (T2) and (T3) as well as the bounds in Theorem 2 below can be interpreted as results analogous to the Sanov theorem in information theory on bounding probability of a non-typical event (cf. for example Cover and Thomas (2006), Section 11.4), as in view of Proposition 1 (i) δ_s may be expressed as $\min_{\beta \in B} 2\sigma^2 KL(\beta \parallel \beta^*)$ for a certain set B such that $\beta^* \notin B$.

The first corollary provides an upper bound on a selection error of the SOS algorithm under simpler conditions. The assumption $r_L^2 = 4r$ is quite arbitrary, but results in the same lower bound for penalty and almost the same bound on error probability as in the Corollary 3 below. Note that boundary values of r_L^2 and r of order $\log p$ are allowed in Corollaries 1–3.

Corollary 1 *Assume (17) and $r_L^2 = 4r$. If for some $a \in (0, 1 - c_1)$ we have (i) $4a^{-1}\sigma^2 \log p \leq r \leq b^2/144 \leq (c_1^2/4)at^{-1}\kappa^4\theta^{*2}_{min}$ and (ii) $r \leq (4c_2/3)t^{-1/2}\kappa^2\delta_s$, then*

$$P(\hat{T}_{SOS} \neq T) \leq 4\exp\left(-\frac{(1-a)r}{2\sigma^2}\right)\left(\frac{\pi r}{2\sigma^2}\right)^{-1/2}.$$

We consider now the results above under stronger conditions. We replace $\kappa = \kappa(T, 3)$ in (17) and the assumption (T1) by smaller $\kappa_t = \kappa(t, 3)$ and additionally assume the following *weak correlation condition*

$$\kappa_t^{-2} \leq 3t^{1/2}, \tag{23}$$

which is weaker than a condition $\kappa_t^{-2} \leq t^{1/2}$ in Theorem 1.1 in Zhou (2009, 2010). Observe that (23) is stronger than inequality (17) with κ_t instead of κ. Indeed, (23) implies in view of definition of s, that $s \leq 4t$. Next, from Proposition 3 we obtain $0 < t^{-1/2}/3 \leq \kappa_t^2 \leq 4\kappa(4t, 0)$, but obviously $\kappa(4t, 0) = 0$ for $4t > n$, hence $4t \leq n$ and $s \leq n$. Moreover, we obtain from (16) that $(c_1^2/4)at^{-1}\kappa_t^4\theta^{*2}_{min} < (4c_2/3)t^{-1/2}\kappa_t^2\delta_s$ as $\delta_s \geq \delta_{4t}$ and $16c_2/(3c_1^2) \geq 1$. Hence the Corollary 1 simplifies to the following corollary.

Corollary 2 *Assume (23) and $r = r_L^2/4$. If for some $a \in (0, 1 - c_1)$ we have $16a^{-1}\sigma^2 \log p \leq r_L^2 \leq b^2/36 \leq c_1^2 at^{-1}\kappa_t^4\theta^{*2}_{min}$, then*

$$P(\hat{T}_{SOS} \neq T) \leq 4\exp\left(-\frac{(1-a)r_L^2}{8\sigma^2}\right)\left(\frac{\pi r_L^2}{8\sigma^2}\right)^{-1/2}.$$

Theorem 1 shows that the SOS algorithm is an improvement of the adaptive and the thresholded Lasso (see Zou, 2006; Huang et al., 2008; Meinshausen and Yu, 2009; Zhou, 2009, 2010; van de Geer et al., 2011) as under weaker assumptions on an experimental matrix than assumed there we obtain much stronger result, namely selection consistency. Indeed, assumptions of Theorem 1 are stated in terms of $\kappa(T, 3)$, δ_s and δ_t instead of $\kappa(t, 3)$, thus allowing for example replication of spurious predictors. Discussion of assumptions of Corollary 2 shows that the original conditions in Zhou (2009, 2010) are stronger than our conditions ensuring screening consistency of the thresholded Lasso. We stress also that our bounds are valid in both cases when the formal or the practical Lasso is used in the screening step. In Section 5 our results will be compared with a corresponding result for MCR.

4.2 Error Bounds for OS

Now we state the corresponding bounds for error probabilities of the OS algorithm in the case of $p \leq n$. We recall that in the case of OS $S_1 = F$. Thus $\mathcal{S}_n = \mathcal{T}_n = \{S_1\}$ and $P(S_1 \notin \mathcal{T}_n) = 0$.

Theorem 2 *If for some $a \in (0,1)$ $\quad a^{-1}\sigma^2 \log(t(p-t)) \le c_2\delta_p$, then*

$$P(\hat{O} \notin O) \le \frac{3}{2}\exp\Big(-\frac{(1-a)c_2\delta_p}{\sigma^2}\Big)\Big(\frac{\pi c_2\delta_p}{\sigma^2}\Big)^{-1/2}.$$

Moreover, (T3) and (T4) of Theorem 1 hold.

Observe that assumptions of Theorem 2 imply that $\delta_p > 0$ which guarantees uniqueness of T in view of (12).

The next corollary is analogous to Corollary 1 and provides an upper bound on a selection error of the OS algorithm under simpler conditions. This bound is more general than in Shao (1998) as we allow for greedy selection (specifically ordering of predictors), $p = p_n \to \infty$, $t = t_n \to \infty$ or GIC penalty may be of order n.

Corollary 3 *If for some $a \in (0, 2c_2)$ $\;4a^{-1}\sigma^2 \log p \le r \le \min\big(at^{-1}\delta_t,\ 2c_2\delta_p\big)$, then*

$$P(\hat{T}_{OS} \ne T) \le 3\exp\Big(-\frac{(1-a)r}{2\sigma^2}\Big)\Big(\frac{\pi r}{2\sigma^2}\Big)^{-1/2}.$$

It is somewhat surprising consequence of the Corollary 1–3 that, from an asymptotic point of view, the selection error of the SOS and OS algorithms, which are versions of a greedy GIC, is not greater than the selection error of a plain, exhaustive GIC. Specifically, if we define the exhaustive GIC selector by

$$\hat{T}_E = \mathrm{argmin}_{J: J \subseteq F, |J| \le p}\{R_J + |J|r\},$$

then it follows from the lower bound in (37) below, that for an arbitrary fixed index $j_0 \notin T$ and $r > 0$ we have

$$P(\hat{T}_E \ne T) \ge P(R_{T\cup\{j_0\}} - R_T > r) \ge \frac{r}{r+\sigma^2}\exp\Big(-\frac{r}{2\sigma^2}\Big)\Big(\frac{\pi r}{2\sigma^2}\Big)^{-1/2}. \tag{24}$$

If the penalty term satisfies $\log p \ll r \ll \min(\delta_t/t,\ \delta_p)$ for $n \to \infty$, then from Corollary 3 and (24) we obtain

$$\overline{\lim_n} \log P(\hat{T}_{OS} \ne T) \le \underline{\lim}_n \log P(\hat{T}_E \ne T). \tag{25}$$

The last inequality indicates that it pays off to apply greedy algorithm in this context as a greedy search dramatically reduces ℓ_0 penalized LS without increasing its selection error.

The bounds on the selection error given in Corollaries 1–3 imply consistency of SOS and OS provided $r_n \to \infty$ and its strong consistency provided $r_n \ge c\log n$ for some $c > 2\sigma^2/(1-a)$. For boundary penalty $r_n = 4a^{-1}\sigma^2 \log p_n$ where $a \in (0, 2c_2)$, we obtain strong consistency of these algorithms if $n^{ca/(1-a)} \le p_n$ for some $c > 0.5$. Comparison of selection errors probabilities of the SOS and OS algorithms for $p < n$ requires further research.

5. Comparison of SOS and MCR

The SOS algorithm also turns out to be a competitor of iterative approaches which require minimization of more demanding LS penalized by quasiconvex functions (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2010a,b; Zhang and Zhang, 2012; Huang and Zhang, 2012; Zhang, 2013; Wang et al., 2014). In this section we compare selection error bounds for SOS and *multi-stage convex relaxation* (MCR) studied in Zhang (2010b, 2013) which is the latest example of this group of algorithms. In Section 8 we compare SOS and MCR in numerical experiments.

5.1 Multi-stage Convex Relaxation Algorithm

Results in Zhang (2013) concern parametrization of the linear model without intercept given in (3). Moreover, coordinates of β are not individually penalized in MCR. In concordance with the discussion below equation (6) this corresponds to $H_0 = \mathbb{I}_n$ and $D = d\mathbb{I}_p$, where $d = \max_{1\le j\le p} ||x_j||$. Obviously,

$$X_0 = H_0 X D^{-1} = X/d, \quad y_0 = y, \quad X\beta^* = \mu = \mu_0 = X_0\theta^*, \quad H_{0J} = H_J, \quad J \subseteq F$$

and $||x_{0j}|| \le 1$. The MCR procedure finds for given $r_Z, b_Z > 0$ approximate solution of the quasiconvex minimization problem

$$\hat{\beta}^{MCR} = d^{-1}\text{argmin}_\theta\big\{||y - X_0\theta||^2 + 2r_Z\sum_{j=1}^{p}(|\theta_j| \wedge b_Z)\big\}. \tag{26}$$

As was shown in Zhang (2010b) a local minimum of (26) could be approximated by the following iterative convex minimization algorithm.

```
Algorithm (MCR)
Input: $y, X$ and $r_Z, b_Z, l$.
    Compute $d$,  $X_0 = X/d$,  $\bar{S} = F$
    for $k = 1, 2, \dots, l$ do
        $\hat{\theta} = \text{argmin}_\theta\{||y - X_0\theta||^2 + 2r_Z|\theta_{\bar{S}}|\}$
        $\bar{S} = \{j \in F : |\hat{\theta}_j| \le b_Z\}$
    endfor
    $S = F \setminus \bar{S}$
Output: $\hat{T}_{MCR} = S$, $\hat{\beta}^{MCR} = \hat{\theta}_S/d$.
```

