Stability Selection

Nicolai Meinshausen and Peter Bühlmann
University of Oxford and ETH Zürich

September 17, 2008

Abstract

Estimation of structure, such as in graphical modeling, cluster analysis or variable selection, is notoriously difficult, especially for high-dimensional data. We introduce the new method of stability selection. It is based on subsampling in combination with (high-dimensional) selection algorithms. As such, the method is extremely general and has a very wide range of applicability. Stability selection provides finite sample control for some error rates of false discoveries and hence a transparent principle to choose a proper amount of regularization for structure estimation or model selection. Maybe even more importantly, results are typically remarkably insensitive to the chosen amount of regularization. Another property of stability selection is the improvement over a prespecified selection method. We prove for randomized Lasso that stability selection will be model selection consistent even if the necessary conditions needed for consistency of the original Lasso method are violated. We demonstrate stability selection for variable selection, Gaussian graphical modeling and clustering, using real and simulated data.

1 Introduction

Estimation of discrete structure, such as graphs or clusters, or model selection is an age-old problem in statistics. It has enjoyed increased attention in recent years due to the massive growth of data across many scientific disciplines. These large datasets often make estimation of discrete structures or model selection imperative for improved understanding and interpretation. Most classical results do not cover the loosely defined case of high-dimensional data, and it is mainly in this area where we motivate the promising properties of our new stability selection.

In the context of regression, for example, an active area of research is to study the $p \gg n$ case, where the number of variables or covariates p outnumber the number of observations n; for an early overview see for example van de Geer and van Houwelingen (2004). In a similar spirit, graphical modeling with many more nodes than sample size has been the focus of recent research, and cluster analysis is another widely used technique to infer a discrete structure from observed data.

Challenges with estimation of discrete structures include computational aspects, since corresponding optimization problems are discrete, as well as determining the right amount of regularization, for example in an asymptotic sense for consistent structure estimation. Substantial progress has been made over the last years in developing computationally tractable methods which have provable statistical (asymptotic) properties, even for the high-dimensional setting with many more variables than samples. One interesting stream of research has focused on relaxations of some discrete optimization problems, for example by ℓ_1-penalty approaches (Donoho and Elad, 2003; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2006; Yuan and Lin, 2007) or greedy algorithms (Freund and Schapire, 1996; Tropp, 2004). The practical usefulness of such procedures has been demonstrated in various applications. However, the general issue of selecting a proper amount of regularization (for the procedures mentioned above and for many others) for getting a right-sized structure or model has largely remained a problem with unsatisfactory solutions.

We address the problem of proper regularization with a very generic subsampling approach (bootstrapping would behave similarly). We show that subsampling can be used to determine the amount of regularization such that a certain familywise error rate for type I multiple testing can be conservatively controlled for finite sample size. Particularly for complex, high-dimensional problems, a finite sample control is much more valuable than an asymptotic statement with the number of observations tending to infinity. Beyond the issue of choosing the amount of regularization, the subsampling approach yields a new structure estimation or model selection scheme which is stable and rather insensitive to the specification of the familywise error rate. For the more specialized case of high-dimensional linear models, we prove what we expect in greater generality: namely that subsampling in conjunction with ℓ_1-penalized estimation requires much weaker assumptions on the design matrix for asymptotically consistent variable selection than what is needed for the (non-subsampled) ℓ_1-penalty scheme. Furthermore, we show that additional improvements can be achieved by randomizing not only via subsampling but also in the selection process for the variables, bearing some resemblance to the successful tree-based Random Forest algorithm (Breiman, 2001). Subsampling (and bootstrapping) has been primarily used so far for asymptotic statistical inference in terms of standard errors, confidence intervals and statistical testing. Our work here is of a very different nature: the marriage of subsampling and high-dimensional selection algorithms yields finite sample familywise error control and markedly improved structure estimation or selection methods.

