Statistica Sinica Preprint No: SS-2017-0139	
Title	Ranking-Based Variable Selection for high-dimensional data
Manuscript ID	SS-2017-0139
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0139
Complete List of Authors	Rafal Baranowski Yining Chen and Piotr Fryzlewicz
Corresponding Author	Piotr Fryzlewicz
E-mail	p.fryzlewicz@lse.ac.uk
Notice: Accepted version subject to English editing.	

Ranking-Based Variable Selection for High-dimensional Data

Rafal Baranowski[1], Yining Chen[1], and Piotr Fryzlewicz[1]

[1]Department of Statistics, Columbia House, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

29 August 2018

Abstract

We propose Ranking-Based Variable Selection (RBVS), a technique aiming to identify important variables influencing the response in high-dimensional data. The RBVS algorithm uses subsampling to identify the set of covariates which non-spuriously appears at the top of a chosen variable ranking. We study the conditions under which such set is unique and show that it can be successfully recovered from the data by our procedure. Unlike many existing high-dimensional variable selection techniques, within all the relevant variables, RBVS distinguishes between the important and unimportant variables, and aims to recover only the important ones. Moreover, RBVS does not require any model restrictions on the relationship between the response and covariates, it is therefore widely applicable, both in a parametric and non-parametric context. We illustrate its good practical performance in a comparative simulation study. The RBVS algorithm is implemented in the publicly available R package **rbvs**.

Key words: Variable screening, subset selection, bootstrap, Stability Selection.

1 Introduction

Suppose Y is a response, covariates $X_1, \ldots, X_p$ constitute the set of random variables which potentially influence Y, and we observe $\mathbf{Z}_i = (Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, independent copies of $\mathbf{Z} = (Y, X_1, \ldots, X_p)$. In modern statistical applications, where p could be very large, even in tens or hundreds of thousands, it is often assumed that there are many variables having no impact on the response. It is then of interest to use the observed data to identify a subset of $X_1, \ldots, X_p$ which affects Y. The so-called variable selection or subset selection problem plays an important role in statistical modelling for the following reasons. First of all, the number of parameters in a model including all covariates can exceed the number of observations when $n < p$, which makes precise statistical inference not possible using traditional methods. Even when $n \geq p$, constructing a model with a small subset of initial covariates can boost the estimation and prediction accuracy. Second,

parsimonious models are often more interpretable. Third, identifying the set of important variables can be the main goal of statistical analysis, which precedes further scientific investigations.

Our aim is to identify a subset of $\{X_1, \ldots, X_p\}$ which contributes to Y, under scenarios in which p is potentially much larger than n. To model this phenomenon, we work in a framework in which p diverges with n. Therefore, both p and the distribution of $\mathbf{Z}$ depend on n and we work with a triangular array, instead of a sequence. To facilitate interpretability, here for each j, what variable X_j represents does not change as p (and n) increases. Our framework includes, for instance, high-dimensional linear and non-linear regression models. Our proposal, termed Ranking-Based Variable Selection (RBVS), can in general be applied to any technique which allows the ranking of covariates according to their impact on the response. Therefore, we do not impose any particular model structure on the relationship between Y and $X_1, \ldots, X_p$. However $\hat{\omega}_j = \hat{\omega}_j(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, $j = 1, \ldots, p$, a chosen measure used to assess the importance of covariates (either joint or marginal) may require some assumptions on the model. The main ingredient of the RBVS methodology is a variable ranking defined as follows.

Definition 1.1. The variable ranking $\mathbf{R}_n = (R_{n1}, \ldots, R_{np})$ based on $\hat{\omega}_1, \ldots, \hat{\omega}_p$ is a permutation of $\{1, \ldots, p\}$ satisfying $\hat{\omega}_{R_{n1}} > \ldots > \hat{\omega}_{R_{np}}$. Potential ties are broken at random uniformly.

A large number of measures can be used to construct variable rankings. In the linear model, the marginal correlation coefficient serves as an example of such a measure. It is the main component of Sure Independence Screening (SIS, Fan and Lv (2008)). Hall and Miller (2009a) consider the generalized correlation coefficient, which can capture (possibly) non-linear dependence between Y and X_j's. Along the same lines, Fan et al. (2011) propose a procedure based on the magnitude of spline approximations of Y over each X_j, aiming to capture dependencies in non-parametric additive models. Fan and Song (2010) extend SIS to a class of generalised linear models (GLMs), using estimates of the maximum marginal likelihood as the measure of association. Cho and Fryzlewicz (2012) consider variable screening based on the tilted correlation, which accounts for high correlations between the variables, when such are present. Li et al. (2012a) utilise the Kendall rank correlation coefficient, which can be applicable when Y is, for example, a monotonic function of the linear combination of $X_1, \ldots, X_p$. Several model-free variable ranking procedures have been also advocated in the literature. Li et al. (2012b) propose to rank the covariates according to their

distance correlation (Székely and Rizzo, 2009) to the response. Zhu et al. (2011) propose to use the covariance between X_j and the cumulative distribution function of Y conditioning on X_j at point Y as the quantity estimated for screening purposes. He et al. (2013) suggest a ranking procedure relying on the marginal quantile utility; Shao and Zhang (2014) introduce a ranking based on the martingale difference correlation. An extensive overview of these and other measures that can be used for variable screening can be found in Liu et al. (2015). In this work we also consider variable rankings based on measures which originally have not been developed for this purpose, e.g. regression coefficients estimated via penalised likelihood minimisation procedures such as Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) or MC+ (Zhang, 2010).

Variable rankings are used for the purpose of so-called variable screening (Fan and Lv, 2008). The main idea behind this concept is that important covariates are likely to be ranked ahead of the irrelevant ones, so variable selection can be performed on the set of the top-ranked variables. Variable screening procedures attained recently considerable attention due to their simplicity, wide applicability and computational gains they offer to practitioners. Hall and Miller (2009a) suggest that variable rankings can be used for the actual variable selection. They propose to construct bootstrap confidence intervals for the position of each variable in the ranking and select covariates for which the right end of the confidence interval is lower than some cutoff, e.g. $p/2$. This principle, as its authors admit, may lead to undesirable high rate of false positives, and the choice of the ideal cutoff might be very difficult in practice, which was the case in our real data study in the supplementary materials. Hall and Miller (2009b) show that various types of the bootstrap are able to estimate the distribution of the ranks consistently. However, they do not prove that their procedure is able to recover the set of the important variables.

Another approach involving subsampling is taken by Meinshausen and Bühlmann (2010), who propose Stability Selection (StabSel), a general methodology aiming to improve any variable selection procedure. In the first stage of the StabSel algorithm, a chosen variable selection technique is applied to randomly picked subsamples of the data of size $\lfloor n/2 \rfloor$. Subsequently, the variables which are most likely to be selected by the initial procedure, i.e. their selection probabilities exceed a prespecified threshold, are taken as the final estimate of the set of the important variables. An appropriate choice of the threshold leads to finite sample control of the rate of false discoveries of a certain type. Shah and Samworth (2013) propose a variant of StabSel with a further improved

error control.

Our proposed method also incorporates subsampling to boost existing variable selection techniques. Conceptually, it is different from StabSel. Informally speaking, RBVS sorts covariates from the most to the least important, while StabSel treats variables as either relevant or irrelevant and equally important in either of the categories. This has several important consequences. First of all, RBVS is able to simultaneously identify subsets of covariates appearing to be important consistently over subsamples. The same is not computationally feasible for Stability Selection, which only analyses the *marginal* distribution of the initial variable selection procedure. The bootstrap ranking approach of Hall and Miller (2009a) relies on *marginal* confidence intervals, thus it can be also regarded as a "marginal" technique. Second, RBVS does not require the choice of a threshold. The main parameters RBVS require are those from the incorporated subsampling procedure (naturally, these are also required by the approaches of Hall and Miller (2009a) and Meinshausen and Bühlmann (2010)), thus appears to be more automatic than both StabSel and the approach of Hall and Miller (2009a).

The key idea behind RBVS stems from the following observation: although some subsets of $\{X_1, \ldots, X_p\}$ containing irrelevant covariates may appear to have a high influence over Y, the probability that they will consistently exhibit this relationship over many subsamples of observations is small. On the other hand, truly important covariates will typically consistently appear to be related to Y, both over the entire sample and over randomly chosen subsamples. This motivates the following procedure. In the first stage, we repeatedly assess the impact of each variable on the response, with the use of a randomly picked part of the data. For each random draw, we sort the covariates in decreasing order, according to their impact on Y, obtaining a ranking of variables. In the next step, we identify the sets of variables which appear in the top of the rankings frequently and we record the corresponding frequencies. Using these, we decide how many and which variables should be selected.

RBVS is a general and widely-applicable approach focusing on the variable selection; it can be used with any measure of dependence between X_j and Y, either marginal or joint, both in a parametric and non-parametric context. The framework does not require Y and X_j's to be scalar, they can also be e.g. multivariate, or be curves or graphs. Furthermore, the covariates that are highly, but spuriously related to the response are intuitively less likely to exhibit relationship to

Y consistently over the subsamples than the important ones, thus our approach is "reluctant" to select irrelevant variables. Finally, the RBVS algorithm is easily parallelizable and adjustable to available computational resources, making it useful in analysis of extremely high-dimensional data sets. Its R implementation is publicly available in the R package **rbvs** (Baranowski et al., 2015).

The rest of the paper is organised as follows. In Section 2, we define the set of important covariates for variable rankings and introduce the RBVS algorithm. We then show that RBVS is a consistent statistical procedure. We also propose an iterative extension of RBVS, which aims to boost its performance in the presence of strong dependencies between the covariates. The empirical performance of RBVS is illustrated in Section 3. All proofs are deferred to the Appendix. Additional numerical experiments and real data analysis could be found in the supplementary materials.

1.1 Motivating examples

To further motivate our methodology, we discuss the following examples.

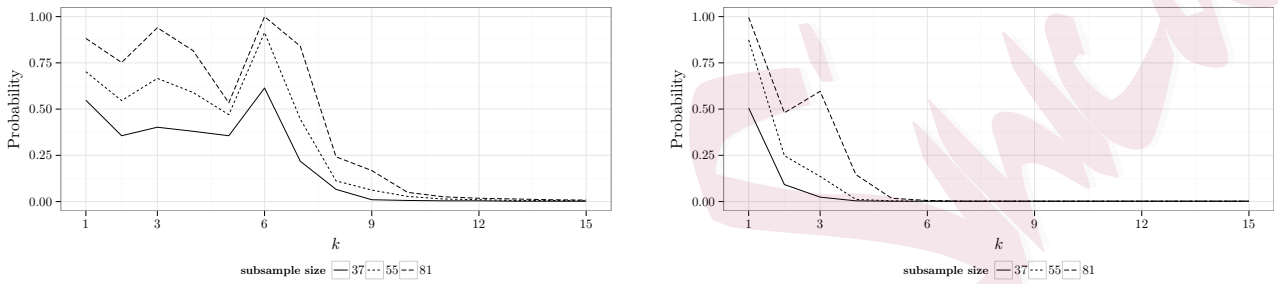
Example 1.1 (riboflavin production with *Bacillus subtils* (Meinshausen and Bühlmann, 2010))**.** The data set consists of the response variable being the logarithm of the riboflavin production rate and transformed expression levels of $p = 4088$ genes for $n = 111$ observations. The aim is to identify those genes whose mutation leads to a high concentration of riboflavin.

Example 1.2 (Fan and Lv (2008))**.** Consider a random sample generated from the linear model $Y_i = 5X_{i1} + 5X_{i2} + 5X_{i3} + \varepsilon_i$, $i = 1, \ldots, n$, where $(X_{i1}, \ldots, X_{ip}) \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$ are independent, $\Sigma_{jk} = 0.75$ for $j \neq k$ and $\Sigma_{jk} = 1$ otherwise. The number of covariates $p = 4088$ and the sample size $n = 111$ are the same as in Example 1.1.