Since $X_0\theta = X_{0S}\theta_S + X_{0\bar{S}}\theta_{\bar{S}}$ and $(I - H_S)X_{0S} = 0$, we obtain

$$||y - X_0\theta||^2 = ||H_S(y - X_{0\bar{S}}\theta_{\bar{S}}) - X_{0S}\theta_S||^2 + ||(I - H_S)(y - X_{0\bar{S}}\theta_{\bar{S}})||^2. \tag{27}$$

Let $\theta_S = W_S^+ Q_S^T(y - X_{0\bar{S}}\theta_{\bar{S}})$, where $X_{0S} = Q_S W_S$, Q_S is an orthogonal matrix, W_S^+ is a pseudoinverse of W_S and Q_S, W_S are computed from the QR or SVD decomposition of X_{0S}. Then θ_S is the LS solution for the response $y - X_{0\bar{S}}\theta_{\bar{S}}$ and predictors X_{0S} and the first term on the right in (27) equals 0. Thus if we set $y_\diamond = (I - H_S)y$ and $X_{\diamond\bar{S}} = (I - H_S)X_{0\bar{S}}$, then

$$||y - X_0\theta||^2 = ||(I - H_S)(y - X_{0\bar{S}}\theta_{\bar{S}})||^2 = ||y_\diamond - X_{\diamond\bar{S}}\theta_{\bar{S}})||^2.$$

It follows that for computing $\hat{\theta}$ in the MCR algorithm, we can use the Lasso and LS subroutines separately as in the following (cf. Zou and Li (2008), Algorithm 2).

```
**Algorithm** *(MCR via Lasso and LS)*
**Input:** $y, X$ and $r_Z, b_Z, l$.
    Compute $d,\ \ X_{\diamond\bar{S}} = X/d,\ \ y_\diamond = y,\ \ S = \emptyset,\ \ \bar{S} = F$
    for $k = 1, 2, \ldots, l$ do
        $\hat{\theta}_{\bar{S}} = \mathrm{argmin}_{\theta_{\bar{S}}}\{||y_\diamond - X_{\diamond\bar{S}}\theta_{\bar{S}}||^2 + 2r_Z|\theta_{\bar{S}}|\}$
        $\hat{\theta}_S = W_S^+ Q_S^T(y - X_{0\bar{S}}\hat{\theta}_{\bar{S}})$, where $X_{0S} = Q_S W_S$
            and $Q_S, W_S$ are computed from the QR or SVD decomposition of $X_{0S}$
        $S = \{j \in F : |\hat{\theta}_j| > b_Z\},\ \ \bar{S} = F \setminus S$
        $X_{\diamond\bar{S}} = X_{0\bar{S}} - Q_S(Q_S^T X_{0\bar{S}}),\ \ y_\diamond = y - Q_S(Q_S^T y)$
    endfor
**Output:** $\hat{T}_{MCR} = S$, $\hat{\beta}^{MCR} = \hat{\theta}_S/d$.
```

In the above algorithm $\hat{\theta}_{\bar{S}}$ is the Lasso estimator for the response $y_\diamond$ and the experimental matrix $X_{\diamond\bar{S}}$ and $\hat{\theta}_S$ is the LS estimator with the experimental matrix X_{0S} and the response equal to residuals of the Lasso fit $y - X_{0\bar{S}}\hat{\theta}_{\bar{S}}$. When one of the iterations returns S such that $|S| > n$ then the LS estimator can be calculated using the SVD decomposition instead of the QR decomposition. The above algorithm allows for usage of one of many implementations of the Lasso and is applied in our numerical experiments in Section 8.

5.2 Error Bound for MCR

In order to compare our results with selection error bounds in Zhang (2013), we restate his result using our notation. The proof of its equivalence with the original form is deferred to the Appendix. We stress that the Zhang's result holds for more general case of sub-Gaussian errors whereas we consider Gaussian errors only. Let $c_3 = 2/49$ and recalling that $\Sigma = X_0^T X_0 = d^{-2} X^T X$ and $\Sigma_J = X_{0J}^T X_{0J}$ we define *sparse eigenvalues* of Σ

$$\lambda_s = \min_{J:|J|\leq s} \lambda_{min}(\Sigma_J) = \min_{\nu:supp(\nu)\leq s} \frac{||X_0\nu||^2}{||\nu||^2} = \kappa^2(s,0),$$

$$\Lambda_s = \max_{J:|J|\leq s} \lambda_{max}(\Sigma_J) = \max_{\nu:supp(\nu)\leq s} \frac{||X_0\nu||^2}{||\nu||^2}.$$

Theorem 3 *(Zhang, 2013) Assume that there exist $s \geq 1.5t$ and $a \in (0,1)$ such that*
(i) (sparse eigenvalue condition) $\Lambda_s/\lambda_{1.5t+2s} \leq 1 + s/(1.5t)$ and
(ii) $c_3^{-1}a^{-1}\sigma^2 \log p \leq r_Z^2 \leq b_Z^2 \lambda_{1.5t+s}^2/81 \leq (18)^{-2}\lambda_{1.5t+s}^2 \theta_{min}^{*2}$,
then for $l > \lfloor 1.24 \ln t \rfloor + 1$ we have

$$P(\hat{T}_{MCR} \neq T) \leq \exp\left(-\frac{(1-a)c_3 r_Z^2}{\sigma^2}\right)\left(\frac{\pi c_3 r_Z^2}{\sigma^2}\right)^{-1/2}.$$

Now we compare Theorem 3 with Corollary 2. Both results assume variants of the beta-min condition and bounds on (restricted or sparse) eigenvalues of Σ, namely the weak correlation condition (23) in Corollary 2 and the sparse eigenvalue condition in Theorem 3, which is similar to *restricted isometry property* described in the Introduction. More specifically, observe that according to (14)

$$0 \leq \lambda_{s'} \leq \lambda_s \leq \Lambda_1 = 1 \leq \Lambda_s \leq \Lambda_{s'} \leq s' \wedge n$$

for $1 \leq s < s' \leq p$ and obviously $\lambda_s = 0$ for $s > n$. Then it follows from the sparse eigenvalue condition that $\lambda_{4.5t} \geq \lambda_{1.5t+2s} > 0$ and thus $4.5t \leq n$ whereas the weak correlation condition stipulates that $4t \leq n$. Whence the condition on correlation of predictors assumed in Theorem 3 is stronger than the corresponding assumption in the Corollary 2, moreover, Corollary 1 allows for replications of spurious predictors. However, from Proposition 3 we have $t^{-1/2}\kappa_t^2 < 4\lambda_{4t} \leq 4\lambda_{3t}$ and thus for the minimal allowed $s = 1.5t$ and disregarding constants, Theorem 3 imposes weaker variant of the beta-min condition. It is worth noting that the considered algorithms as well as the error bounds assuming uniform weak correlation of predictors (Corollary 2 and Theorem 3) do not depend on n. Remaining error bounds require explicitly $s \leq n$.

6. Properties of Post-model Selection Estimators

We list now several properties of post-model selection estimators which follow from the main results. Let $\hat{\mathcal{B}} = \mathcal{B}(\hat{T}, y)$ be any event defined in terms of given selector $\hat{T}$ and y and $\mathcal{B} = \mathcal{B}(T, y)$ be an analogous event pertaining to T and y. Let $\mathcal{B}^c$ and $\hat{\mathcal{B}}^c$ be complements of $\mathcal{B}$ and $\hat{\mathcal{B}}$, respectively. Observe that we have

$$P(\hat{\mathcal{B}}) \leq P(\hat{\mathcal{B}}, \hat{T} = T) + P(\hat{T} \neq T) \leq P(\mathcal{B}) + P(\hat{T} \neq T).$$

Analogously, $P(\hat{\mathcal{B}}^c) \leq P(\mathcal{B}^c) + P(\hat{T} \neq T)$, which implies $P(\mathcal{B}) \leq P(\hat{\mathcal{B}}) + P(\hat{T} \neq T)$. Both inequalities yield

$$|P(\hat{\mathcal{B}}) - P(\mathcal{B})| \leq P(\hat{T} \neq T). \tag{28}$$

In particular, when $\mathcal{B} = \{G > u\}$ and $\hat{\mathcal{B}} = \{\hat{G} > u\}$ and G is some pivotal quantity then (28) implies that $P(\hat{\mathcal{B}})$ is approximated by $P(\mathcal{B})$ uniformly in u. For example, let $\hat{\tilde{\beta}}_T$ denote the LS estimator fitted on T, $h = t + 1$ for parametrization (2) and $h = t$ for parametrization (3) and define

$$f = f(T, y) = \frac{||\tilde{X}_T \hat{\tilde{\beta}}_T^{LS} - \tilde{X}_T \tilde{\beta}_T^{*}||^2/h}{||y - \tilde{X}_T \hat{\tilde{\beta}}_T^{LS}||^2/(n-h)}.$$

Observe that the variable f follows a Fisher-Snedecor distribution $\mathcal{F}_{h,n-h}$. Then the bound on the selection error given in Corollary 1, the assumption $\varepsilon \sim N(0, \sigma^2 \mathbb{I}_n)$ and (28) imply the following corollary.

