EFFICIENT SUPPORT RECOVERY VIA WEIGHTED MAXIMUM-CONTRAST SUBAGGING

JELENA BRADIC

Abstract. In this paper, we study finite sample properties of subagging for non-smooth estimation and model selection in sparse and large-scale regression settings where both the number of parameters and the number of samples can be extremely large. This setup is very different from high-dimensional regression and is such that Lasso estimator might be inappropriate for computational, rather than statistical reasons. We show that subagging of Lasso estimators results in discontinuous estimated support set and is never able to recover sparsity set when at least one of aggregated estimators has probability of support recovery strictly less than 1. Therefore, we propose its randomized and smoothed alternative, which we call weighted maximum-contrast subagging. We develop theory in support of the claim that proposed method has tight error control over both false positives and false negatives, regardless of the size of a dataset. Unlike existing methods, it allows for oracle-like properties, even in cases of non-oracle-like properties of aggregated estimators. Furthermore, we design an adaptive procedure for selecting tuning parameters and appropriate optimal weighting scheme. Finally, we validate our theoretical findings through extensive simulation study and analysis of a part of the million-song-challenge dataset.

1. Introduction

Consider the simple linear model

\[ Y = X\beta^* + \varepsilon, \quad \varepsilon \in \mathcal{N}(0, \text{diag}(\sigma^2)) \]

where the noise level \( \sigma \) is known and \( X \in \mathbb{R}^{n \times p} \), where both \( n \) and \( p \) are very large numbers, possibly of the order of millions and thousands respectively. Sample size \( n \) and feature size \( p \) are such that \( n \gg p \) or \( n \ll p \), with the assumption that \( n \) is always of extremely large size. Important examples of this type include data collected in large-scale energy systems or online advertising, where \( n \) can easily be of the order of a hundred millions or more. In such extremely large-scale settings, the sheer size of the data often prevents implementation of even simple methods. Capturing sparsity patterns of the original data is an important question for which it makes sense, in this specific situation only, to consider approximations to variable selection methods that can be obtained by considering subsamples of size \( N \) that are of much smaller scale than the original \( n \).

Department of Mathematics, UC San Diego.
Assume that vector $\beta^*$ is such that its support set $\text{supp}(\beta^*) = S \subseteq \{1, \cdots, p\}$ describes a particular subset of important variables for linear model (1). Without much loss of generality, we can assume that the columns of $X$ have unit $l_2$ norm. Various regularization methods have been developed to address this problem. A particularly attractive one is Lasso regularization [21], defined as

$$
\hat{\beta}(\lambda_n) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \| Y - X\beta \|_2^2 + \lambda_n \| \beta \|_1 \right\}.
$$

Theoretical properties of Lasso estimator (2) are well understood and have been studied in a variety of high dimensional settings (see for more details [25, 9, 5]). However, in practice there are situations in which usage of Lasso might be statistically desirable, but it’s computation on entire dataset might not be feasible due to resource constraints. In such situations, we might pose a question whether Lasso estimator can be well approximated by bagging a few Lasso estimators computed only over a few small subsets of the full data? In this work, we provide a negative answer to the previous question. We also propose a randomized, weighted and smoothed modification to bagging, inspired by the recent work of [15] and [16]. While the method proposed in [15] requires only partial access to the full data and can be well suited for large-scale regression, in this paper we analyze it’s finite sample theoretical properties and find that, under non-smooth and potentially very large-scale regression assumptions, although it successfully controls for false positives, it fails to control for false negatives. In order to address this issue, we propose a novel method called maximal-contrast selection. Unlike stability selection [16, 20], proposed method does not require variables to be selected in both of the paired-subsamples, but rather in only one randomized subsample which is later tailored to account for uncertainty of the smaller order subsampling.

To the best of our knowledge, almost no effort has been made in studying approximations to Lasso estimator by simple subsampling. Existing work includes [10], in which parametric bootstrap of thresholded lasso estimator is used to stabilize the estimation. However, this method works only for very small $p$ and $N = n$. In [13], authors propose $m$-out-of-$n$ bootstrap for the choice of $\lambda_n$, but only operate in $p \leq n$, $N \geq n/2$ setting. Although, good empirical control over the false negatives, was proposed in [24] with bootstrapping rows and columns of the data simultaneously, no theoretical investigations were given. In the context of multiple hypothesis testing, [17] provides empirical evidence that subagging might have an advantage over bagging in false discovery control. Asymptotic limitations of subagging are underlined in [4] and [8], but its finite sample equivalents, specifically in regression with non-smooth model assumptions, have not been analyzed. In particular, we show that Bolasso [2] fails to even approximately recover the sparsity set of the original data, when some of the subsampled designs are ill conditioned (Theorem 1), whereas the proposed method succeeds in these specific cases (Theorem 3). Asymptotically not negligible mean squared error, as stated in Theorem
3.3 of [8], suggests that there is asymptotically non-zero probability that the subsampled Lasso (sub-Lasso) and Lasso estimators have different sparsity patterns. In this paper, we provide a stronger version of this Theorem, through finite sample analysis and show that given probability is equal to zero (Theorems 5 and 6). Moreover, in Section 5, we show that estimated sparsity set of subagging estimator is a discontinuous function of $\lambda_N$.

Contributions of this paper are three fold. First, we show limitations of using traditional subbagging with regularized methods in achieving exact or approximate support recovery. Interestingly, we show that smoothing effects of bagging, unlike prediction tasks, are not desirable for model selection tasks. Second, we propose a new method based on smoothed, weighted subagging, specifically designed to attain good support recovery in situations where previous methods fail. In this context, our main result shows that Lasso estimator can be stochastically approximated with a aggregated subsampled, weighted and smoothed Lasso equivalents (sub-Lasso). While not-assuming that every sub-Lasso estimator must follow the same sparsity pattern, we show that proposed subagging is able to approximately capture sparse signal set $S$ of the original data. Finally, we provide novel finite sample analysis of the sparsity patterns of the subsampled Lasso (sub-Lasso) and the Lasso estimator. We show that the two estimated sparsity sets never match in cases when $N \ll n$ and classical conditions are possibly violated. We further extend proposed methodology with adaptive choice of tuning parameters which allows relaxation of some of the theoretical conditions.

Let us briefly introduce the notation used throughout the paper. Let $I$ and $J$ be two subsets of $\{1, \cdots, n\}$ and $\{1, \cdots, p\}$ respectively. For any vector $x$, let $x_I$ denote its sub vector with coordinates belonging to $I$, with $x_j$ denoting its $j$-th coordinate. For any matrix $A$, let $a_j$ denote its $j$-th column vector. We let $A_{I,J}$ denote a matrix formed by the rows of $A$ that belong to the set $I$ and columns of $A$ that belong to the set $J$. Shorthand notation $A_{I,J}$ stands for the matrix $A_{I \cap J}$, whereas $A_I$ stands for $A_{I, \{1 \cdots p\}}$. We also use $\phi_{\min}(A) = \lambda_{\min}(A^T A)$ and $\phi_{\max}(A) = \lambda_{\max}(A^T A)$. Note that $\lambda_N$ and $\lambda_n$ denote tuning parameters of the sub-Lasso and Lasso estimators, respectively.

The rest of the paper is organized as follows. Section 2 outlines the inefficiency of classical subagging of Lasso estimators with main results summarized in Theorems 1 and 2. Section 3 introduces novel methodology of weighted small sample bagging of Lasso estimators. Section 4 discusses theoretical findings on support recovery properties of the weighted subagging with main result summarized in Theorem 3. Discussion of difference in estimated sparsity set of the sub-Lasso and Lasso estimators is left for Section 5, in which Theorems 5 and 6 contain main results. Section 6 discusses adaptive techniques regarding the choice of $\lambda_N$ and outlines non-random, optimal weighting scheme. Finally, in Section 7 we provide results of implementation of our method on both simulated and real data.
2. Inefficiency of Subagging

Bootstrap scheme is said to be efficient if it mimics the "behavior" of the original data. In this context, "behavior" can mean many things, like inheriting rates of convergence in central limit theorem or large deviations. In this work, we concentrate on the efficiency of inheriting good variable selection properties. Similar properties have not been investigated in great detail in the existing literature. We use the following definition.

**Definition 1.** (a) Estimator $\hat{S}$ of support set $S$ achieves "exact sparse recovery", if $\hat{S} = S$ with high probability. (b) Estimator $\hat{S}$ of support set $S$ achieves "approximate sparse recovery", if $S \subseteq \hat{S}$ with high probability and $E|\hat{S} \cap S^c|$ is negligible.

Let $\{(Y_i, X_i)\}_{i=1}^n$ be a collection of observations from the simple linear model (1). Let $I_b$ be the $b$-th bootstrap-without-replacement sample of index set $I = \{1, \ldots, n\}$ and $\{I_{b_1}, \ldots, I_{b_M}\}$ a collection of its M disjoint, non-overlapping subsets, ordered in any fashion, each of size $N$. In typical situations, similar to block bootstrap, we will assume that $N = n^\gamma$ for $\gamma \in (0, 1)$. Number of such subsamples, $M$, is determined such that $MN = n$. Without loss of generality, let us denote with pairs $\{(Y_i, X_i)\}_{i \in I_{bm}}$ non-overlapping, bootstrap subsamples of the original data, which have maximal distance from one another. Total number of bootstrap samples $B$, doesn’t have to be exhaustive i.e. $B \ll n!$. Let us denote the model averaging estimator, bagged Lasso estimator [6], as

$$\hat{\beta}(\lambda_N) = \frac{1}{B} \sum_{b=1}^{B} \arg \min_{\beta} \left\{ \frac{1}{N} \sum_{i \in I_{bm}, |I_{bm}| = N} (Y_i - X_i^T \beta)^2 + \frac{1}{N} \|X_i\|_1 \right\}. \quad (3)$$

Given that our focus is on studying approximations to the Lasso estimator, we need to specify conditions under which Lasso estimator has good sparse recovery. Hence, it seems that Irrepresentable condition is unavoidable (see [22] for detailed discussion of the necessity and sufficiency of the following Condition).

**Condition 1 (IR(n)).** If $\hat{S} = \hat{S}(\lambda_n)$ represents estimated sparsity set of the Lasso estimator $\hat{\beta}(\lambda_n)$, then the design matrix $X$ satisfies the following

$$\left\|X_{I, \hat{S}^c}X_{I, \hat{S}}(X_{I, \hat{S}}^T X_{I, \hat{S}})^{-1} \text{sign}(\hat{\beta}(\lambda_n)) \right\|_\infty \leq 1.$$
**Condition 2 (RE(\(\bar{X}, \zeta_N\)))**. For each subsampled matrix \(\bar{X}\), there exists a positive number \(\zeta_N\) such that

\[
\min \left\{ \frac{\|\bar{X}v\|_2}{\sqrt{n}\|v_S\|_2} : |S| \leq s, v \in \mathbb{R}^p, v \neq 0, v \in C(3) \right\} > \zeta_N,
\]

for \(C(a)\) being a cone set defined as \(C(a) = \left\{ v \in \mathbb{R}^p : \sum_{j \in S} |v_j| \leq a \sum_{j \in S} |v_j| \right\}\) and where \(\|\bar{X}v\|_2^2 = \sum_{i \in I_{bm}, |I_{bm}| = N} w_i(X,v)^2\) for any vector \(v \in \mathbb{R}^p\), and realization of weight vector \(w\).

Subscript \(N\) in \(\zeta_N\), denotes the number of distinct rows of the matrix \(\bar{X}\). As long as the tail of the distribution of \(X_{ij}\) decays exponentially, column correlation is not too large and \(N \geq s \log n\), there are no difficulties in establishing such RE condition (see \([19]\) for more details). The following theorem shows inefficiency of classical subagged estimator in support recovery.

**Theorem 1.** Assume that Conditions 1 and 2 are satisfied and that \(p > n\). Then, for every \(\lambda_n\), there exist no sequence of \(\lambda_N^1\) with \(N \leq n\), such that bagged estimator \(\hat{\beta}^b(\lambda_N^1)\) achieves exact or approximate sparse recovery of the Lasso estimator \(\hat{\beta}(\lambda_n)\).

Let us note that, Theorem 1 holds without imposing IR(\(N\)) condition on every subsampled dataset. This contrapositive result is in line with the expected disadvantages of subagging [8], but is somehow surprising. It extends [2] and shows that even bagged Lasso estimator (with \(N = n/2\)) for diverging \(n\) and \(p\) fails to exactly recover sparsity set \(S\), if at least one of IR (\(N\)) conditions is violated. However, for \(p \leq n\) and choice of \(N = n\), situation reverses and bagged estimator recovers exactly the true sparsity set \(S\) with probability very close to 1 (proof is presented in the Appendix). Moreover, in the next result we show that, only for the choices of \(N \leq n/4\), \(\hat{\beta}^b(\lambda_N)\) attains the same predictive properties as guaranteed by Lasso.

**Theorem 2.** Assume that RE(\(X_b, \zeta_n\)) holds, where \(X_b\) is matrix obtained by bootstrap sampling rows of the matrix \(X\). Then the estimator \(\hat{\beta}^b(\lambda_N)\), defined on the bootstrapped sample of size \(n/k\) with \(k \leq 4\), with the choice of \(\lambda_N^1 = k \sigma \sqrt{\log p/n}\), obtains the same \(l_2\) prediction error as Lasso estimator \(\hat{\beta}\).