We consider the variable ranking defined in Definition 1.1, based on the sample marginal correlation coefficient in both examples. This choice is particularly reasonable in Example 1.2, where at the population level the Pearson correlation coefficient is the largest for X_1, X_2 and X_3 which are the only truly important ones. The linear model has been previously used to analyse the riboflavin data set (Meinshausen and Bühlmann, 2010), therefore the sample correlation may be useful in identifying important variables in Example 1.1 too.

Figure 1 demonstrates the "paths" generated by Algorithm 1 introduced in the next section. In both examples, the paths share common features, i.e. the estimated probability is large for the first

few values of k and it declines afterwards. Interestingly, in Example 1.2 the curves reach levels very close to 0 shortly after $k = 3$, which is the number of the important covariates here. Crucially, the subset corresponding to $k = 3$ contains the three first covariates (X_{i1}, X_{i2}, X_{i3}), which are relevant in this example. This observation suggests that such paths as those presented in Figure 1 may be used to identify how many and which variables are important, and hence for the purpose of variable selection.



(a) Example 1.1 (b) Example 1.2

Figure 1: Estimated probabilities corresponding to the k-element sets which appear to be the most highly correlated to the response based on subsamples. On the x-axis, k denotes the number of elements in a set. On the y-axis we have the estimated probability corresponding to the most frequently occurring subset of covariates of size k. The three different lines in each example correspond to a different subsample size used to generate paths details are given in Section 2).

2 Methodology of Ranking-Based Variable Selection

In this section, we introduce the Ranking-Based Variable Selection algorithm and its extension. The main purpose of RBVS is to find the set of *top-ranked* variables, which we formally define.

2.1 Notation

Hereafter, $|\mathcal{A}|$ stands for the number of elements in a set $\mathcal{A}$. For every $k = 0, \ldots, p$ (where p grows with n), we denote $\Omega_{n,k} = \{\mathcal{A} \subset \{1, \ldots, p\} : |\mathcal{A}| = k\}$. For the rest of the paper, we suppress the dependence of $\Omega_{n,k}$ on p (and thus n) for notational convenience, and simply write $\Omega_{n,k} \equiv \Omega_k$. For any $\mathcal{A} \in \Omega_k$, $k = 1, \ldots, p$, we define the probability of its being ranked at the top by a given

ranking method as

$$\pi_n(\mathcal{A}) = \mathbb{P}\left(\{R_{n1}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n), \ldots, R_{n,|\mathcal{A}|}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\} = \mathcal{A}\right). \tag{1}$$

For $k = 0$, we set $\pi_n(\mathcal{A}) = \pi_n(\emptyset) = 1$. Furthermore, for any integer m satisfying $1 \leq m \leq n$, we define

$$\pi_{m,n}(\mathcal{A}) = \mathbb{P}\left(\{R_{n1}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m), \ldots, R_{n,|\mathcal{A}|}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)\} = \mathcal{A}\right). \tag{2}$$

Here we are interested in the probability of being ranked at the top using partial observations. Note that one could think of the random samples in our framework as forming a triangular array, so a double subscript is used in the definition above.

2.2 Definition of a k-top-ranked, a locally-top-ranked, and the top-ranked set

Given a ranking scheme, we define the set of important variables in the context of variable rankings.

Definition 2.1. $\mathcal{A} \in \Omega_k$ (with $k \in \{0, \ldots, p-1\}$) is k-top-ranked if $\limsup_{n\to\infty} \pi_n(\mathcal{A}) > 0$.

Definition 2.2. $\mathcal{A} \in \Omega_k$ is said to be locally-top-ranked if it is k-top-ranked and a $k+1$-top-ranked set does not exist, i.e. $\limsup_{n\to\infty} \pi_n(\mathcal{A}) = 0$ for all $\mathcal{A} \in \Omega_{k+1}$.

Definition 2.3. $\mathcal{A} \in \Omega_k$ is said to be top-ranked if it is locally-top-ranked, and there does not exist any other locally-top-ranked sets $\mathcal{A}' \in \Omega_{k'}$ for any $k' < k$. It is unique when the existence of another top-ranked set $\mathcal{A}' \in \Omega_k$ implies $\mathcal{A} = \mathcal{A}'$.

Some remarks are in order. Firstly, Definition 2.1 formalises the statement that $\mathcal{A}$ appears at the top of the ranking with non-negligible probability. We use limit-supremum in the definitions above as $\lim_{n\to\infty} \pi_n(\mathcal{A})$ in general might not exist. Furthermore, we consider $\limsup_{n\to\infty} \pi_n(\mathcal{A}) > 0$ in Definition 2.1, as in some scenarios it is strictly lower than 1. In Example 1.2, for instance, X_1, X_2, X_3 have equal impact on Y, hence under a reasonable ranking scheme (e.g. via marginal correlations), $\lim_{n\to\infty} \pi_n(\mathcal{A}) = 1/3$ for $k = 1$ and $\mathcal{A} = \{1\}, \{2\}, \{3\}$.

Secondly, it can be shown that locally-top-ranked sets might exist for different values of k in some carefully constructed examples, where k is allowed to grow with n. For instance, suppose that $Y_i = \sum_{j=1}^{\lfloor p/3 \rfloor} 2X_{ij} + \sum_{j=\lfloor p/3 \rfloor + 1}^{\lfloor 2p/3 \rfloor} X_{ij} + \varepsilon_i$, where $(X_{i1}, \ldots, X_{ip}) \sim \mathcal{N}(0, I_p)$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. Then

using marginal correlations, it is easy to see that both $\{1, \ldots, \lfloor p/3 \rfloor\}$ and $\{1, \ldots, \lfloor 2p/3 \rfloor\}$ are locally-top-ranked. Nevertheless, this issue can be handled by picking the smallest k in Definiton 2.3. The appropriateness of this definition is demonstrated in Section 2.3.

Thirdly, although the top-ranked set is unique under our assumptions (see also Section 2.3), this does not imply that other k-top-ranked sets are unique as well. In Example 1.2 again, we observe that $\{1\}, \{2\}, \{3\}$ are 1-top-ranked and $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$ are 2-top-ranked. However, the top-ranked set is unique and equal to $\{1, 2, 3\}$.

Finally, note that for any given $\{\mathbf{Z}_i\}_{i=1}^n$, $1 = \sum_{\mathcal{A} \in \Omega_k} \mathbf{1}_{\{R_{n1}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n), \ldots, R_{nk}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\} = \mathcal{A}\}} \Big| \{\mathbf{Z}_i\}_{i=1}^n$ $= \sum_{\mathcal{A} \in \Omega_k} \mathbb{P}\Big(\{R_{n1}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n), \ldots, R_{nk}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)\} = \mathcal{A} \Big| \{\mathbf{Z}_i\}_{i=1}^n \Big)$. By taking the expection over $\{\mathbf{Z}_i\}_{i=1}^n$ on both sides, we have that $\sum_{\mathcal{A} \in \Omega_k} \pi_n(\mathcal{A}) = 1$ for every k and n, and hence $\max_{\mathcal{A} \in \Omega_k} \pi_n(\mathcal{A}) \geq \frac{1}{\binom{p}{k}}$ for every $k = 1, \ldots, p$. In particular, if p were bounded in n, the top-ranked set (as well as locally-top-ranked sets) would not exist. Therefore, we restrict ourselves to the case of p diverging with n (but allowing for both $p \leq n$ and $p > n$). In Section 3 we show that RBVS works well empirically for p both comparable to and much larger than n.

2.3 Top-ranked set for a class of variable rankings

The top ranked set defined in Definition 2.3 exists for a wide class of variable rankings, as we show in Proposition 2.1 below. Let ω_j, $j = 1, \ldots, p$, be a measure of the contribution of each X_j to the response at the population level. Note that ω_j could depend on the distribution of $\mathbf{Z} = (Y, X_1, \ldots, X_p)$ (therefore on n, as p changes with n), so could in theory change with n. However, we suppress this dependence in the notation for simplicity. Furthermore, let $\hat{\omega}_j = \hat{\omega}_j(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ be an estimator of ω_j. We make the following assumptions.

(C1) $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent. For some $\vartheta > 0$ and any $c_\vartheta > 0$ we have

$$\max_{j=1,\ldots,p} \mathbb{P}\left(|\hat{\omega}_j - \omega_j| \geq c_\vartheta n^{-\vartheta} \right) \leq C_\vartheta \exp\left(-n^\gamma\right),$$

where constants $C_\vartheta, \gamma > 0$ do not depend on n.

(C2) The index set of important variables is denoted as $\mathcal{S} \subset \{1, \ldots, p\}$. $\mathcal{S}$ does not depend on n or p, and could potentially be an empty set.

(C3) For every $a \notin \mathcal{S}$, there exists $\mathcal{M}_a \subset \{1, \ldots, p\} \backslash \mathcal{S}$, such that $a \in \mathcal{M}_a$, the distribution of $\{\hat{\omega}_j\}_{j \in \mathcal{M}_a}$ is exchangeable and $|\mathcal{M}_a| \underset{n}{\to} \infty$.

(C4) There exists $\eta \in (0, \vartheta]$, where ϑ is as in (C1), and $c_\eta > 0$ such that $\min_{j \in \mathcal{S}} \omega_j - \max_{j \notin \mathcal{S}} \omega_j \geq c_\eta n^{-\eta}$ uniformly in n.

(C5) The number of covariates $p \leq C_1 \exp\left(n^{b_1}\right)$, where $0 < b_1 < \gamma$ and γ is as in (C1).

Condition (C1) is a concentration bound which holds for a wide range of measures. A few examples are listed below. The sample correlation coefficient satisfies (C1) when the data follow a multivariate normal distribution (Kalisch and Bühlmann, 2007, Lemma 1), or when $Y, X_1, \ldots, X_p$ are uniformly bounded (Delaigle and Hall, 2012, Theorem 1), which follows from Bernstein inequality. Li et al. (2012a) in their Theorem 2 demonstrate that Kendall's τ meets (C1) under the marginally symmetric condition and multi-modal condition. Distance correlation satisfies (C1) under regularity assumptions on the tails of distribution of X_j's and Y (Li et al., 2012b, Theorem 1). The Lasso and the Dantzig selector (Candes and Tao, 2007) estimates of the regression coefficients in the linear model meet (C1) with additional assumptions on the covariates and the sparsity of the regression coefficients (Lounici, 2008, Theorem 1).

Condition (C2) implies that $|\mathcal{S}|$ is bounded in n, which combined with diverging p implies that the number of important covariates is small. This, combined with Conditions (C3) and (C4) can be seen as a variant of the well-known "sparsity" assumption.

We are interested in the scenarios where there are a few variables with large impact on the response plus many variables with similar impact on the response, where those many variables can only have zero or small impact on the response. Here the first part is characterised by Condition (C3), while the second part is characterised by Condition (C4).

Condition (C3) can be linked to the sparsity assumption which requires that only a few covariates have a significant impact on the response. In our framework, these are $\{X_j\}_{j \in \mathcal{S}}$. For all the remaining covariates, the sparsity may require the regression coefficients corresponding to them to be zero. On the other hand, in (C3), each X_a with $a \notin \mathcal{S}$ may contribute to Y, but, speaking heuristically, it is difficult to select a particular X_a with $a \notin \mathcal{S}$, as many covariates have the same impact on Y. As such, none of these would be included in our framework. We believe that this assumption is likely to be met at least approximately (in the sense that large groups of covariates

exhibit similar small impact on the response), especially for large dimensions p. In addition, we note that Meinshausen and Bühlmann (2010) use the exchangeability assumption on the selection of noise variables. However, it concerns a variable selection procedure, while we impose restrictions on the measure $\hat{\omega}_j$. The main difference between their assumption and (C3) is that they require all covariates to be equally likely to be selected, while we allow for many groups within which each variable has the same impact on Y. In the remaining of the manuscript, we refer to the elements of set $\mathcal{S}$ as "relevant and important" (or just "important") variables, to the covariates with zero impact on the response as "irrelevant" variables, and to the rest as "relevant but unimportant" variables.