Corollary 4 *Assume that conditions of Corollary 1 are satisfied. Then*

$$\sup_{u \in R} |P(\hat{f} \leq u) - P(f \leq u)| \leq 4 \exp\left(-\frac{(1-a)r}{2\sigma^2}\right)\left(\frac{\pi r}{2\sigma^2}\right)^{-1/2}.$$

Note that any a priori upper bound on h in conjunction with Corollary 4 yields an approximate confidence region for $\tilde{\beta}^*_{\hat{T}}$.

Moreover, it follows from the Corollary 7 below that the Lasso estimator has the following estimation and prediction errors

Corollary 5 *Assume that conditions of Corollary 7 are satisfied. Then*

$$||X\hat{\beta} - X\beta^*|| = O_P\big(t_n^{1/2}\kappa_n^{-1}\sqrt{\log p_n}\big), \qquad |D(\hat{\beta} - \beta^*)| = O_P\big(t_n\kappa_n^{-2}\sqrt{\log p_n}\big),$$

where $\kappa_n = \kappa(T_n, 3)$.

Analogous properties of post-selection estimators are given below without proof for $\lambda_n = \lambda_{min}(\Sigma_{T_n})$.

Corollary 6 *(i) Assume that conditions of Corollary 1 are satisfied. Then*

$$||X\hat{\beta}^{SOS} - X\beta^*|| = O_P\big(t_n^{1/2}\big), \qquad |D(\hat{\beta}^{SOS} - \beta^*)| = O_P\big(t_n\lambda_n^{-1/2}\big),$$

(ii) Assume that conditions of Theorem 3 are satisfied. Then

$$||X\hat{\beta}^{MCR} - X\beta^*|| = O_P\big(t_n^{1/2}\big), \qquad |D(\hat{\beta}^{MCR} - \beta^*)| = O_P\big(t_n\lambda_n^{-1/2}\big),$$

In view of the inequality $\kappa_n^2 < \lambda_n$ it is seen that the estimation and prediction rates for the SOS and MCR post-selection estimators are better by the factor $\kappa_n^{-1}\sqrt{\log p_n}$ than the corresponding rates for the Lasso.

7. Error Bounds for the Lasso Estimator

We assume from now on that the general model (1) holds. Let $\mu_0 = H_0\mu$, $\mu_\beta = H_0X\beta = X_0\theta$ for an arbitrary $\beta \in \mathbf{R}^p$ and $\mu_{\hat{\beta}} = H_0X\hat{\beta} = X_0\hat{\theta}$. Moreover, $\Delta = \hat{\theta} - \theta = D(\hat{\beta} - \beta)$ and recall that Δ_J stands for subvector of Δ restricted to coordinates in J and $J_\beta = \text{supp}(\beta) = \{j : \beta_j \neq 0\}$. Finally let $\mathcal{A} = \bigcap_{j=1}^p \{2|x_{0j}^T\varepsilon| \leq r_L\}$ and $\mathcal{A}^c$ be a complement of $\mathcal{A}$. From the Mill inequality (see the right hand side inequality in (37) below) we obtain for $Z \sim N(0,1)$

$$P(\mathcal{A}^c) \leq \sum_{j=1}^p P(2|x_{0j}^T\varepsilon| > r_L) = pP\Big(Z^2 > \frac{r_L^2}{4\sigma^2}\Big) \leq p\exp\Big(-\frac{r_L^2}{8\sigma^2}\Big)\Big(\frac{\pi r_L^2}{8\sigma^2}\Big)^{-1/2}. \qquad (29)$$

As a by-product of the proofs of the theorems above we state in this section a strengthened version of the Lasso error bounds and their consequences.

Theorem 4 *(i) On $\mathcal{A}$ we have*

$$||\mu_0 - \mu_{\hat{\beta}}|| \leq ||\mu_0 - \mu_\beta|| + 3r_L|J_\beta|^{1/2}\kappa^{-1}(J_\beta, 3). \qquad (30)$$

(ii) Moreover, on the set $\mathcal{A} \cap \{\beta : |\Delta| \leq 4|\Delta_J|\}$ we have

$$r_L|\Delta| \leq 2||\mu_0 - \mu_\beta||^2 + 8r_L^2|J_\beta|\kappa^{-2}(J_\beta, 3). \qquad (31)$$

Squaring both sides of (30) yields the following bound

$$||\mu_0 - \mu_{\hat{\beta}}||^2 \leq \Big(||\mu_0 - \mu_\beta|| + \frac{3r_L|J_\beta|^{1/2}}{\kappa(J_\beta, 3)}\Big)^2 = \inf_{a>0}(1+a)\Big(||\mu_0 - \mu_\beta||^2 + \frac{9r_L^2|J_\beta|}{a\kappa^2(J_\beta,3)}\Big),$$

where the equality above is easily seen. Obviously $\kappa(|J_\beta|, 3) \leq \kappa(J_\beta, 3)$, hence (30) is tighter than Theorem 6.1 in Bickel et al. (2009) if we disregard a small difference in normalization of X mentioned in Section 3. Moreover, the bound above is valid for both the practical and the formal Lasso.

Let us note that as β in (30) is arbitrary, the minimum over all $\beta \in \mathbf{R}^P$ can be taken. Analogously we can minimize the right hand side of (31) over all $\beta : |\Delta| \leq 4|\Delta_J|$. Note also that if a parametric model $\mu = \tilde{X}_J\tilde{\beta}_J$ holds, then (33) below implies that indeed a condition $|\Delta| \leq 4|\Delta_J|$ is satisfied. The next corollary strengthens the ℓ_1 estimation error inequality (7.7) and the predictive inequality (7.8) in Theorem 7.2 in Bickel et al. (2009). Note that X below does not need to have normalized columns and the constant appearing in (7.7) and (7.8) in Bickel et al. (2009) is 16.

Corollary 7 *Let β be such that $\mu_0 = \mu_\beta$. Then (31) and (30) have the following form*

$$|\Delta| \leq 8r_L|J_\beta|\kappa^{-2}(J_\beta, 3) \quad \text{and} \quad ||\mu_{\hat{\beta}} - \mu_\beta||^2 \leq 9r_L^2|J_\beta|\kappa^{-2}(J_\beta, 3). \tag{32}$$

Moreover, we have on $\mathcal{A}$ the following bounds.

Corollary 8

$$||\Delta_J|| \leq 3r_L|J_\beta|^{1/2}\kappa^{-2}(J_\beta, 3) \quad \text{and} \quad |\Delta_J| \leq 3r_L|J_\beta|\kappa^{-2}(J_\beta, 3).$$

8. Simulational Study

In this section we investigate the performance of our implementation of SOS and compare it with MCR. We describe the framework of numerical experiments, discuss their results and draw conclusions. More detailed results are presented in Appendix A.4.

8.1 Description of the Experiments

We consider three models with number of potential predictors p exceeding number of observations n. The first model M_1 was analyzed in Zhang (2013). Beside it we introduce two models M_2 and M_3 which seem to fit even more to the *sparse high-dimensional* scenario $t \ll n \ll p$ and are described in Table 1, columns $1-4$. Observe that sparseness of the model measured by ratio p/t increases from 8.3 for M_1 to 100 for M_2 and to 400 for M_3. Corresponding ratios p/n are 2.5, 10 and 20, respectively. Note also that the assumptions of either Corollary 2 or Theorem 3 are not satisfied for M_1 as $4t > n$, whereas two remaining models satisfy $10t \leq n$. In all simulations the $n \times p$ matrix of experiment X with iid standard normal entries is generated and then its columns are normalized to have ℓ_2-norm equal to $\sqrt{n}$. A noise level is specified by $\sigma = 1$. For each replication of the true model, elements of β_T^* are independently generated from uniform distribution with parameters given in the column 5 of Table 1. Such layout resulted in signal to noise ratio $SNR = ||X_T\beta_T^*||/\sqrt{\mathbf{E}||\varepsilon||^2} = ||X_T\beta_T^*||/\sqrt{n}$ and it values averaged over replications are given in column 6 of Table 1.

model	t	n	p	β^*_T	SNR	SOS accuracy	MCR accuracy
M_1	30	100	250	$U(1,10)$	33	72 / 71	56 / 97
M_2	10	100	1000	$U(1,10)/2$	9.5	91 / 88	73 / 82
M_3	5	100	2000	$U(1,10)/3$	4.5	85 / 77	69 / 73

Table 1: Summary of the simulations (details explained in the text).

All computations have been performed using open source software R (see supplemental material at `http://www.mimuw.edu.pl/~pokar/Publications/`) using two frequently used Lasso implementations: `lars` (Efron et al., 2004) and `glmnet` (Friedman et al., 2010). Preliminary experiments indicated that using `lars` yields higher selection accuracies for SOS as well as for MCR than when using `glmnet`; even on grids of order 10^5 the gain in accuracy was around 10%. Moreover, for such dense grids `glmnet` was considerably slower. Thus in main numerical experiments `lars` has been used. We established that accuracy of SOS for all models is the highest when $r \approx 20$ and thus the value of r is fixed at 20. The MCR procedure is implemented via the Lasso and LS as described in Section 5.1. Similarly to Zhang (2013) we fixed number of iterations $l = 8$ for MCR. Thus compared algorithms have mutually corresponding parameters (r_L, b) and (r_Z, b_Z). As in Zhang (2013) we found optimal grid parameters for which selection accuracy is the highest one. In particular we confirmed high selection accuracy for the best parameters shown in Table 1 in Zhang (2013). Namely, the highest selection accuracy of MCR reported there is 93% for penalty and the threshold both equal 0.94 whereas we found selection accuracy 95% for both these parameters equal to 5. The difference is minor taking into account that the original penalty in Zhang (2013) corresponds in our implementation to $2r_Z/\sqrt{n} = r_Z/5$.