### 3. Weighted Maximum-Contrast Subagging

In practice, for large-scale regression with \(n\) of the order of millions or more, subsampling of much smaller order becomes necessary. Previous results show that in such situations, it is not possible to approximate Lasso estimator and recover signal sparsity by simply bagging sub-Lasso estimators. Hence, we propose randomized subsampling alternative, inspired by the recent paper of [15], where the number of distinct original
observations $N \ll n$ is controlled, kept small and is not considered random. Unfortunately, Lasso estimator derived from such subsamples, although computationally feasible (as they shrink $n$-scale computation to a very small $N$ scale one) is potentially very statistically unstable (especially as IR conditions of size $N$ aren’t required). In order to address this, we propose a hybrid of randomized aggregation and maximal-contrast selection, which we call Weighted Maximum-Contrast Subagging. Even though proposed method is based on subsamples of order much smaller than $n$, it approximates sparse recovery of a Lasso estimator, without requiring IR conditions on each of the subsamples.

Define weighted sub-Lasso estimator $\hat{\beta}_m^b(\lambda N, w)$ as follows

$$\hat{\beta}_m^b(\lambda N, w) = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i \in I_{bm} \mid |I_{bm}| = N} w_i (Y_i - X_i^T \beta)^2 + \lambda N \| \beta \|_1 \right\},$$

where $w \in \mathbb{R}^N$ is a vector of random weights and $I_{bm}$ random permutation of the index set $I$. Since $N \ll n$, possible model selection inconsistencies of each of sub-Lasso estimators might be carried over to the aggregated estimator. Therefore, we propose smoothing out the selection set, by aggregation of subsampled selections sets, which we call ”maximum-contrast” selection. With that in mind, for each realization of random weight vector $w$, let

$$\hat{S}_m^b(\lambda N, w) = \{1 \leq j \leq p : \hat{\beta}_m^b,j(\lambda N, w) \neq 0\},$$

denote a set of selected variables for each of the sub-Lasso estimators. In order to contrast the selected sets, we define

$$\hat{S}^b(\lambda N) = \bigcap_{m=1}^M \bigcup_w \hat{S}_m^b(\lambda N, w)$$

and use similarity measure between these bootstrap sets to determine which variables are going to be selected. Let us denote with $p_j^* = P(j \in \hat{S}^b(\lambda N))$, probability that variable $j$ belongs to the set $\hat{S}^b(\lambda N)$. Due to subsampling, estimated sets $\hat{S}^b(\lambda N)$ may exhibit significant variability which may lead to severe overdispersion. Hence, we propose the following minimax risk estimator of the selection probability $p_j^*$

$$\pi^*_j(\lambda N) = \frac{1}{\sqrt{B} (\sqrt{B} + 1)} \sum_{b=1}^B 1 \left\{ j \in \hat{S}^b(\lambda N) \right\} + \frac{1}{2} \frac{1}{\sqrt{B} + 1}.$$  

Compared to the MLE estimator $B^{-1} \sum_{b=1}^B 1 \left\{ j \in \hat{S}^b(\lambda N) \right\}$, proposed estimator $\pi^*_j$ has larger bias but much smaller variance in two situations: for small values of $B \leq 40$ and all $p_j^*$, or all values of $p_j^*$’s that are close to 0.5. By reducing the effects of observations belonging to the tail, this estimator is less sensitive to overdispersion and more robust
than MLE in the presence of subsampling effects. Therefore, $\pi_j^*$ increases the chance that variable $j$ has to be selected. Estimator of the support set is $\hat{S}_\tau(\lambda_N) = \{1 \leq j \leq p : \pi_j^*(\lambda_N) \geq \tau\}$, with appropriately chosen thresholding parameter $\tau$.

Next, we define weighted maximum-contrast bagged estimator $\hat{\beta}_a(\lambda_N)$ as

$$ (6) \quad \hat{\beta}_a(\lambda_N) = \frac{1}{BM} \sum_{b,m} \mathbb{P}_w^* \hat{\beta}_b^a(\lambda_N, w) $$

if $j \in \hat{S}_\tau(\lambda_N)$ and zero otherwise. Here $\mathbb{P}_w^*$ is the empirical average with respect to all $w$ realizations. Proposed estimator (6) is Monte Carlo approximation of the theoretical weighted small-sample bagged estimator $\tilde{\beta}_a^*(\lambda_N) = (E^*[\hat{\beta}_{\hat{S}_\tau(\lambda_N)}(\lambda_N)]^T, 0^T)^T$, with $E^*$ denoting expectation with respect to randomization, subsampling and bootstrapping. Let us discuss choices of tuning parameters $\lambda_N, B, M, K$ in Section 6.

3.1. Choice of the weight vector $w$. Note that each of the sub-Lasso estimators $\hat{\beta}_m^b(\lambda_N, w)$ can be recasted as $n \times p$ Lasso problem

$$ \hat{\beta}_m^b(\lambda_N, w) = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n (\tilde{Y}_i - \tilde{X}_i \beta)^2 + \lambda_N \|\beta\|_1, $$

where the new design matrix is equal to $\tilde{X}_i^T = [X_{i_1}, \ldots, X_{i_1}, \ldots, X_{i_N}, \ldots, X_{i_N}] \in \mathbb{R}^{p \times n}$, with rows $X_{i_j}$ repeated $w_{ij}$ times, each for $j \in 1, \ldots, N$ and $I_m^b = \{i_1, i_2, \ldots, i_N\}$.

Similarly, vector $\tilde{Y}_i^T = [Y_{i_1}, \ldots, Y_{i_1}, \ldots, Y_{i_N}, \ldots, Y_{i_N}] \in \mathbb{R}^n$ is a vector of responses with each response $Y_{i_j}$ being repeated $w_{ij}$ times.

Proposed weighted subsampling, unlike Efron’s bootstrap [12] which has random number of distinct observations, adds a constraint that the number of distinct observations $N$ has to be to fixed, non-random and much smaller then $n$. This choice of $N$ alleviates computational costs, whereas random weights $w$ are designed to allow theoretical advantages by making ghost subsamples of provisional size $n$. Hence, we propose weighted bootstrap alternative as in [15], with

$$ (7) \quad w^* = (w_{i_1}, \ldots, w_{i_N}, 0, \cdots, 0), \text{ s.t. } \sum_{j=1}^N w_{ij} = n, w_{ij} \geq 1, w_{ij} \in \mathbb{Z}. $$

Note that randomness is double in this proposal, both as the ”placement” of non-zero elements of $w^*$ and as their actual (random) values. Hence, proposed method is equivalent to a randomized delete-$(n-N)$- jackknife [11]. Unlike traditional jackknife, it has randomized weights $w_i$, non necessarily equal to $n/N$ and a very large size of the deleted set i.e. $(n-N)/n \to 1$. The following condition is needed for subsequent theoretical analysis.
Condition 3. Let \( w = (w_1, \ldots, w_N) \) be a vector of exchangeable random weights, that is independent of the data, such that \( \sum_{i=1}^N w_i = n \) and \( P(w_1 > 0) = 1 \). Moreover, let \( w_{2,\min}^2 = \int_0^\infty P_w^2(\min_{1 \leq i \leq N} w_i > t)dt < \infty \) and \( \mathbb{E}_w \|w\|_\infty < \infty \).

Condition 3 is inspired by similar conditions imposed for weighted bootstraps [18]. Unlike those, it does not require for \( \frac{1}{N} \sum_{i=1}^N (w_i - 1)^2 \) to converge to a positive constant \( c \), independent of \( n \). This condition is violated given that \( c \to 0 \) for \( N \ll n \), as seen through previous jackknife comparison. Weights are chosen to have larger variance, in order to allow for smaller choices of tuning parameters \( \lambda_N \). Let us present a few examples of random weighting schemes that satisfy proposed structure. We leave the proposition of optimal and deterministic scheme for Section 6.

**Example 1. (Multinomial weights)** Traditional multinomial weights can easily be adapted as in [15], with vector \( (w_1, \ldots, w_N) \), drawn from Multinomial distribution \( \mathcal{M}_N(n, \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)) \), with \( N \ll n \). Then previous Condition 3 is easily satisfied.

**Example 2. (Coupling Poisson and Multinomial)** Consider the following game, where we throw \( n \) balls into \( N \ll n \) urns. If \( Z_j \) denote the number of balls successfully thrown in an urn labeled \( j \), then \( Z_1, \ldots, Z_N \) are independent random variables, with Poisson distribution with rate \( n/N \) and \( (Z_1, \ldots, Z_N) \mid \sum_{i=1}^N Z_i = k \) has Multinomial distribution \( \mathcal{M}_N(k, \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)) \). Let \( (w_1, \ldots, w_N) \) denote occupation numbers of the urns. If we have successfully thrown all \( n \) balls with \( \sum_{i=1}^N Z_i = n \) then let \( w_{ij} = Z_j \). If on the other hand, \( \sum_{i=1}^N Z_i < n \) then choose uniformly at random, \( n - \sum_{i=1}^N Z_i \) balls and throw them in urns by choosing urns independently with probability \( 1/N \). Let \( (w_1, \ldots, w_N) \) be new occupation numbers of the urns.

**Example 3. (Balanced Pólya scheme)** Consider a Pólya urn scheme with an urn containing \( n \) balls with \( N \) different colors. At an initial step, let the number of balls of each color be equal to \( n/N \). Balanced Pólya bootstrap is an evolution of the number of balls within each color over discrete time steps that guarantees that the total number of all balls in an urn does not change (it is kept at \( n \)) and the number within each color changes. At each step, a ball is drawn uniformly at random from an urn and its color is noted. If a ball of color \( j \) (\( j = 1, \ldots, N \)) is drawn at a stage, it is returned at urn together with \( R_{jk} \) balls of color \( k \). It is customary to represent these changes with a replacement matrix and we propose the following

\[
R = \begin{pmatrix}
-1 & B_{12} & \cdots & B_{1N} \\
B_{21} & -1 & \cdots & B_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
B_{N1} & B_{12} & \cdots & -1
\end{pmatrix},
\]

for \( B_{jk} \) being Bernoulli random variables with \( p_{jk} \in (0, 1] \), such that only one in row \( j \) takes value 1 (sum of row elements must be kept at zero to preserve the total number of balls in an urn) and \( \sum_{k=1,k\neq j}^N p_{jk} = p_{jj} \). Then, by using classical arguments it can be
shown that the number of balls within each color is a Markov chain with a stationary Multinomial distribution \( \mathcal{M}_N(n, (\frac{1}{N}, \cdots, \frac{1}{N})) \). By Borel-Cantelli arguments, it can be seen that this scheme will also guarantee that with probability 1 each color is represented in the evolution of the scheme after some number of draws, and Condition 3 is satisfied.

4. Theoretical Properties

In this section, we explore finite sample variable selection property of the proposed weighted scheme (7). We provide novel insights into finite sample equivalents of the asymptotic bias of subagging (as defined in [8]) and show the way weights and smoothing are instrumental in overcoming it. If irrepresentable condition IR(N) is assumed to hold for every subsampled dataset, it is easy to verify that each of the sub-Lassos perfectly recovers the set \( S \) with high probability. Since proposed methodology requires small sampling with possibly \( N \ll p \), requiring IR(N) condition is undesirable. Hence, a more interesting scenario is allowing for possible violations of such small-sampled irrepresentable conditions. Major part of this section is devoted to studying this particular case and showing that weighted maximal-contrast subagging (6) does not require such condition to achieve approximate sparse recovery (according to Definition 1).

In situations when IR(N) condition is possibly violated, we show that for every realization of random weights \( w \) and \( \lambda_N \geq \lambda_n \)

\[
P(\bar{\beta}_{m,j}^b(\lambda_N, w) = 0) < P(\bar{\beta}_j(\lambda_n) = 0), \quad j \in S^c,
\]

(see Theorem 3). Therefore, there is little hope that each of the sub-Lasso estimators performs the same as Lasso estimator (2). However, the expectation is that by controlling out of sample properties and aggregating information over multiple subsamples, we can approximate Lasso.

**THEOREM 3.** Let Conditions 2 and 3 be satisfied. Assume that the distribution of \( \{1(j \in \cup_w \hat{S}_{\tau,m}^b(\lambda_N, w)), s + 1 \leq j \leq p\} \) is exchangeable for \( \lambda_N \) as in (9) and that \( \|x_{\hat{S}_m}^b\|_\infty \leq c_2 \sqrt{N/\log p\|w\|_\infty} \), for some constant \( c_2 > 1 \) and all \( 1 \leq j \leq p \). Then, there exists bounded positive constant \( c_N > 1 \) such that, for

\[
\frac{1}{c_N} \sqrt{\log p/n} \leq \lambda_N \leq c_N \sqrt{\log p/n}
\]

the following holds \( P(S \subseteq \hat{S}_{\tau}(\lambda_N)) \rightarrow 1 \) and \( E[S^c \cap \hat{S}_{\tau}(\lambda_N)] \rightarrow 0 \) simultaneously for all \( \tau >(2 + \sqrt{B})/(2 + 2\sqrt{B}) \).