1.1 Preliminaries and examples

In general, let β be a p-dimensional vector, where β is sparse in the sense that $s < p$ components are 0. In other words, $\|\beta\|_0 = s < p$. Denote the set of non-zero values by $S = \{k : \beta_k \neq 0\}$ and the set of variables with vanishing coefficient by $N = \{k : \beta_k = 0\}$. The goal of structure estimation is to infer the set S from noisy observations.

As a first supervised example, consider data $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ with univariate response variable Y and p-dimensional covariates X. We typically assume some independence structure among the data. The vector β could be the coefficient vector in a linear model

$$Y = X\beta + \varepsilon, \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)$, X is the $n \times p$ design matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the random noise whose components are independent, identically distributed. Thus, inferring the set S from data is the well-studied variable selection problem in linear regression. A main stream of classical methods proceeds to solve this problem by penalizing the negative log-likelihood with the ℓ_0-norm $\|\beta\|_0$ which equals the number of non-zero components of β. The computational task to solve such an ℓ_0-norm penalized optimization problem becomes quickly infeasible if p is getting large, even when using efficient branch and bound techniques. Alternatively, one can relax the ℓ_0-norm by the ℓ_1-norm penalty. This leads to the Lasso estimator (Tibshirani, 1996; Chen et al., 2001) estimator:

$$\hat{\beta}^\lambda = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter and we typically assume that the covariates are on the same scale, i.e. $\|X_k\|_2 = \sum_{i=1}^n (X_k^{(i)})^2 = 1$. An attractive feature of Lasso is its computational feasibility for large p since the optimization problem in (2) is convex. Furthermore, the Lasso is able to do variable selection by shrinking certain estimated coefficients exactly to 0 and hence, we can estimate the set S of non-zero β coefficients by $\hat{S}^\lambda = \{k;\ \hat{\beta}_k^\lambda \neq 0\}$ which involves convex optimization only. Substantial understanding has been gained over the last few years about consistency of such Lasso model selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2006; Yuan and Lin, 2007), and we present the details in Section 3.1. Among the challenges are the issue of choosing a proper amount of regularization λ for consistent model selection and the fact that restrictive design conditions are needed for asymptotically recovering the true set S of relevant covariates.

A second example is on unsupervised Gaussian graphical modeling. The data is

$$X^{(1)}, \ldots, X^{(n)} \text{ i.i.d. } \sim \mathcal{N}_d(\mu, \Sigma). \tag{3}$$

The goal is to infer conditional dependencies among the d variables or components in $X = (X_1, \ldots, X_d)$. It is well-known that X_j and X_k are conditionally dependent given all other components $\{X_{(\ell)};\ \ell \neq j, k\}$ if and only if $\Sigma_{jk}^{-1} \neq 0$, and we then draw an edge between nodes j and k in a corresponding graph (Lauritzen, 1996). The structure estimation is thus on the index set $\mathcal{G} = \{(j,k);\ 1 \leq j < k \leq d\}$ which has cardinality $p = \binom{d}{2}$ (and of course, we can represent $\mathcal{G}$ as a $p \times 1$ vector) and the set of relevant conditional dependencies is $S = \{(j,k) \in \mathcal{G};\ \Sigma_{jk}^{-1} \neq 0\}$. Similarly to the problem of variable selection in regression, ℓ_0-norm methods are computationally very hard and become very quickly infeasible for moderate or large values of d. A relaxation with ℓ_1-type penalties has also proven to be useful in this context (Meinshausen and Bühlmann, 2006). A recent proposal is the graphical Lasso (Friedman et al., 2007):

$$\hat{\Theta}^\lambda = \text{argmin}_{\Theta \text{ nonneg.def.}} \{-\log(\det(\Theta)) + \text{tr}(S\Theta) + \lambda \sum_{j<k} |\Theta_{jk}|\}. \tag{4}$$

This amounts to an ℓ_1-penalized estimator of the Gaussian log-likelihood, partially maximized over the mean vector μ, when minimizing over all nonnegative definite symmetric matrices. The estimated graph structure is then $\hat{S}^\lambda = \{(j,k) \in \mathcal{G};\ \hat{\Theta}_{jk}^\lambda \neq 0\}$ which involves convex optimization only and is computationally feasible for large values of d.