As a measure of performance of both algorithms we present in columns $7-8$ of Table 1 a percent of correct screening and percent of correct selection separated by the slash that is $100 \times \hat{P}(T \subseteq S)$ / $100 \times \hat{P}(\hat{T} = T)$. In simulations for the SOS algorithm, we used as a screening set $S = S_0 = \{j : |\hat{\theta}_j| > b\}$, since a double-pass screening S_1 does not lead to significant improvement of selection accuracy. Similarly, for MCR we considered as a screening set $S = \{j : |\hat{\theta}_j| > b_Z\}$ after the first iteration of the algorithm. Knowledge of both screening and selection errors allows us to estimate errors pertaining to ordering and greedy selection for SOS as well as advantage of MCR over the thresholded Lasso. Note that algorithms behave differently in that whereas for MCR probability of correct selection is larger than that of screening after the first iteration, the opposite is true for SOS. Both measures for all grid parameters are reported in Appendix A.4.

All results are based on $N = 5000$ replicates as for estimation a success probability $\pi \approx 0.75$ (corresponding crudely to our selection accuracies) in N Bernoulli experiments with prescribed error $\eta = 0.01$ and confidence level $1 - \gamma = 0.9$, we need $N \approx \pi(1 - \pi)\eta^{-2}(\Phi^{-1}(1 - \gamma/2))^2 \approx 5000$, where Φ^{-1} denotes the quantile function of the standard normal distribution.

8.2 Conclusions from the Experiments

Computing time of both SOS and MCR is dominated by calls to the `lars` function which is used to compute the Lasso, and as MCR uses $l = 8$ calls of this function and SOS only one, so MCR is around eight time slower than SOS.

For model M_1, MCR is substantially more precise then SOS in selecting the true subset of variables: 97% versus 71%. Recall that the highest accuracy given in Zhang (2013) is 93%. The SOS selection error is mostly due to the screening error of the Lasso as in the case of relatively large number of true predictors compared to n, the Lasso finds it difficult to filtering in all of them.

For models M_2 and M_3, SOS is more precise than MCR by approximately 5%. We note that optimal grid penalty r_L for SOS and MCR coincide whereas the threshold b is approximately twice as large for MCR as for SOS. As the results for SOS are better in these cases it turns out that thresholding the Lasso, ranking the remaining estimators and optimizing GIC in the nested family is superior to MCR iterations performed on the same initial Lasso estimator.

In conclusion, if we expect large number of genuine predictors compared to sample size, MCR is preferable, but for the sparse high-dimensional scenario SOS may be faster and more accurate.

For practical model selection we recommend the following easily achievable strategy. After performing the Lasso, we look at the paths of parameters and choose only those whose magnitude is substantially larger than others. This yields screening set S on which LS is computed, and then screened regressors are ordered according to their $|t|$ statistics from the fit. Finally we look for an 'elbow' of R_J in the nested family of the models $J \in \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, S\}$ which determines a cut-off point.

9. Concluding Remarks

We introduce the three-step SOS algorithm for a linear model selection. The most computationally demanding part of the method is screening of predictors by the Lasso. Ordering and greedy GIC could be computed using only two QR decompositions of X_{0S_1}. In the paper we give non-asymptotic upper bounds on error probabilities of each step of SOS in terms of the Lasso and GIC penalties (Theorem 1). As corollaries we obtain selection consistency for different (n, p) scenarios under conditions which are needed for screening consistency of the Lasso (Corollaries 1-2). The SOS algorithm is an improvement of the new version of the thresholded Lasso (Zhou, 2009, 2010) and turns out to be competitive for MCR, the latest quasiconvex penalized LS (Zhang, 2010b, 2013). The condition on correlation of predictors assumed there seems to be stronger than ours, whereas the beta-min condition may be weaker (compare discussion of Corollary 2 and Theorem 3). Theoretical comparison of SOS and MCR, in general, requires comparing λ_{3t} and $\kappa^2(T, 3)$ and remains an open problem. In simulations for the sparse high-dimensional scenario, SOS was faster and more accurate than MCR. For a traditional setting when $n > p$ we give Sanov-type bounds on error probabilities of the OS algorithm (Theorem 2). It is surprising consequence of Theorems 1-2 that the selection error of greedy GIC is asymptotically not larger than of exhaustive GIC, see formula (25). Comparison of selection errors probabilities of the SOS and OS algorithms for $p < n$ requires further research.

It is worth noticing that all results are proved for general form of the Lasso defined in (6), which encompasses two versions of the estimator: algorithm used in practice as well as its formal version.

Acknowledgments

We appreciate comments of two referees which greatly contributed to improving of the original manuscript.

Appendix A: Proofs and Supplemental Tables.

In the Appendix we provide all proofs and supplemental tables for numerical experiments.

A.1 Proofs for Section 3.

Proof of Proposition 1. We have

$$2\sigma^2 KL(\tilde{\beta}_T^*||\tilde{\beta}_J) = 2\sigma^2 \mathbf{E}_{\tilde{\beta}_T^*}\left(\frac{||y-\tilde{X}_J\tilde{\beta}_J||^2 - ||y-\tilde{X}_T\tilde{\beta}_T^*||^2}{2\sigma^2}\right) = ||\tilde{X}_T\tilde{\beta}_T^* - \tilde{X}_J\tilde{\beta}_J||^2.$$

The last expression is symmetric with respect to $\tilde{\beta}_T^*$ and $\tilde{\beta}_J$, thus $KL(\tilde{\beta}_T^*||\tilde{\beta}_J) = KL(\tilde{\beta}_J||\tilde{\beta}_T^*)$ and the second equality in (i) follows. For the proof of the first equality in (i) observe that $\delta(T||J) = \min_{\tilde{\beta}_J} ||\tilde{X}_T\tilde{\beta}_T^* - \tilde{X}_J\tilde{\beta}_J||^2$. The equality in (ii) follows from (10), the inequality there follows from Rayleigh-Ritz theorem. ■

Proof of Proposition 3. We can assume that $c \geq 1$. Consider a model J and a vector ν such that $J \supseteq \mathrm{supp}(\nu)$ and $|J| = (\lfloor c \rfloor + 1)s$ and $\kappa^2(\lfloor c \rfloor + 1)s, 0) = \nu^T\Sigma\nu/\nu^T\nu$. Sort coordinates of ν in nonincreasing order $|\nu_{j_1}| \geq |\nu_{j_2}| \ldots \geq |\nu_{j_{(\lfloor c \rfloor+1)s}}|$ and let $J_0 = \{j_1, \ldots, j_s\}$. Then we have $|J_0| = s$, $|\nu_{\bar{J}_0}| \leq \lfloor c \rfloor |\nu_{J_0}| \leq c|\nu_{J_0}|$ and $(\lfloor c \rfloor + 1)\nu_{J_0}^T\nu_{J_0} \geq \nu^T\nu$. Thus

$$\kappa^2(s,c) \leq \frac{\nu^T\Sigma\nu}{\nu_{J_0}^T\nu_{J_0}} \leq (\lfloor c \rfloor + 1)\frac{\nu^T\Sigma\nu}{\nu^T\nu} = (\lfloor c \rfloor + 1)\kappa^2((\lfloor c \rfloor + 1)s, 0)$$

and the conclusion follows. ■

Proof of Proposition 4. Assume by contradiction that there are two different true models T_1, T_2 such that $T_i = \mathrm{supp}(\beta_i) = \mathrm{supp}(\theta_i)$ for some different $\beta_i = D\theta_i$, $i = 1, 2$ and $\mu_0 = X_0\theta_1 = X_0\theta_2$. It is enough to prove that assumptions imply $\gamma(T_1, 1)\gamma(T_2, 1) = 0$, where $\gamma(J, c) = \inf\{||X_0\theta_J - X_0\theta_{\bar{J}}||, |\theta_J| = 1, |\theta_{\bar{J}}| \leq c\}$ as in view of (13) and Schwarz inequality $\kappa(J, c)/\sqrt{|J|} \leq \gamma(J, c)$. Define a vector θ with support equal to $T_1 \cup T_2$ in such a way that $\theta_{T_1 \cap T_2} = \theta_{T_1 \cap T_2, 1} - \theta_{T_1 \cap T_2, 2}$, $\theta_{T_1 \setminus T_2} = \theta_{T_1 \setminus T_2, 1}$ and $\theta_{T_2 \setminus T_1} = \theta_{T_2 \setminus T_1, 2}$. As assumptions on T_1 and T_2 are symmetric we may assume that $|\theta_{T_1 \setminus T_2}| \geq |\theta_{T_2 \setminus T_1}|$ and let $\theta^o = \theta/|\theta_{T_1}|$. Then $|\theta^o_{T_1}| = 1$ and $|\theta^o_{\bar{T}_1}| = |\theta^o_{T_2 \setminus T_1}| \leq 1$. Moreover, $X\theta^o_{T_1} = X\theta^o_{\bar{T}_1}$ which yields $\gamma(T_1, 1) = 0$. ■

Proof of Proposition 5. To prove (i) observe that (11) and (14) imply for $j \in T$

$$\kappa^2(T, 3) \leq \kappa^2(T, 0) \leq \theta_j^{*-2}\delta(T \parallel T \setminus \{j\}).$$

For (ii) we have

$$\begin{aligned}
\kappa^2(t,3)/4 &\leq \kappa^2(4t,0) = \min_{J:|J|\leq 4t} \lambda_{min}(\Sigma_J) \leq \min_{J:J\supseteq T,|J|\leq 4t} \lambda_{min}(\Sigma_J) \\
&= \min_{J:J\not\supseteq T,|J\cup T|\leq 4t} \lambda_{min}(\Sigma_{J\cup T}) \leq \theta^{*-2}_{min} \min_{J:J\not\supseteq T,|J\cup T|\leq 4t} \delta(T||J) \\
&\leq \theta^{*-2}_{min} \min_{j\in T,J\supseteq T,|J|\leq 4t} \delta(T||J\setminus\{j\}) = \theta^{*-2}_{min}\delta(T,4t),
\end{aligned}$$

where the first inequality follows from the Proposition 3 and the third from (11). ■

A.2 Proofs for Section 6.