Studying estimator (6) is complicated by the fact that we allow deviations of the small sampled IR(N) condition. Proof of the Theorem 3 is based on a tight control of the out-of-sample properties of each of the sub-Lassos and a novel method of finding appropriate range for the \( \lambda_N \). This proof is based on allowing possible small deviations in
prediction rates and finding the smallest such deviation that allows good variable selection properties. Assumption of exchangeability of indices \( j \) over the set \( \bigcup_w \mathcal{S}_m^w(\lambda_N,w) \), is actually an assumption on the appropriate sets of probability distributions for the weight vectors \( w \). It is easy to see that such a condition is not too restrictive for Examples 1 through 3 of Section 2.2 (all of which avoid trivial cases when exchangeability condition might be violated).

Theorem 3 suggests that even when sub-Lassos have probability of support recovery smaller than 1/2, proposed method \( 6 \) boosts naive aggregation and increases probability of support recovery to be very close to 1. Sub-Lasso estimators cannot select all features as signal variables; their “goodness” is controlled through assumptions \( \text{RE}(\hat{X},\zeta_N) \), which guarantee that the size of the estimated set is not too large. Note that these are the only conditions required for each subsample (as \( \text{IR}(n) \) condition is for complete data only).

Although this work is mainly concerned with variable selection, let us provide a results on the effect the weighting scheme \( w \) has on prediction properties. Interestingly, it shows that the weights \( w \) are necessary for controlling out-of-sample properties. We look at both previously introduced minimal eigenvalue condition \( \text{RE}(\hat{X},\zeta_N) \) and its relaxation of the following kind: here exists a positive number \( \zeta_N \) such that \( \min\{n^{-1/2}E_w\|\tilde{X}v\|_2/\|v_S\|_2 : \|S\| \leq s, v \in \mathbb{R}^p, v \neq 0, v \in C(3)\} > \zeta_N \); which we call assumption \( \text{ERE}(\hat{X},\zeta_N) \). This relaxation is related but not equivalent to the \( \text{RE} \) condition on sub matrices \( X_{I_m} \). Since \( \text{ERE} \) holds for all data perturbations, it implies \( \text{RE}(\hat{X},\zeta) \). By noticing that \( \|X_{I_m}v\|_2 \leq E_w\|\tilde{X}v\|_2 \) we can put forward the following set of inequalities: \( \text{RE}(\hat{X},\zeta_N) \leq \text{RE}(X_{I_m},\zeta_N) \leq \text{ERE}(\hat{X},\zeta_N) \leq \text{RE}(X,\zeta_N) \).

**Theorem 4.** Let \( \|X_{I_m}^2\|_\infty \leq c_2 \sqrt{N/\log p} \), for some constant \( c_2 > 1 \) and all \( 1 \leq j \leq p \). Then, with probability of at least \( 1 - p^{1-c} \), \( c > 1 \): (i) If condition \( \text{ERE}(\hat{X},\zeta_N) \) holds for all \( m \) and \( b \), then \( \|X_{I_m}(\beta^*-\hat{\beta}_m^b(\lambda_N,k))\|_2^2 \leq 16\lambda_N^2sn\zeta_N^2a_N^2 \) for \( c_1 \sigma \sqrt{\log p/n} \leq \lambda_N \leq c_2\sigma \sqrt{\log p/N} \), for some positive constant \( a_N \) such that \( P(\min_{1 \leq i \leq N} w_i \geq a_N^{-1}E_w\|\sqrt{w}\|_\infty) \rightarrow 1 \). (ii) If Condition 2 holds, then \( \|X_{I_m}(\beta^*-\hat{\beta}_m^b(\lambda_N,k))\|_2^2 \leq 16\lambda_N^2sn\zeta_N^{-2} \), for \( c_1 \sigma \sqrt{\log p/n} \leq \lambda_N \).

The upper bound on the constant \( a_N \) varies with the distribution of \( w \). Its minimum is reached when the minimum of the weight vector \( w \) is maximized i.e. for equally weighted \( w \). In such cases, it is easy to show that \( a_N \leq \sqrt{N/n} \) and that the upper bound in (i) and (ii) match despite different \( \text{RE} \) assumptions.

**Remark 1.** In contrast to existing work on Lasso estimator, where prediction bounds are obtained for \( \lambda_N > \sigma \sqrt{\log p/N} \), Theorem 4 holds for much smaller choices of the tuning parameter \( \lambda_N \). By appropriately weighting each row of the subsample matrix.
we were able to obtain good prediction properties of subsampled estimators for choices of parameter $\lambda_N$ of much smaller order than $\sigma\sqrt{\log p/n}$. Hence, results of Theorem 4 suggest that existence of $w$ is crucial in eliminating false negatives (see Theorem 6).

5. Sparsity Approximation

As an important aspect of an efficiency of bootstrap subsampling, we explore how close connected are sub-Lasso $\hat{\beta}_m^{b}(\lambda_N, w)$ (4) and Lasso $\hat{\beta}(\lambda_n)$ (2) estimators in terms of their sparsity patterns. To the best of our knowledge, such an important question was not investigated in existing literature. We aim at providing novel insights by showing that the two estimated sparsity sets have strictly nested structure and are almost never of the same size. Main results are summarized in Theorems 5 and 6.

DEFINITION 2. We say that the sparsity pattern of an estimator $\hat{\beta}$ is optimal for an estimator $\hat{\eta}$, if non-zero part of the estimator $\hat{\beta}$ satisfies Karush-Kuhn-Tucker optimality conditions of the non-zero part of the estimator $\hat{\eta}$.

We find values of the tuning parameters $\lambda_N$ and $\lambda_n$, which allow sparsity pattern of the sub-Lasso to match the one of the Lasso estimator. We begin by showing in Lemma 2 (in the Appendix) that the signal variables identified by sub-Lasso estimator (4) lie in the feasible set of the Lasso problem (2). Moreover, we show that sub-Lasso estimator requires much smaller regularization (compared to the expected order of $\sqrt{\log p/N}$) in order to reach the number of identified signal variables similar to Lasso. This phenomenon complements the recent work of [14] in which the authors argue empirically that higher correlations, as expected by subsampling the rows of data matrix, induce smaller tuning parameter.

THEOREM 5. Let Conditions 7, 2 and 3 hold. Then, sparsity pattern of each of the sub-Lasso estimators (4), with $\lambda_N \geq c_1 \sqrt{\log p/n}$, $c_1 > 0$, with probability $1 - p^{1-c}$, $c > 1$, is optimal for Lasso estimator (2) with $\lambda_n$ satisfying

$$\lambda_n \geq \lambda_N \left( 2 + \phi_{\min}^{-1}(X_{\tilde{S}_m, \tilde{S}_m}) \right) + \sigma \sqrt{\log p (n^{-1/2} + 2n^{-3/2} \zeta_N^{-1/2} s^{1/2} \phi_{\max}^{1/2}(X))}.$$  

Note that the choice of $\lambda_N$ depends on the $w$ as well, through the last term in the previous inequality. Without evaluating prediction on the testing data, sub-Lasso and Lasso estimators would not be close and in case of not valid small sample irrepresentable condition, the gap could be enormous. The core of the proof of Theorem 5 relies on evaluating finite sample bounds for the out-of-sample properties for each of the sub-Lasso estimators. One of the consequences of the proof is that out of sample prediction property of the sub-Lasso estimators is linearly inflated by $\lambda_{\max}(X_{km}^r X_{km})$, i.e by the conditioning of the new data. This phenomenon was separately noticed in [7] through detailed simulation investigations. Since new data in our bagging procedure has much
more rows then columns, \( n - N \gg p \), out of sample prediction is fairly good. This is the key of good properties of the proposed small sample bagging methodology.

In the remainder of this section, we show contrapositive result; that without requiring irrepresentable condition on each of the small samples, it is not possible to obtain exact sparse recovery.

**Theorem 6.** Let Conditions \(1, 2\) and \(3\) hold. Then, sparsity pattern of Lasso estimator \((2)\), for \(\lambda_n \geq c_1 \sigma \sqrt{\log p / n}\) for some \(c_1 > 1\) is, with probability \(1 - p^{1-c}\), \(c > 0\), optimal for each of the sub-Lasso estimators \((4)\), with

\[
\lambda_N \geq n \lambda_n \left(2 + \phi_\min^{-1}(X_{\hat{m}b}^b, \hat{\beta}_{mb}^b, \hat{S})\right) + \sigma \sqrt{\log p \left(1 + 2n^{-1}c_1^{-1/2}s^{1/2} \phi_{\max}^{1/2}(X)\right)}.
\]

**Remark 2.** Let us denote with \(f_{\lambda_N} : \mathcal{P} \rightarrow \mathbb{R}_+\) a set function that maps the set to its cardinality defined as \(f_{\lambda_N}(S) = |S|\), where \(S \subset \mathcal{P}\) and \(\mathcal{P}\) denotes sets of all possible subsets of \(\{1, \ldots, p\}\). Then, as a consequence of the previous Theorems \(\ref{3}\) and \(\ref{6}\) we conclude that \(f_{\lambda_N}(\cup \hat{S}_m^b(\lambda_N, w))\), as a function of \(\lambda_N\), has a discontinuity of the first order with probability close to 1. Its jump discontinuity lies in the interval \(\left(\frac{1}{c} \lambda_n - \frac{c}{\sqrt{n}} n c \lambda_n + c^2 c_1\right)\) for some small but possibly random, positive constants \(c\) and \(c_1\). Therefore, "maximal-contrast" selection acts as a smoothing operator and allows us to recover sparsity set \(S\) optimally.

### 6. Implementation

In this section, we discuss extensions of the proposed method and propose adaptive choice of \(\lambda_N\) along with weight vector \(w\) which, remarkably, allows only 3 repetitions of reweighing and 2 bootstrap resampling total.

6.1. **Adaptive warm-start choice of tuning parameters.** Since we seldom know in advance a good value for \(\lambda_N\), we often compute a sequence of solutions for a (typically) decreasing sequence of values \(\lambda_{N1} > \lambda_{N2} > \cdots > \lambda_{Nq}\). Warm-start or continuation methods use the solution at \(\lambda_i\) as an initial guess for the solution at \(\lambda_{i+1}\) and often yield better computational efficiency. When sample size of the original data \(n\) is too large, plain Lasso estimator \((2)\) cannot be efficiently computed and we have no relative comparison on how to choose appropriate sequence of \(\lambda_N\). With careful analysis of all the factors of Theorems \(\ref{5}\) and \(\ref{6}\) we design the following adaptive, warm-start scheme for choice of tuning parameter \(\lambda_N\).

For \(b\) and given sample index partition \(\{I_{mb}^b\}_{m=1}^M\), initialize \(\hat{A}_1 = \{1, \ldots, p\}\) and for \(b > 1\) define \(\hat{A}_b = \{j : \hat{\beta}_{mb}^{-1} \neq 0\}\). Compute \(\kappa_{mb}^b = \lambda_{\min}^{-1}(X_{\hat{m}b}^b, A_{\hat{m}b}, \hat{A}_b)\) and \(\kappa_m = \kappa(X_{\hat{m}b}^b, A_{\hat{m}b}, I_{mb}^b, \hat{A}_b)\). From the design of the subbagging, we know that \(\kappa_{mb}^1\) and \(\kappa_m^1\) will be finite but overinflated, as all variables are included in the model. Due to usage of warm up techniques for choices of \(\hat{A}_b\), both quantities will decrease in the future iterations and have small fluctuation (of the order of \(n^{-1/2}\)) around their "true" value.
\[ \lambda_N^{-1}(X^T_p S X_p, S) \] and \( \kappa(X^T_p S X_p, S) \). Hence, instead of doing extensive grid search for the optimal choice of \( \lambda_N \), we propose to only consider

\[ \lambda_N \in \left\{ \frac{z}{2 + \kappa^b m} \sqrt{\kappa^b m |\hat{A}_b| \log p}, z \in \{1, 2, \ldots, 8\} \right\}. \]

Constant 8 here guaranties good properties of the main Lasso estimator by choosing its tuning parameter of the order of \( 8\sqrt{\log p/n} \) (see [5]) and governs the upper bound on the choice of \( \lambda_N \). RHS of previous inequality is always smaller or equal than that of \( \lambda_{1N} \) and intuitively, we expect the difference between its two subsequent bootstrap ranges to shrink as \( n^{-1/2} \). Therefore, our procedure does not need for more than just two bootstrap permutations of the whole data (see Example 2 of Section 7).

### 6.2. Optimal Weighting Scheme.

Previously, we have introduced stochastic weight vector \( w \) in order to imitate nonparametric bootstrap realizations as closely as possible. We also saw that assuming \( \mathbf{RE} \) conditions on the whole and small subsample data was unavoidable. \( \mathbf{RE} \) conditions are established for several broad classes of random matrices, ranging from those with dependent entries, including random matrices with subgaussian rows and non-trivial covariance structure, to the matrices with independent and uniformly bounded entries [19, 26]. Such structure would be desirable at the population level. Unfortunately, interplay of population conditions corresponding to the whole data and to the weighted subsamples is not quite interpretable, hence we propose alternating the choice of the weight vector \( w \). For every \( b \) and \( m \), we propose choosing \( w \) which maximizes \( \zeta_N \), such that \( w_1 + \cdots + w_N = n \) and \( w_1, \ldots, w_N \in \mathbb{Z}_+^N \).