Furthermore, in Condition (C4), we assume that there is a gap between $\min_{j\in\mathcal{S}} \omega_j$ and $\max_{j\notin\mathcal{S}} \omega_j$, which separates the important variables from the remaining (i.e. irrelevant, and relevant but unimportant) ones. This gap is allowed to decrease slowly to zero. Conditions (C1) and (C4) together imply that the ranking based on $\hat{\omega}_j$ has the sure independence screening property (Fan and Lv, 2008).

Finally, Condition (C5) restricts the maximum number of covariates, but allows the ultra-high dimensional setting where the number of covariates grows exponentially with n^{b_1} for some $b_1 > 0$.

Proposition 2.1. *Let $\mathbf{R}_n$ be a variable ranking based on $\hat{\omega}_j$, $j = 1, \ldots, p$, given in Definition 1.1. Under conditions (C1)–(C5), the unique top-ranked set defined in Definition 2.3 exists and equals $\mathcal{S}$.*

Proposition 2.1 can be applied to establish a link between the top-ranked set and the set of the important variables understood in a classic way. Consider the following linear regression model $Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$, where β_j's are unknown regression coefficients, X_j's - random predictors and ε is an error term. In this model, the top-ranked set could coincide with $\{k : \beta_k \neq 0\}$. To see that, we consider the variable ranking based on $\hat{\omega}_j = \widehat{\mathrm{Cor}}\,(Y, X_j)$, which satisfies (C1) when $(Y, X_1, \ldots, X_p)$ is e.g. Gaussian (Kalisch and Bühlmann, 2007). Condition (C3) is met when e.g. $\widehat{\mathrm{Cor}}\,(Y, X_j) = \rho$ for some $\rho \in (-1, 1)$ and all j such that $\beta_j = 0$, and $p \underset{n}{\to} \infty$. Imposing some restrictions on the correlations between the covariates, we also guarantee that (C4) holds. Finally, provided that $p \underset{n}{\to} \infty$ no faster than as indicated in (C5), Proposition 2.1 would then imply that $\{k : \beta_k \neq 0\}$ is the unique top-ranked set.

Nevertheless, we would emphasize that top-ranked set is only about both relevant and important variables with respect to the chosen measure. Relevant but unimportant variables (unimportant via exchangeability, in the sense of (C3), so not necessarily *small* in the traditional sense) will not be included in the top-ranked set. For instance, in the setting of Example 1.2, but with $Y_i = 5X_{i1} + 5X_{i2} + 5X_{i3} + \sum_{j=\lceil p/2\rceil+1}^{p} \beta X_{ij} + \varepsilon_i$ and $|\beta| < 5$, the top-ranked set via marginal correlations would still be $\{1, 2, 3\}$, even though thanks to the covariance structure in the covariates, for all $j = 1, 2, 3, \lceil p/2\rceil + 1 \ldots, p$, $\mathrm{Cor}(Y, X_j)$ is non-zero. For other work to overcome the issue of small relevant covariates, see Barut et al. (2016). In particular, Barut et al. (2016) also deals with the issue of marginally uncorrelated covariates, which we aim to address by proposing an iterative approach in Section 2.7. See also our simulation examples in this direction in Section 3.

2.4 Main idea of Ranking-Based Variable Selection

Now assume the existence and uniqueness of the top-ranked set $\mathcal{S}$, to construct an estimate of $\mathcal{S}$, we introduce the estimators of $\pi_{m,n}(\mathcal{A})$ defined by (2) using a variant of the m-out-of-n bootstrap (Bickel et al., 2012).

Definition 2.4. Fix $m \in \{1, \ldots, n\}$, $B \in \mathbb{N}$ and set $r = \lfloor n/m \rfloor$. For any $b = 1, \ldots, B$, let $I_{b1}, \ldots, I_{br}$ be mutually exclusive subsets of $\{1, \ldots, n\}$ of size m, drawn uniformly from $\{1, \ldots, n\}$ without replacement. Assume that the sets of subsamples are independently drawn for each b. For any $\mathcal{A} \in \Omega_k$, we estimate $\pi_{m,n}(\mathcal{A})$ by the fraction of subsamples in which $\mathcal{A}$ appeared at the top of the ranking, i.e.

$$\hat{\pi}_{m,n}(\mathcal{A}) = B^{-1} \sum_{b=1}^{B} r^{-1} \sum_{j=1}^{r} \mathbf{1}_{\left\{\mathcal{A} = \left\{R_{n,1}(\{\mathbf{Z}_i\}_{i\in I_{bj}}), \ldots, R_{n,|\mathcal{A}|}(\{\mathbf{Z}_i\}_{i\in I_{bj}})\right\}\right\}}.$$

In general $\pi_{m,n}(\mathcal{A})$ can be different from $\pi_n(\mathcal{A})$, however, we will show in Section 2.6 that $\pi_{m,n}(\mathcal{A})$ and $\pi_n(\mathcal{A})$ are similar (in term of their magnitudes) for the same subsets, provided that m is not too small. This combined with some bounds on the estimation accuracy of $\hat{\pi}_{m,n}(\mathcal{A})$ will imply that $\hat{\pi}_{m,n}(\mathcal{A})$ can be used to find k-top-ranked sets from the data. In practice the number of elements in $\mathcal{S}$ is typically unknown, thus we need to consider subsets of any size in our estimation procedure. From our argument above, for n sufficiently large, the top-ranked set $\mathcal{S}$, provided existence and uniqueness, will have to be one of the following sets for a particular

$k \in \{0, 1, \ldots, p-1\}$, where

$$\mathcal{A}_{k,m} = \text{argmax}_{\mathcal{A} \in \Omega_k} \pi_{m,n}(\mathcal{A}). \tag{3}$$

We define the correspondingly sample version of $\mathcal{A}_{k,m}$ as

$$\hat{\mathcal{A}}_{k,m} = \text{argmax}_{\mathcal{A} \in \Omega_k} \hat{\pi}_{m,n}(\mathcal{A}). \tag{4}$$

To motivate the use of resampling scheme, we remark that some irrelevant covariates (i.e. those having zero impact on the response) can spuriously exhibit large empirical impact on the response, especially when $p \gg n$. The resampling based set probability estimation could provide more stable estimates to help discover variables which non-spuriously appear at the top of the analysed rankings. Moreover, to understand the importance of the parameter B introduced in Definition 2.4, we note that $\max_{\mathcal{A} \in \Omega_k} \hat{\pi}_{m,n}(\mathcal{A}) \geq (Br)^{-1}$. For moderate sample sizes, r may not be large, while we expect the majority of $\pi_{m,n}(\mathcal{A})$'s to be small, even smaller than $1/r$. In this situation, the estimation error of $\max_{\mathcal{A} \in \Omega_k} \hat{\pi}_{m,n}(\mathcal{A})$ with $B = 1$ for is expected to be high and estimate of $\hat{\mathcal{A}}_{k,m}$ could be inaccurate. A moderate value of B aims to bring $\hat{\mathcal{A}}_{k,m}$ closer to its population counterpart $\mathcal{A}_{k,m}$. The theoretical requirements on m and B are given in Section 2.6; our guidance for the choice of m and B in practice is provided in Section 3.3.

In practice, we do not know the size of the top-ranked set $s = |\mathcal{S}|$, so it should be estimated as well. One possibility is to apply hard thresholding rule and set $\hat{s}_\zeta = \min\left\{k : \hat{\pi}_{m,n}\left(\hat{\mathcal{A}}_{k+1,m}\right) \leq \zeta\right\}$, where $\zeta > 0$ is a pre-specified threshold. This approach could be justified by the existence of the asymptotic gap between $\pi_{m,n}(\mathcal{A}_{s+1,m})$ and $\pi_{m,n}(\mathcal{A}_{s,m})$. However, the magnitude of this difference is typically unknown and can be rather small, which makes the choice of ζ difficult. As an alternative, we propose to estimate s by

$$\hat{s} = \text{argmin}_{k=0,\ldots,k_{\max}-1} \frac{\hat{\pi}^\tau_{m,n}(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})}, \tag{5}$$

for some pre-specified $\tau \in (0, 1]$, and some pre-specified large integer $k_{\max}$. The intuition of this choice is explained as follows. Note that

$$\frac{\hat{\pi}^\tau_{m,n}(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})} = \left(\frac{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})}\right)^\tau \left(\frac{1}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})}\right)^{1-\tau}.$$

When $\tau = 1$, we look for k where $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m})$ declines in proportion the most drastically. For a general τ, in essence, we look for k that is a trade-off between the most drastically decline in proportion and the hard thresholding rule (by not permitting $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})$ to be too small). Furthermore, since we assume that $|\mathcal{S}|$ is much smaller than p, it is computationally more efficient to optimize over $\{0, \ldots, k_{\max}\}$ instead of $\{0, \ldots, p-1\}$ in (5). In Section 2.6, we show that this approach leads to consistent estimation of $\mathcal{S}$.

2.5 The Ranking-Based Variable Selection algorithm and its computational cost

The RBVS algorithm consists of the four main steps. Its pseudocode is described in Algorithm 1. In Step 1, we draw subsamples from the data using the subsampling scheme introduced in Definition 2.4. In Step 2, for each subsample drawn we calculate the estimates of ω_j's based on the subsamples I_{bl}, and sort the sample measures $\{\hat{\omega}_j(\{\mathbf{Z}_i\}_{i\in I_{bl}})\}_{j=1}^{p}$ in non-increasing order to find $\mathbf{R}_n(\{\mathbf{Z}_i\}_{i\in I_{bl}})$ defined in Definition 1.1. In Step 3, for each $k = 1, \ldots, k_{\max}$ we find $\hat{\mathcal{A}}_{k,m}$, the k-element set the most frequently occurring in the top of $\mathbf{R}_n(\{\mathbf{Z}_i\}_{i\in I_{bl}})$, for all $b = 1, \ldots, B$ and $l = 1, \ldots, r$. In Step 4, probabilities $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})$ are used to find $\hat{s}$, the estimate of the size of the top-ranked set, and $\hat{\mathcal{S}} = \hat{\mathcal{A}}_{\hat{s},m}$ is returned as the final estimate of $\mathcal{S}$.

Algorithm 1 Ranking-Based Variable Selection algorithm

Input: Random sample $\mathbf{Z}_i = (Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, subsample size m with $1 \le m \le n$, positive integers $k_{\max}$, B, and $\tau \in (0, 1]$.

Output: The estimate of the set of important variables $\hat{\mathcal{S}}$.

procedure RBVS($\mathbf{Z}_1, \ldots, \mathbf{Z}_n, m, B, k_{\max}, \tau$)

Step 1 Let $r = \lfloor n/m \rfloor$. For each $b = 1, \ldots, B$, draw uniformly without replacement m-element subsets $I_{b1}, \ldots, I_{br} \subset \{1, \ldots, n\}$.

Step 2 Calculate $\hat{\omega}_j(\{\mathbf{Z}_i\}_{i\in I_{bl}})$ and the corresponding variable ranking $\mathbf{R}_n(\{\mathbf{Z}_i\}_{i\in I_{bl}})$ for all $b = 1, \ldots, B$, $l = 1, \ldots, r$ and $j = 1, \ldots, p$.

Step 3 For $k = 1, \ldots, k_{\max}$, find $\hat{\mathcal{A}}_{k,m}$ given by (4) and compute $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})$.