We now proceed to prove Theorem 4 and its corollaries. The following modified version of Lemma 1 in Bunea et al. (2007) holds.

Lemma 1 *(i) We have on $\mathcal{A}$ for an arbitrary $\beta \in \mathbf{R}^p$ and $J = \{j : \beta_j \neq 0\}$*

$$||\mu_0 - \mu_{\hat{\beta}}||^2 + r_L|\Delta| \leq ||\mu_0 - \mu_\beta||^2 + 4r_L|\Delta_J|. \tag{33}$$

(ii) Moreover, we have

$$||\mu_0 - \mu_{\hat{\beta}}||^2 \leq ||\mu_0 - \mu_\beta||^2 + 3r_L|\Delta_J|. \tag{34}$$

Proof. It follows from (6) that

$$||H_0(\varepsilon + \mu - X\hat{\beta})||^2 + 2r_L|D\hat{\beta}| \leq ||H_0(\varepsilon + \mu - X\beta)||^2 + 2r_L|D\beta|.$$

Equivalently, as H_0 is symmetric and idempotent, we get

$$||H_0(\mu - X\hat{\beta})||^2 \leq ||H_0(\mu - X\beta)||^2 + 2\varepsilon^T H_0 X(\hat{\beta} - \beta) + 2r_L(|D\beta| - |D\hat{\beta}|).$$

Thus we obtain *the basic inequality*

$$||\mu_0 - \mu_{\hat{\beta}}||^2 \leq ||\mu_0 - \mu_\beta||^2 + 2\varepsilon^T X_0(\hat{\theta} - \theta) + 2r_L(|\theta| - |\hat{\theta}|).$$

On $\mathcal{A}$ we have $|2\varepsilon^T X_0(\hat{\theta} - \theta)| \leq 2\max_j |x_{0j}^T\varepsilon||\hat{\theta} - \theta| \leq r_L|\hat{\theta} - \theta|$ and whence on this set

$$||\mu_0 - \mu_{\hat{\beta}}||^2 + r_L|\hat{\theta} - \theta| \leq ||\mu_0 - \mu_\beta||^2 + 2r_L(|\hat{\theta} - \theta| + |\theta| - |\hat{\theta}|).$$

Note that for $j \notin J$ $|\hat{\theta}_j - \theta_j| + |\theta_j| - |\hat{\theta}_j| = 0$ and thus

$$||\mu_0 - \mu_{\hat{\beta}}||^2 + r_L|\hat{\theta} - \theta| \leq ||\mu_0 - \mu_\beta||^2 + 2r_L(|\hat{\theta}_J - \theta_J| + |\theta_J| - |\hat{\theta}_J|).$$

Thus (i) follows from triangle inequality and (ii) from (i) in view of $|\hat{\theta}_J - \theta_J| \leq |\hat{\theta} - \theta|$. ■

Proof of Theorem 4. Proof of (i). Let $J = J_\beta$ and $\kappa = \kappa(J,3)$. We consider two cases: (a) $|\Delta| > 4|\Delta_J|$ and (b) $|\Delta| \leq 4|\Delta_J|$. In the case (a) it follows from (33) that stronger inequality $||\mu_0 - \mu_{\hat{\beta}}|| \leq ||\mu_0 - \mu_\beta||$ holds. When (b) is satisfied we have $|\Delta_{\bar{J}}| \leq 3|\Delta_J|$ and it follows from the definition of κ that $\kappa^2 ||\Delta_J||^2 \leq ||X_0 \Delta||^2 = ||\mu_{\hat{\beta}} - \mu_\beta||^2$ and thus

$$||\Delta_J|| \leq ||\mu_{\hat{\beta}} - \mu_\beta|| \kappa^{-1}. \tag{35}$$

Using (35) and Jensen inequality we get

$$|\Delta_J| \leq |J|^{1/2} ||\mu_{\hat{\beta}} - \mu_\beta|| \kappa^{-1}. \tag{36}$$

It follows now from (34), (36) and triangle inequality that

$$||\mu_0 - \mu_{\hat{\beta}}||^2 \leq ||\mu_0 - \mu_\beta||^2 + 3 r_L |J|^{1/2} \kappa^{-1} (||\mu_0 - \mu_{\hat{\beta}}|| + ||\mu_0 - \mu_\beta||)$$

and whence

$$(||\mu_0 + \mu_{\hat{\beta}}|| + ||\mu_0 - \mu_\beta||)(||\mu_0 - \mu_{\hat{\beta}}|| - ||\mu_0 - \mu_\beta||) \leq 3 r_L |J|^{1/2} \kappa^{-1} (||\mu_0 - \mu_{\hat{\beta}}|| + ||\mu_0 - \mu_\beta||)$$

from which the conclusion follows.

Proof of (ii). Define $m = ||\mu_0 - \mu_\beta||$, $\hat{m} = ||\mu_0 - \mu_{\hat{\beta}}||$ and $c = 2 r_L |J|^{1/2} \kappa^{-1}$. Using (33), (36) which holds provided $|\Delta| \leq 4|\Delta_J|$, and triangle inequality we get

$$\hat{m}^2 + r_L |\Delta| \leq m^2 + 2c(\hat{m} + m) \leq 2m^2 + c^2 + \hat{m}^2 + c^2,$$

from which the desired bound follows. ■

Proof of Corollary 8. The proof follows from inequality (35), (36) and the second inequality in Corollary 7. ■

A.3 Proofs for Section 4.

The next lemma states bounds on upper tail of χ^2_k distribution

Lemma 2 *Let W_k denote variable having χ^2_k distribution.(i) (Gordon, 1941 and Mill, 1926) We have for $k = 1$ and $x > 0$*

$$w_{xk} l_{xk} \leq P(W_k \geq x) \leq w_{xk}, \tag{37}$$

where $w_{xk} = e^{-x/2} (\frac{x}{2})^{k/2-1} \Gamma^{-1}(\frac{k}{2})$ and $l_{xk} = \frac{x}{x-k+2}$.
(ii) (Inglot and Ledwina, 2006) Let $k > 1$ and $x > k - 2$. Then

$$w_{xk} \leq P(W_k \geq x) \leq w_{xk} l_{xk}. \tag{38}$$

Proof. We provide the unified reasoning for both cases. For $x > 0$ and $k \in \mathbf{Z}$ let $I_k(x) = \int_x^\infty t^{(k/2)-1} e^{-t/2}\, dt$. Integration by parts yields

$$I_k(x) = 2x^{(k/2)-1} e^{-x/2} + (k-2) I_{k-2}(x). \tag{39}$$

It is easy to see that the following inequalities hold for $x > 0$ and $k \in \mathbf{Z}$

$$0 \leq I_{k-2}(x) \leq I_k(x)/x. \tag{40}$$

We treat cases $k = 1$ and $k > 1$ separately, as $k = 1$ is the only integer for which the second term on the RHS of (39) is negative. Dividing both sides of (39) by $2^{k/2}\Gamma(k/2)$, noting that the LHS is then $P(W_k \geq x)$ and using (40) we have for $k = 1$ and $x > 0$

$$P(W_k \geq x) \leq e^{-x/2}\Big(\frac{x}{2}\Big)^{-1/2}\Gamma^{-1}\Big(\frac{1}{2}\Big)$$

and

$$P(W_k \geq x) \geq e^{-x/2}\Big(\frac{x}{2}\Big)^{-1/2}\Gamma^{-1}\Big(\frac{1}{2}\Big)\Big(1 - \frac{1}{1+x}\Big),$$

which proves (37). Analogously for $k = 2, 3, \ldots$ we obtain from (39) inequalities proved by Inglot and Ledwina (2006)

$$P(W_k \geq x) \leq e^{-x/2}\Big(\frac{x}{2}\Big)^{k/2-1}\Gamma^{-1}\Big(\frac{k}{2}\Big)\Big(1 + \frac{k-2}{x-k+2}\Big)$$

for $x > k-2$, and for $x > 0$

$$P(W_k \geq x) \geq e^{-x/2}\Big(\frac{x}{2}\Big)^{k/2-1}\Gamma^{-1}\Big(\frac{k}{2}\Big),$$

which proves (38). ∎

Now we state the main lemma from which Theorems 1 and 2 follow. Let us recall that $c_1 = (3 + 6\sqrt{2})^{-1}$ and $c_2 = (6 + 4\sqrt{2})^{-1}$. Define $\mathcal{T}_n^o = \mathcal{T}_n \setminus \{T\}$ and observe that for OS algorithm we have $P(S_1 \notin \mathcal{T}_n) = 0$ and as $p \geq t+1$, $\mathcal{T}_n = \mathcal{T}_n^o = \{F\}$, so $|\mathcal{T}_n^o| = 1$.

