Abstract We introduce extensions of Stability Selection, a method to stabilise variable selection methods introduced by Meinshausen and Bühlmann (2010). We propose to apply a base selection method repeatedly to subsamples of the observations and to subsets of the covariates under scrutiny, and to select covariates based on their selection frequency. We analyse the effects and benefits of these extensions. Our analysis generalizes the theoretical results of Meinshausen and Bühlmann (2010) from the case of half-samples to subsamples of arbitrary size. We study, in a theoretical manner, the effect of taking random covariate subsets using a simplified score model. Finally we validate these extensions on numerical experiments on both synthetic and real datasets, and compare the obtained results in detail to the original stability selection method.

Keywords Variable selection · Stability Selection · Subsampling

1 Introduction

1.1 Motivation

In many applications a very large number of covariates are observed, of which only a few carry information about an outcome of interest. Variable selection techniques aim at identifying such relevant covariates (for a review see Guyon 2006). Usually, variable selection aims at one of two goals: to identify informative covariates in order to get scientific insight into the data and the process that generated the outcome; or to use the covariates identified as relevant in order to predict the outcome. Even though we give some prediction results in Section 4.2 this work primarily focuses on the identification of informative covariates for the sake of interpretation of the data. We consider variable selection (also called feature selection in computer science-related communities) as a part of the broader field of dimensionality reduction.

Many variable selection methods share the common drawback of being unstable with respect to small changes of the data: if one estimates the set of relevant covariates on different sets of observations coming from the same source, the result can vary significantly. While this is not necessarily of concern if prediction is the goal, it makes the identification of relevant covariates very difficult. One approach to overcome this problem is Stability Selection (Meinshausen and Bühlmann 2010). It consists of applying a variable selection method to randomly chosen subsamples of the observations, and counting for each covariate the number of times it was selected. The final selection is obtained by picking only those covariates whose selection frequency exceeds a certain threshold. This threshold can be chosen such that the expected number of false positive selections is guaranteed to be below a chosen value (Meinshausen and Bühlmann 2010 Section 2.4).
1.2 Contributions

In the remainder of the paper we will refer to the variable selection method that is repeatedly applied to data subsamples as the base method. Similarly to [Meinshausen and Bühlmann (2010), we construct a method that can be applied on top of an arbitrary base method, which is considered a black box.

We propose to extend the central idea of Stability Selection in two natural directions. First, Meinshausen and Bühlmann (2010) use random samples containing half of the observations of the full dataset. Instead, we choose some integer \( L > 1 \) and draw subsamples of size \( \frac{1}{L} \) of the full sample size. More precisely, we randomly partition the observations into disjoint subsamples, extending the approach of Complementary Pairs Stability Selection - CPSS (Shah and Samworth 2013). We investigate, in a theoretical manner, the behaviour of the expected number of false positive selections depending on the number of subsamples. In addition, we perform extensive simulation studies to compare the number of correct variables recovered for artificial and semi-synthetic datasets.

Secondly, Meinshausen and Bühlmann (2010) remarked from empirical comparisons that Stability Selection can be improved by randomising the base method. We propose a randomization by simply applying the base method to random subsets of the covariates. Because the effect of doing so depends on the base method being used, it is difficult to analyse it, in a theoretical manner, in full generality. In this work, we restrict our theoretical analysis to a simplified toy model, in which we assume that for each covariate there exists a latent score reflecting its informativeness about the outcome of interest. Furthermore, we assume that the base method has access to noisy observations of these scores, and outputs the covariate with the largest observed score. We investigate how the estimation of the index of the covariate with the maximal observed score is influenced by the size of the subset of covariates used. Besides the theoretical analysis of this toy model, we performed simulation studies similar to the ones in the investigation of the subsampling of observations.

We call the method that combines the two extensions proposed extended Stability Selection. It can be seen as applying the base method to randomly chosen subsets of the observations and covariates and finally ranking covariates by their selection frequency.

There are two parallel goals for this extension. The first goal is to improve the precision of the selection, that is to reduce the number of falsely chosen noise covariates. The second goal is to reduce the computational complexity of Stability Selection. Indeed, each call of the base method is restricted to a subset of observations and covariates; not only does this significantly reduce the computation cost of a single call, but also the memory requirements. This is particularly appealing if the base method needs to load the data that it operates on into memory, which can be infeasible for large data matrices, but easy for smaller submatrices. Furthermore, this naturally allows for parallelization of the method, since these submatrices can be processed independently.

1.3 Overview of results

Concerning the subsampling of observations, we obtain a bound on the expected number of false positives, depending on the size of the subsamples. This bound sharply generalizes Theorem 1 of Meinshausen and Bühlmann (2010) and Theorem 1 of Shah and Samworth (2013). Our results suggest that there is a trade-off between improving the selection of covariates for each individual subsample by using a smaller number of larger subsamples, and improving the final selection by averaging over a larger number of smaller, independent subsamples. This finding is in line with general insights on subsampling methods (Politis et al., 1999). Even though our empirical comparison shows only small differences, a significant advantage of our proposed subsampling extension is that it has much less computational and memory requirements compared to the original Stability Selection or CPSS.

For the randomization of the base method obtained by taking subset of covariates, the theoretical analysis of our simplified score model shows that under certain assumptions, there exists an optimal size for the randomly chosen subsets of covariates. Our empirical results support this finding: such randomization generally improves the performance, unless the subset size is too small.

1.4 Organization of the paper

In Section 2 we give a detailed description of the algorithm proposed, including details about its implementation in R. The theoretical analysis is presented in Section 3. We motivate the use of small subsamples of observations in Section 3.1 and investigate the randomization of the base method in Section 3.2. Experimental Results are given in Section 4, where we measure the performance of the algorithm in selecting informative covariates in Section 4.1 and apply our method in an image classification setting in Section 4.3. We conclude our work in Section 5 with a summary and a discussion.
2 Methods

2.1 Description of the algorithm

The basic idea of extended Stability Selection is to partition the observations into $L$ disjoint subsamples and to apply a randomized base method to each subsample. We randomize the base method by applying it to subsets of the covariates.

In the sequel we assume to be given a dataset containing $N$ observations of $D$ covariates and an outcome of interest. We choose $T$, the number of times we repeat the random partitioning of the data and a threshold $\tau \in (0, 1)$ that indicates the fraction of subsamples of observations in which a covariate needs to be chosen in order to enter the final selection. The number of observations and covariates that we apply our base method on is determined by the crucial parameters $L$ and $V$. Hereby we use subsamples of observations of size $\lfloor N / V \rfloor$ and randomly choose in each subsample $\lfloor D / L \rfloor$ covariates. Finally we denote by $\text{Base}(S^{(i)}, F^{(j)})$ the output of the base selection method applied to the covariates $F^{(j)}$ of observations $S^{(i)}$. We give the pseudocode of the method we propose below. Note that we recover the original Stability Selection (more precisely, the CPSS algorithm of Shah and Samworth [2013]) for $L = 2, V = 1$.

Parameters:
- Number of iterations $T$
- Number of subsamples of observations per iteration $L$
- Number of subsets of covariates per iteration $V$
- Threshold $\tau \in (0, 1)$

Input: $N$ observations of $D$ covariates

Initialization: scores $Q_d = 0, d = 1, \ldots, D$

for $t = 1$ to $T$

Draw $L$ disjoint random subsamples $S^{(1)}, \ldots, S^{(L)}$ of size $\lfloor N / V \rfloor$ from $S = \{1, \ldots, N\}$ without repetition.

Draw $V$ random subsets of the covariates indices of size $\lfloor D / L \rfloor$ without repetition.

for $i = 1, \ldots, L; j = 1, \ldots, V$

$G := \text{Base}(S^{(i)}, F^{(j)})$

for all $d \in G$

$Q_d := Q_d + 1 / LT$

end for

end for

return set of indices $\hat{S} := \{ d : Q_d \geq \tau \}$.

For computational reasons we use in our experiments a variation of the algorithm described above: Instead of taking random subsets of the covariates we use random partitions of the covariates.