Hence, \( w \) increases the minimal eigenvalue of the sub matrix \( \tilde{X}^T \tilde{X} \) and at the same time imitates bootstrap realizations. Unfortunately, this is an integer optimization problem which is NP hard. However, the following proposition shows us that with simple relaxation of \( w_1, \ldots, w_N \in \mathbb{R}_+^N \), the problem can be recasted as a semi-definite program (SDP) and hence efficiently implemented.

**Theorem 7.** The problem of maximizing the minimum eigenvalue of \( \tilde{X}^T \tilde{X} \) under conditions that \( 1^T w = n \) and \( w \in \mathbb{R}_+^N \) is equivalent to solving the following SDP

\[
\max_{s.t.} \quad t \\
\sum_{j=1}^N w_j x_i^T x_j \geq tI \\
\sum_{j=1}^N w_j = n, w \geq 0
\]

This proposition enables us to design an adaptive choice of weight vector \( w \), in a way that increases the minimal eigenvalue of \( \tilde{X}^T \tilde{X} \) compared to \( X_{t_m}^T X_{t_m} \). Let \( k = 1, \ldots, K \) denote the number of realizations of vector \( w \). For \( k = 1 \), choose equal weights of size \( n/N \). For next iterations \( k=2, \) and for fixed \( b \) and fixed sample splitting,
take $J_k = \{ j : \hat{\beta}_{m,j}^b (\lambda_N, k - 1) \neq 0 \}$ and $\hat{s} = \# \{ j : \hat{\beta}_{m,j}^b (\lambda_N, k - 1) \neq 0 \}$ and choose $w_k = [w]$ such that

$$\max \quad t$$

$$\text{s.t.} \quad \sum_{j=1}^N \| \sqrt{w_j} x_{i,j,k} \|_2 > t \sqrt{ns}$$

$$1^T w = n, w \in \mathbb{R}_+^N,$$

where $x_{i,j,k}$ denotes $i,j$-th row of matrix $X_{I_{bm}}^a$ with columns (i.e. indices) in $J_k$. Previous stochastic optimization problem, given by Theorem 7, should have been solved for all possible sparsity patterns $v \neq 0$ such that $\|v_{J_k}\|_1 \leq 3 \|v_{J_k}\|_1$ and $\sum_{j=1}^N \| \sqrt{w_j} x_{i,j,k} v \|_2 > t \sqrt{\|v_{J_k}\|_2}$. However, this is computationally less attractive, so we relax it to take into account the sparsity structure of the previous step estimator and look only at those sub matrices. By doing so, we are making a tradeoff between computational efficiency and theoretical requirements.

Problem (12) is a regular SDP that can be plugged in any of the existing SDP-solvers. The problem is not of large size, as $|I_{bm}| = N \ll n$, and therefore computational costs of repeating it are not substantial. Since there exists a unique optima of the previous problem, there is no need for $K$ to be larger then 3. Step two can still potentially have too large $\hat{s}$, hence to correct for it - we allow one additional $k$ iteration.

7. Simulated and Real Data Examples

Simulated and real data are analyzed in order to underline strengths of the proposed weighted maximal-contrast subagging. We generated 3 synthetic datasets from a simple linear model (1) with varying sample size $n$, feature size $p$, signal strength and correlation structure in the design matrix $X$. All the choices of the tuning parameters were followed by the adaptive results.

7.1. Example 1. In this example, we measure performance of the proposed method through false positive(FP) and false negative(FN) error control. In order to compare performance between proposed method and plain Lasso estimator, sample size $n$ was chosen such that it has large, but computationally feasible value. Therefore, we fix $n = 10000$ and $p = 500$ and generate design matrix $X$ from multivariate normal distribution with correlation matrix $\rho \{i-j\}$ with choices of $\rho$: 0.8(a), 0.2(b), 0.2(c), 0.8(d). We vary signal to noise ratio by setting the first 30 parameters to be equal to 1 in (a) and (b) and to 0.3 in (c) and (d) and the rest to be equal zero. We contrast the results of traditional Lasso (2), traditional subagging (3) and proposed weighted maximal-contrast subagging (6) with varying choices of parameters $\gamma$ and fixed number of bootstrap permutations and weight approximations $B = 2, K = 3$. We repeat all procedure 100 times.

Results are summarized in Figures 1 and 2. Figure 1 outlines the difference among the three methods with respect to the true and false positives, by plotting average selection probability for the signal and noise variables, against the choice of size of
the subsamples $N$. Methods are color coded as follows: blue is for subagging, green for Lasso and pink for the proposed weighted maximal-contrast subagging. In these examples, we explore worst case behavior of the proposed methods and chose $\tau = 1$ in order to do so.

We expect Lasso to be able to select the sparse set in example (b) but to fail to some extent in other examples; examples (a) and (d) are designed to violate the Condition 1, whereas examples (c) and (d) decrease signal to noise ratio. Figure 1 confirms all of the above and illustrates much more. Surprisingly, subagging fails to perform variable selection at all; it selects almost all features as important in all 4 settings. We discover that weighted maximal-contrast subagging results in better than expected properties. It manages to control the number of FP in all 4 cases at fast logarithmic rate, with respect to the size of subsamples. The number of FN depends on the choice of $\tau$, but for most cases with $\tau \sim 1/2$ it keeps FN at a low rate, even for subsamples of size $\sqrt{n}$. As expected, FN deteriorates with smaller signal strength, requiring larger subsample sizes for obtaining low rate of FNs. For example (b), both subsample size of 75 with $\tau \sim 1/2$ and subsample size of 87 with $\tau \sim 3/4$, guarantee that both FP and FN will be close to zero. In the case of example (d), the equivalent requirements for subsample size are larger with 130 and 250 respectively. On the other hand, examples (a) and (d) show how additional strong feature correlation helps weighted maximal-contrast subagging to control FN with little or no increase in FP.

Figure 2 zooms into the left column plots of Figure 1 and outlines the true positives (TP) of weighted maximal-contrast subagging. We show box plots of the selection probabilities for each of the signal variables individually with $N = 100$. Although in each individual run, sub-Lasso estimator missed a few of the first 30 important variables, given that we have averaged among many small randomized Lassos and perturbed the data, final estimator was able to pick up those ”loose” variables in all four setups. As long as $\tau$ was bigger then 1/2, all of the significant variables were picked up. In that sense, the proposed method advances the finite sample selection property of plain Lasso estimator, given that it keeps FN= 0 even when Condition 1 is violated.

7.2. Example 2. We are also interested in testing the sensitivity of the proposed method on the choice of the number of bootstrap permutations $B$. For that purpose, we generate synthetic data from simple linear model with a different number of bootstrap data permutations. Model specifications are equivalent to the ones of Example 1 (b), with small column correlations and strong signal. As measure of performance, we contrast mean selection frequency of the first 30 important variables with a different number of bootstrap replications. Figure 3 summarizes our findings and reports average selection frequency and its 5% confidence interval for 4 different bootstrap replications: $B= 1, 2, 5$ and 10. As expected, the bigger the number of bootstrap replications, the smaller is the variability in estimating $\pi_j$. Interestingly, $B = 2$ was sufficient to guarantee $\pi_j > 1/2$. Even for $B = 1$ this seems to be true in the example considered but
Figure 1. Sample average of the probability of True and False Positive selection of the Lasso (green), subagging (blue) and weighted maximum-contrast subagging with $\tau = 1$ (pink) and $\tau = 0.5$ (black).

the variability is significantly larger then for $B \geq 2$, hence making general conclusion seem inappropriate.
Figure 2. Distribution of selection probabilities $\pi_i^*$ in weighted maximum-contrast subagging among the first important 30 variables with $n = 10000, p = 500$
Figure 3. Variability of selection probabilities $\pi_j^*$ among signal variables with change in number of data perturbations $(B = 1, B = 2, B = 5$ blue, green and pink respectively). Dashed lines represent corresponding 5% confidence intervals.

Example 3. We complement the first two examples with the following study where we tested for sensitivity of the proposed method to the change in choice of $\gamma$, $n$ and $p$ simultaneously. For this study, we fix $B = 1, M = 1, K = 1$ in order to see the possible worst-case behaviors, and fixed design setting with column correlation of 0.3.

Figure 4 shows variability in sampling frequencies. In most modern application we don’t have control over the size of $\gamma$ or don’t know the size of the whole sample (hence we only know that $\gamma$ is potentially small). We consider four different scenarios where $n = 500, 1000, p = 250, 500$ and $\gamma = 0.75, 0.8$. First and third scenario use simple equal weighting scheme $=n/N$, whereas all others use the optimal weights of Algorithm 1. We show that, even with relatively small subsampled data sets, proposed method preforms reasonably well (see Figures 4(b).,d)). Note that implemented maximum-contrast selection in Figures 4(a) and (c), corresponds to paired stability selection of [20] (since $M = 1$ and $K = 1$). Hence, we observe that paired stability selection underperforms for small subsamples and small $n/p$ ratios.

Million Songs Dataset. The Million Song Dataset, developed with support of NSF, is new, freely-available collection of audio features and metadata for a million contemporary popular music tracks. The goal of this project is development of smarter platforms for enjoying music, apps that deliver enhanced listening experiences and tools for detailed musical analysis. In order to get a sense of the size of this project, note that 1% subset of this data is 1.8 GB and is substantial on its own. Hence, Columbia University together with The Echo Nest and LabROSA have started a Million Songs Challenge (http://labrosa.ee.columbia.edu/millionsong/) with the goal of finding best solutions to the number of problems related to this dataset, few of which are: dividing a song into meaningful segments, associating appropriate keywords to some specific sound segment, predicting tags based on clusters of song titles, etc.
One such task is the prediction of an album’s year of release. Using special audio features like segments_timbre, covariance of every song and possibly some other information like energy, danceability, etc, our goal is to try to understand which one of them has the highest correlation with the year or decade at which the song has been released. A simplified dataset for that task is available on the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD).

Even simplified, this data is still huge. It consists of 515345 different track instances and 90 audio features for each track. Simply applying Lars package is not possible, especially given the longitudinal nature of the data. We split the data into training (first 463,715 examples) and testing subsets (last 51,630 examples) in order to avoid the producer effect, by making sure no song from a given artist ends up in both train and the test set. We apply the proposed adaptive weighted maximum-contrast subagging on the training data with generalized linear regression model with link function $g(\mu_k(X_i)) = \beta_k^T X_i$ and $\mu_k = P(Y = k)$ for $k = \{1, \cdots, 83\}$. We model the probability of $i$-th observation belonging to $k$-th year of release using logit function $p_k = P(Y_i = k|X_i) = \frac{e^{\beta_0 + \sum_{l=1}^{83} \beta_l^T X_{il}}}{\sum_{l=1}^{83} e^{\beta_l^T X_{il}}}$ with $\beta_0$ denoting an intercept. Choice on the tuning parameters are
B = 2, K = 3 for a sequential grid of values of \( \lambda_N \) of Section 4. Among a few possible values for \( \lambda_N \), we pick the one that minimizes 5-fold cross validation statistics.

Error prediction rate for each year of release \( k = 1, \cdots, 83 \) is defined as

\[
\text{err}_k = \frac{1}{n_1} \sum_{i=1}^{n_1} I\{Y_i \neq \hat{k}_i\}, \quad \hat{k}_i = \begin{cases} 
  k, & \hat{p}_{ik} \geq c \\
  0, & \hat{p}_{ik} < c 
\end{cases}
\]

for some threshold \( c \) usually taken to be larger than 0.012 and \( n_1 \) the size of the testing dataset. Since we have 83 years of releases, we summarize our findings across decades of track’s release in Table 1. Due to size of the dataset, plain Lasso estimator could not be implemented, hence we compare performance of the proposed weighted maximum-contrast subagging (WMCSub) with the plain subagging (Sub) in terms of their prediction errors and estimated sparsity sets. For each of the decades, we report the median misclassification rate and its standard deviation, of each year of release in that decade (column 2 of Table 1).

**Table 1. Summary of UCI Million Songs Dataset results by decades.**

| Decade of a track release | WMCSub err (SD) | Sub err (SD) | WMCSub \( FN+FP \) (SD) | Sub \( FN+FP \) (SD) |
|---------------------------|----------------|-------------|-------------------------|---------------------|
| 2001-2010                 | 21% (0.66%)    | 18% (0.26%) | 13 (0.16)               | 55 (4.16)           |
| 1990-1999                 | 20% (0.76%)    | 17% (0.26%) | 14 (0.17)               | 50 (5.61)           |
| 1980-1989                 | 21% (0.65%)    | 18% (0.25%) | 12 (0.15)               | 54 (3.86)           |
| 1970-1979                 | 23% (0.52%)    | 20% (0.32%) | 8 (0.21)                | 47 (6.92)           |
| 1960-1969                 | 21% (0.73%)    | 19% (0.27%) | 6 (0.30)                | 43 (7.46)           |
| 1950-1959                 | 24% (0.95%)    | 22% (0.43%) | 9 (0.18)                | 48 (5.78)           |
| 1940-1949                 | 28% (1.51%)    | 24% (0.75%) | 13 (0.17)               | 53 (5.21)           |
| 1930-1939                 | 32% (1.84%)    | 27% (0.84%) | 17 (0.22)               | 57 (4.33)           |
| 1922-1929                 | 37% (2.12%)    | 31% (1.21%) | 20 (0.45)               | 60 (4.98)           |

First, we observe differences in rates of a successful classification, with recent decades showing much better results compared to the early beginnings of the 20th century. This may potentially be explained by the unbalanced nature of the data, with recent years having much more observations than the early 20th century. Second, we observe that prediction rates of the two methods are very close to each other but that the selected features are not. Subbagging selects almost all the features all the time, whereas the proposed method finds smaller subsets.