Lemma 3 *(T1) If $r_L^2 \leq b^2/36 \leq c_1^2 t^{-1}\kappa^4\theta_{min}^{*2}$, then*

$$P(S_1 \notin \mathcal{T}_n) \leq p\exp\Big(-\frac{r_L^2}{8\sigma^2}\Big)\Big(\frac{\pi r_L^2}{8\sigma^2}\Big)^{-1/2}.$$

(T2) If $s \leq n$, then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \notin O_{S_1}) \leq \frac{3}{2}|\mathcal{T}_n^o|t(s-t)\exp\Big(-\frac{c_2\delta_s}{\sigma^2}\Big)\Big(\frac{\pi c_2\delta_s}{\sigma^2}\Big)^{-1/2}.$$

(T3) If for some $a \in (0,1)$ $r \leq at^{-1}\delta_t$, then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| < t) \leq \frac{t}{2}\exp\Big(-\frac{(1-a)^2\delta_t}{8\sigma^2}\Big)\Big(\frac{\pi(1-a)^2\delta_t}{8\sigma^2}\Big)^{-1/2}.$$

(T4) Assume that $r/\sigma^2 \geq 2$ and $(r/\sigma^2) - \log(r/\sigma^2) \geq 2\log p$. Then

$$P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| > t) \leq (p-t)(s-t)\exp\Big(-\frac{r}{2\sigma^2}\Big)\Big(\frac{\pi r}{2\sigma^2}\Big)^{-1/2}.$$

Proof. Observe that we may assume that $t > 0$ in proofs of $(T2) - (T3)$ as for $t = 0$ probabilities appearing in those parts are 0 and the conclusions are trivially satisfied.

Proof of (T1). It follows from (29) or equivalently from Lemma 2 that it is enough to prove that $\{S_1 \in \mathcal{T}_n\} \supseteq \mathcal{A}$ that is that on $\mathcal{A}$ we have

$$T \subseteq S_1 \qquad \text{and} \qquad |S_1| \le t + \lfloor \sqrt{t}\kappa^{-2} \rfloor. \tag{41}$$

For parametric models $\mu_\beta = \mu_0$ and from (33) we have $|\Delta| \le 4|\Delta_T|$ or equivalently $4|\Delta_{\bar{T}}| \le 3|\Delta|$, which together with the first part of (32) yields

$$|\Delta_{\bar{T}}| \le 6 r_L t \kappa^{-2}. \tag{42}$$

From the assumption $6r_L \le b$ and (42) we obtain $|S_0 \setminus T| < |\Delta_{\bar{T}}|/b \le t\kappa^{-2}$, $|S_0| < t(1+\kappa^{-2})$ and $B < b\sqrt{t(1+\kappa^{-2})}$. Using this and the first part of Corollary 8 we have $||\Delta_T|| + B < \theta^*_{min}$ or

$$||\Delta_T||^2 < (\theta^*_{min} - B)^2.$$

Indeed, from Corollary 8, the fact that $\kappa \le 1$ and the assumption of the lemma, respectively, we have

$$\begin{aligned} ||\Delta_T|| + B < 3r_L t^{1/2}\kappa^{-2} + b\sqrt{t(1+\kappa^{-2})} \le 0.5 b t^{1/2}\kappa^{-2}(1 + 2\sqrt{\kappa^4 + \kappa^2}) \\ \le 0.5(1+2\sqrt{2}) b t^{1/2}\kappa^{-2} = (6c_1)^{-1} b t^{1/2}\kappa^{-2} \le \theta^*_{min}. \end{aligned}$$

Evidently, $|T \setminus S_1|(\theta^*_{min} - B)^2 \le ||\Delta_T||^2 < (\theta^*_{min} - B)^2$ and thus we have $T \subseteq S_1$ on $\mathcal{A}$. But $S_1 \subseteq S_0$, hence $|S_0| \ge t$ and $B \ge b t^{1/2}$. Thus using (42) again, we have $|S_1 \setminus T| < |\Delta_{\bar{T}}|/B \le t^{1/2}\kappa^{-2}$. Hence $|S_1 \setminus T| \le \lfloor t^{1/2}\kappa^{-2} \rfloor$ and we obtain (41).

Proof of (T2). Let for $J_1 \in \mathcal{S}_n \setminus \mathcal{T}_n$ and $J_2 \in \mathcal{T}_n$ $W_{J_1 J_2} = \varepsilon^T(\tilde{H}_{J_1} - \tilde{H}_{J_1 \cap J_2})\varepsilon$, $\sigma^2 W_{J_2 J_1} = \varepsilon^T(\tilde{H}_{J_2} - \tilde{H}_{J_1 \cap J_2})\varepsilon$ and $\sigma Z_{J_1} = \tilde{\beta}^{*T}_T \tilde{X}^T_T (I - \tilde{H}_{J_1})\varepsilon / \sqrt{\delta_{J_1}}$, where $\delta_{J_1} = \delta(T \parallel J_1)$. Then we have that $W_{J_1 J_2} \sim \chi^2_d$, where $d \le |J_1 \setminus J_2|$, $W_{J_2 J_1} \ge 0$ and $Z_{J_1} \sim N(0,1)$. We will use a popular decomposition of a difference between sums of squared residuals

$$\begin{aligned} R_{J_1} - R_{J_2} &= \tilde{\beta}^{*T}_T \tilde{X}^T_T (I - \tilde{H}_{J_1}) \tilde{X}_T \tilde{\beta}^*_T + 2\tilde{\beta}^{*T}_T \tilde{X}^T_T (I - \tilde{H}_{J_1})\varepsilon \\ &+ \varepsilon^T (I - \tilde{H}_{J_1})\varepsilon - \varepsilon^T (I - \tilde{H}_{J_2})\varepsilon \\ &= \delta_{J_1} + 2\sqrt{\delta_{J_1}}\sigma Z_{J_1} - \sigma^2 W_{J_1 J_2} + \sigma^2 W_{J_2 J_1} \\ &\ge \delta_{J_1}\Big(1 + \frac{2\sigma Z_{J_1}}{\sqrt{\delta_{J_1}}} - \frac{\sigma^2 W_{J_1 J_2}}{\delta_{J_1}}\Big). \end{aligned}$$

For fixed $S \in \mathcal{T}^o_n$ let $\bar{j} = S \setminus \{j\}$. Then we have from (9)

$$\begin{aligned} \{S_1 \in \mathcal{T}^o_n, \hat{O} \notin O_{S_1}\} &\subseteq \bigcup_{S \in \mathcal{T}^o_n} \bigcup_{j_1 \in T} \bigcup_{j_2 \in S \setminus T} \{R_{\bar{j}_1} \le R_{\bar{j}_2}\} \\ &\subseteq \bigcup_{S \in \mathcal{T}^o_n} \bigcup_{j_1 \in T} \bigcup_{j_2 \in S \setminus T} \Big\{ - \frac{2\sigma Z_{\bar{j}_1}}{\sqrt{\delta_{\bar{j}_1}}} + \frac{\sigma^2 W_{\bar{j}_1 \bar{j}_2}}{\delta_{\bar{j}_1}} \ge 1 \Big\}, \end{aligned}$$

where $Z_{\bar{j}_1} \sim N(0,1)$ and $W_{\bar{j}_1\bar{j}_2} \sim \chi^2_d$, with $d \leq 1$. Thus it follows that for $W = Z^2$ denoting r.v. with χ^2_1 distribution, we get

$$\begin{aligned} P(S_1 \in \mathcal{T}_n^o, \hat{O} \notin O_{S_1}) &\leq \sum_{S \in \mathcal{T}_n^o} \sum_{j_1 \in T} \sum_{j_2 \in S \setminus T} P\Big(- \frac{2\sigma Z_{\bar{j}_1}}{\sqrt{\delta_{\bar{j}_1}}} + \frac{\sigma^2 W_{\bar{j}_1\bar{j}_2}}{\delta_{\bar{j}_1}} \geq 1 \Big) \\ &\leq \sum_{S \in \mathcal{T}_n^o} \sum_{j_1 \in T} \sum_{j_2 \in S \setminus T} \Big(P\Big(- \frac{2\sigma Z_{\bar{j}_1}}{\sqrt{\delta_{\bar{j}_1}}} \geq c \Big) + P\Big(\frac{\sigma^2 W_{\bar{j}_1\bar{j}_2}}{\delta_{\bar{j}_1}} \geq 1 - c \Big) \Big) \\ &\leq |\mathcal{T}_n^o| t(s-t) \Big(\frac{1}{2} P\Big(Z^2 \geq \frac{c^2 \delta_s}{4\sigma^2} \Big) + P\Big(W \geq \frac{(1-c)\delta_s}{\sigma^2} \Big) \Big), \end{aligned}$$

where $j_1 \in T$ and $j_2 \in S \setminus T$ are fixed and we used $\delta_{\bar{j}_1} \geq \delta_s$. Choosing c such that $c^2/4 = 1-c$ that is $c = 1 - 2c_2$ in view of Lemma 2 we get the desired bound.