2.2 Comparison to previous work

Statistical methods can be applied to subsamples of the data in various ways. A classical way in this context is the bootstrap (Efron [1979]), where subsamples are drawn with replacement. In contrast, Stability Selection and our extension follow the idea of subsampling without replacement (Politis et al. [1999]) and are strongly related to subagging (Bühlmann and Yu [2002]).

Several approaches have been developed to combine variable selection and subsampling of observations. Sauerbrei and Schumacher (1992) investigated bootstrapping variable selection methods in the Cox Regression Model. Further, several methods have been proposed where a predictor that does variable selection intrinsically is applied to subsamples of the data. The selection obtained is then used as input to the final prediction method. Bi et al. (2003) used linear Support Vector Machines (SVMs) to select variables and train a kernel SVM on them. Similarly, in Random Lasso (Wang et al. [2011]) the importance of covariates is estimated by applying Lasso to bootstrap subsamples of the data. Afterwards Lasso is applied to a random selection of the covariates, the probability to be included is proportional to the measure obtained in the first step.

Meinshausen and Bühlmann (2010) proposed Stability Selection, a method that combines subsampling with variable selection. In their analysis the authors give a bound for the number of false positives depending on the regularization parameter. Their work was the basis for numerous follow up studies by theoreticians and practitioners.

Shah and Samworth (2013) introduced Complementary Pairs Stability Selection (CPSS), a variant of Stability Selection which uses not only subsamples of size $\lfloor N / V \rfloor$, but also its complements. They loosen the assumptions of the original method and give bounds on both errors of the selection procedure, false positives and false negatives. Our work is an extension of their results, as we apply our base method to complementary subsamples as well.

Stability Selection has been applied in various disciplines. Fields of application include genome-wide association studies (Alexander and Lange [2011], He and Lin [2011]), biomarker discovery (He and Yu [2010]) and inference of gene regulatory networks (Haury et al. [2012]).
The idea of repeatedly applying a statistical method to subsets of the covariates has been investigated before. One famous example is Random Forest (Breiman, 2001) where decision trees are built on subsets of the covariates. Each decision tree can be regarded as a variable selection method as well. Recently, Hinton et al. (2012) remarked that omitting randomly chosen covariates in the training of a neuronal network improves classification accuracy on test data drastically. A different approach to reduce the dimensionality of the problem is to cluster the covariates and apply a variable selection method to cluster representatives (Bühlmann et al., 2013).

2.3 Base methods

In Stability Selection any variable selection method can be used as a base method. In this subsection we describe variable selection methods in general and give details about two selection methods that we used as a base method for extended Stability Selection.

Variable selection methods can be classified into filters, wrapper and embedded methods (Guyon, 2006). Computationally most efficient are usually filter methods, which perform a variable selection independently of the specific statistical treatment that might be applied afterwards. Examples are methods based on simple univariate correlation between covariates and the outcome of interest or Mutual Information. Wrapper methods evaluate the relevance of a subset of covariates using the output of the ensuing statistical treatment (typically regression or discrimination) computed on the subset of covariates only. Embedded methods perform variable selection and prediction at the same time.

To assess the performance of the proposed methods, we choose two popular methods as base method: CMIM (Fleuret, 2004) from the class of filters and Lasso (Tibshirani, 1996) from the class of embedded methods. Lasso was also used as a base method in the original work of Meinshausen and Bühlmann (2010).

2.3.1 Conditional Mutual Information Maximization (CMIM)

Intuitively, the aim of the CMIM algorithm (Fleuret, 2004) is to find a subset of covariates of given size $K$ that maximizes the amount of information that the selected covariates $X_{\nu(1)}, \ldots, X_{\nu(K)}$ contain about the outcome $Y$. It usually finds a good trade-off between relevance and redundancy of the selected covariates and is much faster than many competing Mutual Information based variable selection methods.

The notion of information is quantified through the Mutual Information (see, e.g., Cover and Thomas, 2006, Chapter 2).

To make the computation feasible, the CMIM algorithm does not look directly for the set of covariates that globally maximizes the Mutual Information with the target, but performs greedy selection by iteratively selecting the covariate that has the largest mutual information with the target, conditional to the set of already selected covariates. Furthermore, the latter quantity is approximated by a simpler upper bound, namely, the minimal information about the target that a candidate covariate adds to any of the already selected covariates, taken individually (rather than jointly). The final algorithm takes the following simple form:

$$
\nu(1) = \text{Arg Max}_{d \in \{1, \ldots, D\}} I(X_d; Y); \\
\nu(k) = \text{Arg Max}_{d \in \{1, \ldots, D\}} \min_{k \leq d} I(X_d; Y | X_{\nu(k-1)}),
$$

where $I(X; Y)$ denotes the Mutual Information of $X$ and $Y$ and $I(X; Y | Z)$ the Mutual Information between $X$ and $Y$ given $Z$. To speed up the computation, one can use a fast implementation of the algorithm (Fleuret, 2004).

2.3.2 Lasso

The Lasso problem (Tibshirani, 1996) is to find an $\ell_1$ regularized solution for the least squares problem in the linear model. It can be stated as

$$
\hat{\beta}_\lambda = \text{Arg Min}_{\beta \in \mathbb{R}^D} \{||Y - X\beta||_2^2 + \lambda|\beta|\}
$$

(1)

where $\lambda \in \mathbb{R}_{\geq 0}$ is some regularization parameter.

Thanks to the geometric properties of the $\ell_1$-Norm, solutions of the Lasso problem tend to be sparse, meaning that only a few coefficients of $\beta$ are non-zero. This property allows us to use Lasso solvers as variable selection methods, and to define $\tilde{S}_\lambda := \{i : \beta_\lambda^{(i)} \neq 0\}$ for any $\lambda > 0$. Additionally to the selection of relevant covariates, a solution $\hat{\beta}_\lambda$ of the Lasso problem can be used to predict an outcome for new data. Examples for popular Lasso solvers are lars (Hastie and Efron, 2012) and glmnet (Friedman et al., 2010).

3 Analysis

In this section, we analyse the effect of the size of the observations subsamples as well as the effect of taking random subsets of covariates on the performance of the final method.
In the first subsection we focus on subsampling the observations. We show that the expected number of errors made by the final method can be bounded and we specify the notion of an error. This part of the analysis extends the results in [Shah and Samworth (2013)] to an arbitrary number of subsamples of the observations.

In the second subsection, we motivate, in a theoretical manner, the use of subsets of the covariates. We analyse a simplified toy model and show in particular that the performance of a variable selection method can improve when it is applied to a random subset of covariates rather than directly to all covariates.

3.1 Subsampling of observations

We assume to be given a dataset with \( N \) observations and \( D \) covariates and a base variable selection method. The \( N \) observations are always assumed to be drawn i.i.d. from an underlying, unknown generating distribution.

We now draw \( L \) disjoint subsamples of observations of equal size without repetition and repeat this drawing \( T \) times. We apply the base method to each of these subsamples of observations, count the number of times each covariate was selected and select the covariates with selection frequency above a certain threshold. This naturally extends the notion of Complementary Pairs Stability Selection [Shah and Samworth (2013)] (which is obtained in the case \( L = 2 \)).

On the \( L \times T \) subsamples that we obtain, the selection frequency of covariate \( d \) is

\[
P_{SFS}^L(d) := \frac{1}{TL} \sum_{t=1}^{T} \sum_{\ell=1}^{L} 1\{d \in S_{L,\tau}^{\text{base}}(\ell, t)\}
\]

where \( S_{L,\tau}^{\text{base}}(\ell, t) \) is the output of the the base selection procedure applied to the \( \ell \)-th subsample of observations in the \( t \)-th drawing.