A third example is on unsupervised clustering. The data is $X^{(1)}, \ldots, X^{(n)}$ i.i.d., where $X^{(i)}$ is a d-dimensional variable. When partitioning the samples into clusters, we can encode this information with an index set $\mathcal{C} = \{(i,j);\ 1 \le i < j \le n\}$ of cardinality $p = \binom{n}{2}$ and a corresponding $p \times 1$ β-vector whose entries are 1 if the corresponding sample indices belong to the same cluster and zero otherwise. Thus, the true clustering is given by $S = \{(i,j) \in \mathcal{C};\ \beta_{(i,j)} = 1\}$. For a fixed number of clusters, sparsity in the sense that $s = |S|$ is small implies that the clusters are balanced with about equally many members per cluster.

The structure of the paper is as follows. The generic stability selection approach, its familywise type I multiple testing error control and some representative examples from high-dimensional linear models, Gaussian graphical models and clustering are presented in Section 2. A detailed asymptotic analysis of Lasso and randomized Lasso for high-dimensional linear models is given in Section 3 and more numerical results are described in Section 4. After a discussion in Section 5, we collect all the technical proofs in the Appendix.

2 Stability selection

Stability selection is not a new model selection technique. Its aim is rather to enhance and improve existing methods. First, we give a general description of stability selection and we present specific examples and applications later.

For a generic structure estimation or model selection technique, we have a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines the amount of regularization. This tuning parameter could be the penalty parameter in ℓ_1-penalized regression, see (2), or in Gaussian graphical modeling, see (4); or it may be number of steps in forward variable selection or Orthogonal Matching Pursuit (Mallat and Zhang, 1993) or the number of iterations in Matching Pursuit (Mallat and Zhang, 1993) or Boosting (Freund and Schapire, 1996); a large number of steps of iterations would have an opposite meaning from a large penalty parameter, but this doesn't cause conceptual problems. For every value $\lambda \in \Lambda$, we obtain a structure estimate $\hat{S}^\lambda \subseteq \{1, \ldots, p\}$. It is then of interest to determine whether there exists an $\lambda \in \Lambda$ such that $\hat{S}^\lambda$ is identical to S with high probability and how to achieve that right amount of regularization.

2.1 Stability paths

We motivate the concept of stability paths in the following, first for regression. Stability paths are derived from the concept of regularization paths. A regularization path is given by the coefficient value of each variable over all regularization parameters: $\{\hat{\beta}_k^\lambda;\ \lambda \in \Lambda,\ k = 1, \ldots, p\}$. Stability paths (defined below) are, in contrast, the *probability* for each variable to be selected when randomly resampling from the data. For any given regularization parameter $\lambda \in \Lambda$, the selected set $\hat{S}^\lambda$ is implicitly a function of the samples $I = \{1, \ldots, n\}$. We write $\hat{S}^\lambda = \hat{S}^\lambda(I)$ where necessary to express this dependence.

Definition 1 (Selection probabilities) *Let I be a random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement. For every set $K \subseteq \{1, \ldots, p\}$, the probability of being in the selected*