Step 4 Find $\hat{s} = \operatorname{argmin}_{k=0,\ldots,k_{\max}-1} \frac{\hat{\pi}^{\tau}_{m,n}(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})}$.

return $\hat{\mathcal{S}} = \hat{\mathcal{A}}_{\hat{s},m}$.

end procedure

We now investigate the computational complexity of Algorithm 1. Denote by $c(n, p)$ the computational cost of evaluating $\hat{\omega}_j$ for all $j = 1, \ldots, p$ using n observations. Firstly, performing B times of random partition of n observations into r subsets (each of size m and dimension p) takes $O(Bn)$ operations, while finding all $\hat{\omega}_j$'s for all Br different subsets takes $c(m, p) \times Br$ manipula-

tions. Next, evaluating the rankings based on each subset (only for those $k_{\max}$ highest ones) takes $O(p + k_{\max}\log(k_{\max}))$ operations via the selection algorithm and QuickSort partition scheme, so doing it for all Br subsets takes $O\big((p + k_{\max}\log(k_{\max}))Br\big)$ operations. Moreover, Step 3 can be performed in $O(Brk_{\max}^2)$ basic operations (NB. see the supplementary materials for more information). Finally, the remaining step requires $O(k_{\max})$ operations. Consequently, the total computational complexity of Algorithm 1 is $c(m,p) \times Br + O(\max\{p, k_{\max}^2\}Br)$. For our recommended choice of $k_{\max}$ and m, see Section 3.3.

2.6 Theoretical results

Under the theoretical framework below, we show that Algorithm 1 recovers the top-ranked set given by Definition 2.3 with probability tending to 1 as $n \to \infty$. We make the following assumptions.

(A1) $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent. For some $\vartheta > 0$ and any $c_\vartheta > 0$ we have that for any n,

$$\max_{j=1,\ldots,p} \mathbb{P}\left(|\hat{\omega}_j(\mathbf{Z}_1, \ldots, \mathbf{Z}_m) - \omega_j| \geq c_\vartheta m^{-\vartheta}\right) \leq C_\vartheta \exp\left(-m^\gamma\right),$$

where constants $C_\vartheta, \gamma > 0$ and m (as a function of n) is specified in Assumption (A3) below.

(A2) There exist constants $C_1 > 0$ and $0 < b_1 < \gamma$ with γ as in (A1) s.t. $p \leq C_1 \exp\left(n^{b_1}\right)$.

(A3) The subsample size m goes to infinity at rate n^{b_2}, with $0 < b_2 < 1$ and $\gamma b_2 - b_1 > 0$, where γ is as in (A1) and b_1 as in (A2).

(A4) The index set of important variables is denoted as $\mathcal{S} \subset \{1, \ldots, p\}$. $\mathcal{S}$ does not depend on n (or p). Denote $s = |\mathcal{S}|$. For every $a \notin \mathcal{S}$, there exists $\mathcal{M}_a \subset \{1, \ldots, p\}\backslash\mathcal{S}$, such that $a \in \mathcal{M}_a$, the distribution of $\{\hat{\omega}_{j,m}\}_{j\in\mathcal{M}_a}$ is exchangeable and $\min_{a\notin\mathcal{S}} |\mathcal{M}_a| \geq C_3 n^{b_3}$ with $C_3 > 0$ and $b_3/2 < 1 - b_2 < b_3$, where b_2 from (A3).

(A5) There exists $\eta \in (0, \vartheta]$, where ϑ is as in (C1), and $c_\eta > 0$ such that $\min_{j\in\mathcal{S}} \omega_j - \max_{j\notin\mathcal{S}} \omega_j \geq c_\eta m^{-\eta}$ uniformly in n. (Here, m, as in (A3), depends solely on n.)

(A6) The number of random draws B is bounded in n.

(A7) The maximum subset size $k_{\max} \in [s, C_4 n^{b_4}]$ with $C_4 > 0$ and b_4 satisfying $b_3 > b_4$, where b_3 is as in (A4).

Assumptions (A1), (A2), (A4) and (A5) can be seen as natural extensions or restatements of (C1) – (C5), to the case when $\hat{\omega}_j$'s are evaluated with m out of n observations only. They are formally repeated here for the sake of clarity. Note that the last part of (A4) implies a lower bound on p ($\geq C_3 n^{b_3}$).

Assumption (A3) establishes the required size of the subsample size m. It implies that both $n/m \underset{n}{\to} \infty$ and $m \underset{n}{\to} \infty$. Such condition is common in literature on bootstrap resampling and U-statistics, see for instance Bickel et al. (2012), Götze and Račkauskas (2001) or Hall and Miller (2009b). Finally, (A6) and (A7) impose conditions on B and $k_{\max}$ respectively.

Theorem 2.1. *Suppose that assumptions (A1)–(A7) hold. Write* $\hat{\mathcal{S}} = \hat{\mathcal{A}}_{\hat{s},m}$, *where* $\hat{\mathcal{A}}_{\hat{s},m}$ *is given by* (4) *and* (5). *Then, for any* $\tau \in (0, 1]$, *there exists a constants* $\beta, C_\beta > 0$ *such that* $\mathbb{P}\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) = o\left(\exp\left(-C_\beta n^\beta\right)\right) \underset{n}{\to} 0$.

The above theorem states that $\hat{\mathcal{S}}$ obtained by RBVS is a consistent estimator of the top-ranked set $\mathcal{S}$, with $\mathbb{P}\left(\hat{\mathcal{S}} = \mathcal{S}\right)$ goes to one at an exponential rate. Its proof can be found in the Appendix. Some empirical evidence is provided in Section 3.

2.7 Iterative extension of RBVS

In the presence of strong dependence between covariates, measure $\hat{\omega}_j$ may fail to detect some important variables. For instance, a covariate may be jointly related but marginally unrelated to the response (see Fan and Lv (2008), Barut (2013) or Barut et al. (2016)). Under such a setting, the estimated top-ranked set may only contain a subset of the important variables. To overcome this problem, we propose IRBVS, an iterative extension of Algorithm 1. The pseudocode of IRBVS is given in Algorithm 2. In each iteration, IRBVS removes the linear effect on the response of the variables found at the previous iteration. Therefore, it is applicable when the relationship between Y and X_j's is at least approximately linear. Nevertheless, it is possible to extend our methodology further. For instance, Barut (2013) and Barut et al. (2016) demonstrate how to remove the impact of a given set of covariates on the response in generalised linear models.

Iterative extensions of variable screening methodologies are frequently proposed in the literature, see for instance Fan and Lv (2008), Zhu et al. (2011) or Li et al. (2012a). A practical advantage of the IRBVS algorithm over its competitors is that it does not require the specification of the

Algorithm 2 Iterative Ranking-Based Variable Selection algorithm

Input: Random sample $\mathbf{Z}_i = (Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, subsample size m with $1 \leq m \leq n$, positive integers $k_{\max}$, B , and $\tau \in (0, 1]$.

Output: The estimate of the set of important variables $\hat{\mathcal{S}}$.

procedure IRBVS($\mathbf{Z}_1, \ldots, \mathbf{Z}_n, m, B, k_{\max}, \tau$)

Initialise $\hat{\mathcal{S}} = \emptyset$.

repeat

Step 1 Let $(Y_1^*, \ldots, Y_n^*)'$ and $(X_{1j}^*, \ldots, X_{nj}^*)'$ (for $j = 1, \ldots, p$) be the residual vectors left after projecting $(Y_1, \ldots, Y_n)'$ and $(X_{1j}, \ldots, X_{nj})'$ onto the space spanned by the covariates with indices in $\hat{\mathcal{S}}$. (NB. for any $j' \in \mathcal{S}$, $(X_{1j'}^*, \ldots, X_{nj'}^*)' = \mathbf{0}$.) Set $\mathbf{Z}_i^* = (Y_i^*, X_{i1}^*, \ldots, X_{ip}^*)$ for $i = 1, \ldots, n$.

Step 2 Calculate $\hat{\mathcal{S}}^* = \text{RBVS}(\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*, m, B, k_{\max}, \tau)$.

Step 3 Set $\hat{\mathcal{S}} := \hat{\mathcal{S}}^* \cup \hat{\mathcal{S}}$.

until $\hat{\mathcal{S}}^* = \emptyset$

return $\hat{\mathcal{S}}$.

end procedure

number of variables added at each iteration or the total number of iterations. Moreover, IRBVS appears to offer better empirical performance than other iterative methods such as ISIS (Fan and Lv, 2008); see Section 3.

2.8 Relations to some selected existing methodology

In this section, we provide a brief overview of the differences between Algorithm 1, Stability selection of Meinshausen and Bühlmann (2010) and the bootstrap ranking approach of Hall and Miller (2009a).

2.8.1 Stability selection (StabSel)

Denote the selection probabilities by $\pi_j = \mathbb{P}\left(j \in \hat{S}^\lambda\right)$, $j = 1, \ldots, p$, where $\hat{S}^\lambda$ is the set of variables selected by a chosen variable selection technique with its tuning parameter set to λ. The aim of StabSel is two-fold: first, to select covariates that the initial procedure selects with a high probability; and second, to bound the average number of false positives (denoted by $\mathrm{E}\, V$) below some prespecified level $\alpha > 0$. For this purpose, Meinshausen and Bühlmann estimate π_j's and select variables for which $\hat{\pi}_j > \pi$, where $\pi \in (1/2, 1)$ is a pre-specified threshold. To control $\mathrm{E}\, V$, one can set λ such that $|\hat{S}^\lambda| \leq q$, where $q \in \{1, \ldots, p\}$ depends on π and α and is adjusted to ensure $\mathrm{E}\, V \leq \alpha$. The exact formula for q and other possible ways of controlling $\mathrm{E}\, V$ are given in Meinshausen and Bühlmann (2010).

In contrast to StabSel, which needs a variable selection procedure, RBVS selects variables based on a variable ranking. In particular, in our approach we consider joint probabilities $\pi_{m,n}(\mathcal{A})$, while in StabSel only marginal probabilities are used. The estimates of the joint probabilities can be used to determine the number of important covariates at the top of the variable ranking, without the specification of a threshold, as we demonstrate in Section 2.6. Consequently, we believe that RBVS can be seen as more automatic and "less marginal" than StabSel.

2.8.2 The bootstrapped rankings

Let r_{nj} be the position of the jth covariate in the variable ranking $\mathbf{R}_n = (R_{n1}, \ldots, R_{np})$. Mathematically, assuming there is no tie, $r_{nj} = l$ if and only if when $R_{nl} = j$. To identify important covariates based on $\mathbf{R}_n$, Hall and Miller (2009a) compute $[r_{nj}^-, r_{nj}^+]$, two-sided, equal tiled, percentile-method bootstrap confidence intervals for r_{nj} at a significance level α. A variable is considered to be influential when r_{nj}^+ is lower than some prespecified cutoff level c, for instance $c = p/2$. The number of variables selected by the procedure of Hall and Miller (2009a) depends therefore on α and c and "marginal" confidence intervals $[r_{nj}^-, r_{nj}^+]$. By contrast, RBVS is based on the joint probabilities $\pi_{m,n}(\mathcal{A})$ and does not require the specification of a threshold or a significance level.