Thresholding this quantity, we obtain the output of the Stability Selection procedure. For any \( \tau \in (0, 1) \) we define

\[
S_{L,\tau}^{SFS} := \{d : P_{SFS}^L(d) \geq \tau\}.
\]

To evaluate the performance of the method, we need to define the set of covariates that we would like to be excluded from our selection, the unwanted covariates. Since the base method is treated here as a black box and otherwise unspecified, we have to trust that, on average (over a random i.i.d. sample), the base method selects relevant covariates more frequently than irrelevant ones. Consider virtually drawing an independent, i.i.d. set of observations of size \( \lfloor \frac{N}{2} \rfloor \) and denote the random output of the base method on this sample by \( S_{L}^{\text{base}} \). Then we define for each covariate \( d \)

\[
p_L(d) := \mathbb{P} \{d \in S_{L}^{\text{base}}\},
\]

its probability to be selected by the base method using a sample of size \( \lfloor \frac{N}{2} \rfloor \).

The quantities \( p_L(d) \) give us a yardstick to rank the covariates by relevance; accordingly, for any given threshold \( \theta \in (0, 1) \) we define

\[
A_{\theta,L} := \{d : p_L(d) \leq \theta\}
\]

the set of unwanted covariates at base selection probability lower than \( \theta \). In this definition, and the assumption that the probability of selection under the base method reflects the true relevance, we follow the general approach of [Shah and Samworth (2013)], the relation to the assumptions of [Meinshausen and Bühlmann (2010)] is discussed below. Observe that, since each subsample appearing in the definition \( P_{SFS}^L(d) \) is individually an i.i.d. sample of size \( \lfloor \frac{N}{2} \rfloor \), it follows that \( P_{SFS}^L(d) \) is an unbiased estimate of \( p_L(d) \).

The following theorem bounds the ratio of the expected number of unwanted covariates selected by \( S_{L,\tau}^{SFS} \) compared to the expected total number of unwanted covariates, as well as compared to the expected number of unwanted covariates selected by the base method applied on a single subsample. A corresponding result for the wanted covariates missed out is available in the second part of the theorem.

**Theorem 1** Let \( L \in \mathbb{N} \), \( \ell_0 \in \{1, \ldots, L\} \) and \( \tau \in (0, 1) \). Denote \( p_0 := \frac{\ell_0}{L} \), \( D(p,q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \) and denote the constants \( C(p_0, \tau, L) := \left(\frac{1-\rho_0+\frac{1}{\rho_0}}{\rho_0-\tau+\frac{1}{\rho_0}}\right) \) and \( C'(p_0, \tau, L) := \left(\frac{1-\rho_0+\frac{1}{\rho_0}}{\rho_0-\tau+\frac{1}{\rho_0}}\right) \forall p, q \in (0, 1) \). Depending on the choice of \( \theta, \tau \) and \( \ell_0 \) we have four cases:

- **If** \( \theta < \tau \):
  - **for any** \( \ell_0 \) s.t. \( \lfloor L\theta \rfloor \leq \ell_0 \leq \lfloor L\tau \rfloor \):
    \[
    \mathbb{E} \left[ \left| S_{L,\tau}^{\text{SFS}} \cap A_{\theta,L} \right| \right] \geq C(p_0, \tau, L, \exp(-LD(p_0, \theta))) - (2)
    \]
  - **for any** \( \ell_0 \) s.t. \( \lfloor L\theta \rfloor + 1 \leq \ell_0 \leq \lfloor L\tau \rfloor \):
    \[
    \mathbb{E} \left[ \left| S_{L,\tau}^{\text{SFS}} \cap A_{\theta,L} \right| \right] \leq C(p_0, \tau, L, \exp(-LD(p_0, \theta))) \notag - (3)
    \]
  - Similarly, if \( \tau < \theta \):
corollary 1: suppose we are given a set of noise covariates $\mathcal{N}$ and a set of signal covariates. assume that all noise covariates have the same probability to be selected by the base method. assume further that the base variable selection method has a larger probability to select any wanted covariate than random guessing. we denote $q = E\left[|S^\text{base}\cap \mathcal{N}|\right]$. then for any $\tau > \frac{q}{D}$:

$$E\left[|S^{\mathcal{SF},S}_L\cap \mathcal{N}|\right] \leq \min_{\frac{q}{D} \leq \ell_0 \leq \lfloor \tau L \rfloor} \left(1 - \frac{\ell_0 - 1}{L} \right) \exp \left(-LD\left(\frac{\ell_0}{L}, \frac{q}{D}\right)\right).$$

(6)

if we choose $\ell_0 = L = 2$ and we use that $\exp(-2D(1, \theta)) = \theta^2$ we recover the order $O(q^2/D)$ of the bound of [10] theorem 1 as well as the constraint $\tau > 1/2$.

in [6] we minimized the bound over the allowed values of $\ell_0$, which is merely a technical parameter in the bound. in order to make the bound more readable, we can pick a specific value of $\ell_0$ as follows. as the exponential term in the bound is monotonically decreasing in $\ell_0$, the largest allowed value $\ell_0 = \lfloor \tau L \rfloor$ seems a natural choice. however, one should ensure that the multiplicative term in front does not become too large. for this, choose $\tau$ and $L$ such that $\lfloor \tau L \rfloor$ is an integer; then we have $\frac{\ell_0 - 1}{L} = \frac{\tau - 1}{\tau} = \tau - 1$. in this case, corollary 1 simplifies as follows.

corollary 2: under the assumptions of corollary 1 if $\tau L$ is an integer and if we choose $\ell_0 = \tau L$, then we have as a direct consequence of corollary 1 for any $\tau > \frac{q}{D}$:

$$E\left[|S^{\mathcal{SF},S}_L\cap \mathcal{N}|\right] \leq (L(1 - \tau) + 1) \exp \left(-LD\left(\tau, \frac{q}{D}\right)\right).$$

(7)

3.2 Randomization by taking subsets of covariates

it is not possible to study the effect of taking random subsets of covariates on the selection probabilities in as much generality as we studied the effect of observation subsampling in the previous section, as this effect depends prominently on the specific base procedure used (see section 2.3). here, we focus on score-based base selection procedures (such as cmim, see section 2.3.1), which we analyse through a strongly simplified model. (in the case of the lasso base selection procedure [10] studied a related form of randomization, which consists in attributing random weights to covariates.)

we model a given iteration in an iterative score-based selection procedure as follows. assume that each covariate has an underlying score $Q_d$ reflecting its true informativeness; only an estimation $\hat{Q}_d$ of that score is available, which we assume to follow the simple additive
We expect that the amplitude of the random estimation noise \( \epsilon \) will typically depend on the size of the observations subsample and thus on the parameter \( L \). However in this section we consider \( L \) as fixed and therefore omit it from the notation from now on.

The base procedure then outputs the covariate with the largest estimated score. We want to compare this base procedure to the randomized base procedure consisting in first picking at random a subset of \( D' < D \) covariates, and returning out of those the one with the largest estimated score.

Similarly to what was considered in the previous section, we define unwanted covariates as those having true scores below a certain threshold \( \theta \):

\[
A_{D,\theta} := \{ 1 \leq d \leq D : Q_d \leq \theta \};
\]

we also denote \( A_{D,\theta}^c := \{ 1, \ldots, D \} \setminus A_{D,\theta} \). Denote \( p(d) \) and \( p^{\text{rand}}(d) \) the probability of selecting covariate \( d \) using the deterministic and the randomized base procedure respectively. It is desirable for these to be as large as possible for \( d \in A_{D,\theta}^c \), so that we can compare the two base procedures by means of the sum of these probabilities, i.e., for the deterministic base procedure,

\[
\sum_{d \in A_{D,\theta}^c} p(d) = \mathbb{P} \left[ \hat{D} \in A_{D,\theta}^c \right],
\]

where \( \hat{D} \) denotes the index of the covariate selected by the deterministic base procedure.

In the following theorem we analyse the behaviour of the latter quantity as \( D \) grows large. The main theoretical finding of this subsection is that, under certain circumstances concerning the distribution of the estimation noise, in an asymptotic sense \( \hat{D} \) will be determined only by the estimation error, and not by the true score. In other words, if \( D \) grows too large, the deterministic selection resembles picking at random.

This therefore supports the principle of taking subsets of covariates in Stability Selection, since when the total number of covariates \( D \) is large enough, the probability of correct selection will be higher when taking a (random) covariate subset of size \( D' < D \). This phenomenon is illustrated by a small simulation example at the end of the present section.