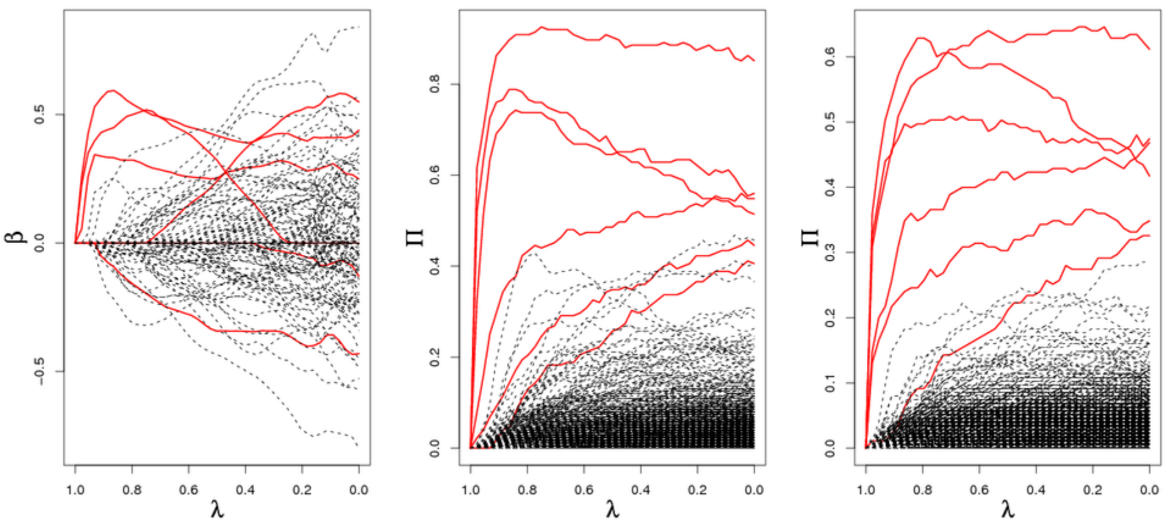

Figure 1: *Left: The Lasso path for the vitamin gene-expression dataset. The paths of the 6 non-permuted genes are plotted as solid, red lines, while the paths of the 4082 permuted genes are shown as broken, black lines. Selecting a model with all 6 unpermuted genes invariably means selecting a large number of irrelevant noise variables. Middle: the stability path of Lasso. The first 4 variables chosen with stability selection are truly non-permuted variables. Right: The stability path for the 'randomized Lasso' with weakness $\alpha = 0.2$, introduced later. Now all 6 non-permuted variables are chosen before any noise variable enters the model.*

set $\hat{S}^\lambda(I)$ is

$$\hat{\Pi}_K^\lambda \;=\; P^*\big(K \subseteq \hat{S}^\lambda(I)\big). \tag{5}$$

Remark 1 *The probability P^* in (5) is with respect to both the random subsampling (and other sources of randomness if $\hat{S}^\lambda$ is a randomized algorithm, see Section 3.1).*

Remark 2 *The sample size of $\lfloor n/2 \rfloor$ is chosen as it resembles most closely the bootstrap (Freedman, 1977; Bühlmann and Yu, 2002), while allowing computationally efficient implementation. But the approach can be extended to work with different subsample sizes.*

For every variable $k = 1, \ldots, p$, the stability path is given by the selection probabilities $\hat{\Pi}_k^\lambda$, $r \in \Lambda$. It is a complement to the usual path-plots that show the coefficients of all variables $k = 1, \ldots, p$ as a function of the regularization parameter. It can be seen in Figure 1 that this simple path plot is potentially very useful for improved model selection for high-dimensional data.

2.2 Example I: Variable selection in regression

We apply stability selection to the Lasso defined in (2). We work with a gene expression dataset for illustration which is kindly provided by DSM Nutritional Products (Switzerland). For $n = 115$ samples, there is a continuous response variable measuring the logarithm of riboflavin (vitamin B2) production rate of Bacillus Subtilis, and we have $p = 4088$ continuous covariates measuring the logarithm of gene expressions from essentially the whole genome of Bacillus Subtilis. Certain mutations of genes are thought to lead to higher vitamin concentrations and the challenge is to identify those relevant genes via a linear regression analysis. That is, we consider a linear model as in (1) and want to infer the set $S = \{k;\ \beta_k \neq 0\}$.

To see how Lasso and the related stability path cope with noise variables, we randomly permute all but 6 of the 4088 gene expression across the samples, using the same permutation to keep the dependence structure between the permuted gene expressions intact. The Lasso path $\{\hat{\beta}^\lambda;\ \lambda \in \Lambda\}$ is shown in the left panel of Figure 1, as a function of the regularization parameter λ (rescaled so that $\lambda = 1$ is the minimal λ-value for which the null model is selected and $\lambda = 0$ amounts to the Basis Pursuit solution). Three of the "relevant" (unpermuted) genes stand out, but all remaining three variables are hidden within the paths of noise (permuted) genes. The middle panel of Figure 1 shows the stability path. At least four relevant variables stand out much clearer now than they did in the regularization path plot. The right panel shows the stability plot for randomized Lasso which will be introduced in Section 3.1: now all 6 unpermuted variables stand above the permuted variables and the separation between (potentially) relevant variables and irrelevant variables is even better.