2.8.3 Computational complexity of the related methods

Table 1 summarizes computational complexity of Algorithm 1 (with $m = \lfloor n/2 \rfloor$) and its competitors: SIS (Fan and Lv, 2008) and StabSel (Meinshausen and Bühlmann, 2010). For reference, we include the computational complexity of the k-fold cross-validation (k-fold CV), which is frequently used to find optimal parameters for e.g. Lasso, MC+ or SIS. The computational complexity of the method proposed by Hall and Miller (2009a) is comparable to StabSel, hence omitted in this comparison. In theory, SIS requires the least computational resources, especially in the case of $p \gg n$. Simple k-fold cross-validation has the second lowest computational complexity. StabSel in the case of $n > \sqrt{p}$ is theoretically quicker than RBVS, however, the common factor $B \times c\,(n/2, p)$ typically dominates both $O(Bp)$ and $O(\max\{p, n^2\})$, and our experience suggests that StabSel and RBVS usually take similar amount of computational resources.

k-fold CV	SIS	StabSel	RBVS
$k \times c\left(\frac{(k-1)n}{k}, p\right)$	$O(np) + k \times c\left(\frac{(k-1)n}{k}, \frac{n}{\log(n)}\right)$	$B \times c\left(\frac{n}{2}, p\right) + O(Bp)$	$B \times c(\frac{n}{2}, p) + O(\max\{n^2, p\}B)$

Table 1: Computational complexity of Algorithm 1 and its competitors. The cost of the base learner in relation to the sample size n and the number of variables p is denoted by $c(n, p)$; B is the number of subsamples used in StabSel and RBVS. Parameters for SIS, StabSel, RBVS are set to the recommended values. For SIS, we assume that k-fold CV is used after the screening step.

3 Simulation study

To facilitate comparison among different methods, we focus on linear models in this section. We also provide two real data examples in the supplementary materials.

3.1 Simulation models

Model (A) Taken from Fan and Lv (2008): $Y_i = 5X_{i1} + 5X_{i2} + 5X_{i3} + \varepsilon_i$, where $(X_{i1}, \ldots, X_{ip})$ are i.i.d. observations from $\mathcal{N}(0, \Sigma)$ distribution and ε_i follow $\mathcal{N}(0, 1)$ distribution. The covariance matrix satisfies $\Sigma_{jj} = 1$, $j = 1, \ldots, p$, $\Sigma_{jk} = \rho$, $|\rho| < 1$ for $k \neq j$. This is a relatively easy setting, where all important X_j's are "visible" to any reasonable marginal approach as they are the most highly correlated to Y at the population level.

Model (B) Factor model taken from Meinshausen and Bühlmann (2010): $Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$, where X_{ij}'s follow the factor model $X_{ij} = \sum_{l=1}^{K} f_{ijl}\varphi_{il} + \theta_{ij}$, with f_{ijl}, φ_{il}, θ_{ij}, ε_i i.i.d. $\mathcal{N}(0, 1)$. We set $K = 2, 10$. In addition, the number of $\beta_j \neq 0$ is set to $s = 5$, with their indices drawn uniformly without replacement, and with their values i.i.d. uniformly distributed on $[0, 5]$. In this model some of the non-zero regression coefficients are potentially small, thus the corresponding covariates might be difficult to detect.

Model (C) Modified from Model **Model (A)**: same covariate and noise structure as **Model (A)**, but with $Y_i = 5X_{i1} + 5X_{i2} + 5X_{i3} + \sum_{j=\lceil \frac{p}{2} \rceil + 1}^{p} \beta X_{ij} + \varepsilon_i$, where we set $\beta = 2^{-2}, 2^{-1}, 2^0, 2^1$. Here we have the important variables (i.e. the top-ranked set is $\{1, 2, 3\}$), the relevant but unimportant variables (i.e. $\left\{\lceil \frac{p}{2} \rceil + 1, \ldots, p\right\}$), as well as the irrelevant ones (i.e. $\left\{4, \ldots, \lceil \frac{p}{2} \rceil\right\}$). The challenge is to select only the important ones. Here we are interested in the behaviour of RBVS as β gets closer to 5 (so the problem becomes harder).

Model (D) Modified from Fan and Lv (2008):

$$Y_i = 5X_{i1} + 5X_{i2} + 5X_{i3} - 15\sqrt{\rho}X_{i4} + \sum_{j=\lceil \frac{p}{2} \rceil + 1}^{p} 5p^{1/2}X_{ij} + \varepsilon_i,$$

where $(X_{i1}, \ldots, X_{ip})$ are i.i.d. observations from $\mathcal{N}(0, \Sigma)$ and ε_i follow $\mathcal{N}(0, 1)$ distribution. The covariance Σ is as in **Model (A)**, except that $\Sigma_{4,k} = \Sigma_{j,4} = \sqrt{\rho}$ for $k, j = 1, 2, 3, 5, \ldots, p$. The challenge of this model is two-folded: first, X_{i4} has a large contribution to Y_i but it is marginally unrelated to the response; second, similar to **Model (C)**, there are both important and unimportant relevant variables, and our aim is to recover only the important ones, i.e. the top-ranked set.

3.2 Simulation methods

We have applied RBVS and IRBVS with the absolute values of the following measures: Pearson correlation coefficient (PC) (Fan and Lv, 2008), the regression coefficients estimated via Lasso (Tibshirani, 1996), the regression coefficients estimated via MC+ algorithm (Zhang, 2010). The performance of RBVS and IRBVS with Lasso is typically slightly worse than that of MC+ in our numerical experiments, so is not reported here. More comprehensive numerical results can be found in Baranowski (2016).

For competitors, we consider the standard MC+ estimator defined as

$$\hat{\beta}_{pen} = \operatorname{argmin}_{\beta}(n^{-1}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \sum_{j=1}^{p}\operatorname{pen}(|\beta_j|),$$

where $\operatorname{pen}(t) = \lambda\int_0^t \max\{0, (1 - x/(\gamma\lambda))\}\,dx$ and $\lambda, \gamma > 0$ are tuning parameters. Here λ chosen via 10-fold cross-validation, and $\gamma = 3$ as in Breheny and Huang (2011). We also consider StabSel, where we set the tuning parameters as per the recommendation of Meinshausen and Bühlmann (2010).

The final group of the techniques included in our comparison consists of SIS and its iterative extension ISIS (Fan and Lv, 2008) (and with MC+ after the screening stage). For the SIS method, we have considered both the standard version of Fan and Lv (2008) based on the marginal sample correlations (MSC), and a more recent version of Chang et al. (2013) based on the marginal empirical likelihood (MEL). Note that the standard ISIS procedure did not perform well in our experiments,

as it was selecting a very large number of false positives, therefore we apply a modified version of ISIS which involves certain randomisation mechanism (Saldana and Feng, 2018).

We use implementations of MC+ algorithm from the R package **ncvreg** (Breheny and Huang, 2011). For SIS based methods we use the R package **SIS** (Saldana and Feng, 2018).

3.3 Choice of parameters of the (I)RBVS algorithm

RBVS involves the choice of several parameters, namely B, m, $k_{\max}$ and τ.

The B parameter has been introduced to decrease the randomness of the method. Naturally, the larger the value of B, the less the algorithm depends on a particular random draw. However, the computational complexity of RBVS increases linearly with B. In the simulation study, we take $B = 50$. Our experience suggests that little will be gained in terms of the performance of RBVS for a much larger B.

The problem of the choice of the subsample size m is more challenging. In Section 2.6, we require $m \to \infty$ at an appropriate rate, which is, however, unknown. In the finite-sample case m cannot be too small, as it is unlikely that $\mathbf{R}_n$ based on a small sample could give a high priority to the important variables. On the other hand, when m is too large (i.e. close to n), subsamples largely overlap. In practical problems, we propose to choose $m = \lfloor n/2 \rfloor$. See also our additional simulation study in the supplementary materials, which confirms that this choice results in good finite-sample properties of the RBVS-based methods.

From our experience, $k_{\max}$ has limited impact on the outcome of RBVS, as long as it is not too small. In all simulations conducted, $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})$ given by (4) reaches and stays at the level of $1/(Br)$ for some $k \leq n$, so we recommend $k_{\max} = \min\{n, p\}$. Finally, our experience also suggests that RBVS is not very sensitive to the choice of τ as well, as long as it is not too close to zero. Here we simply take $\tau = 0.5$.

3.4 Results

Our results are reported in Tables 2–5, in terms of the average number of False Positives (FP), False Negatives (FN), total errors (FP+FN), and the estimated $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S})$, i.e. probability (Pr) of correct estimation of the top-ranked set.

		SIS		StabSel		RBVS		ISIS		IRBVS	
	MC+	MSC	EML	PC	MC+	PC	MC+	MSC	EML	PC	MC+
$n = 100$ $p = 100$ $\rho = 0$											
FP	.18	.00	.00	.18	.02	.03	.00	.32	.26	.04	.01
FN	.00	.00	.00	.00	.00	.10	.00	.00	.00	.08	.00
FP+FN	.18	**.00**	**.00**	.18	.02	.13	**.00**	.32	.26	.12	.01
Pr	.88	**1.00**	**1.00**	.82	**.98**	.92	**1.00**	.91	.92	.94	**.99**
$n = 100$ $p = 1000$ $\rho = 0$											
FP	.92	.02	.05	.34	.01	.00	.00	.07	.06	.00	.00
FN	.00	.00	.00	.01	.00	.30	.00	.00	.00	.20	.00
FP+FN	.92	.02	.06	.34	.01	.31	**.00**	.07	.06	.20	**.00**
Pr	.70	**.99**	**.98**	.70	**.99**	.84	**1.00**	.94	**.95**	.93	**1.00**
$n = 100$ $p = 100$ $\rho = 0.75$											
FP	.00	.00	.25	.40	.03	.02	.00	.18	.11	.05	.00
FN	.00	.00	.18	.04	.00	1.23	.00	.00	.00	1.00	.00
FP+FN	**.00**	**.00**	.43	.44	.03	1.25	**.00**	.18	.11	1.05	**.00**
Pr	**1.00**	**1.00**	.84	.64	**.97**	.49	**1.00**	.94	**.95**	.62	**1.00**
$n = 100$ $p = 1000$ $\rho = 0.75$											
FP	.00	.00	2.29	.70	.00	.00	.00	.08	.11	.04	.00
FN	.00	.00	1.16	.20	.00	2.12	.03	.00	.12	1.71	.03
FP+FN	**.00**	**.00**	3.45	.90	**.00**	2.12	.04	.08	.22	1.75	.04
Pr	**1.00**	**1.00**	.25	.43	**1.00**	.17	**.98**	.94	.93	.40	**.98**

Table 2: **Model (A)**: the average number of False Positives (FP), False Negatives (FN), total errors (FP+FN), and the estimated probability (Pr) of the correct selection of the top-ranked set (i.e. $P(\hat{\mathcal{S}} = \mathcal{S})$), calculated over 200 realisations. Bold: within 10% of the lowest value of FP+FN (or within 5% of the highest value of Pr).

		SIS		StabSel		RBVS		ISIS		IRBVS	
	MC+	MSC	EML	PC	MC+	PC	MC+	MSC	EML	PC	MC+
$n = 100$ $p = 100$ $K = 2$											
FP	.12	.08	.04	.18	.00	.00	.00	.26	.21	.04	.00
FN	.14	.88	.91	2.04	.20	3.38	.28	.16	.15	1.22	.28
FP+FN	.26	.97	.96	2.22	**.20**	3.38	.28	.41	.36	1.26	.28
Pr	**.81**	.34	.34	.00	**.82**	.00	**.79**	.76	**.78**	.60	**.79**
$n = 100$ $p = 1000$ $K = 2$											
FP	.40	.22	.32	.36	.00	.01	.00	.06	.08	.04	.00
FN	.24	1.65	1.84	2.60	.35	3.69	.39	.30	.36	1.51	.32
FP+FN	.65	1.87	2.16	2.96	**.35**	3.70	.39	**.35**	.43	1.55	**.32**
Pr	.65	.06	.04	.00	**.70**	.00	.68	**.72**	.67	.48	**.72**
$n = 100$ $p = 100$ $K = 10$											
FP	.00	.04	.02	.19	.00	.01	.01	.18	.19	.08	.02
FN	.22	.86	.84	1.95	.15	3.01	.19	.12	.12	.93	.17
FP+FN	.22	.89	.86	2.14	**.16**	3.02	.20	.30	.32	1.00	**.18**
Pr	.78	.36	.38	.02	**.84**	.00	**.82**	**.84**	.80	.64	**.82**
$n = 100$ $p = 1000$ $K = 10$											
FP	.02	.08	.14	.33	.00	.00	.00	.07	.04	.02	.00
FN	.26	1.52	1.59	2.27	.20	3.33	.22	.16	.18	.88	.18
FP+FN	.28	1.60	1.74	2.60	.20	3.34	.22	.22	.22	.89	**.18**
Pr	.78	.14	.12	.00	**.82**	.00	**.81**	.80	.80	.69	**.84**

Table 3: **Model (B)**: the average number of False Positives (FP), False Negatives (FN), total errors (FP+FN), and the estimated probability (Pr) of the correct selection of the top-ranked set (i.e. $P(\hat{\mathcal{S}} = \mathcal{S})$), calculated over 200 realisations. Bold: within 10% of the lowest value of FP+FN (or within 5% of the highest value of Pr).