**Theorem 2.** Consider a sequence of models of the form (8), a fixed number \( \theta \), and the following assumptions:

(i) The true scores \( Q_d \) belong to the bounded interval \([0, M]\).

(ii) The noise variables \( \epsilon_i \) are independent and identically distributed, and their marginal distribution belongs to the maximal domain of attraction (MDA) of a Fréchet(\( \alpha \)) distribution, for some \( \alpha > 0 \).

(iii) As \( D \to \infty \), \( \frac{|A_{D,\theta}|}{D} \to \eta \in [0, 1] \), where \( A_{D,\theta} \) is defined by (9).

Then for \( d_D = \text{Arg Max} \hat{Q}_d \), it holds that

\[
\lim_{D \to \infty} \mathbb{P} \left[ d_D \in A_{D,\theta} \right] \to \eta.
\]

*Interpretation of the theorem.* Consider for comparison the blind strategy of simply drawing a covariate uniformly at random among \( D \), regardless of observed scores. Then the probability for this covariate to lie in \( A_{D,\theta} \) is obviously \( |A_{D,\theta}|/D \). Thus, the theorem states that as \( D \) grows, the strategy consisting in picking the largest observed scores will asymptotically not be any better than the blind strategy.

*Comments on assumption (iii).* This assumption concerns the asymptotic repartition of the true scores \( Q_d \), \( 1 \leq d \leq D \), as \( D \) grows. It is quite weak, and allows for the family of true scores to depend on \( D \), provided this assumption remains satisfied. In particular, we can apply the theorem if the true scores are themselves random, and such that the assumption is satisfied a.s. In that situation, the theorem can be applied conditionally to the true score sequence and the conclusion will hold (for conditional incorrect selection probabilities) for almost all realizations of this sequence, and therefore also in expectation over the scores (i.e. for unconditional probabilities).

A simple and intuitive instance of the above is when the true scores \( Q_i \) are themselves assumed to be i.i.d. random draws following some (arbitrary) distribution on the interval \([0, M]\). In that case, assumption (iii) is satisfied a.s. with \( \eta = \mathbb{P} \{ Q_1 \leq \theta \} \) by the law of large numbers. Additionally, if the true scores are modeled as random i.i.d., the randomized procedure is equivalent to the base procedure with \( D \) simply replaced by \( D' \). In this sense, the conclusion of the theorem applies to the randomized selection, as well. Denoting \( E_D := \mathbb{P} \{ \hat{D} \in A_{D,\theta} \} \) in this setting, we have clearly \( E_1 = \mathbb{P} \{ Q_1 \leq \theta \} = \eta \) as well as \( \lim_{D \to \infty} E_D = \eta \) by the theorem. On the other hand, it is easy to see that \( E_D > \eta \) for any \( D > 1 \) (for any \( D > 1 \), selecting the covariate with largest observed score must be at least slightly better than random guessing). We conclude that \( E_D \) is not monotone in \( D \), and that it must attain a minimum value for some finite \( D_{\text{opt}} > 1 \). (In the simulations shown at the end of the section, we actually see that \( E_D \) is unimodal.) The same applies to \( q^{\text{rand}}_D \), and we conclude that if \( D > D_{\text{opt}} \), then it brings an advantage to select covariates out of random subsets of size smaller \( D' \) (the optimal size being \( D' = D_{\text{opt}} \)). To sum up the finding in an (apparently) paradoxical
statement: as $D$ grows too large, the deterministic selection behaves more randomly than the randomized selection using $D' < D$.

Comments on assumption (ii). The independence assumption is needed to apply classical results of extreme value theory. It is arguably unrealistic, and made here in order to illustrate the phenomenon in the simplest setting possible. We note that some extreme value results are also available under weak dependence models (Leadbetter et al., 1983, Chapter 3), so that this assumption might be relaxed somewhat, though this is out of the scope of the present work.

The assumption that the noise marginal distribution belongs to MDA(Fr´echet) means that it is heavy-tailed. This includes Cauchy, Student’s t, Pareto, Burr and Loggamma distributed errors (Embrechts, 1997, Table 3.4.2). This assumption is reasonable if one considers that the estimated scores could reasonably be based on $t$-statistics estimated from a limited number of observations. In contrast, the next result shows that this phenomenon does not occur for Gaussian distributed errors:

**Theorem 3** Consider a sequence of models of the form $\hat{d}_D$, a fixed number $\theta$, and the following assumptions:

(i) the noise variables $\varepsilon_i$ are independent and identically distributed with normal distribution.

(ii) $\lim_{D \to \infty} \frac{|A_{D,\theta}|}{D} > 0$, where $A_{D,\theta}$ is defined by Equation (9).

Then for any $\theta' < \theta$:

$$\lim_{D \to \infty} P[\hat{d}_D \in A_{D,\theta'}] = 0.$$  

Finally we illustrate by a simulation study how the error probability $P[\hat{d}_D \in A_{D,\theta}]$ depends on the distribution of the noise $\varepsilon_D$. We simulated 10000 noisy scores according to Equation (8) for each of the five noise distributions Gaussian, Cauchy, Student’s t with 3, 5 and 10 degrees of freedom. We count how often the largest noisy score comes from an uninformative true score and plot this frequency against the total number of scores $D$ in Figure 1. We see that for the Gaussian distribution, the probability to select a uninformative covariate is monotone decreasing to zero as $D$ increases. In contrast, for heavy tailed distributions such as Cauchy and Student’s t, the error probability decreases until reaching a minimum and increases afterwards. For the Student’s t distribution, the error probability behaves as for the Cauchy distribution, but grows more slowly.

**Fig. 1** Dependence of the average error probability on $D$ for different noise distributions

**4 Experimental results**

In this section we evaluate the performance of our method in two different settings and for several datasets. In the first setting, we consider the number of informative covariates selected in a controlled framework where the ground truth is known. In the second setting, we consider a real-data classification task and evaluate the effect of variable selection on the prediction performance of a learning algorithm using only the selected covariates.

**4.1 Identification of informative covariates**

**Setting.** In this subsection, the base method is the Lasso, and the evaluation criterion is the number of truly informative covariates among the top $k = 20$ selected, when ranked by selection frequency. (This criterion is also known as precision@k in the information retrieval literature.) To compute this criterion, we need knowledge of the ground truth, i.e., which are the informative covariates. For this, we generate the output variable of interest according to a known linear model:

$$Y = X\beta + \varepsilon,$$  

with $X \in \mathbb{R}^{N \times D}$, $Y \in \mathbb{R}^N$, $\beta \in \mathbb{R}^D$ and $\varepsilon \in \mathbb{R}^N$. Here $\varepsilon \sim \mathcal{N}(0, 1)$ is a noise vector and $\beta$ is a vector of coefficients, such that only 20 coefficients are non-zero. Their indices are chosen randomly and the coefficients
are generated from a $U[0,1]$ distribution. This setting is similar to the one considered by Meinshausen and Bühlmann (2010). We consider several different settings for the design matrix $X$, corresponding to controlled simulated situations or to real data. Except otherwise specified, each experiment is performed for $N = 500$ observations and $D = 1000$ covariates.

- 4 Blocks: The covariates are divided into 4 blocks with correlation inside but not among the blocks. The covariates follow a multivariate normal distribution $\mathcal{N}_D(0, \Sigma)$, where $\Sigma_{i,j} = 0.5 \times 1 \{i = j \mod 10\}$
- Toeplitz design: The correlation between two covariates is higher the closer their indices are. The covariates follow a multivariate Normal distribution $\mathcal{N}_D(0, \Sigma)$, with $\Sigma_{i,j} = 0.99^{i-j}$
- Toeplitz (grouped predictors): As Toeplitz design, but the indices of the informative covariates consist of 5 groups of 4 indices, each drawn uniformly in the interval $[100g - 20, 100g + 20]$ where $g$ is the group number. Therefore the informative covariates exhibit a cluster structure.
- 10 factors: Each covariate $X_d$ is generated as a linear combination of unknown latent factors

$$X_d = \sum_{i=1}^{10} f_{d,i} \Phi_i + \nu_d \forall d = 1, \ldots, D$$

where the latent factors $\Phi_i$ and the noise $\nu_d$ follow a standard normal distribution. The factor loading coefficients $f_{d,i}$ are fixed for any given realization of the dataset and are drawn beforehand from a $\mathcal{N}(0,1)$ distribution.
- Correlated informative covariates, independent noise: The covariates follow a multivariate normal distribution $\mathcal{N}_D(0, \Sigma)$, with $\Sigma_{i,j} = 0.9$ for all indices $i,j$ of informative covariates and $\Sigma_{i,j} = 0$ elsewhere.
- Vitamin data: A gene expression dataset considered by Meinshausen and Bühlmann (2010) (with $N = 115, D = 4088$.)