Choosing the right regularization parameter is very difficult for the original path. The prediction optimal and cross-validated choice include too many variables, as shown in (Meinshausen and Bühlmann, 2006; Leng et al., 2006), and the same effect can be observed in this example, where 14 permuted variables are included in the model chosen by cross-validation. We will discuss model selection for stability paths below. Figure 1 motivates that choosing the right regularization parameter is much less critical for the stability path and that we have a better chance to select truly relevant variables.

2.3 Stability selection

In a traditional setting, model selection would amount to choosing one element of the set of models

$$\{\hat{S}^\lambda;\ \lambda \in \Lambda\}, \tag{6}$$

where Λ is again the set of considered regularization parameters, which can be either continuous or discrete. There are typically two problems: first, the correct model S might not be a member of (6). Second, even if it is a member, it is typically very hard for high-dimensional data to determine the right amount of regularization λ to select exactly S, or to select at least a close approximation.

With stability selection, we do not simply select one model in the list (6). Instead the data are perturbed (for example by subsampling) many times and we choose all structures or variables that occur in a large fraction of the resulting selection sets.

Definition 2 (Stable variables) *For a cutoff π_{thr} with $0 < \pi_{thr} < 1$ and a set of regularization parameters Λ, the set of stable variables is defined as*

$$\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda \geq \pi_{thr}\}. \tag{7}$$

We keep variables with a high selection probability and disregard those with low selection probabilities. The exact cutoff π_{thr} with $0 < \pi_{thr} < 1$ is a tuning parameter but the results vary surprisingly little for sensible choices in a range of the cutoff. Neither do results depend strongly on the choice of regularization λ or the regularization region Λ. See Figure 1 for an example. We present some guidance on how to choose the cutoff parameter and the regularization region Λ below.

2.4 Choice of regularization and error control

A natural goal in recovery of the set S is to include as few variables of the set N of noise variables as possible. The choice of the regularization parameter is hence crucial. An advantage of our stability selection is that the choice of the initial set of regularization parameters Λ has typically not a very strong influence on the results, as long as Λ is varied with reason. Another advantage, which we focus on below, is the ability to choose this set of regularization parameters in a way that guarantees, under stronger assumptions, a certain bound on the expected number of false selections.

Definition 3 (Additional notation) *Let $\hat{S}^\Lambda = \cup_{\lambda\in\Lambda}\hat{S}^\lambda$ be the set of selected structures or variables if varying the regularization λ in the set Λ. Let q_Λ be the average number of selected variables, $q_\Lambda = E^*(|\hat{S}^\Lambda(I)|)$, where the expectation E^* is with respect to random subsampling. Define V to be the number of falsely selected variables with stability selection,*

$$V = \big|N \cap \hat{S}^{stable}\big|.$$

In general, it is very hard to control V, as the distribution of the underlying estimator $\hat{\beta}$ depends on many unknown quantities. Exact control is only possible under some simplifying assumptions.

Theorem 1 (Error control) *Assume that the distribution of $\{1_{\{k\in\hat{S}^\lambda\}}, k \in N\}$ is exchangeable for all $\lambda \in \Lambda$. Also, assume that the original procedure is not worse than random guessing, i.e. for any $\lambda \in \Lambda$,*

$$\frac{E(|S \cap \hat{S}^\lambda|)}{E(|N \cap \hat{S}^\lambda|)} \geq \frac{|S|}{|N|}. \tag{8}$$

The number V of falsely selected variables is then bounded by

$$E(V) \quad \leq \quad \frac{1}{2\pi_{thr} - 1}\,\frac{q_\Lambda^2}{p}, \tag{9}$$

where q_Λ is described in Definition 3.