		SIS		StabSel		RBVS		ISIS		IRBVS	
	MC+	MSC	EML	PC	MC+	PC	MC+	MSC	EML	PC	MC+
$n = 100$ $p = 100$ $\rho = 0$ $\beta = 0.25$											
FP	7.65	1.06	.95	.24	.04	.02	.01	3.46	3.50	.04	.01
FN	.00	.00	.00	.00	.00	.26	.00	.00	.00	.24	.00
FP+FN	7.65	1.06	.95	.24	.04	.28	**.01**	3.46	3.50	.28	**.01**
Pr	.24	.76	.80	.78	.96	.88	**.99**	.32	.33	.88	**.99**
$n = 100$ $p = 100$ $\rho = 0$ $\beta = 0.5$											
FP	11.96	4.25	4.04	.32	.06	.02	.00	4.36	4.17	.08	.01
FN	.00	.00	.00	.00	.00	.59	.00	.00	.00	.50	.00
FP+FN	11.96	4.25	4.04	.32	.06	.61	**.00**	4.36	4.17	.58	.01
Pr	.12	.38	.38	.72	.94	.74	**1.00**	.23	.25	.76	**.99**
$n = 100$ $p = 100$ $\rho = 0$ $\beta = 1$											
FP	19.63	11.44	11.10	.44	.06	.02	.00	5.34	5.13	.04	.00
FN	.00	.00	.00	.18	.00	2.06	.34	.00	.00	1.78	.34
FP+FN	19.63	11.44	11.11	.62	**.06**	2.08	.34	5.34	5.14	1.83	.34
Pr	.00	.00	.01	.54	**.94**	.14	**.90**	.06	.07	.31	**.90**
$n = 100$ $p = 100$ $\rho = 0$ $\beta = 2$											
FP	34.10	15.37	15.10	.78	.14	.00	.00	5.94	5.79	.00	.00
FN	.04	.30	.30	1.70	1.59	2.83	2.97	.94	1.00	2.82	2.97
FP+FN	34.15	15.67	15.40	2.48	**1.73**	2.83	2.97	6.88	6.79	2.83	2.97
Pr	.00	.00	.00	.04	**.06**	.00	.00	.00	.00	.00	.00

Table 4: **Model (C)**: the average number of False Positives (FP), False Negatives (FN), total errors (FP+FN), and the estimated probability (Pr) of the correct selection of the top-ranked set (i.e. $P(\hat{\mathcal{S}} = \mathcal{S})$), calculated over 200 realisations. Bold: within 10% of the lowest value of FP+FN (or within 5% of the highest value of Pr).

		SIS		StabSel		RBVS		ISIS		IRBVS	
	MC+	MSC	EML	PC	MC+	PC	MC+	MSC	EML	PC	MC+
$n = 100$ $p = 100$ $\rho = 0.5$											
FP	.07	1.99	1.17	.40	.03	.00	.01	3.27	3.35	.10	.01
FN	.00	.06	.44	.90	.00	2.58	.00	.00	.02	.38	.00
FP+FN	.07	2.05	1.60	1.30	.03	2.58	**.01**	3.27	3.36	.48	**.01**
Pr	**.95**	.56	.44	.20	**.97**	.02	**.99**	.24	.27	.78	**.99**
$n = 100$ $p = 1000$ $\rho = 0.5$											
FP	.00	.02	.03	.88	.28	.02	.00	.10	.10	.05	.02
FN	2.27	2.59	2.72	2.98	.00	3.00	.00	.00	.02	.69	.00
FP+FN	2.27	2.62	2.75	3.86	.28	3.02	**.00**	.10	.12	.74	.02
Pr	.06	.00	.00	.00	.76	.00	**1.00**	.92	.92	.72	**.98**
$n = 100$ $p = 100$ $\rho = 0.75$											
FP	.00	1.14	.52	.47	.04	.00	.00	2.62	2.75	.06	.01
FN	1.04	.06	.84	1.24	.00	2.80	.00	.00	.02	.40	.00
FP+FN	1.04	1.21	1.35	1.71	.04	2.80	**.00**	2.63	2.76	.46	.01
Pr	.39	.62	.27	.12	**.96**	.01	**1.00**	.31	.30	.80	**.99**
$n = 100$ $p = 1000$ $\rho = 0.75$											
FP	.00	.16	.02	.86	2.31	.00	.01	.10	.08	.05	.02
FN	3.00	2.69	2.86	2.98	.00	3.00	.02	.00	.01	.82	.00
FP+FN	3.00	2.85	2.88	3.85	2.31	3.00	**.02**	.10	.08	.87	**.02**
Pr	.00	.00	.00	.00	.07	.00	**.98**	.92	.92	.68	**.98**

Table 5: **Model (D)**: the average number of False Positives (FP), False Negatives (FN), total errors (FP+FN), and the estimated probability (Pr) of the correct selection of the top-ranked set (i.e. $P(\hat{\mathcal{S}} = \mathcal{S})$), calculated over 200 realisations. Bold: within 10% of the lowest value of FP+FN (or within 5% of the highest value of Pr).

Overall, in all the settings we consider here, RBVS and IRBVS with a proper choice of measurement (such as with MC+) typically offer similar and sometimes better performance than their competitors such as StabSel. In general, RBVS and IRBVS tend to improve the performance of the base learners (such as Lasso or MC+). Moreover, the iterative extension, IRBVS, in many cases is able to detect variables overlooked by the pure RBVS, especially with PC.

In **Model (C)** for fixed n and p, when $|\beta|$ is small to moderate (i.e. $\beta \in \{0.25, 0.5\}$), both RBVS and IRBVS are able to frequently recover the top-ranked set. Nevertheless, as the value of $|\beta|$ increases, the difference between the important and unimportant relevant variables becomes smaller, making it harder to estimate the top-ranked set. When $\beta = 2$, these algorithms (as well as all their competitors) would fail completely. Not surprisingly, both RBVS and IRBVS tend to include no variable in the estimated top-ranked set, as all variables appear to be quite similar to each other in terms of their coefficients using PC or MC+.

In contrast, MC+, SIS and ISIS perform poorly In **Model (C)** (even when $|\beta|$ is very small), as well as in **Model (D)**, due to the presence of unimportant but relevant variables. Thus they are not suitable for recovering the top-ranked set in these settings. Though StabSel MC+ is also very competitive in **Model (A)**–**Model (C)**, it appears to perform considerably worse than RBVS MC+ or IRBVS MC+ in **Model (D)**, especially when p is large and the covariates are highly-correlated.

Finally, we note that as long as the covariates are not too highly-correlated, the performance of IRBVS is relatively robust against the choice of the measure used in the procedure. Therefore, we recommend to adjust this choice to the available computational resources and the size of the data. In particular, for large data sets ($p > 10000$, $n > 500$), we recommend using IRBVS PC, which is extremely fast to compute with the R package **rbvs**. Nevertheless, penalisation-based methods such as MC+ typically offer better performance, so should be used as the base measure in the case of moderate data size.

A Proofs

A.1 Proof of Proposition 2.1

Proof. Firstly, we show that $\pi_n(\mathcal{S})$ tends to 1. Denote by $\mathcal{E} = \{\min_{j\in\mathcal{S}} \hat{\omega}_j > \max_{j\notin\mathcal{S}} \hat{\omega}_j\}$. If there is no tie, $\mathcal{E}$ is equivalent to $\left\{\{R_{n1},\ldots,R_{ns}\} = \mathcal{S}\right\}$, i.e. all indices from $\mathcal{S}$ are ranked in front of those do not belong to $\mathcal{S}$. Otherwise, $\{\min_{j\in\mathcal{S}} \hat{\omega}_j > \max_{j\notin\mathcal{S}} \hat{\omega}_j\}$ implies that $\left\{\{R_{n1},\ldots,R_{ns}\} = \mathcal{S}\right\}$. Using (C4) we have

$$\pi_n(\mathcal{S}) \geq \mathbb{P}(\mathcal{E}) \geq \mathbb{P}\left(\max_{j=1,\ldots,p} |\hat{\omega}_j - \omega_j| < \epsilon\right),$$

where $\epsilon = \frac{c_\eta n^{-\eta}}{2}$. Application of Bonferroni's inequality yields that

$$\mathbb{P}\left(\max_{j=1,\ldots,p} |\hat{\omega}_j - \omega_j| < \epsilon\right) \geq 1 - p \sup_{j=1,\ldots,p} \mathbb{P}\left(|\hat{\omega}_j - \omega_j| \geq \epsilon\right).$$

The last term is of order $1 - O\left(\exp\left(-n^\gamma\right)\right)$ (since $b_1 < \gamma$), which tends to 1 as $n \to \infty$. This proves that $\mathcal{S}$ is a s-top-ranked set, where $s = |\mathcal{S}|$.

Secondly, consider any $\mathcal{A} \in \Omega_{s+1}$. We will prove that $\pi_n(\mathcal{A}) \underset{n}{\to} 0$. Note that $\mathcal{E}$ implies that $\mathcal{S} \subset \mathcal{A}$, as all indices from $\mathcal{S}$ are ranked in front of those do not belong to $\mathcal{S}$. Thus, it suffices to only consider the case of $\mathcal{S} \subset \mathcal{A}$ in which $\mathcal{A} \setminus \mathcal{S}$ has only one element, which we denote by a. Suppose there is no tie in the ranking, on the event $\mathcal{E}$, $\mathbb{P}\left(\{\min_{j\in\mathcal{A}} \hat{\omega}_j > \max_{j\notin\mathcal{A}} \hat{\omega}_j\} \cap \mathcal{E}\right) = \mathbb{P}\left(\{\hat{\omega}_a > \max_{j\notin\mathcal{A}} \hat{\omega}_j\} \cap \mathcal{E}\right)$. To bound $\mathbb{P}\left(\hat{\omega}_a > \max_{j\notin\mathcal{A}} \hat{\omega}_j\right)$, we observe that $\mathbb{P}\left(\hat{\omega}_a > \max_{j\notin\mathcal{A}} \hat{\omega}_j\right) \leq \mathbb{P}\left(\hat{\omega}_a > \max_{j\in\mathcal{M}_a\setminus\{a\}} \hat{\omega}_j\right)$. Using the exchangeability assumption (C3), we have that the values of $\mathbb{P}\left(\hat{\omega}_{j^*} > \max_{j\in\mathcal{M}_a\setminus\{j^*\}} \hat{\omega}_j\right)$ are the same for every $j^* \in \mathcal{M}_a$ (i.e. any element in $\{\hat{\omega}_j\}_{j\in\mathcal{M}_a}$ are equally likely to be the largest). Since $\sum_{j^*\in\mathcal{M}_a} \mathbb{P}\left(\hat{\omega}_{j^*} > \max_{j\in\mathcal{M}_a\setminus\{j^*\}} \hat{\omega}_j\right) \leq 1$, we have that $\mathbb{P}\left(\hat{\omega}_a > \max_{j\in\mathcal{M}_a\setminus\{a\}} \hat{\omega}_j\right) \leq \frac{1}{|\mathcal{M}_{\{a\}}|} \underset{n}{\to} 0$. Consequently,

$$\pi_n(\mathcal{A}) \leq \mathbb{P}\left(\hat{\omega}_a > \max_{j\notin\mathcal{A}} \hat{\omega}_j\right) + \mathbb{P}\left(\mathcal{E}^c\right) \leq \mathbb{P}\left(\hat{\omega}_a > \max_{j\in\mathcal{M}_a\setminus a} \hat{\omega}_j\right) + \mathbb{P}\left(\mathcal{E}^c\right) \underset{n}{\to} 0.$$

Otherwise, if there are ties in the ranking, since we break the ties at random uniformly, it follows from the exchangeability assumption that we are equally likely to pick any index from $\mathcal{M}_a$, given that we have picked one of them. Thus we can argue in a similar manner to show that $\pi_n(\mathcal{A}) \leq 1/|\mathcal{M}_a| + \mathbb{P}\left(\mathcal{E}^c\right) \underset{n}{\to} 0$, i.e. $\mathcal{S}$ is always locally-top-ranked.