Let us comment on the structure of the selected sets and a pattern that appears consistent across decades. Interestingly, we find that the first 12 timbre-features are being selected repeatedly for different years of song releases, whereas amongst the last 72 timbre-correlation features, only a small portion is selected. Figure 5 tries to capture this consistency in selection across tracks from different decades. Variability within
This important paradigm extends well beyond this current work and if left for future
formations of the regularized empirical risk minimization, designed to keep computational
is of interest and fast answers are of great significance.

ular importance when dealing with extremely large datasets in which sparse recovery
recover sparsity set in the context of high dimensional linear models. This is of partic-
subagging of Lasso estimators and demonstrated ability of this method to effectively
and noise variables. We have presented new, weighted and smooth nonparametric
strap aggregation and showed its inability to successfully differentiate between signal

sixties and seventies have only a few of those present. Nonetheless, relative
difference in the sizes of the coefficients among the first 12 and the last 72 correlation
features is quite significant.

8 . Discussion

In this paper, we have outlined the causes for failure of the nonparameteric boot-
strap aggregation and showed its inability to successfully differentiate between signal
and noise variables. We have presented new, weighted and smooth nonparametric
subagging of Lasso estimators and demonstrated ability of this method to effectively
recover sparsity set in the context of high dimensional linear models. This is of partic-
ular importance when dealing with extremely large datasets in which sparse recovery
is of interest and fast answers are of great significance.

In general, the proposed method can be viewed as a sequence of stochastic approxi-
mations of the regularized empirical risk minimization, designed to keep computational
costs low, yet maintain small approximation error. These two conflicting requirements
lie in parallel to Neyman-Pearson paradigm, where Type I and Type II error cannot
be simultaneously satisfied. Type I error is now replaced with a requirement of low
computational costs and Type II error with a requirement of small statistical error.
This important paradigm extends well beyond this current work and if left for future
explorations.
EFFICIENT SUPPORT RECOVERY VIA WEIGHTED MAXIMUM-CONTRAST SUBAGGING

ACKNOWLEDGMENTS

The author would like to thank Jianging Fan and Dimitris Politis for fruitful discussions.

APPENDIX: DETAILED PROOFS

This Appendix contains details of the proofs of Theorems 1, 2 and 3, whereas the supplement contains all the details of the proofs of Theorems 4, 5, 6 and 7.

A. Preliminary Lemma

**Lemma 1.** Let \( \mathbf{w} = (w_1, \ldots, w_N) \) be a vector of exchangeable random variables that satisfies Condition [3]. Let \( P \) and \( P^* \) stand for the probability measures generated by the error vector \((\varepsilon_1, \ldots, \varepsilon_N)\) and generated jointly by the weights \( w_1, \ldots, w_N \) and the errors \( \varepsilon_1, \ldots, \varepsilon_N \). Then for a sequence \( u_n \) of real numbers bounded from below with \( n^{-1/2} \) it holds

\[
\mathbb{P}^* \left( \frac{1}{n} \sum_{i \in I^b_m} w_i |X_{ij} - \varepsilon_i| > u_n \right) \leq 2w_{2, \min} \exp \left\{ -\frac{n^2 u_n^2 / 2}{\sigma^2 \mathbb{E}_w |w_1|^2 N x^2 N + \sigma^2 u_n \mathbb{E}_w \|w\|_\infty x^N} \right\},
\]

where \( x^N := \max\{\|x^j_m\|_\infty : I^b_m \subset \{1, \ldots, n\}, |I^b_m| = N\} \) and \( x^2 N := \max\{\|x^j_m\|_2 : I^b_m \subset \{1, \ldots, n\}, |I^b_m| = N\} \).

**Proof of Lemma [7]** Let \( \mathbf{R} \) denote a random permutation uniformly distributed over the set of all permutations of \( 1, \ldots, N \) and let \( \mathbf{w} = (w_1, \ldots, w_N) \) be a vector of exchangeable random variables that satisfy Condition [3]. Note that \( X_1, \ldots, X_N \) are interdependent of \((\mathbf{w}, \mathbf{R})\). Define \( \mathbf{S} \) to be a random permutation of \( \{1, \ldots, N\} \) by requiring that \( w_{S(1)} \geq w_{S(2)} \geq \cdots \geq w_{S(N)} \) and if \( w_{S(j)} = w_{S(j+1)} \) then \( S(j) < S(j+1) \). This is one possible definition that is ambiguous to the presence of ties. Random permutation \( \mathbf{R} \) is independent of \((\mathbf{w}, \mathbf{S})\). Notice that \( \mathbb{P}^* = P_\varepsilon \times P_w \). By exchangeability of vector \( \mathbf{w} \) we have

\[
\mathbb{P}^* \left( \frac{1}{n} \sum_{i \in I^b_m} w_i |X_{ij} - \varepsilon_i| > u_n \right) = \mathbb{E}_w P_\varepsilon \left( \frac{1}{n} \sum_{i \in I^b_m} w_i |X_{ij} - \varepsilon_i| > u_n \right)
\]

\[
= \mathbb{E}_w P_\varepsilon \left( \frac{1}{n} \sum_{i \in I^b_m} w_i |X_{R(i)j} - \varepsilon_{R(i)}| > u_n \right) = \mathbb{E}_w P_\varepsilon \left( \frac{1}{n} \sum_{i=1}^N w_{S(i)} |X_{R(i)j} - \varepsilon_{R(i)}| > u_n \right)
\]

Since \( \mathbf{R} \circ \mathbf{S} \) is distributed as \( \mathbf{R} \) and independent of \( \mathbf{S} \) we have

\[
\mathbb{E}_w P_\varepsilon \left( \frac{1}{n} \sum_{i \in I^b_m} w_i |X_{ij} - \varepsilon_i| > u_n \right) = \mathbb{E}_w P_\varepsilon \left( \frac{1}{n} \sum_{i=1}^N w_{S(i)} |X_{R(i)j} - \varepsilon_{R(i)}| > u_n \right)
\]
The right hand side of previous equation is upper bounded by
\[
\mathbb{E}_w P_\epsilon \left( \frac{1}{n} \int_0^\infty \left| \sum_{i=1}^n 1\{w_{S(i)}^* \geq t\} t X_{R(i)j}^* \epsilon_{R(i)} \right| dt > u_n \right) \\
= \mathbb{E}_w P_\epsilon \left( \frac{1}{n} \int_0^\infty \sum_{i \leq n; \#\{w_{S(i)}^* \geq t\} = N} t X_{R(i)j}^* \epsilon_{R(i)} \right| dt > u_n \right)
\]

Then, the following chain of inequalities hold as an upper bound of the right hand side of the last equation
\[
\sum_{\text{all possible} \ w} P_w \left( w_{S(1)} = w_1, \ldots, w_{S(N)} = w_N \right) \sum_{i=1}^N w_{S(i)} = n \left( \frac{1}{n} \sum_{i=1}^N w_i \left| X_{R(i)j}^* \epsilon_{R(i)} \right| \geq u_n \right) \\
\leq \int_0^\infty \left( \sum_{i=1}^N 1\{w_{S(i)} > t\} = N \left( \sum_{i=1}^N w_{S(i)} = n \right) \right) P_\epsilon \left( \frac{1}{n} \sum_{i=1}^N t \left| X_{R(i)j}^* \epsilon_{R(i)} \right| \geq u_n \right) dt \\
\leq \int_0^{n/N} \prod_{i=1}^N P_w \left( w_{S(i)} > t \right) P_\epsilon \left( \frac{1}{n} \sum_{i=1}^N t \left| X_{R(i)j}^* \epsilon_{R(i)} \right| \geq u_n \right) dt \\
\leq \int_0^{n/N} P_w \left( \min_{1 \leq i \leq N} w_{S(i)} > t \right) P_\epsilon \left( \frac{1}{n} \sum_{i=1}^N t \left| X_{R(i)j}^* \epsilon_{R(i)} \right| \geq u_n \right) dt
\]

Let us denote with \( p_t \) tail probability on the right hand side of previous equation, i.e.
\[
p_t = P_\epsilon \left( \frac{1}{n} \sum_{i=1}^k t \left| X_{R(i)j}^* \epsilon_{R(i)} \right| \geq u_n \right)
\]

By direct application of tail inequality for Gaussians we have the following simple bound
\[
(14) \quad p_t \leq 2 \exp \left\{ - \frac{n^2 u_n^2}{2 \sigma^2 t^2 \sum_{i=1}^N X_{R(i)j}^2 + 2 \sigma^2 u_n nt \max_{1 \leq i \leq N} \left( X_{R(i)j} \right) } \right\},
\]

By applying Caushy Shwarz and Jensen inequalities, equation \((13)\) can be upper bounded with
\[
(15) \quad \sqrt{\int_0^{n/N} \frac{p_t^2 \left( \min_{1 \leq i \leq N} w_{S(i)} \right) \leq t \right) dt} \sqrt{\int_0^\infty p_t^2 dt}
\]
To bound last term we can use (14) to conclude that
\[
\int_0^\infty p_t^2 dt \leq 4 \exp \left\{ -\frac{n^2 u_n^2}{\sigma^2 \sum_{i=1}^N \mathbb{E}_w [u_i^2] X_{\mathbf{R}(i)}^2 + \sigma^2 u_n n \mathbb{E}_w \max_{1 \leq i \leq N} (W_i X_{\mathbf{R}(i)})} \right\}
\leq 4 \exp \left\{ -\frac{n^2 u_n^2}{\sigma^2 \mathbb{E}_w [u_i^2] N \|x^{(2)}_R\|_{\infty,2} + \sigma^2 u_n n \mathbb{E}_w [\|w\|_{\infty}] \|x^{(2)}_R\|_{\infty}} \right\}
\]
where \( x^{(2)}_R \) stands for the \( j \)-th column of matrix \( X_{\mathbf{R}(t_n^b)} \), \( \|x^{(2)}_R\|_{\infty,2} \) for the maximum of the squared and \( \|x^{(2)}_R\|_{\infty} \) maximum of the absolute values of elements of the \( n \)-vector \( x^{(2)}_R \). Then, by observing simple relations \( \|x^{(2)}_R\|_{\infty} \leq x^{(2)}_N \) := \( \max \{\|x^{(2)}_R\|_{\infty} : R \in \text{set of all permutations of} \{1, \cdots, N\}\} \), \( \|x^{(2)}_R\|_{\infty,2} \leq x^{(2)}_N \) := \( \max \{\|x^{(2)}_R\|_{\infty,2} : R \in \text{set of all permutations of} \{1, \cdots, N\}\} \) for (15) is bounded from above with
\[
2w_{\text{min}} \exp \left\{ -\frac{n^2 u_n^2}{\sigma^2 \mathbb{E}_w [u_i^2] N x^{(2)}_N + \sigma^2 u_n n \mathbb{E}_w [\|w\|_{\infty}] x^{(2)}_N} \right\}
\]
where in the last step we used \( w_{\text{min}} = \sqrt{\int_0^\infty P_w^2 (\min_{1 \leq i \leq N} w_i > t) dt} \).

\[ \square \]

B. Proofs of Section 2

Major part of the proof of Theorem 1 lies in the following two simple lemmas and major results of Theorems 5 and 6 analyzing sparsity patterns of the sub-lasso and Lasso estimators.

B.1. Preliminary Lemmas. Let \( \hat{\beta}^b_m \) be the sub-lasso estimator and \( \hat{\beta} \) Lasso estimator throughout this section.

Lemma 2. For all \( j \) for which \( \hat{\beta}^b_{m,j}(\lambda_N) = 0 \) and \( |\sum_{i \in I_m} (Y_i - X_i^T \hat{\beta}^b_{m}(\lambda_N)) X_{ij}| \leq n(\lambda_n - \lambda_N) \), we have \( \hat{\beta}_{j}(\lambda_n) = 0 \).

Proof of Lemma 2 Since both optimization problems are convex it suffices to analyze their respective Karush-Kuhn-Tucker (KKT) optimality conditions. For all \( j \in \{1, \cdots, p\} \) such that \( \beta_j^* = 0 \), Lasso estimator needs to satisfy \( |\sum_{i=1}^n 2(Y_i - X_i^T \hat{\beta}(\lambda_n)) X_{ij}| \leq n \lambda_n \), where KKT for sub-lasso estimator are \( |\sum_{i \in I_m} 2w_i(Y_i - X_i^T \hat{\beta}^b_{m}(\lambda_N, k)) X_{ij}| \leq n \lambda_N \). Then we are left to check if under conditions of the lemma sub-lasso estimator \( \hat{\beta}^b_m \) satisfies KKT condition of Lasso estimator. That calculation follows easily from the following inequality
\[
\frac{1}{n} \left| \sum_{i=1}^n 2(Y_i - X_i^T \hat{\beta}^b_{m}) X_{ij} \right| \leq \frac{1}{n} \left| \sum_{i \in I_m} 2w_i(Y_i - X_i^T \hat{\beta}^b_{m}) X_{ij} \right| + \frac{1}{n} \left| \sum_{i \notin I_m} 2(Y_i - X_i^T \hat{\beta}^b_{m}) X_{ij} \right| \leq \lambda_n
\]
where the first term in the rhs is bounded by $\lambda_N$ and the second with $\lambda_n - \lambda_N$ (by the assumption of the Lemma).