The vector $Y$ is generated by the linear model given in Equation (10). (The original labels of the semi-synthetic vitamin dataset is discarded.) The signal to noise ratio is adjusted such that $\text{Var}(X \beta)/\text{Var}(\varepsilon) = 2$.

**Experimental protocol.** For each design, we compare the proposed methods with several choices of parameters, the usual Stability Selection and the Lasso as reference methods. Let us denote an instance of the proposed method by $SFS(L, V)$ with $L, V$ as in Section 2.1. We consider the choices $SFS(L, 1)$ for $L \in \{2,4,8\}$ to investigate the effect of the subsample size of observations, and $SFS(2, V)$ for $V \in \{1,2,4,8\}$ to investigate the effect of the amount of randomization of the base method by taking covariate subsets. Note that the $SFS(2, 1)$ is the standard Stability Selection.

We studied the performance of the different methods for a range of possible regularization parameters of the base method. More specifically, rather than comparing the methods for a grid of fixed values of the regularization parameter $\lambda$ of Lasso, see Equation (1), for each realization of the data, we used the values $\lambda_0$ such that exactly $q \in \{1,\ldots, 100\}$ covariates are selected by the base method. (In case several such values exist, the largest one is picked.) This approach seemed more fair and in line with the analysis of Section 3 in the sense that the number of covariates selected by the baseline, $q = \mathbb{E}[[S_{L,V}^{\text{base}}]]$, is kept constant across choices of $(L, V)$. For each value of $q$, we report the average precision@$k$, i.e., how many informative covariates are selected amongst the top $k = 20$ covariates ranked by their selection frequency. For standalone Lasso, we ranked the covariates by their estimated regression coefficients. The performance reported in Figure 2 and 3 is an average over 100 repetitions of each experiment.

**Results.** A first finding is that Stability Selection does not systematically outperform standalone Lasso, which appears to contradict the results of Meinshausen and Bühlmann (2010). The reason for this is that here are some substantial differences between our evaluation setting and theirs; this is discussed in more detail below. To summarize, we find that standalone Lasso recovers informative covariates more successfully when it is allowed to select a number of covariates higher than the true actual number of informative ones, and to rank them by magnitude of their estimated coefficients. In other words, relevant covariates might at first not be included in the selection set along the Lasso path, but once they are, they tend to get estimated coefficients larger than those of noisy variables. This possibility was not considered by Meinshausen and Bühlmann (2010) due to their stricter error criterion. Correspondingly, in the regime where $q$ is markedly larger than the true number of informative covariates, the success of stability selection over standalone Lasso is not systematic and appears to depend on the setting. A qualitative observation is that stability selection appears to be more successful for situations with limited or short-range dependence (Toeplitz design, Toeplitz grouped, independent noise, realistic datasets with presumably weak dependence) than with systematic dependence structure (factorial design, block design).

As a second finding, we observe that (in a majority of observed cases), if for a fixed value of $q$ standard Stability Selection $SFS(2, 1)$ improves on plain Lasso, then the extended methods we propose generally lead to a further improvement. Naturally, and as expected

---

1 We used the R-package LARS [Hastie and Efron 2012] as Lasso implementation.
from the analysis and discussion in Section 3 eventually the performance deteriorates again if the size of the subsamples of observations $N/L$ becomes too small to allow a reasonable estimation, or if the number of covariates $D/V$ used in the randomized base procedure is too low (experiments not shown). Overall, these results vindicate the relevance of extended stability selection, both for reasons of potential increase in performance, and of efficient scalability via possible parallelization over standard Stability Selection, whenever the latter itself improves on the base method.

Comparison to the setting of Meinshausen and Bühlmann (2010). We now discuss in some detail and motivate the differences in the evaluation criteria used by us and by the authors of the original work on Stability Selection. They reported two evaluation measures: the first was the probability that the top $\ell$ covariates by selection frequency are all relevant (for a fixed $\ell$ equal to some proportion such as 10% of the truly relevant variables). The second was the number of noise variables included in the top-$\ell$ selection, where $\ell$ is chosen (separately in each data realization) so that the selection contains at least a fixed proportion such as 20% of the true variables. In experiments using these criteria, we found that the first criterion was only poorly informative, in the sense that it was often very close to either 0 or 1, not providing a very clear contrast between the methods. Also, we found that the second criterion was in many cases subject to a very large expectation and variance, so that we were also wary about its relevance. For instance, assume that in order to select 5 relevant covariates, Method 1 also selects on average 100 noise covariates, but Method 2 only 80. We doubt that this measure provides a meaningful comparison criterion of the methods. For this reason, we chose precision@k as the performance measure; this criterion has the advantage of being stable, comparable across settings, in plausible relation to intended applications, and is standard in information retrieval. Additionally, for Stability Selection Meinshausen and Bühlmann (2010) only reported results for a specific value of $q$, namely $q \approx \sqrt{0.8D}$, and we preferred to report performance over a range of values of $q$.

A second important difference is that we evaluate standalone Lasso differently. For evaluating standalone Lasso performance, Meinshausen and Bühlmann (2010) ranked covariates by their order of selection along the “Lasso path”, i.e., sequential selection order as the parameter $\lambda$ decreases (for the sake of simplicity of the present discussion, we ignore the fact that some covariates can drop out of the Lasso selection as $\lambda$ decreases). In contrast, we rank the covariates by the magnitude of their estimated Lasso coefficients (for each value of $q$ taken separately).

Overall, adhering to the spirit of Meinshausen and Bühlmann (2010) for precision@20 would correspond, roughly speaking, to reporting this error criterion for Lasso only evaluated at $q = 20$, and for the stability selection methods only evaluated at $\tilde{q} \approx \sqrt{0.8D} \approx 28$ (for $D = 1000$); this puts stability selection in a very favourable light, in our opinion, our protocol presents a more complete and fairer overall picture.

4.2 Application: Image classification.

In this subsection, we use variable selection as a preprocessing step to solve a classification problem. We investigate whether stabilising the variable selection method with extended Stability Selection improves the classification error rate compared with applying the variable selection method to the whole dataset.

Dataset Description. For prediction we used a subset of 6000 images of handwritten digits from the MNIST dataset (LeCun et al., 1998). Covariates were computed using feature extractors from the a collaborative machine learning framework (The MASH project) where external contributors can directly submit feature extractors of their choice by uploading C++ code through a web interface. A total of 48416 covariates were computed on each image from the contributed feature extractors. These covariates are heterogeneous because they come from different contributors, generally exhibit strong correlations, do not have a sparse structure, and many of them contain at least some information about the class label. Here, the goal is not to strictly identify informative covariates, but to reduce the dimensionality of the problem in order to apply a learning and prediction method afterwards. This way, the computational complexity and memory requirements of the learning algorithm can be reduced considerably. Note that in these experiments we are not aiming at outperforming state-of-the-art classification results on MNIST, but we wish to evaluate the effect of the proposed extensions to Stability Selection on prediction performance, as compared to using standard Stability Selection with the same base method.