The involved exchangeability assumption is perhaps stronger than one would wish, but there does not seem to be a way of getting error control in the same generality without making similar assumptions. For regression in (1), the exchangeability assumption is fulfilled if the design is random and the distribution of $\{X_k, k \in N\}$ is exchangeable. Independence of all variables in N is a special case. More generally, the variables could have a joint normal distribution with

$\text{Cov}(X_k, X_l) = \rho$ for all $k, l \in N$ with $k \neq l$ and $0 < \rho < 1$. For real data, we have no guarantee that the assumption is fulfilled but the numerical examples in Section 4 show that the bound holds up very well.

Note also that the assumption of exchangeability is only needed to prove Theorem 1. All other benefits of stability selection shown in this paper do not rely on this assumption. Besides exchangeability, we needed another, quite harmless, assumption, namely that the original procedure is not worse than random guessing. One would certainly hope that this assumption is fulfilled. If it is not, the results below are still valid with slightly weaker constants. The assumption seems so weak, however, that we do not pursue this further.

The threshold value π_{thr} is a tuning parameter whose influence is very small. For sensible values in the range of, say, $\pi_{thr} \in (0.6, 0.9)$, results tend to be very similar. Once the threshold is chosen at some default value, the regularization region Λ is determined by the desired error control. Specifically, for a default cutoff value $\pi_{thr} = 0.9$, choosing the regularization parameters Λ such that say $q_\Lambda = \sqrt{0.8\,p}$ will control $E(V) \leq 1$; or choosing Λ such that $q_\Lambda = \sqrt{0.8\,\alpha\,p}$ controls the familywise error rate (FWER) at level α, i.e. $P(V > 0) \leq \alpha$. Of course, we can proceed the other way round by fixing the regularization region Λ and choosing π_{thr} such that $E(V)$ is controlled at the desired level.

Without stability selection, the regularization parameter λ invariably has to depend on the unknown noise level of the observations. The advantage of stability selection is that (a) exact error control is possible, and (b) the method works fine even though the noise level is unknown. This is a real advantage in high-dimensional problems with $p \gg n$, as it is very hard to estimate the noise level in these settings.

Pointwise Control. For some applications, evaluation of subsampling replicates of $\hat{S}^\lambda$ are already computationally very demanding for a single value of λ. If this single value λ is chosen such that some overfitting occurs and the set $\hat{S}^\lambda$ is rather too large, in the sense that it contains S with high probability, the same approach as above can be used and is in our experience very successful as results typically do not depend strongly on the utilized regularization λ. See the example below for graphical modelling. Setting $\Lambda = \{\lambda\}$, one can immediately transfer all results above to the case of what we call here pointwise control. For methods which select structures or incrementally, i.e. for which $\hat{S}^\lambda \subseteq \hat{S}^{\lambda'}$ for all $\lambda \geq \lambda'$, pointwise control and control with $\Lambda = [\lambda, \infty)$ are equivalent since $\hat{\Pi}_k^\lambda$ is then monotonically increasing with decreasing λ for all $k = 1, \ldots, p$.

2.5 Example II: Graphical modeling

Stability selection is also promising for graphical modelling. Here we focus on Gaussian graphical models as described in Section 1.1 around formula (3) and (4).

The pattern of non-zero entries in the inverse covariance matrix Σ^{-1} corresponds to the edges between the corresponding pairs of variables in the associated graph and is equivalent to a non-zero partial correlation (or conditional dependence) between such pairs of variables (Lauritzen, 1996).