**Lemma 3.** For all $j$ such that $\hat{\beta}_j(\lambda_n) = 0$ and $|\sum_{i \in I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}| \leq \lambda_N - n\lambda_n$, then $\hat{\beta}^b_{m,j}(\lambda_N, k) = 0$.

**Proof of Lemma 3.** Let us assume that the KKT conditions of the plain Lasso estimator, computed over the complete dataset hold i.e. that for those $j$ such that $\hat{\beta}_j = 0$, $\frac{1}{n} \sum_{i=1}^{n} 2(Y_i - X_i \hat{\beta})X_{ij} \leq \lambda_n$. Note that

$$\frac{1}{n} \sum_{i=1}^{n} 2(Y_i - X_i \hat{\beta})X_{ij} \geq \left| \frac{1}{n} \sum_{i \in I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij} \right| - \frac{1}{n} \sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}$$

By analyzing two cases individually: Case (I): $|\sum_{i \in I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}| \leq |\sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}|$, and Case (II): $|\sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}| \geq |\sum_{i \in I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij}|$, we have

$$\frac{1}{n} \sum_{i \in I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij} \leq \frac{1}{n} \sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij} + \lambda_n$$

to hold in both cases. With it we can then see that

$$\frac{1}{n} \sum_{i \in I^b_n} 2w_i(Y_i - X_i \hat{\beta})X_{ij} \leq \frac{1}{n} \sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij} + n\lambda_n$$

since each $w_i \leq n$ for all $i$ almost surely. Hence to show that $\hat{\beta}$ satisfies KKT for sub-Lasso as well, we need $\frac{1}{n} \sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta})X_{ij} \leq \frac{\lambda_N - \lambda_n}{n}$.

**B.2. Proof of Theorem 1.** Previous Lemmas are stated for general sub-Lasso estimator but can be easily adapted to the bagged estimator. With their help and results of Theorems 5 and 6 we are ready to finalize the proof of Theorem 1. Equivalent of Lemma 2 requires $|\sum_{i \notin I^b_n} (Y_i - X_i \hat{\beta}(\lambda_N^1))X_{ij}| \leq n\lambda_n - N\lambda^1_N$ to hold, whereas equivalent of Lemma 3 requires $|\sum_{i \notin I^b_n} 2(Y_i - X_i \hat{\beta}(\lambda_n)X_{ij}) \leq N\lambda_N - n\lambda_n$ to hold. The proof follows easily as a consequence of results obtained in Theorems 3, 5 and 6.

First, on the subject of approximating sparse recovery, we conclude that it is not possible as a consequence of the proof of Theorem 3. For bagged estimator, there exists no feasible $q$ that is different from zero which solves the equivalent of (24). Equivalent of (24) would require, on one side $q > \sqrt{\log p / N}$ and on the other $q < \sqrt{\log p / n}$. For $N \ll n$ this is not possible as $\sqrt{\log p / N} > \sqrt{\log p / n}$. For a special case of $N = n/k$, only the choice of $k = 1$ and fixed, not divergent $s$ and $p$, will allow both conditions to be satisfied.
Second, on the subject of the exact sparse recovery, as a consequence of previous equivalent of Lemma 2 and Theorem 5 we have

\[(16) \quad \lambda_n^{-1}(1 + \frac{N}{n} + \lambda_n^{-1}(X_p^T \hat{A} X_p)) + \sigma \sqrt{\frac{\log p}{n}} \left( 1 + 2 \sqrt{\frac{s \lambda_{\text{max}}(X^T \hat{X})}{\sqrt{n} \zeta_N}} \right) \leq \lambda_n \]

As a consequence of equivalent of Lemma 3 and Theorem 6 we have

\[(17) \quad \lambda_n(2 + \lambda_n^{-1}(X_p^T \hat{A} X_p)) + \sigma \sqrt{\frac{n - N}{n}} \left( 1 + 2 \sqrt{\frac{s \lambda_{\text{max}}(X^T X)}{\sqrt{\zeta_n}}} \right) \leq \frac{N}{n} \lambda_n \]

If \( N \ll n \) then, from the above contradictory conditions we can see that subagged estimator can not have the exact same sparsity of the Lasso estimator,

\[P(\exists \lambda_n \geq 0, \exists \lambda_N \geq 0 : \hat{S}(\lambda_n) = \hat{S}^b(\lambda_N)) = 0.\]

For a special case of \( N = n/k \) we see that the only choice of \( k = 1 \) and fixed, not divergent \( s \) and \( p \) will allow equations (16) and (17) to be satisfied up to a constant. That is, there exist two constants \( 0 < c < \infty \) and \( 0 < c_1 < \infty \) such that for the choice of \( \lambda_n = c \lambda^1_N \geq c_1 n^{-1/2} \) (i.e. result of [2] only holds for fixed not divergent \( n \))

**B.3. Proof of Theorem 2.** If RE condition holds on the bootstrapped data matrix, then the result of this Proposition follows by repeating the steps of proof of Theorem 4 with simplification of no weighting scheme \( w \) to obtain

\[\left\| \sum_{j \in S} |\hat{\beta}_{m,j} - \beta^*_j| \right\|_2 \leq \frac{\left\| X_{ib} (\hat{\beta}_m - \beta^*) \right\|_2}{\zeta_N \sqrt{n}}.\]

Following the steps parallel to those in [3], we can obtain the predictive bounds of the order of \( s \lambda_N / \zeta^2_n \), for \( N = n/k \) and \( \lambda_N \geq 2\sigma \sqrt{2k \log p/n} \). From the classical results on Lasso prediction bounds, we know that optimal \( \lambda_n \geq 2\sigma \sqrt{2 \log p/n} \). The statement of the theorem follows if we are able to bound the following expression

\[\delta_n = \left| \frac{\lambda_n}{\zeta^2_n} - \frac{\lambda_n}{\zeta^2_n} \right|\]

Let us write \( |\zeta^2_n - \zeta^2_n| = \zeta^2_n - \zeta^2_n = \epsilon_n \) for some \( \epsilon_n \geq 0 \). Then for \( \lambda_N \leq 2\sigma \sqrt{2k \log p/n} \) we have

\[\delta_n = \left| \frac{\lambda_n}{\zeta^2_n - \epsilon_n} - \frac{\lambda_n}{\zeta^2_n} \right| < \frac{2\sigma \sqrt{\log p/n}}{\zeta^2_n} \left( \sqrt{k} \frac{1}{1 - \frac{\epsilon_n}{\zeta^2_n}} - 1 \right)\]
Now we claim that if \( k \leq 4 \) then \( \sqrt{k \frac{1-\epsilon_n^2}{\zeta_n^2}} - 1 \leq C \) for some bounded constant \( C > 1 \).

This claim is equivalent to claiming that \( \frac{1+\epsilon_n^2}{1-\epsilon_n^2} \leq C \), that is \( \epsilon_n/\zeta_n^2 \leq \frac{C-1}{C+1} \). But, from \( \text{RE}(N) \) condition we know that \( 0 \leq \epsilon_n/\zeta_n^2 < \eta \) for some constant \( \eta < 1 \). Hence constant \( C > 1 \) that satisfies above properties is \( \left( \eta + 1 \right) / \left( 1 - \eta \right) \). Therefore, we can conclude that \( \delta_n \leq \frac{2C \sigma}{\zeta_n} \sqrt{\log p_n} \).

\[ \square \]

### C. Proof of Theorem 3

Main ingredient of the proof is based on the intermediary results (of significant importance on its own), stated in Theorems 5 and 6. First part of the statement follows by utilizing Theorem 6 in order to conclude that \( S \subseteq \hat{S}_\tau \) with probability close to 1. Unfortunately, as conditions of Theorems 5 contradict those of Theorem 6 we cannot easily apply them to conclude the second part of the statement of Theorem 3. Therefore, we develop a new method for finding the optimal value of the tuning parameter \( \lambda_N \). It is based on the optimal bias-variance tradeoff, where bias is replaced with variable selection error and variance with prediction error. It allows good, but not the best prediction properties while obtaining desired variable selection properties. For that end, we will split the proof into two parts. The first is based on bounding the number of false positives and second is based on finding the optimal choice of \( \lambda_N \) that implies results of Theorem 6 and bounds false positives simultaneously.

#### C.1. Bounding False Positives

Assume that the distribution of \( \{1(j \in S^b_m(\lambda_N), j \in S^c\}, \) with \( S^b_m(\lambda_N) = \cup_w S^b_m(\lambda_N, w) \) is exchangeable for all \( \lambda_N \). Also, assume that the approximate randomized lasso procedure is not worse than a random guess, i.e. for

\[
|N|E \left[ |S \cap \hat{S}_m^b(\lambda_N)| \right] \geq |S|E \left[ |N \cap \hat{S}_m^b(\lambda_N)| \right]
\]

then the expected number of falsely selected variables is bounded by

\[
E[|N \cap \hat{S}_\tau|] \leq \frac{2\sqrt{B}(E[\hat{S}_1(\lambda_n)])^M}{p^{M-1}(1 + \sqrt{B})},
\]

for all choices of \( \tau > (2 + \sqrt{B})/(2 + 2\sqrt{B}) \).

**Proof**

Let us fix \( b \) and denote sets \( I^b_m \) with \( I_m \) (i.e. random split of the samples \( \{1, \ldots, n\} \) into \( M \) disjoint subsets each of size \( n^{\gamma} \)). Then, similar to [20] we can define a binary random variable \( H^K_m = 1\{j \subseteq \cap_{m=1}^M \hat{S}_m(\lambda_N)\} \) for all variables \( j \subseteq \{1, \ldots, p\} \). Then minimax selection probability is expressed as a function of simultaneous selection probability

\[
\pi^*_j(\lambda_n) := \frac{\sqrt{B}}{1 + \sqrt{B}} \mathbb{P}^*(j \subseteq \cap_{m=1}^M \hat{S}_m(\lambda_N)) + \frac{1}{2(1 + \sqrt{B})} = \frac{\sqrt{B}}{1 + \sqrt{B}} E(H^*_j | Z) + \frac{1}{2(1 + \sqrt{B})}
\]
where the probability \( P^* \) is with respect to the random sample splitting. Here \( Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \) denotes the whole original sample of size \( n \). Then, for \( p^*_j(\lambda_N) = P^*(j \subseteq \cap_{m=1}^M \hat{S}_m(\lambda_N)) \) we have

\[
P(\pi^*_j(\lambda_n) \geq \tau) = P\left(p^*_j(\lambda_N) \geq \frac{1 + \sqrt{B}}{\sqrt{B}} \left( \tau - \frac{1}{2(1 + \sqrt{B})} \right) \right)
\]

Then, \( E(H_j^\lambda | Z) = E(p^*_j(\lambda_n)) = P(H_j^\lambda = 1) \leq \prod_{m=1}^M P(K \subseteq \hat{S}_m(\lambda_n)) = P(j \subseteq \hat{S}_1(\lambda_n))^M \). By Markov inequality we know that \( P(p^*_j(\lambda_n) \geq \xi) \leq E(p^*_j(\lambda_n))/\xi \leq P(j \subseteq \hat{S}_1(\lambda_n))^M/\xi \). By arguments of Theorem 1 in [16] we have that \( P(j \subseteq \hat{S}_1(\lambda_n)) \leq E|\hat{S}_1(\lambda_n)|/p \), and hence for a threshold \( \tau \geq 1/2 + \frac{1}{2(1 + \sqrt{B})} \) we have

\[P(\pi^*_j(\lambda_n) \geq \tau) \leq \frac{\sqrt{B}E|\hat{S}_1(\lambda_n)|^M}{(1 + \sqrt{B})(\tau - \frac{1}{2(1 + \sqrt{B})})^p^M} \leq 2 \frac{\sqrt{B} E|\hat{S}_1(\lambda_n)|^M}{1 + \sqrt{B} p^M - 1}.
\]

This leads to the conclusion that \( E[|S^c \cap \hat{S}|] = \sum_{j \in S^c} P(\max_{\lambda_n} \pi^*_j(\lambda_n) \geq \tau) \leq 2 \frac{\sqrt{B} E|\hat{S}_1(\lambda_n)|^M}{1 + \sqrt{B} p^M - 1} \).