Proof of (T3). Reasoning as previously we have for $\bar{j} = T \setminus \{j\}$

$$\{S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| < t\} \subseteq \bigcup_{S \subset T} \{R_S + r|S| \leq R_T + r|T|\} \subseteq \bigcup_{j \in T} \{R_{\bar{j}} \leq R_T + rt\}.$$

Thus in view of Lemma 2 and the assumption $rt < a\delta_t$ we obtain

$$\begin{aligned} P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| < t) &\leq \sum_{j \in T} P(R_{\bar{j}} \leq R_T + rt) \\ &\leq \sum_{j \in T} P\Big(- 2\sigma Z_{\bar{j}} \geq \sqrt{\delta_{\bar{j}}} \Big(1 - \frac{rt}{\delta_{\bar{j}}} \Big) \Big) \\ &\leq tP\Big(- 2\sigma Z \geq \sqrt{\delta_t} \Big(1 - \frac{rt}{\delta_t} \Big) \Big) \\ &= \frac{t}{2} P\Big(W \geq \frac{1}{4\sigma^2} \delta_t \Big(1 - \frac{rt}{\delta_t} \Big)^2 \Big) \\ &\leq \frac{t}{2} \exp\Big(- \frac{(1-a)^2 \delta_t}{8\sigma^2} \Big) \Big(\frac{\pi(1-a)^2 \delta_t}{8\sigma^2} \Big)^{-1/2}. \end{aligned}$$

Proof of (T4). Observe first that for $m > 0$

$$\begin{aligned} &P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| = t + m) \\ &\leq P(R_{T \cup \{j_1, \ldots, j_m\}} + (t+m)r \leq R_T + tr \text{ for some } j_1, \ldots, j_m \in F \setminus T) \\ &\leq \binom{p-t}{m} P(\sigma^2 W_m \geq mr) \leq \frac{(p-t)^m}{m!} P(\sigma^2 W_m \geq mr) = B_m, \end{aligned}$$

where $W_m \sim \chi^2_m$. This follows since for any fixed $J = T \cup \{j_1, \ldots, j_m\}$ we have $R_T - R_J \sim \sigma^2 \chi^2_d$, where $d \leq m$ and $W_d \leq W_m$ in stochastic order. We will show that under conditions given in (T4) $B_m \geq B_{m+1}$ for any $m = 1, 2, \ldots$ thus yielding

$$P(S_1 \in \mathcal{T}_n, \hat{O} \in O_{S_1}, |\hat{T}| \geq t + m) \leq (s - t - m + 1) B_m,$$

which for $m = 1$ coincides with the desired inequality. Let $Q_m = B_m / B_{m+1}$, $\bar{r} = r/\sigma^2$ and observe that for $m > 1$ we have in view of (38) (note that $m\bar{r} \geq m - 2$ as $\bar{r} \geq 2$)

$$Q_m \geq \frac{m+1}{p} e^{\bar{r}/2} \Big(\frac{m}{m+1} \Big)^{m/2-1} \frac{1}{\big((m+1)\bar{r}/2\big)^{1/2}} \frac{\Gamma((m+1)/2)}{\Gamma(m/2)} \frac{(m+1)\bar{r} - m + 1}{(m+1)\bar{r}}.$$

Using the inequality for gamma functions (cf. formula 2.2 in Laforgia, 1984)

$$\Gamma\Big(\frac{m+1}{2}\Big)\Big/\Gamma\Big(\frac{m}{2}\Big) \geq \Big(\frac{m-1/2}{2}\Big)^{1/2}$$

we have that

$$Q_m \geq \exp\Big\{\frac{\bar{r}}{2} - \frac{1}{2}\log\bar{r} - \log p\Big\} f_1(m,\bar{r}),$$

where

$$f_1(m,\bar{r}) = \Big(\frac{m}{m+1}\Big)^{m/2-1}(m+1)^{1/2}2^{1/2}\Big(\frac{m-1/2}{2}\Big)^{1/2}\frac{(m+1)\bar{r}-m+1}{(m+1)\bar{r}}.$$

Thus in order to show that $Q_m \geq 1$ for $m > 1$ in view of assumptions it is enough to show that $f_1(m,\bar{r}) > 1$. As $f(m,\cdot)$ is increasing, it suffices to check that $f_1(m,2) > 1$. Let $f_2(m) = (\frac{m-1/2}{m+1})^{(m-1)/2}(\frac{m+3}{2})$. We have $f_1(m,2) > f_2(m)$ and $f_2(2) > 1$ thus it is enough to show that f_2 is increasing. Let

$$f_3(m) = \log(2f_2(m)) = \frac{m-1}{2}\log\frac{m-1/2}{m+1} + \log(m+3).$$

We have that

$$\begin{aligned} f_3'(m) &= \frac{1}{2}\log\frac{m-1/2}{m+1} + \frac{m-1}{2}\frac{m+1}{(m-1/2)}\frac{3}{2(m+1)^2} + \frac{1}{m+3} \\ &\geq \frac{1}{2}\frac{-3}{-3+2(m+1)} + \frac{3(m-1)}{4(m-1/2)(m+1)} + \frac{1}{m+3}, \end{aligned}$$

where the last inequality follows from $\log(1+x) > x/(1+x)$ for $x > -1$. As $1/(m+3) \geq 3/(-6+2(m+1))$ it follows that $f_3' > 0$ which implies that f_3 and thus f_2 is increasing. ■

Proof of Theorem 1. The result readily follows from Lemma 3. For (T1) we observe that

$$-\frac{r_L^2}{8\sigma^2} + \log p \leq -\frac{(1-a)r_L^2}{8\sigma^2}$$

is equivalent to $8\sigma^2 a^{-1}\log p \leq r_L^2$. Similar reasoning yields (T4). Consider derivation of (T2). From the bound

$$|\mathcal{T}_n^o| = |\mathcal{T}_n| - 1 = \sum_{k=1}^{s-t}\binom{p-t}{k} \leq (p-t) + \ldots + \frac{(p-t)^{s-t}}{(s-t)!} \leq \frac{(p-t)^{s-t}}{(s-t)!}(s-t)$$

it follows that $|\mathcal{T}_n^o|t(s-t) \leq (p-t)^{s-t}t(s-t) \leq p^{s-t}t(s-t)$. Thus the bound in (T2) will follow from $-c_2\delta_n/\sigma^2 + (s-t)\log p + \log(s-t) + \log t \leq -c_2(1-a)\delta_s/\sigma^2$ which is implied by $(s-t+2)\log p \leq c_2 a\delta_s/\sigma^2$. For (T3) we observe that

$$-\frac{(1-a)^2\delta_t}{8\sigma^2} + \log t \leq -\frac{(1-a)^3\delta_t}{8\sigma^2}$$

is equivalent to $8\sigma^2\log t \leq (1-a)^2 a\delta_t$. ■

Proof of Corollary 1. We proceed by showing that assumptions (i) and (ii) imply all assumptions of Theorem 1. We first note that (i) with the assumption $r_L^2 = 4r$ is stronger than the assumption in Theorem 1 (T1). Next, observe that condition

$$4a^{-1}\sigma^2 \log p \leq (4c_2/3)t^{-1/2}\kappa^2\delta_s \tag{43}$$

is stronger than the assumption in Theorem 1 (T2). Indeed, as $\kappa \leq 1 \leq t$ we have

$$s - t + 2 = \lfloor t^{1/2}\kappa^{-2} \rfloor + 2 \leq t^{1/2}\kappa^{-2} + 2 \leq 3t^{1/2}\kappa^{-2}.$$

Obviously, left inequalities in (i) and (ii) imply (43). Moreover, the assumption of Theorem 1 (T4) is satisfied. Furthermore, from the first $\kappa - \delta$ inequality (15) and assumption $a \in (0, 1 - c_1)$ we obtain that (i) is stronger than both conditions in Theorem 1 (T3).

In order to justify the conclusion, in view of the fact that $e^{-(1-a)x}(\pi x)^{-1/2}$ is decreasing function of $x > 0$, it is enough to show that the expressions in the exponents of the bounds (19) and (20) are larger than $r/(2\sigma^2)$ that is a value in the exponents of the bounds (18) and (21) . In the case of (19) the condition is equivalent to $r \leq 2c_2\delta_s$, which is implied by (ii). In the case of (20) the ensuing inequality is implied by $r \leq ((1-a)^2/4)\kappa^2\theta^{*2}_{min}$ which in turn is implied by (i) as $a \in (0, 1 - c_1)$. ■

Proof of Theorem 2. Let us recall that for OS algorithm we have $P(S_1 \not\in \mathcal{T}_n) = 0$ and $|\mathcal{T}_n^o| = 1$, so the results follow from Lemma 3 analogously to Theorem 1. ■

Proof of Corollary 3. We proceed as in the proof of Corollary 1. The following condition

$$4a^{-1}\sigma^2 \log p \leq 2c_2\delta_s. \tag{44}$$

is stronger than the assumption in Theorem 2. The assumption imply (44) and the assumption of (T4). Furthermore, from the first $\kappa - \delta$ inequality (15) and assumption $a \in (0, 2c_2)$ we obtain that the assumption is stronger than both conditions in (T3).

Next we show that the powers in the exponents of the bounds (19) and (20) are larger than $r/(2\sigma^2)$. In the case of (19) the condition is equivalent to $r \leq 2c_2\delta_s$ which is implied by the assumption. In the case of (20) the ensuing inequality is implied by $r \leq ((1-a)^2/4)\delta_t$, which is implied by $r \leq at^{-1}\delta_t$ because for $a \in (0,1)$ a condition $a \leq (1-a)^2/4$ is equivalent to $a \in (0, 2c_2)$. ■

A.4 Proof for Section 5.