There has been interest recently in using ℓ_1-penalties for model selection in Gaussian Graphical models due to their computational efficiency for moderate and large graphs (Meinshausen and

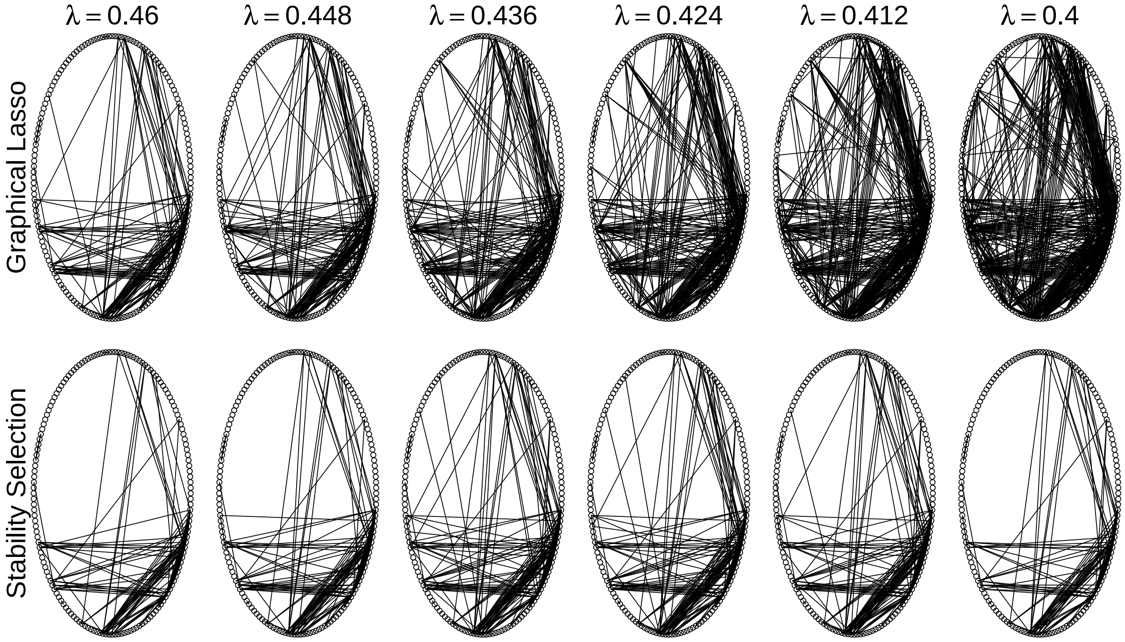

Figure 2: *Vitamin gene-expresssion dataset. The regularization path of graphical lasso (top row) and the corresponding point-wise stability selected models (bottom row).*