Thirdly, for every $k' = 1\ldots, s-1$, we show that there exists some $\mathcal{A} \in \Omega_{k'}$ such that $\limsup_{n\to\infty} \pi_n(\mathcal{A}) > 0$. Note that

$$\sum_{\{\mathcal{A}:\ \mathcal{A}\in\Omega_{k'} \text{ and } \mathcal{A}\subset\mathcal{S}\}} \pi_n(\mathcal{A}) \geq \mathbb{P}\left(\min_{j\in\mathcal{S}} \hat{\omega}_j > \max_{j\notin\mathcal{S}} \hat{\omega}_j\right) \underset{n}{\to} 1$$

from our previous argument. However, there are $\binom{s}{k'}$ elements in $\{\mathcal{A}:\ \mathcal{A}\in\Omega_{k'} \text{ and } \mathcal{A}\subset\mathcal{S}\}$, so

$$\max_{\{\mathcal{A}:\ \mathcal{A}\in\Omega_{k'} \text{ and } \mathcal{A}\subset\mathcal{S}\}} \limsup_{n\to\infty} \pi_n(\mathcal{A}) \geq \frac{1}{\binom{s}{k'}}.$$

This implies that $\mathcal{S}$ is indeed a top-ranked set.

Finally, the uniqueness of $\mathcal{S}$ (among those in Ω_s) follows from the fact that $\pi_n(\mathcal{S}) \underset{n}{\to} 1$ and $\sum_{\mathcal{A}\in\Omega_s} \pi_n(\mathcal{A}) = 1$. □

A.2 Auxiliary lemmas and proof of Theorem 2.1

A.2.1 Auxiliary lemmas

Lemma A.1 (Theorem 1 of Hoeffding (1963))**.** *Let W be a binomial random variable with the probability of success π and r trials. For any $1 > t > \pi$, we have $\mathbb{P}(W \geq rt) \leq \left(\frac{\pi}{t}\right)^{rt}\left(\frac{1-\pi}{1-t}\right)^{r(1-t)}$. Moreover, for any $0 < t < \pi$, $\mathbb{P}(W \leq rt) \leq \left(\frac{\pi}{t}\right)^{rt}\left(\frac{1-\pi}{1-t}\right)^{r(1-t)}$.*

Lemma A.2. *Let $a_1, \ldots, a_l$ be non-negative numbers s.t. $\sum_{i=1}^{l} a_i \leq 1$ and $\max a_i \leq t$ for some $\frac{1}{l} \leq t \leq 1$. Let $N \in \mathbb{N}$ be the minimum integer such that there exist mutually exclusive sets $I_1, \ldots, I_N \subset \{1, \ldots, l\}$ with $\sum_{i\in I_j} a_i \leq t$ and $\bigcup_{j=1}^{N} I_j = \{1, \ldots, l\}$. Then, $N \leq \lfloor \frac{2}{t} \rfloor + 1$.*

Proof. Since N is the smallest possible integer, there must be at most one $j \in \{1, \ldots, N\}$ with $\sum_{i\in I_j} a_i \leq t/2$. Otherwise, such two sets could be combined, leading to a smaller N. So for all other $j \in \{1, \ldots, N\}$, we have that $\sum_{i\in I_j} a_i > t/2$. Consequently, $(N-1)t/2 \leq \sum_{i=1}^{l} a_i \leq 1$. This implies that $N \leq \lfloor \frac{2}{t} \rfloor + 1$. □

Lemma A.3. *Let be $\Omega \subset \Omega_k$ for some $k = 1, \ldots, p-1$, $m \leq n$, $B \geq 1$, and t_1, t_2 satisfying $\max_{\mathcal{A}\in\Omega} \pi_{m,n}(\mathcal{A}) \leq t_2 < t_1 < 1$. Then $\mathbb{P}(\max_{\mathcal{A}\in\Omega} \hat{\pi}_{m,n}(\mathcal{A}) \geq t_1) \leq \frac{3B}{t_2}\left[\left(\frac{t_2}{t_1}\right)^{t_1}\left(\frac{1-t_2}{1-t_1}\right)^{1-t_1}\right]^r$, where $\pi_{m,n}(\mathcal{A})$, $\hat{\pi}_{m,n}(\mathcal{A})$ are defined by (2) and Definition 2.4, respectively.*

Proof. Denote by $\mathcal{A}^1, \ldots, \mathcal{A}^l$ all the elements of Ω. Applying Lemma A.2 we find a partition $I_1, \ldots, I_N$ such that $\max_{j=1,\ldots,N} \sum_{i \in I_j} \pi_{m,n}(\mathcal{A}^j) \leq t_2$ and $N \leq \frac{2}{t_2} + 1$. Using the union bound, we have that

$$\mathbb{P}\left(\max_{i=1,\ldots,l} \hat{\pi}_{m,n}(\mathcal{A}^i) \geq t_1 \right) \leq N \max_{j=1,\ldots,N} \mathbb{P}\left(\sum_{i \in I_j} \hat{\pi}_{m,n}(\mathcal{A}^i) \geq t_1 \right).$$

Note that when $B = 1$, $r \sum_{i \in I_j} \hat{\pi}_{m,n}(\mathcal{A}^i)$ is a binomial random variable, where there are r trials, each with the probability of success $p_j^* = \sum_{i \in I_j} \pi_{m,n}(\mathcal{A}^i)$. We could conclude from Lemma A.1 that

$$\mathbb{P}\left(\sum_{i \in I_j} \hat{\pi}_{m,n}(\mathcal{A}^i) \geq t_1 \right) \leq \left[\left(\frac{p_j^*}{t_1} \right)^{t_1} \left(\frac{1 - p_j^*}{1 - t_1} \right)^{1 - t_1} \right]^r \leq \left[\left(\frac{t_2}{t_1} \right)^{t_1} \left(\frac{1 - t_2}{1 - t_1} \right)^{1 - t_1} \right]^r,$$

where we used the fact that $\left(\frac{x}{t_1} \right)^{t_1} \left(\frac{1-x}{1-t_1} \right)^{1-t_1}$ is increasing for $x \in [0, t_1]$. When $B = 1$, the above displayed equation, combined with $N \leq \frac{3}{t_2}$ gives that

$$\mathbb{P}\left(\max_{i=1,\ldots,l} \hat{\pi}_{m,n}(\mathcal{A}^i) \geq t_1 \right) \leq \frac{3}{t_2} \left[\left(\frac{t_2}{t_1} \right)^{t_1} \left(\frac{1 - t_2}{1 - t_1} \right)^{1 - t_1} \right]^r.$$

Finally, when $B > 1$, $r \sum_{i \in I_j} \hat{\pi}_{m,n}(\mathcal{A}_j)$ is a sample average of B (not necessarily independent) binomial random variables. Since the average of a collection of non-negative numbers is always no greater than its maximum, we could simply use the union bound again to establish that

$$\mathbb{P}\left(\max_{i=1,\ldots,l} \hat{\pi}_{m,n}(\mathcal{A}^i) \geq t_1 \right) \leq \frac{3B}{t_2} \left[\left(\frac{t_2}{t_1} \right)^{t_1} \left(\frac{1 - t_2}{1 - t_1} \right)^{1 - t_1} \right]^r.$$

□

A.2.2 Proof of Theorem 2.1

Proof of Theorem 2.1. For notational convenience, define $\hat{\omega}_{j,m} = \hat{\omega}_j(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$, $\delta = \pi_{m,n}(\mathcal{S})$ and $\theta = \max_{\mathcal{A} \not\subset \mathcal{S}, |\mathcal{A}| \leq k_{\max}} \pi_{m,n}(\mathcal{A})$, where $\pi_{m,n}(\cdot)$ is given by (2). We start from showing that δ and θ are well-separated for sufficiently large n.

Take $\epsilon = \frac{c_\eta m^{-\eta}}{2}$. Using (A1) and (A5) combined with a simple Bonferroni's inequality, we get $\delta \geq \mathbb{P}\left(\max_{j=1,\ldots,p} |\hat{\omega}_{j,m} - \omega_j| < \epsilon \right) \geq 1 - C_\epsilon p \exp(-m^\gamma)$ for some constant $C_\epsilon > 0$. In views of (A2)

and (A3), since here we assume that $\gamma b_2 > b_1$, we get that $\delta = 1 - O(\exp(-n^{\gamma b_2}))$, which tends to one as $n \to \infty$.

For every $\mathcal{A} \in \Omega_k$ with $k \leq k_{\max}$ that contains at least one $a \in \mathcal{A} \setminus \mathcal{S}$, if there is no tie in the ranking of $\{\hat{\omega}_{j,m}\}_{1 \leq j \leq p}$, we have that

$$\begin{aligned} \pi_{n,m}(\mathcal{A}) &= \mathbb{P}\left(\min_{j \in \mathcal{A}} \hat{\omega}_{j,m} > \max_{j \notin \mathcal{A}} \hat{\omega}_{j,m}\right) \leq \mathbb{P}\left(\hat{\omega}_{a,m} > \max_{j \in \mathcal{M}_a \setminus \mathcal{A}} \hat{\omega}_{j,m}\right) \\ &\leq \frac{1}{|\mathcal{M}_a| - k_{\max}} \leq \frac{1}{\min_{a \notin \mathcal{S}} |\mathcal{M}_a| - k_{\max}} \\ &\leq \frac{1}{C_3 n^{b_3} - C_4 n^{b_4}}, \end{aligned} \tag{6}$$

where $\mathcal{M}_a$ is as in (A4). Here we utilized the exchangeability of $\{\hat{\omega}_{j,m}\}_{j \in \mathcal{M}_a \setminus \mathcal{S}}$ together with (A4) and (A7). Even if there are ties, we still have that $\pi_{n,m}(\mathcal{A}) \leq 1/(C_3 n^{b_3} - C_4 n^{b_4})$ due to the exchangeability and since we break the ties uniformly at random. See also the previous proof of Proposition 2.1 for a similar argument but with a more detailed explanation. Notice that (6) does not depend on $\mathcal{A}$ or a, so the inequality $\pi_{n,m}(\mathcal{A}) \leq 1/(C_3 n^{b_3} - C_4 n^{b_4})$ holds for every $\mathcal{A} \in \Omega_k$ with $k \leq k_{\max}$ and $\mathcal{A} \setminus \mathcal{S} \neq \emptyset$. As such, we conclude that $\theta = \max_{\mathcal{A} \not\subset \mathcal{S}, |\mathcal{A}| \leq k_{\max}} \pi_{m,n}(\mathcal{A}) = O(n^{-b_3})$.