Proof of Theorem 3. Let $v_j^T = x_j^T(I - H_T)$ for $j \not\in T$ and 0 otherwise and $u_j^T = e_j^T(X_T^T X_T)^{-1}X_T^T$ for $j \in T$ and 0 otherwise, where e_j is the unit vector having 1 as the jth coordinate. Let

$$\mathcal{A} = \Big\{\forall j \in F \quad |v_j^T\varepsilon| < \frac{2r_Z}{7}\,, \, |u_j^T\varepsilon| < \frac{2r_Z}{7\lambda_t}\Big\}.$$

Using the left part of the assumption (ii), we observe that the following statement, which is equivalent of Lemma 3 in Zhang (2013) in the case of Gaussian errors, holds

$$P(\mathcal{A}^c) \leq \exp\Big(\frac{-c_3(1-a)r_Z^2}{\sigma^2}\Big)\Big(\frac{c_3\pi r_Z^2}{\sigma^2}\Big)^{-1/2}. \tag{45}$$

Then the proof of Theorem 3 follows the lines of the original proof in Zhang (2013), but just before the end we simplify the condition $l > l_0 + 1$, noting that

$$l_0 = \frac{\ln t}{2\ln(\lambda_{1.5t+s} b_Z/(6r_Z))} \leq \frac{\ln t}{2\ln(1.5)} < 1.24 \ln t.$$

In order to prove (45) observe that for $j \notin T$ $\mathrm{var}(v_j^T \varepsilon) = \sigma^2 x_j^T (I - H_T) x_j \leq \sigma^2$ and $W_j = (v_j^T \varepsilon)^2 / \mathrm{var}(v_j^T \varepsilon) \sim \chi_1^2$. Thus using Mill's inequality (37) we have

$$P\Big(|v_j^T \varepsilon| \geq \frac{2r_Z}{7}\Big) \leq P\Big(W_j \geq \frac{2c_3 r_Z^2}{\sigma^2}\Big) \leq \exp\Big(\frac{-c_3 r_Z^2}{\sigma^2}\Big)\Big(\frac{c_3 \pi r_Z^2}{\sigma^2}\Big)^{-1/2}. \qquad (46)$$

Using the same reasoning for $j \in T$ with $\mathrm{var}(u_j^T \varepsilon) = \sigma^2 e_j^T (X_T^T X_T)^{-1} e_j \leq \sigma^2 \lambda_t^{-1}$ and $\tilde{W}_j = (u_j^T \varepsilon)^2 / \mathrm{var}(u_j^T \varepsilon) \sim \chi_1^2$, we have

$$P\Big(|u_j^T \varepsilon| \geq \frac{2r_L}{7\sqrt{\lambda_t}}\Big) \leq P\Big(\tilde{W}_j \geq \frac{2c_3 r_Z^2}{\sigma^2}\Big) \leq \exp\Big(\frac{-c_3 r_Z^2}{\sigma^2}\Big)\Big(\frac{c_3 \pi r_Z^2}{\sigma^2}\Big)^{-1/2}. \qquad (47)$$

From (46) and (47) we obtain with $c = 2c_3 r_Z^2/\sigma^2$

$$P(\mathcal{A}^c) \leq \sum_{j \in T} P(\tilde{W}_j \geq c) + \sum_{j \notin T} P(W_j \geq c) \leq p \exp\Big(\frac{-c_3 r_Z^2}{\sigma^2}\Big)\Big(\frac{c_3 \pi r_Z^2}{\sigma^2}\Big)^{-1/2}.$$

Finally, we observe that inequality

$$-c_3 r_Z^2/\sigma^2 + \log p \leq -(1-a) c_3 r_Z^2/\sigma^2$$

is equivalent to the left part of the assumption (ii) of the theorem $c_3^{-1} a^{-1} \sigma^2 \log p \leq r_Z^2$, thus yielding (45). ■

A.4 Tables for Section 8.

$r_L \setminus b$	1.3	1.9	2.5	3.1	3.7
0.01	74.9 / 55.9	73.8 / 65.3	72.5 / 70.5	70.8 / 70.3	68.7 / 68.4
1.0	74.9 / 57.9	73.8 / 67.0	72.5 / 71.0	70.8 / 70.5	68.7 / 68.6
2.5	74.7 / 60.7	73.6 / 68.7	**72.3 / 71.3**	70.6 / 70.4	68.6 / 68.6
5.0	74.0 / 64.7	72.8 / 70.2	71.6 / 71.1	69.5 / 69.5	67.6 / 67.6
10.0	70.1 / 67.2	68.4 / 68.1	66.5 / 66.5	64.2 / 64.2	62.0 / 62.0

Table 2: Screening / selection accuracy of SOS for $M_1,\ r = 20$.

$r_L \setminus b$	0.6	0.9	1.2	1.5	1.8
5.0	95.7 / 74.0	94.2 / 78.8	92.6 / 83.9	90.6 / 86.0	88.5 / 85.7
10.0	95.5 / 78.1	94.2 / 83.4	92.5 / 86.7	90.5 / 87.1	87.8 / 85.4
15.0	94.7 / 82.0	93.1 / 85.9	**91.3 / 87.5**	89.1 / 86.6	86.1 / 84.2
20.0	93.1 / 85.1	91.2 / 86.7	89.1 / 86.4	86.1 / 84.2	83.6 / 82.2
30.0	87.6 / 84.7	85.1 / 83.1	82.2 / 80.9	78.4 / 77.4	75.2 / 74.4

Table 3: Screening / selection accuracy of SOS for $M_2,\ r = 20$.

$r_L \setminus b$	0.4	0.8	1.2	1.6	2.0
2.5	93.0 / 69.4	90.1 / 70.0	86.4 / 74.4	82.4 / 75.5	78.3 / 74.3
5.0	93.0 / 69.7	90.1 / 71.6	86.4 / 75.3	82.4 / 76.0	78.2 / 74.7
10.0	92.5 / 70.0	89.4 / 72.8	85.6 / 76.3	81.9 / 76.2	77.8 / 74.8
15.0	91.7 / 71.2	88.6 / 74.8	**84.9 / 76.9**	80.4 / 76.0	76.5 / 74.2
25.0	88.7 / 74.8	84.8 / 76.8	80.5 / 76.0	76.0 / 73.7	72.2 / 71.0
35.0	82.0 / 76.1	77.4 / 74.2	73.2 / 71.6	68.7 / 67.9	64.5 / 64.2

Table 4: Screening / selection accuracy of SOS for $M_3,\ r = 20$.

$r_Z \setminus b_Z$	4.0	5.0	6.0	7.0
0.5	67.5 / 63.0	63.3 / 90.9	57.8 / 95.6	50.7 / 94.2
2.5	67.5 / 75.0	63.3 / 94.1	57.8 / 96.3	50.7 / 94.6
5.0	66.2 / 84.5	61.9 / 95.4	**56.0 / 96.9**	48.7 / 94.8
10.0	60.8 / 90.4	55.2 / 96.5	**49.0 / 96.9**	41.8 / 94.2
20.0	43.8 / 93.9	37.3 / 96.8	31.4 / 96.0	25.6 / 90.0
30.0	28.0 / 94.2	23.0 / 95.3	18.9 / 90.0	15.1 / 78.8

Table 5: Screening / selection accuracy of MCR for $M_1,\ l = 8$.

$r_Z \setminus b_Z$	2.5	3.0	3.5	4.0
2.5	82.5 / 40.0	76.6 / 72.2	70.3 / 79.9	63.7 / 76.5
5.0	82.0 / 49.5	76.0 / 76.2	69.8 / 80.8	63.3 / 76.3
10.0	80.9 / 64.2	75.3 / 80.2	68.9 / 81.1	62.1 / 75.2
15.0	78.5 / 72.7	**72.8 / 81.9**	66.5 / 80.4	59.6 / 73.2
20.0	75.6 / 76.8	69.7 / 81.5	63.3 / 78.0	56.5 / 70.9
25.0	71.6 / 78.1	65.2 / 79.6	59.0 / 74.0	52.8 / 67.1

Table 6: Screening / selection accuracy of MCR for $M_2,\ l = 8$.

$r_Z \setminus b_Z$	1.3	1.95	2.6	3.25
5.0	85.8 / 0.2	79.0 / 28.5	71.4 / 67.1	63.1 / 68.5
10.0	85.5 / 1.6	78.0 / 45.8	70.7 / 71.1	62.4 / 68.4
15.0	84.6 / 7.7	77.4 / 58.5	**69.4 / 72.5**	61.2 / 66.9
20.0	82.9 / 22.2	75.5 / 66.9	67.1 / 72.2	59.2 / 65.0
25.0	80.2 / 40.4	72.2 / 71.4	64.5 / 70.6	56.7 / 62.6
30.0	77.1 / 53.6	69.1 / 71.2	61.2 / 67.1	54.0 / 59.3
40.0	67.1 / 64.1	60.0 / 64.3	53.1 / 58.5	46.4 / 50.9

Table 7: Screening / selection accuracy of MCR for $M_3,\ l = 8$.

References

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

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

F. Bunea, M. H. Wegkamp, and A. Auguste. Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, 136: 4349–4364, 2006.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 37:169–194, 2007.

G. Casella, F. Giron, M. Martinez, and E. Moreno. Consistency of Bayesian procedures for variable selection. *Annals of Statistics*, 37:1207–1228, 2009.

J. Chen and Z. Chen. Extended Bayesian Information Criterion for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, New York, 2006.

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

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.

J. Huang, S. Ma, and C.H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

T. Inglot and T. Ledwina. Asymptotic optimality of new adaptive test in regression model. *Annales de l'Institut Henri Poincare. Probability and Statistics*, 42:579–590, 2006.

A. Laforgia. Further inequalities for the gamma function. *Mathematics of Computation*, 42:597–600, 1984.

S. Luo and Z. Chen. Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, 143:494–504, 2013.

J. Mairal and B. Yu. Complexity analysis of the Lasso regularization path. *ArXiv*, 2012.

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

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

B.M. Pötscher and U. Schneider. Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electronic Journal of Statistics*, 5:1876–1934, 2011.

C.R. Rao and Y. Wu. Strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76:369–374, 1989.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35: 1012–1030, 2007.

J. Shao. Convergence rates of the Generalized Information Criterion. *Journal of Nonparametric Statistics*, 9:217–225, 1998.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

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

S. van de Geer and P. Bühlmann. On the conditions used to prove pracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *ArXiv*, 2014.

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.

T. Zhang. Multistage convex relaxation for feature selection. *Bernoulli*, 19:2277–2293, 2013.

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

H. Zheng and W. Loh. Consistent variable selection in linear models. *Journal of the American Statistical Association*, 90:151–156, 1995.

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

S. Zhou. Thresholded Lasso for high dimensional variable selection and statistical estimation. *ArXiv*, 2010.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533, 2008.