Experimental setup and results. As base method, we use CMIM (Conditional Mutual Information Maximization, Fleuret, 2004). As described in Section 2.3, CMIM iteratively selects covariates updating in each iteration a score for each covariate. To speed up computation, we limited the number of iterations of CMIM to 10 or 100 (Beinrucker et al., 2012a) and denoted the resulting
Extensions of Stability Selection using subsamples of observations and covariates

Fig. 2 Comparison of plain LASSO and Stability Selection with varying numbers of disjoint subsamples of the observations (corresponding to line types). We plot the average number of informative covariates among the top 20 scored, depending on the number of covariates selected by the base method, figures correspond to designs.

Fig. 3 As Figure 2 but with varying numbers of disjoint subsets of covariates.
Tables 1 and 2 The effect of the size of covariate subsets on prediction error. (selection with SFS\(_{10}(L,1)\) and CMIM, prediction with Adaboost.MH with various numbers of iterations)

| # Iterations | SFS\(_{10}(2,V)\), \(V = 1\) | SFS\(_{10}(2,V)\), \(V = 10\) |
|--------------|----------------------------|----------------------------|
| 50           | 8.1 (0.2)                  | 7.7 (0.3)                  |
| 100          | 5.0 (0.2)                  | 4.8 (0.2)                  |
| 200          | 3.3 (0.1)                  | 2.9 (0.1)                  |
| 400          | 2.8 (0.1)                  | 2.7 (0.2)                  |
| 800          | 2.4 (0.1)                  | 2.5 (0.2)                  |
| 1600         | 2.0 (0.1)                  | 2.0 (0.2)                  |

| # Iterations | SFS\(_{10}(L,1), L = 1\) | SFS\(_{10}(L,1), L = 10\) |
|--------------|--------------------------|---------------------------|
| 50           | 8.1 (0.2)                | 8.0 (0.2)                 |
| 100          | 5.0 (0.2)                | 5.2 (0.3)                 |
| 200          | 3.3 (0.1)                | 3.2 (0.2)                 |
| 400          | 2.8 (0.1)                | 2.7 (0.2)                 |
| 800          | 2.4 (0.1)                | 2.3 (0.2)                 |
| 1600         | 2.0 (0.1)                | 1.9 (0.1)                 |

5 Discussion/Conclusion

This work contributes to the development of new variants of Stability Selection. As a first step, we investigated a variation of Stability Selection where we choose smaller subsamples of the observations. We show that the error bounds of the preceding investigations [Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013] can be generalized to that case. These new bounds give insights into the effect of the subsample size on the selection performance and help practitioners to use a version of Stability Selection that suits their needs.

As a second step, we investigated the effect of randomising an arbitrary base procedure in the Stability Selection framework. This randomization consists in applying the base procedure for each subsample of the observations to subsets of the covariates only. We motivate this randomization using a simplified score model, in which we show that applying the base procedure to a too large set of covariates can decrease the probability of correctly recovering informative covariates. This makes the procedure more similar to random guessing, whereas restricting the search to a random subset increases the probability of correct recovery. We expect that this second step of extending the original method is particularly appealing for practitioners in high dimensional settings, where the number of covariates largely exceeds the number of observations.

Finally, we demonstrated the applicability of our extensions of Stability Selection by performing experiments on simulated and semi-realistic datasets. Our results suggest that our variations improve on Stability Selection whenever Stability Selection improves on the base method applied to the whole dataset.

It remains an open task to determine precise conditions under which a variable selection method can be improved by Stability Selection. Even though our analysis gave first insights on the dependence of the error probability on the size of subsamples of observations, a
rule for the optimal choice of the subsample size is left for further work. The same holds for the optimal choice of the size of subsets of covariates, even though in practice in very high dimensional problems this choice might be directed by computational constraints.

Acknowledgements We are extremely grateful to Nicolai Meinshausen and Peter Bühlmann for communicating to us the R-code used by Meinshausen and Bühlmann (2010) as well as for numerous discussions. We are indebted to Richard Samworth and Rajen Shah for numerous discussions and for hosting the first author during part of this work. We thank Maurilio Gutzeit for helping us with part of the numerical experiments.

A Proofs of theoretical results

A.1 Proofs of Section 3.1

To prove Theorem 1 and Corollary 2 we need some notation and two Lemmas. We define

\[ P^{\text{simult}}_{L, \ell_0}(d) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{d \in S^{\text{base}}(\ell, t)\} \geq \ell_0 \]

the ratio of repetitions out of \( T \) where covariate \( d \) has been selected in at least \( \ell_0 \) subsamples simultaneously.

Lemma 1 (Relation of \( P^{\text{simult}} \) and \( P^{\text{SFS}} \)) It holds for any \( d \in F \):

\[ \left( \frac{L-\ell_0+1}{L} \right) P^{\text{simult}}_{L, \ell_0}(d) + \frac{\ell_0-1}{L} \geq P^{\text{SFS}}_{L, 1}(d). \]

Proof. We have for all repetitions of drawings of subsamples \( t = 1, \ldots, T \):

\[ \frac{1}{L} \sum_{t=1}^{T} \mathbb{1}\{d \in S^{\text{base}}(\ell, t)\} \]

\[ \leq \left( \frac{\ell_0-1}{L} \right) \mathbb{1}\{t \in \{\ell_0 \leq t \leq L\}\} \leq \ell_0 - 1 \]

\[ + \mathbb{1}\{\sum_{t=1}^{L} \mathbb{1}\{d \in S^{\text{base}}(\ell, t)\} \geq \ell_0\}. \]

Averaging over the repetitions \( t = 1, \ldots, T \), we obtain

\[ P^{\text{SFS}}_{L, 1}(d) \leq \frac{\ell_0-1}{L} \mathbb{1}\{1 - P^{\text{simult}}_{L, \ell_0}(d)\} + \frac{\ell_0-1}{L} \mathbb{1}\{P^{\text{simult}}_{L, \ell_0}(d)\} \]

\[ = \left( \frac{L-\ell_0+1}{L} \right) P^{\text{simult}}_{L, \ell_0}(d) + \frac{\ell_0-1}{L}. \]

Lemma 2 (Exponential inequality for \( P^{\text{simult}} \)) The following inequality holds for any \( d \in F, \xi > 0, \) and \( \ell_0 \in \{1, \ldots, L\} \) such that \( p_0 := \frac{1}{L} \geq p_L(d) \):

\[ P\left[P^{\text{simult}}_{L, \ell_0}(d) \geq \xi\right] \leq \frac{1}{\xi} \exp\left(-LD(p_0, p_L(d))\right). \]

Proof. We have

\[ E\left[P^{\text{simult}}_{L, \ell_0}(d)\right] = \mathbb{E}\left[\sum_{t=1}^{L} \mathbb{1}\{d \in S^{\text{base}}(\ell, t)\} \geq \ell_0\right] \]

\[ = \mathbb{P}[\text{Bin}(L, p_L(d)) \geq \ell_0] \leq \exp\left(-LD(p_0, p_L(d))\right). \]

The first equality is valid because the \( L \) random subamples of observations are disjoint. Therefore, their joint distribution is the same as that of \( L \) independent samples of size \( \frac{L}{T} \); thus \( (S^{\text{base}}(\ell, t))_{1 \leq t \leq L} \) has the same distribution as \( L \) independent copies of the variable \( S^{\text{base}}_L \). The last inequality is the Chernoff binomial bound. Using Markov’s inequality we get (11).