Next, to fix ideas, take $\Delta = (b_2 + b_3 - 1)/2$ (NB. $\Delta > 0$ from (A4)), $t_1 = n^{(-b_3 + \Delta)/2}$ and $t_2 = t_1^2$. Note that for sufficiently large n we always have $\theta < t_1^2 < t_1 < \frac{1}{2} < \delta$. Now define events

$$\begin{aligned} \mathcal{E}_k &= \left\{ \max_{\mathcal{A} \in \Omega_k, \mathcal{A} \not\subset \mathcal{S}} \hat{\pi}_{m,n}(\mathcal{A}) < t_1 \right\}, \text{ for } k = 1, \ldots, k_{\max}, \\ \mathcal{B} &= \left\{ \hat{\pi}_{m,n}(\mathcal{S}) > \frac{1}{2} \right\}, \\ \mathcal{E} &= \mathcal{B} \cap \bigcap_{k=1}^{k_{\max}} \mathcal{E}_k. \end{aligned}$$

We will demonstrate that $\mathbb{P}(\mathcal{E}) \underset{n}{\to} 1$ at an exponential rate, and with $\hat{\mathcal{A}}_{\hat{s},m} = \mathcal{S}$ on the event $\mathcal{E}$.

To prove the first claim, when $B = 1$, for sufficiently large n, we could use Lemma A.1 and the fact that $1 - \delta = O(\exp(-n^{\gamma b_2})) \underset{n}{\to} 0$ to bound $\mathbb{P}(\mathcal{B}^c)$ by

$$\mathbb{P}(\mathcal{B}^c) \leq \left[\left(\frac{\delta}{0.5} \right)^{0.5} \left(\frac{1 - \delta}{0.5} \right)^{0.5} \right]^r \leq [2(1 - \delta)]^{0.5r} \leq \exp\left(-C' n^{\gamma b_2 (1 - b_2)/2} \right) \tag{7}$$

for some $0 < C' < 1$. When $B > 1$, since $\hat{\pi}_{m,n}(\mathcal{S})$ is the average of B copies of the that with $B = 1$, using the Bonferroni bound, we have that $\mathbb{P}(\mathcal{B}^c) \leq B \exp\left(-C' n^{\gamma b_2 (1-b_2)/2}\right)$. Moreover, by Lemma A.3,

$$\mathbb{P}(\mathcal{E}_k^c) \leq \frac{3B}{t_2}\left[\left(\frac{t_2}{t_1}\right)^{t_1}\left(\frac{1-t_2}{1-t_1}\right)^{1-t_1}\right]^r = \frac{3B}{t_1^2}\left[\left(\frac{t_1}{1+t_1}\right)^{t_1}(1+t_1)\right]^r. \tag{8}$$

Take the logarithm of $\left(\frac{t_1}{1+t_1}\right)^{t_1}(1+t_1)$. After simple algebra we get $t_1 \log\left(\frac{t_1}{1+t_1}\right) + \log(1+t_1) = t_1 \log\left(1 - \frac{1}{1+t_1}\right) + \log(1+t_1)$, which can be bounded using (A6) and $\log(1+x) \leq \frac{2x}{2+x}$ for $x \in (-1, 0)$ and $\log(1+x) \leq \frac{x}{2}\frac{2+x}{1+x}$ for $x \geq 0$ (Topsøe, 2004). Putting things together, we have that $t_1 \log\left(1 - \frac{1}{1+t_1}\right) + \log(1+t_1) \leq -t_1 \frac{(2-t_1-2t_1^2)}{2(1+t_1)(1+2t_1)} \leq -\frac{t_1}{6}$. Here we also used the fact that the function $h(x) = \frac{(2-x-2x^2)}{2(1+x)(1+2x)}$ is decreasing for $x \in [0,1]$, $h(\frac{1}{2}) = \frac{1}{6}$ and $t_1 = n^{(-b_3+\Delta)/2} < \frac{1}{2}$. This applied to (8) yields

$$\mathbb{P}(\mathcal{E}_k^c) \leq \frac{3B}{t_1^2}\exp\left(\frac{-rt_1}{6}\right) < \exp\left(-C'' n^{1-b_2-b_3/2}\right), \tag{9}$$

with positive constant C'', for sufficiently large n. (A4) implies that the right hand side of the above inequality goes to 0 because (A4) says that $1 - b_2 - b_3/2 > 0$. It follows from (7), (9) and (A7) that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\geq 1 - k_{\max}\exp\left(-C'' n^{1-b_2-b_3/2}\right) - \exp\left(-C' n^{\gamma b_2(1-b_2)/2}\right) \\
&\geq 1 - C_4 n^{b_4}\exp\left(-C'' n^{1-b_2-b_3/2}\right) - \exp\left(-C' n^{\gamma b_2(1-b_2)/2}\right) \\
&\geq 1 - \exp\left(-C_\beta n^\beta\right)
\end{aligned}$$

for some $\beta \in (0,1)$ and $C_\beta > 0$, for sufficiently large n. Therefore, $\mathbb{P}(\mathcal{E}) \underset{n}{\to} 1$.

The remaining arguments used in the proof are valid on $\mathcal{E}$ with a sufficiently large n. Notice that from $1/2 > t_1$ one concludes that $\hat{\mathcal{A}}_{s,m} = \mathcal{S}$, where $\hat{\mathcal{A}}_{s,m}$ is given by (4), hence showing $\hat{s} = s$ proves $\hat{\mathcal{S}} = \mathcal{S}$. Denote $T_k = \frac{\hat{\pi}^\tau_{m,n}(\hat{\mathcal{A}}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k,m})}$, then from definition, $\hat{s} = \operatorname{argmin}_{k=0,1,\ldots,k_{\max}} T_k$. Three cases are considered.

- For every $k = 0, \ldots, s-1$, the event $\{\{R_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_m), \ldots, R_{n,s}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)\} = \mathcal{S}\}$ implies that the index set $\{R_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_m), \ldots, R_{n,k+1}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)\}$ (i.e. of size $k+1$) must be one

of the elements in $\{\mathcal{A} \in \Omega_{k+1} : \mathcal{A} \subset \mathcal{S}\}$. Consequently,

$$\sum_{\{\mathcal{A}\in\Omega_{k+1}:\mathcal{A}\subset\mathcal{S}\}} \hat{\pi}_{m,n}(\mathcal{A}) \geq \hat{\pi}_{m,n}(\mathcal{S}).$$

The facts that $\Big|\{\mathcal{A} \in \Omega_{k+1} : \mathcal{A} \subset \mathcal{S}\}\Big| = \binom{s}{k+1}$ and $\hat{\pi}_{m,n}(\mathcal{S}) > \frac{1}{2}$ imply that

$$\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m}) \geq \max_{\{\mathcal{A}\in\Omega_{k+1}:\mathcal{A}\subset\mathcal{S}\}} \hat{\pi}_{m,n}(\mathcal{A}) \geq \frac{\hat{\pi}_{m,n}(\mathcal{S})}{\binom{s}{k+1}} \geq \frac{1}{2\binom{s}{k+1}},$$

and hence $T_k \geq \frac{1}{2\binom{s}{k+1}}$. for $k = 0, \ldots, s-1$.

- Directly from the definition of the events $\mathcal{E}_s$ and $\mathcal{B}$, we bound $T_s \leq 2t_1^\tau$.

- $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m}) \geq \frac{1}{Br}$ for any k. To see this, note that $\sum_{\mathcal{A}\in\Omega_{k+1}} \hat{\pi}_{m,n}(\mathcal{A}) = 1$. Picking $\hat{\mathcal{A}}_{k+1,m} \in \operatorname{argmax}_{\mathcal{A}\in\Omega_{k+1}} \hat{\pi}_{m,n}(\mathcal{A})$ would mean that $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m}) > 0$, because otherwise it would imply $\sum_{\mathcal{A}\in\Omega_{k+1}} \hat{\pi}_{m,n}(\mathcal{A}) = 0$, leading to a contradiction. Now that $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m}) > 0$, it must be the case that $\hat{\pi}_{m,n}(\hat{\mathcal{A}}_{k+1,m}) > \frac{1}{Br}$, according to Definition 2.4. Thus $T_k \geq \frac{1}{t_1(Br)^\tau}$ for every $k = s+1, \ldots, k_{\max}$.

To prove $T_k > T_s$ for $k = 0, \ldots, s-1$, it is sufficient to demonstrate that $\frac{1}{2\binom{s}{k+1}} > 2t_1^\tau$, which is true for sufficiently large n, as $t_1 \underset{n}{\to} 0$ and $\max_{k=0,\ldots,s-1} \binom{s}{k+1}$ is bounded. Similarly, to claim that $T_s < T_k$ for $k = s+1, \ldots, k_{\max}$, we need to show $2t_1^\tau < \frac{1}{t_1(Br)^\tau}$, which amounts to $2t_1^{1+\tau} < \frac{1}{(Br)^\tau}$, or $2^{1/\tau} t_1^{1+1/\tau} < \frac{1}{Br}$. This is true for sufficiently large n, because $t_1^2 = n^{-b_3+\Delta}$, $Br = O(n^{1-b_2})$ and $b_2 + b_3 - \Delta > 1$ from (A4).

Therefore T_k is necessarily minimised at $k = s$ over $\mathcal{E}$ for sufficiently large n, meaning that $\hat{s} = s$, which finishes the proof. □

Acknowledgement

We thank Jinyuan Chang, Cheng Yong Tang and Yichao Wu for kindly providing us with the code for implementing the marginal empirical likelihood screening method. We also thank the associate editor and the referee for their comments and suggestions. The third author's research is supported by the Engineering and Physical Sciences Research Council grant no. EP/L014246/1.

References

R. Baranowski. *On variable selection in high dimensions, segmentation and multiscale time series*. PhD thesis, The London School of Economics and Political Science (LSE), 2016.

R. Baranowski, P. Breheny, and I. Turner. *rbvs: Ranking-Based Variable Selection*, 2015. URL `https://CRAN.R-project.org/package=rbvs`. R package version 1.0.2.

A. E. Barut. *Variable Selection and Prediction in High Dimensional Problems*. PhD thesis, Princeton University, 2013.

E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *J. Amer. Statist. Assoc.*, 111:1266–1277, 2016.

P. J. Bickel, F. Götze, and W. R. van Zwet. *Resampling fewer than n observations: gains, losses, and remedies for losses*. Springer, 2012.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, 5(1):232–253, 2011.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

J. Chang, C. Y. Tang, and Y. Wu. Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, 41:2123–2148, 2013.

H. Cho and P. Fryzlewicz. High dimensional variable selection via tilting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74:593–622, 2012.

A. Delaigle and P. Hall. Effect of heavy tails on ultra high dimensional variable ranking methods. *Statist. Sinica*, 22:909–932, 2012.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38:3567–3604, 2010.

J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106:544–557, 2011.

F. Götze and A. Račkauskas. Adaptive choice of bootstrap sample sizes. *Lect. Notes-Monograph Ser.*, 31:286–309, 2001.

P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.*, 18, 2009a.

P. Hall and H. Miller. Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.*, 37:3929–3959, 2009b.

X. He, L. Wang, and H. G. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.*, 41:342–369, 2013.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, 2007.

G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *Ann. Statist.*, 40: 1846–1877, 2012a.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.*, 107:1129–1139, 2012b.

J. Liu, W. Zhong, and R. Li. A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58:1–22, 2015.

K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72: 417–473, 2010.

D. F. Saldana and Y. Feng. Sis: An r package for sure independence screening in ultrahigh-dimensional statistical models. *J. Stat. Softw.*, 82, 2018.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 75:55–80, 2013.

X. Shao and J. Zhang. Martingale difference correlation and its use in high dimensional variable screening. *J. Amer. Statist. Assoc.*, 109:1302–1318, 2014.

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *Ann. Appl. Stat.*, 3:1236–1265, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58:267–288, 1996.

F. Topsøe. Some bounds for the logarithmic function. *Inequal. Theory Appl.*, 4:137–151, 2004.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38: 894–942, 2010.

L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.*, 106:1464–1475, 2011.