C.2. Optimal choice of \( \lambda_N \). Let us show that proposed aggregated sub-Lasso estimators are better than random guessing in the sense of (18). As expected, such property will not be achieved for all values of \( \lambda_N \), as for quite small values of \( \lambda_N \), too many variables (possibly even \( p \)) might be selected. By analyzing equation (18) and utilizing results of Theorem 6 we infer that the sequence of \( \lambda_N \) that achieves control of false positives and allows results of Theorem 6 to hold, is the optimum of the following stochastic optimization problem

\[
\min_q \quad q
\]

\[
s.t. \quad E|\hat{S}_m(\lambda_N)| \leq \frac{1}{2} p^{1-1/M}
\]

\[
P(A_q(\lambda_N)) \geq 1 - \delta
\]

\[
\lambda_N > 0, q \geq 0, \delta > 0
\]

where the events \( A_q(\lambda_N) = \cap_{j=1}^p \{ 2 \frac{1}{n} \sum_{i \in I_m} w_i |\varepsilon_i X_{ij}| \leq \lambda_N - q \} \). Although the problem (19) is stated in terms of \( q \), we will show that the optimal value of \( q \) leads to the optimal value of \( \lambda_N \).

Let us give a few comments on the stochastic optimization problem. The first constraint bounds the number of selected variables and comes as a consequence of Theorem 6 and condition (18). It is sufficient to guarantee that sub-Lasso estimators are better than random guessing (i.e. that (18) is satisfied) even when IR(N) conditions are not fulfilled. The second constraint restricts our attention to a sequence of random coverage sets \( A_q(\lambda_N) \). As Theorem 6 controls variable selection properties and first constraint
intrinsically controls predictive properties, they cannot be satisfied simultaneously on sets \( A_0(\lambda_N) \). For \( q = 0 \) best prediction is still achievable but variable selection (Theorem 5) is not. Hence, we need to allow for possible deviation of the smallest sets \( A_0(\lambda_N) \) by allowing small perturbations of size \( q \). Our goal is to find the smallest possible perturbation \( q \) which allow high probability bounds on the selection of the false negatives and simultaneously controls the size of the selected sets in sub-lasso estimators to allow small number of false positives.

Let us represent conditions of stochastic problem (19) in a concise way. Note that from optimality conditions of each of the sub-Lasso problems we can see that whenever \( \beta_{m,j}^b(\lambda_N, k) \neq 0 \)

\[
\frac{1}{n} \sum_{i \in I_m^b} w_i^k \|(Y_i - X_i^T \hat{\beta}_{m}^b(\lambda_N, k))X_{ij}\|_\infty \geq \frac{\lambda_N}{2}.
\]

Then, on the event \( A_q \) from Theorem 4 we have the following

\[
|\hat{S}_m^b(\lambda_N, k)| \leq \frac{4}{n^2 \lambda_N^2} \sum_{j \in \hat{S}_m^b(\lambda_N, k)} \sum_{i \in I_m^b} w_i^k \left( X_i^\beta* - X_i^\beta_{m}^b(\lambda_N, k) \right)X_{ij}^2
\]

\[
\leq \frac{4 \lambda_{\max}(X^T I_m^b X I_m^b)}{\lambda_N^2 n^2} \| \hat{X}_k(\beta^* - \beta_{m}^b(\lambda_N, k)) \|_2^2
\]

and \( \frac{1}{n} \| \hat{X}_k(\beta^* - \beta_{m}^b(\lambda_N, k)) \|_2^2 \leq \frac{(16 \lambda_N^2 + q^2)s}{\xi_0^2} \). Hence, from all above and tower property of expectations we can conclude that

\[
E|\hat{S}_m^b(\lambda_N)| \leq \frac{(64 \lambda_N^2 + 4q^2)sK}{N \lambda_N^2 N^2} E^*(\lambda_{\max}(X^T I_m^b X I_m^b)).
\]

Now we are left to evaluate the size of sets \( A_q \). This step of the proof is based on a Bernstein’s inequality for exchangeable weighted bootstrap sequences contained in an intermediary result Lemma 1. From Lemma 1 we have, \( P \left( | \sum_{i \in I_m^b} \left( w_i^k X_{ij} \varepsilon_i \right) > nt \sigma \sqrt{2} \right) \) is less then or equal to

\[
2 \exp \left\{ - \frac{nt^2}{\mathbb{E}[w_i^2] \| \hat{X}_m \|_\infty^2 + t \mathbb{E}[\| w \|_\infty] \| \hat{X}_m \|_\infty} \right\},
\]

Furthermore, for \( \mathbb{E}[w_i^2] < \infty \) and \( \mathbb{E}[\| w \|_\infty] < \infty \) there exist \( c_1 \) and \( c_2 \) such that if \( c_2 \sqrt{\log p/N} \geq t \geq c_1 \sqrt{\log p/n} \) and \( \| \hat{X}_m \|_\infty \leq c_2 \sqrt{N/(\log p \mathbb{E}[\| w \|_\infty])} \) then there exists a positive constant \( c > 1 \)

\[
P \left( \frac{1}{n} \left| \sum_{i \in I_m^b} w_i^k X_{ij} \varepsilon_i \right| > t \sigma \sqrt{2} \right) \leq 2 \exp \left\{ - c \frac{nt^2}{\| \hat{X}_m \|_\infty^2} \right\}.
\]
For a choice of $t$ of $\lambda_N - q$ we then have $P(A_q(\lambda_N)) \geq 1 - 2 \exp \{ -cn(\lambda_N - q)^2/\|x_{I_{bm}}^j\|_{\infty,2} \} \to 1$. Hence the solution to the stochastic optimization problem (19) also solves

\begin{equation}
\min_{q \geq 0} \quad P(A_q(\lambda_N)) \geq 1 - 2 \exp \{ -cn(\lambda_N - q)^2/\|x_{I_{bm}}^j\|_{\infty,2} \}
\end{equation}

for some constants $c_1 < c_2$. The RHS of the last constraint inequality is a consequence of Theorem 4 (which is used numerous times in the steps of the proof). First constraint of the above problem can be reformulated as

\[ \lambda_N^2 > \frac{32q^2}{N^2} E^* \lambda_{\text{max}} (X_{I_{bm}}^T X_{I_{bm}}) \left( 1 - \frac{16sp^{1/M}}{N^2 \lambda_{\text{max}} (X_{I_{bm}}^T X_{I_{bm}})} \right)^{-1} \]

Then we can see that the optimal value of $q$ is of the order of $c_3 \sigma \sqrt{\log p/n}$ for

\[ c_3 = \sqrt{\frac{n - N}{n} \frac{\sqrt{\zeta_N} + 2s \lambda_{\text{max}} (X_{I_{bm}}^T X_{I_{bm}})}{2} \sqrt{\zeta_N}} < 1. \]

Notice that such value allows sets $A_q(\lambda_N)$ to have large coverage probability. Such optimal choice of $q$ results in the choice of the optimal value of the tuning parameter $\lambda_N$ as follows $(c_1 - c_3)\sigma \sqrt{\log p/n} \leq \lambda_N \leq (c_2 - c_3)\sigma \sqrt{\log p/n}$, satisfies all previous constraints, hence leading to the statement of the Theorem.

\[ \square \]

**D. Proof of Theorem 4**

Although sub-Lasso can be reformulated as modified Lasso estimator, sparse oracle inequalities for Lasso cannot straightforwardly apply as the error terms in the modified problem are not independently distributed anymore. Main results of Theorem 4 are proved using the novel, preliminary result of Lemma 1. From the definition of sub-Lasso problem we have for every realization of random weights $w$, $\frac{1}{n} \sum_{i \in I_{bm}, |I_{bm}| = N} w_i^k (Y_i - X_i \hat{\beta}_m^b (\lambda_N, k))^2 + 2\lambda_N \| \hat{\beta}_m^b (\lambda_N, k) \|_1 \leq \frac{1}{n} \sum_{i \in I_{bm}, |I_{bm}| = N} w_i^k (Y_i - X_i \beta)^2 + 2\lambda_N \| \beta \|_1$ holds for any value of $\beta$. For simplicity of notation we have suppressed the dependence $\hat{\beta}_m^b$ of
\( \lambda_N \) and \( k \). By using \( Y_i = X_i\beta^* + \varepsilon_i \), and setting \( \beta = \beta^* \), previous becomes equivalent to

\[
\frac{1}{n} \sum_{i \in I_m^h \cap |I_m^h| = N} w_i^k (X_i\beta^* - X_i\hat{\beta}_m^b)^2 \leq 2 \frac{\|\hat{\beta}_m^b - \beta^*\|_1}{n} \max_{1 \leq j \leq p} \left( \sum_{i \in I_m^h \cap |I_m^h| = N} w_i^k |\varepsilon_i X_ij| \right) + 2\lambda_N \|\beta^*\|_1 + 2\lambda_N \|\hat{\beta}_m^b\|_1
\]

In the rest of the proof we are providing evidence to a result more stronger then the one in the statement of the Theorem 2. This is done, as such a result is needed for Theorem 1. Consider the event \( A_q(\lambda_N) = \bigcap_{j=1}^p \left\{ 2\frac{1}{n} \sum_{i \in I_m^h \cap |I_m^h| = N} w_i^k |\varepsilon_i X_ij| \leq \lambda_N - q \right\} \), for \( q < \lambda_N \). Using the fact that \( w_i \geq 1 \) for all \( i \in I_m^h \) we have the following

\[
\frac{1}{n} \min_{1 \leq i \leq N} w_i \|X I_m^h (\beta^* - \hat{\beta}_m^b)\|_2^2 + \lambda_N \|\hat{\beta}_m^b - \beta^*\|_1
\]

\[
\leq \frac{1}{n} \sum_{i \in I_m^h \cap |I_m^h| = N} w_i (X_i\beta^* - X_i\hat{\beta}_m^b)^2 + \lambda_N \|\hat{\beta}_m^b - \beta^*\|_1
\]

\[
\leq (2\lambda_N - q) \|\hat{\beta}_m^b - \beta^*\|_1 + 2\lambda_N \|\beta^*\|_1 - 2\lambda_N \|\hat{\beta}_m^b\|_1
\]

which leads to the first conclusion. From the previous result we can also see that \( \lambda_N \|\hat{\beta}_m^b - \beta^*\|_1 \leq (4\lambda_N - q) \sum_{j \in S} |\hat{\beta}_{m,j}^b - \beta_{j}^*| \leq (4\lambda_N - q) \sqrt{s} \sum_{j \in S} |\hat{\beta}_{m,j}^b - \beta_{j}^*|^2 \) leading to the conclusion that \( \hat{\beta}_m^b - \beta^* \in \mathbb{C}(3) \) and using the expected Restricted Eigenvalue Condition on the set \( \Omega^\ast \) we have

\[
\sqrt{\sum_{j \in S} |\hat{\beta}_{m,j}^b - \beta_{j}^*|^2} \leq \mathbb{E}_w \| \tilde{X}(\hat{\beta}_m^b - \beta^*) \|_2 / (\zeta_N \sqrt{n}).
\]

Here \( \mathbb{E}_w \) denotes expectation taken with respect to probability measure generated by \( w \). Using Jensen’s inequality for concave functions and independence of the weighting scheme \( w \) of vectors \( X_i \), we have

\[
\mathbb{E}_w \| \tilde{X}(\hat{\beta}_m^b - \beta^*) \|_2 \leq \mathbb{E}_w [\tilde{X}] (\hat{\beta}_m^b - \beta^*) \|_2 \leq \mathbb{E}_w \| \tilde{X} \|_\infty \| X I_m^h (\hat{\beta}_m^b - \beta^*) \|_2 .
\]

Combining previous inequalities we have

\[
\frac{1}{n} \min_{1 \leq i \leq N} w_i \|X I_m^h (\hat{\beta}_m^b - \beta^*)\|_2 \leq (4\lambda_N - q) \sqrt{s} \mathbb{E}_w \| \tilde{X} \|_\infty \| X I_m^h (\hat{\beta}_m^b - \beta^*) \|_2 / \sqrt{n} \zeta_N .
\]

leading to \( \|X I_m^h (\beta^* - \hat{\beta}_m^b)\|_2 \leq (4\lambda_N - q) \sqrt{s} \mathbb{E}_w \| \tilde{X} \|_\infty / \zeta_N (\min_{1 \leq i \leq N} w_i) \). Hence, if we define \( a_N \) as such that event \( \{ \min_{1 \leq i \leq N} w_i \geq \mathbb{E}_w \| \tilde{X} \|_\infty / a_N \} \) has probability close
to 1, then
\[
\|X_{I^b_m}(\beta^* - \hat{\beta}_m^b)\|^2_2 \leq \frac{(4\lambda_N - q)^2sn}{\zeta^2 N} a_N^2.
\]
Size of set \(\mathcal{A}_q(\lambda_N)\) can be deduced from Lemma 1 and Theorem 4 and is hence omitted. \(\square\)

E. Proofs of Section 5

E.1. Proof of Theorem 5. Using results of Lemma 2 it is sufficient to show that for every sub-Lasso estimator \(\hat{\beta}_m^b\) the event \(\Omega_n = \{\frac{1}{n}\sum_{i \in I^b_m}(Y_i - X_i\beta_m^b)X_i^T I \leq \lambda_n - \lambda_N^b\}\) happens with large probability. In other words, sub-Lasso will be one of the possible optima for Lasso if we can evaluate its behavior on the data on which it was not built on.