Proof of Theorem 1. We relate \( P^{\text{SFS}} \) to \( P^{\text{simult}} \) and apply an exponential inequality on \( P^{\text{simult}} \). For any \( d \in A_{\theta, L} \), it holds by definition of \( A_{\theta, L} \) and the assumptions on \( p_0 \) that \( p_L(d) \leq \theta = p_0 \), hence it holds by Lemma 1 and Lemma 2 that

\[ P\left[P^{\text{SFS}}_{L, 1}(d) \geq \tau\right] \leq \frac{1}{\tau - p_0 + L^{-1}} \exp(-LD(p_0, p_L(d))) \]

where we have used \( \xi := \frac{L-\ell_0+1}{L} \). This result generalizes Shah and Samworth (2013) Lemma 5). Hence

\[ E\left[\left|S_{L, \tau}^{\text{SFS}} \cap A_{\theta, L}\right|\right] \]

\[ = \frac{1}{\left|A_{\theta, L}\right|} \sum_{d \in A_{\theta, L}} P\left[P^{\text{SFS}}_{L, 1}(d) \geq \tau\right] \]

\[ \leq \frac{1}{\tau - p_0 + L^{-1}} \left|A_{\theta, L}\right| \sum_{d \in A_{\theta, L}} \exp(-LD(p_0, p_L(d))). \]

Since \( x \rightarrow \exp(-LD(p_0, x)) \) is non-decreasing, we obtain the first part of the result by upper bounding for all \( d \in A_{\theta, L} \):

\[ \exp(-LD(p_0, p_L(d))) \leq \exp(-LD(p_0, \theta)). \]

For the second part, we use the upper bound

\[ \exp(-LD(p_0, p_L(d))) = \frac{\exp(-LD(p_0, p_L(d)))}{p_L(d)} \]

\[ \leq \frac{\exp(-LD(p_0, \theta))}{\theta} p_L(d), \]

since the function \( x \rightarrow \exp(-LD(p_0, x)) \) can be shown to be non-decreasing for \( x \leq p_0 - L^{-1} \). Finally, summing over \( d \in A_{\theta, L} \), observe

\[ \sum_{d \in A_{\theta, L}} p_L(d) = E\left[\sum_{d \in A_{\theta, L}} \mathbb{1}\{d \in S^{\text{base}}_L\}\right] = E\left[\left|A_{\theta, L} \cap S^{\text{base}}_L\right|\right], \]

leading to the desired conclusion. Equations (8) and (9) can be proved similarly.
Proof of Corollary 1. This follows the same argument as in [Shah and Samworth (2013)]. If the variable selection was completely at random, the marginal selection probability of any given covariate would be $\frac{1}{L}$, where we recall $q = \mathbb{E}[S^*_{max}]$ is the average number of covariates selected by the base method. As we assume that the selection probability of a signal covariate is better than random; it entails that for any $d \in N^C$, we must have $p_d(L) > \frac{1}{L}$. Conversely, as all noise covariates have the same probability to be selected by the base method, one has $p_d(L) < \frac{1}{L}$ for any $d \in N$. Therefore, with $\theta := \frac{2}{L}$, we must have $A_{\theta,L} = N$ and $A^0_{\theta,L} = N^C$. Inequality (2) therefore implies (6), wherein we have taken a minimum over the range of $\ell_0$ allowed in Theorem 1.

A.2 Proofs for Section 3.2

Proof of Theorem 2. We can first bound the error probability from above by omitting $Q_d$:

$$P\left[\hat{d}_D \in A_{D,\theta}\right] = P\left[\max_{d \in A_{D,\theta}} \hat{Q}_d > \max_{d \in A_{D,\theta}} \hat{Q}_d \right] \leq P\left[\max_{d \in A_{D,\theta}} (Q_d + \varepsilon_d) > \max_{d \in A_{D,\theta}} (Q_d + \varepsilon_d) \right] \leq P\left[\max_{d \in A_{D,\theta}} (\theta + \varepsilon_d) > \max_{d \in A_{\theta,\theta}} (\theta + \varepsilon_d) \right] = P\left[\arg\max_{d \in \{1, \ldots, D\}} \varepsilon_d \in A_{D,\theta}\right],$$

as $D \to \infty$. If $\eta = 0$, the conclusion is therefore established; in the remainder of the proof we hence assume $\eta > 0$. We defer to the end of the proof the case $\eta = 1$ and assume for now that $\eta \in (0, 1)$. Then $\frac{|A_{D,\theta}|}{D} \to \eta \in (0, 1)$ implies both $|A_{D,\theta}| \to \infty$ and $|A_{D,\theta}| \to \infty$, as well as $|A^0_{\theta,\theta}| \to \gamma := 1 - \eta$. We return to the error probability and bound it from below by using $Q_d \geq 0$ for $d \in A_{D,\theta}$ and $Q_d \leq M$ for $d \in A_{D,\theta}$:

$$P\left[\hat{d}_D \in A_{D,\theta}\right] \geq P\left[\max_{d \in A_{D,\theta}} \varepsilon_d > M + \max_{d \in A^0_{\theta,\theta}} \varepsilon_d \right].$$

(12)

Since the distribution of $\varepsilon_i$ belongs to MDA($\text{Fréchet}(\alpha)$), from classical results of extreme value theory [Embrechts (1997) Theorem 3.3.7] we know that there exists a slow varying function $L$ so that, if we denote $G(x) := x^{1/\alpha} L(x)$, then

$$\frac{\max_{d \in A_{D,\theta}} \varepsilon_d}{G(|A_{D,\theta}|)} \to \text{Fréchet}(\alpha)$$

and

$$\frac{\max_{d \in A^0_{\theta,\theta}} \varepsilon_d}{G(|A^0_{\theta,\theta}|)} \to \text{Fréchet}(\alpha),$$

in the sense of convergence in distribution, as $D \to \infty$. Following on (12):

$$P\left[\hat{d}_D \in A_{D,\theta}\right] \geq P\left[\max_{d \in A_{D,\theta}} \varepsilon_d \geq \frac{M}{G(|A_{D,\theta}|)} + \frac{G(|A^0_{\theta,\theta}|) \max_{d \in A^0_{\theta,\theta}} \varepsilon_d}{G(|A^0_{\theta,\theta}|)} \right].$$

As $L$ is slowly varying, we have $\lim_{x \to \infty} \frac{L(x)}{x^{1/\alpha}} \to 1$ uniformly for $x$ belonging to a bounded interval of the positive real axis [Embrechts (1997) Theorem A 3.2]. We deduce

$$\frac{G(A_{D,\theta})}{G(|A_{D,\theta}|)} \sim \frac{|A_{D,\theta}|}{|A_{D,\theta}|} \leq L\left(\frac{|A_{D,\theta}|}{|A_{D,\theta}|}\right) \to \gamma^{1/\alpha},$$

(15)

as $D \to \infty$. We apply Slutsky’s theorem [Embrechts (1997) Example A 2.7] to Equations (13) and (15) to obtain

$$\frac{G(A_{D,\theta})}{G(|A^0_{D,\theta}|)} \to \text{Fréchet}(\alpha, \gamma^{1/\alpha})$$

in distribution, where Fréchet($\alpha, c$) denotes the Fréchet($\alpha$) distribution rescaled by a factor $c > 0$.

Further, slow variation of $L$ implies that $L(x)$ is asymptotically negligible with respect to any power function, so that

$$\frac{M}{G(|A_{D,\theta}|)} \to 0, \quad \text{as } D \to \infty.$$ (16)

As the maxima in $\max_{d \in A_{D,\theta}} \varepsilon_d$ and $\max_{d \in A^0_{\theta,\theta}} \varepsilon_d$ are taken over disjoint sets of independent random variables, they are independent. Since they converge marginally in distribution, they also converge jointly and therefore their difference converges [Embrechts (1997) Theorem A 2.6]. Combining with (16) and using Slutsky’s theorem again, we conclude that

$$\frac{\max_{d \in A_{D,\theta}} \varepsilon_d}{G(|A_{D,\theta}|)} - \frac{M}{G(|A_{D,\theta}|)} \leq G(|A_{D,\theta}|) \to G(|A_{D,\theta}|) \to G(|A_{D,\theta}|)$$

converges in distribution to the difference of two independent Fréchet distributed random variables. This convergence implies the convergence of the c.d.f. for all continuity points [Embrechts (1997) Equation A.1]. As the limiting distribution is continuous, we finally obtain

$$\lim_{D \to \infty} \inf \mathbb{P}\left[\hat{d}_D \in A_{D,\theta}\right] \geq \mathbb{P}\left[|F - \gamma^{1/\alpha} F'| > 0\right],$$

where $F, F'$ are independent Fréchet($\alpha$) random variables. To identify the value of this lower bound, observe that it is also the limiting value of

$$P\left[\arg\max_{d \in \{1, \ldots, D\}} \varepsilon_d \in A_{D,\theta}\right] = \frac{|A_{D,\theta}|}{D}.$$

Indeed, it suffices to repeat the above argument, except for skipping inequality (12). Hence this limiting value is exactly equal to $\eta$.