To that end, let us introduce notation of the following kind. Let
\[
\frac{1}{n} \sum_{i \notin I^b_m} X_{ij}(Y_i - X_i\hat{\beta}_m^b) = X_{m,j}^T (Y_{m} - X_{m} \hat{\beta}_m^b)
\]
where \(I^b_m\) was used as a short hand notation for the compliment of \(I^b_m\). With \(\hat{\beta}_m^b\), a set of non-zero coefficient of \(\beta_m^b\), let us denote with \(\hat{\beta}_m^b = \hat{\beta}_m^b \cup \hat{\beta}_m^s\), where \(\hat{\beta}_m^s\) is a set of non-zero coefficient of Lasso estimator \(\hat{\beta}\). Let \(P_{\hat{\beta}}\) be a projection operator into the space spanned by all variables in the set \(\hat{\beta}\), that is
\[
P_{\hat{\beta}} = X_{m,\hat{\beta}}(X_{m,\hat{\beta}}^T X_{m,\hat{\beta}})^{-1} X_{m,\hat{\beta}}.
\]
Then, we can split the term \(X_{m,j}^T (Y_{m} - X_{m} \hat{\beta}_m^b)\) into two terms
\[
X_{m,j}^T (I - P_{\hat{\beta}})(Y_{m} - X_{m} \hat{\beta}_m^b) + X_{m,j}^T P_{\hat{\beta}}(Y_{m} - X_{m} \hat{\beta}_m^b).
\]
Controlling the size of the set \(\Omega_n\) is equivalent to upper bounding the RHS of the previous equation.

E.1.1. Step I: Controlling \(X_{m,j}^T P_{\hat{\beta}}(Y_{m} - X_{m} \hat{\beta}_m^b)\). This term is upper bounded by
\[
X_{m,j}^T X_{m,j} (X_{m,\hat{\beta}}^T X_{m,\hat{\beta}})^{-1} (\text{sign}_m^b) \lambda_N \leq \lambda_N \|X_{m,j}^T X_{m,j} (X_{m,\hat{\beta}}^T X_{m,\hat{\beta}})^{-1} X_{m,j}\|_\infty.
\]
Last term in the RHS is bounded by
\[
\lambda_N \|X_{m,j}^T X_{m,j}\|_\infty + \lambda_N \|X_{m,j} (X_{m,\hat{\beta}}^T X_{m,\hat{\beta}})^{-1} - (X_{m,\hat{\beta}}^T X_{m,\hat{\beta}})^{-1} X_{m,j}\|_\infty.
\]
Since irrepressible condition holds on the whole data then the first term in bounded by \(\lambda_N\). The second term of the above equation is bounded from above by
\[
\lambda_N \|X_{m,j}^T X_{m,j}\|_\infty^2 - (X_{m,j}^T X_{m,j})^{-1} \|X_{m,j}^T X_{m,j}\|_2
\]
where \( \| \| \) stands for operator norm of a matrix. Note that by definition of operator norm we have that for two semi-positive definite matrices \( A, B \)
\[
\| A^{-1} - B^{-1} \| = \max\{ |\lambda_{\min}(A^{-1} - B^{-1})|, |\lambda_{\max}(A^{-1} - B^{-1})| \}.
\]
Using Weyl's inequalities \(^3\) we have
\[
\frac{1}{\lambda_{\max}(A)} - \frac{1}{\lambda_{\max}(B)} \leq \lambda_{\max}(A^{-1} - B^{-1}) \leq \frac{1}{\lambda_{\min}(A)} - \frac{1}{\lambda_{\min}(B)},
\]
\[
\frac{1}{\lambda_{\min}(A)} - \frac{1}{\lambda_{\max}(B)} \leq \lambda_{\max}(A^{-1} - B^{-1}) \leq \frac{1}{\lambda_{\min}(A)} - \frac{1}{\lambda_{\min}(B)},
\]
where we used \( \lambda_{\max}(-B^{-1}) = -\lambda_{\min}(B^{-1}) = -\lambda_{\max}^{-1}(B) \) and \( \lambda_{\max}(A^{-1}) = \lambda_{\max}^{-1}(A) \). Therefore,
\[
\| A^{-1} - B^{-1} \| \leq \frac{1}{\lambda_{\min}(A)} - \frac{1}{\lambda_{\min}(B)} \leq \frac{1}{\lambda_{\min}(A)}
\]
Using \( B = X^T \hat{A} X \hat{A} \) and \( A = X^T \hat{p}_m A X \hat{p}_m \hat{A} \) we can see
\[
\left\| (X^T \hat{p}_m A X \hat{p}_m \hat{A})^{-1} - (X^T \hat{A} X \hat{A})^{-1} \right\| \leq \lambda_{\min}^{-1}(X^T \hat{p}_m A X \hat{p}_m \hat{A}).
\]
We are left to bound the last term on the RHS of (26). As \( X^T \hat{A} X \hat{A} \) is the projection of \( X \) into the space spanned by columns of \( X \hat{A} \) we have that \( \| X^T \hat{p}_m A X \hat{p}_m \hat{A} \|_2 \leq \| X^T \hat{A} X \hat{A} \|_2 \leq 1 \).

E.1.2. Step II: Controlling \( X^T \hat{p}_m \hat{A}(I - P_A)(Y \hat{p}_m - X \hat{p}_m \hat{b}) \). This term is then upper bounded by
\[
\sup_{A; |A| \leq 2m} \sup_{j \notin A} \left| X^T \hat{p}_m \hat{A}(I - P_A)\varepsilon \right|.
\]
Right hand side of the previous inequality is then bounded by
\[
\max_{j \in \{1, \ldots, p\}} \left| X^T \hat{p}_m \hat{A}\varepsilon \right| + \sup_{A; |A| \leq 2m} \| P_A \varepsilon \|_2.
\]
Since \( \varepsilon \) are i.i.d. normal \( \mathcal{N}(0, \sigma^2) \) we have that there exists a constant \( c > 0 \) such that
\[
P\left( \frac{n}{n - N} \max_{j \in \{1, \ldots, p\}} \left| X^T \hat{p}_m \hat{A}\varepsilon \right| \geq 2\sigma \sqrt{\frac{\log p}{n - N}} \right) \leq \exp\{-c \log p\}
\]
Using similar arguments as Lemma 6 in \(^{10}\) we have that
\[
P\left( \frac{n}{n - N} \sup_{A; |A| \leq 2m} \| P_A \varepsilon \|_2 \geq 2\sigma \sqrt{\frac{4\delta \log p}{n - N}} \right) \leq \exp\{-c \log p\}
\]
(exact steps omitted). Combining all the previous results of Step I and Step II we have that

$$X_{\mathbf{p}_m, j}^T (Y_{\mathbf{p}_m} - X_{\mathbf{p}_m} \hat{\beta}_{\mathbf{p}_m}) \leq \lambda_N (1 + \lambda^{-1}_{\min} (X_{\mathbf{p}_m, A}^T X_{\mathbf{p}_m, A})) + \sigma \sqrt{\frac{\log p}{n}} \left( 1 + 2 \sqrt{\frac{s \lambda_{\max}(\hat{X}^T \hat{X})}{\sqrt{\zeta_n}}} \right)$$

with probability of $1 - p^{1-c}$, $c \geq 1$.

\[\square\]

**E.2. Proof of Theorem 6.** To wrap up the proof we need to show that the condition of Lemma 2 holds with high probability for the Lasso estimator $\hat{\beta}$. To that end let us define an event $\Omega_n = \{ \frac{1}{n} \sum_{i \notin \mathcal{I}_m} 2(Y_i - X_i \hat{\beta})X_i | \leq \frac{\lambda_n}{n} \}$ and introduce notation of

$$\frac{1}{n} \sum_{i \notin \mathcal{I}_m} X_{ij}(Y_i - X_i \hat{\beta}) = X_{\mathbf{p}_m, j}^T (Y_{\mathbf{p}_m} - X_{\mathbf{p}_m} \hat{\beta}).$$

By repeating similar decomposition analysis developed in Theorem 5 (direct steps omitted), and using the intermediary result of the following Lemma.

**Lemma 4.** If $\text{RE}((\hat{X}, \zeta_N)$ condition holds almost surely, and $\lambda_N \geq c_1 \sigma \sqrt{\log p/n}$, then $|\hat{S}_n^b(\lambda_N)| \leq \hat{s} := C s \lambda_{\max}(\hat{X}^T \hat{X})/(n \zeta_N)$ with probability $1 - p^{1-c}$, $c > 1$.

**Proof of Lemma 4.** Result of Lemma 4 follows by repeating steps of Theorem 3 and by repeating similar steps as in [5] hence it is omitted. \[\square\]

Hence,

$$X_{\mathbf{p}_m, j}^T (Y_{\mathbf{p}_m} - X_{\mathbf{p}_m} \hat{\beta}) \leq \lambda_n (1 + \lambda^{-1}_{\min} (X_{\mathbf{p}_m, A}^T X_{\mathbf{p}_m, A})) + \sigma \sqrt{\frac{\log p}{n}} \left( 1 + 2 \sqrt{\frac{s \lambda_{\max}(\frac{1}{n} X^T X)}{\sqrt{\zeta_n}}} \right)$$

with probability $1 - p^{1-c}$, $c > 1$. Hence, $P(\Omega_n) \geq 1 - p^{1-c}$ if

$$\lambda_n (2 + \lambda^{-1}_{\min} (X_{\mathbf{p}_m, A}^T X_{\mathbf{p}_m, A})) + \sigma \sqrt{\frac{\log p}{n}} \left( 1 + 2 \sqrt{\frac{s \lambda_{\max}(\frac{1}{n} X^T X)}{\sqrt{\zeta_n}}} \right) \leq \frac{\lambda_N}{n}.$$

\[\square\]

**F. Proof of Theorem 7**

Using representation $\tilde{X}_k^T \tilde{X}_k = \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^T$ where $\tilde{x}_i$ is the $i$th row of $\tilde{X}_k$. However, note that the first $w_1$ rows of $\tilde{X}_k$ are identical, so we add the same term $w_1$ times. After this, we add a different row $w_2$ times, and so on. We can then rewrite

$$\tilde{X}_k^T \tilde{X}_k = \sum_{j=1}^N w_j x_{ij} x_{ij}^T$$
where $x_{ij}^T$ is the $ij$th row of $X$. The problem of maximizing the minimum eigenvalue of $\tilde{X}_k^T \tilde{X}_k$ can thus be recast as follows:

$$\max \lambda_{\min} \left( \sum_{j=1}^{N} w_j x_{ij} x_{ij}^T \right)$$

$$\text{s.t.} \quad \sum_{j=1}^{N} w_j = n, \quad w \in \mathbb{Z}^N_+.$$ 

Let us relax the problem by replacing $\mathbb{Z}^N_+$ with $\mathbb{R}^N_+$. We know, from [23] and [1], that 1) the minimum eigenvalue function is concave, and 2) a convex function of an affine function is convex. It follows that the objective function of this optimization problem is concave in $w$. Since the constraints are linear, the optimization problem is concave (and thus has a unique maximum).

In fact, by [23] we can recast the problem as the following semi-definite program:

$$\max \ t$$

$$\text{s.t.} \quad \sum_{j=1}^{N} w_j x_{ij} x_{ij}^T \succeq tI$$

$$\sum_{j=1}^{N} w_j = n, \quad w \succeq 0.$$

References

[1] J. C. Allwright. On maximizing the minimum eigenvalues of a linear combination of symmetric matrices. *SIAM J. Matrix Anal. Appl.*, 10(3):347–382, July 1989.

[2] F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. *ArXiv e-prints*, April 2008.

[3] R. Bhatia. *Matrix Analysis*. Springer-Verlag, New York, 1997.

[4] P. Bickel, F. Gtze, and W.R. van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, page 131, 1997.

[5] Peter J. Bickel, Ya acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.*, 37(4):1705–1732, 2009.

[6] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, August 1996.

[7] Peter Bühlmann and Jacopo Mandozzi. High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Technical Report*, 2012.

[8] Peter Bühlmann and Bin Yu. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2002.

[9] Florentina Bunea, Alexandre Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[10] A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[11] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7:1–26, 1979.

[12] B. Efron. Model selection, estimation, and bootstrap smoothing. *Technical Report Stanford University*, (262), 2012.

[13] P. Hall, E. R. Lee, and B.U. Park. Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica*, (19):449–471, 2009.
[14] M. Hebiri and J. C. Lederer. How Correlations Influence Lasso Prediction. *ArXiv e-prints*, April 2012.

[15] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A Scalable Bootstrap for Massive Data. *ArXiv e-prints*, December 2011.

[16] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B*, 72(4):417–473, 2010.

[17] Dimitris Politis. Bagging multiple comparisons from microarray data. In Ion Mandoiu, Raj Sunderraman, and Alexander Zelikovsky, editors, *Bioinformatics Research and Applications*, volume 4983 of *Lecture Notes in Computer Science*, pages 492–503. Springer Berlin / Heidelberg, 2008.

[18] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993.

[19] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, August 2010.

[20] Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages no–no, 2012.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

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

[23] L. Vandenberghe and S. Boyd. Applications of semidefinite programming. *Appl. Numer. Math.*, 29(3):283–299, 1999. Proceedings of the Stieltjes Workshop on High Performance Optimization Techniques (HPOPT '96) (Delft).

[24] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *Ann. Appl. Stat.*, 5(1):468–485, 2011.

[25] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[26] S. Zhou. Restricted Eigenvalue Conditions on Subgaussian Random Matrices. *ArXiv e-prints*, December 2009.

Jelena Bradic, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, email: jbradic@ucsd.edu