Finally, for the case $\eta = 1$, observe that the above argument remains valid provided $|A_{D,\theta}| \to \infty$. Even if this is not the case (i.e. $|A_{D,\theta}|$ remains bounded), then the conclusion would be a fortiori true since we could replace $A_{D,\theta}$ by a slightly larger set of cardinality $\ln(D)$ (say), which can only decrease the lower bound while still obtaining the above limiting value.

Proof of Theorem 3. To show the convergence of the error probability we use similar arguments as in the proof of Theorem 2. From classical results of extreme value theory for independent normal random variables $\varepsilon_k$ [Embrechts (1997) Example 3.3.29] it holds that

$$\max_{k=1,\ldots,k} \varepsilon_k \Rightarrow \text{Gumbel} \quad \text{as } k \to \infty.$$
in distribution, where \( a_k := \frac{1}{\sqrt{2\ln k}} \) and \( b_k := \sqrt{2\ln k} - \frac{\ln(4+) + \ln(\ln k)}{2\sqrt{2\ln k}} \). Denote \( k_D := |A_{D',\theta}| \) and \( \ell_D := |A'_{D',\theta}| \).

From now on all limiting statements will be meant as \( D \to \infty \).

Assumption (ii) of the Theorem states that \( \lim \inf \frac{\ell_D}{k_D} := \eta > 0 \), in particular \( \ell_D \to \infty \). On the other hand, since \( A_{D',\theta} \subseteq A_{D',\theta} \), we have \( \lim \sup \frac{k_D}{\ell_D} \leq 1 - \eta \), as well as \( \lim \sup \frac{k_D}{\ell_D} \leq \frac{1}{\eta} =: \gamma \).

Since \( A_{D',\theta} \subseteq A_{D',\theta} \), denoting \( \Delta := \theta - \theta' > 0 \) we can bound the error probability from above as follows:

\[
\Pr \left[ \hat{d}_D \in A_{D',\theta} \right] \\
\leq \Pr \left[ \max_{d \in A_{D',\theta}} (Q_d + \epsilon_d) > \max_{d \in A_{D',\theta}} (Q_d + \epsilon_d) \right] \\
\leq \Pr \left[ \max_{d \in A_{D',\theta}} \epsilon_d > \max_{d \in A_{D',\theta}} \epsilon_d + \Delta \right] \\
\leq \Pr \left[ \frac{a_{k_D}}{a_{\ell_D}} \frac{\max_{d \in A_{D',\theta}} \epsilon_d - b_{k_D}}{a_{k_D}} - \frac{\max_{d \in A'_{D',\theta}} \epsilon_d - b_{\ell_D}}{a_{\ell_D}} > \frac{b_{k_D} - b_{\ell_D} + \Delta}{a_{\ell_D}} \right].
\]

We treat the different terms in the above upper bound. First, we have

\[
\lim \sup \frac{a_{k_D}}{a_{\ell_D}} = \lim \sup \sqrt{\frac{2\ln(k_D)}{2\ln(\ell_D)}} = \lim \sup \sqrt{\frac{\ln \left( \frac{k_D}{\ell_D} \right)}{\ln(\ell_D)}} \\
= \lim \sup \frac{\ln \left( \frac{k_D}{\ell_D} \right) + \ln(\ell_D)}{\ln(\ell_D)} \\
\leq 1.
\]

Noting that \( b_k = \sqrt{2\ln k} + o(1) \), we have

\[
\frac{b_{k_D} - b_{\ell_D} + \Delta}{a_{\ell_D}} = -2 \left( \ln \frac{k_D}{\ell_D} \right) \left( \frac{\sqrt{\ln(\ell_D)}}{\ln(k_D) + \sqrt{\ln(\ell_D)}} \right) + \Delta \sqrt{2\ln k_D} + o(\sqrt{\ln(\ell_D)}) \\
\geq \Delta \sqrt{2\ln k_D} + o(\sqrt{\ln(\ell_D)}).
\]

To check that the last inequality holds, note that if \( \ln \frac{k_D}{\ell_D} \leq 0 \) it is trivial, on the other hand we have \( \lim \sup \left( \frac{\ln \frac{k_D}{\ell_D}}{\ln \frac{k_D}{\ell_D}} \right) = \ln(1) = 0 \).

We deduce that for any \( B > 0 \), for \( D \) large enough, both \( \frac{a_{k_D}}{a_{\ell_D}} \leq 2 \) and \( \frac{b_{k_D} - b_{\ell_D} + \Delta}{a_{\ell_D}} > B \) hold, and we have

\[
\Pr \left[ \hat{d}_D \in A_{D',\theta} \right] \leq \Pr \left[ \frac{\max_{d \in A_{D',\theta}} \epsilon_d - b_{k_D}}{a_{k_D}} - \frac{\max_{d \in A'_{D',\theta}} \epsilon_d - b_{\ell_D}}{a_{\ell_D}} > B \right].
\]

By similar arguments as in the proof of Theorem 1 the latter upper bound converges to \( \Pr[2G - G' > B] \), where \( G, G' \) are two independent Gumbel random variables. As \( B \) is arbitrary we come to the announced conclusion.

References

The MASH project. http://www.mash-project.eu.

D. H. Alexander and K. Lange. Stability selection for genome-wide association. Genetic epidemiology, 35(7):722–728, 2011.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the 25th international conference on Machine learning, pages 33–40. ACM, 2008.

A. Beinrucker, U. Dogan, and G. Blanchard. Early stopping for mutual information based feature selection. In 2012 21st International Conference on Pattern Recognition (ICPR), pages 975–978, 2012a.

A. Beinrucker, U. Dogan, and G. Blanchard. A simple extension of stability feature selection. In Pattern Recognition, volume 7476 of Lecture Notes in Computer Science, pages 256–265. Springer, 2012b.

J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. JMLR, 3:1229–1243, 2002.

L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

P. Bühlmann and B. Yu. Analyzing bagging. Ann. Statist., 30(4):927–961, 2002.

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. J. Statist. Plann. Inference, 143(11):1835–1858, 2013.

T. M. Cover and J. A. Thomas. Elements of information theory. Wiley-Interscience, second edition, 2006.

B. Efron. Bootstrap methods: another look at the jackknife. Ann. Statist., 7(1):1–26, 1979.

P. Embrechts. Modelling extremal events: for insurance and finance, volume 33. Springer, 1997.

G. Escudero, L. Marquez, and G. Rigau. Boosting applied to word sense disambiguation. Machine Learning: ECML 2000, pages 129–141, 2000.

F. Fleuret. Fast binary feature selection with conditional mutual information. J. Mach. Learn. Res., 5:1531–1555, 2004.

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

I. Guyon. Feature extraction: foundations and applications, volume 207. Springer, 2006.

T. Hastie and B. Efron. LARS: Least Angle Regression, Lasso and Forward Stagewise, 2012. URL http://CRAN.R-project.org/package=lars R package version 1.1.

A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. Tigress: trustful inference of gene regulation using stability selection. BMC systems biology, 6(1):145, 2012.

Q. He and D.-Y. Lin. A variable selection method for genome-wide association studies. Bioinformatics, 27(1):1–8, 2011.

Z. He and W. Yu. Stable feature selection for biomarker discovery. Computational biology and chemistry, 34(4):215–225, 2010.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and L. J. Deng. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

M. R. Leadbetter, G. Lindgren, and H. Rootzén. Extremes and related properties of random sequences and processes. Springer Series in Statistics. Springer, 1983.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

N. Meinshausen and P. Bühlmann. Stability selection. Journal of the Royal Statistical Society, Series B, 72(4):417–
D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling.* Springer Series in Statistics. Springer, 1999.

W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: application to the cox regression model. *Statistics in medicine*, 11(16):2093–2109, 1992.

R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

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(1):55–80, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

S. Wang, B. Nan, S. Rosset, and J. Zhu. Random Lasso. *Ann. Appl. Stat.*, 5(1):468–485, 2011.