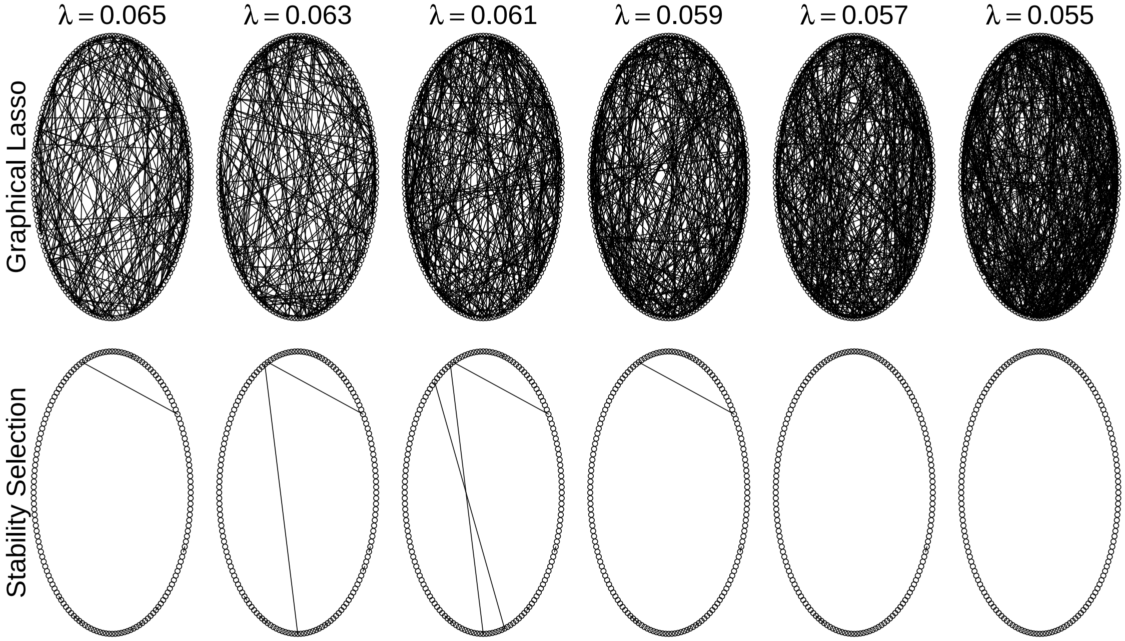

Figure 3: *The same plot as in Figure 2 but with the variables (expression values of each gene) permuted independently. The empty graph is the true model. With stability selection, only a few errors are made, as guaranteed by the made error control.*

Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2007; Banerjee and El Ghaoui, 2008; Bickel and Levina, 2008; Rothman et al., 2008). Here we work with the graphical Lasso of (Friedman et al., 2007), as applied to the data from 160 randomly selected genes from the vitamin gene-expression dataset (without the response variable) introduced in Section 2.2. We want to infer the set of non-zero entries in the inverse covariance matrix Σ^{-1}. Part of the resulting regularization path of the graphical Lasso showing graphs for various values of the regularization parameter λ, i.e. $\{\hat{S}^\lambda;\ \lambda \in \Lambda\}$ where $\hat{S}^\lambda = \{(j,k);\ (\hat{\Sigma}^{-1})^\lambda_{jk} \neq 0\}$, are shown in the first row of Figure 2. For reasons of display, variables (genes) are ordered first using hierarchical clustering and are symbolized by nodes arranged in a circle. Stability selection is shown in the bottom row of Figure 2. We pursue a pointwise control approach. For each value of λ, we select the threshold π_{thr} so as to guarantee $E(V) \leq 30$, that is we expect fewer than 30 wrong edges among the 12720 possible edges in the graph. The set $\hat{S}^{stable}$ varies remarkably little for the majority of the path and the choice of q (which is implied by λ) does not seem to be critical, as already observed for variable selection in regression.

Next, we permute the variables (expression values) randomly, using a different permutation for each variable (gene). The true graph is now the empty graph. As can be seen from Figure 3, stability selection selects now just very few edges or none at all (as it should). The top row shows the corresponding graphs estimated with the graphical Lasso which yields a much poorer selection of edges.

2.6 Example III: Clustering

We show now stability selection for clustering which has been briefly introduced in Section 1.1. There are potentially many applications and we cannot exhaustively discuss all possibilities, but the main point is to show that stability selection can work nicely with a wide variety of procedures. Here, we take K-means clustering with the Hartigan-Wong algorithm (Hartigan, 1975) as the underlying clustering scheme. In Figure 4, we show the outcome for a 2-dimensional example as a function of the number of cluster centers. The underlying 2-dimensional data $X^{(i)} \in \mathbb{R}^2$ for $i = 1, \ldots, 102$, were generated independently for $\sigma = 1/4$ as follows:

$$\begin{aligned} X^{(i)} &\sim \mathcal{N}_2((1,1), \sigma^2 \mathbf{1}), \qquad i = 1, \ldots, 50, \\ X^{(i)} &\sim \mathcal{N}_2((-1,-1), \sigma^2 \mathbf{1}), \qquad i = 51, \ldots, 100, \end{aligned}$$

and $X^{(101)} = (-1.5, 1.5)$ and $X^{(102)} = c(1.5, -1.5)$. The first 100 data samples form thus two obvious clusters, and the two remaining data points are outliers in the sense that they do not belong to either of the two obvious clusters. As can be seen in Figure 4, the results of K-means depend heavily on how many cluster centers are chosen. The number of cluster centers is thus an important regularization parameter. Even if this parameter is picked in an optimal way, we cannot recover the true underlying structure of two obvious clusters and two outliers (in the sense described above). The two outliers are assigned to parts of the two main clusters if using 3 or 4 cluster centers. The two main clusters are broken apart if choosing 3 or more cluster centers and there is no entirely satisfactory solution.

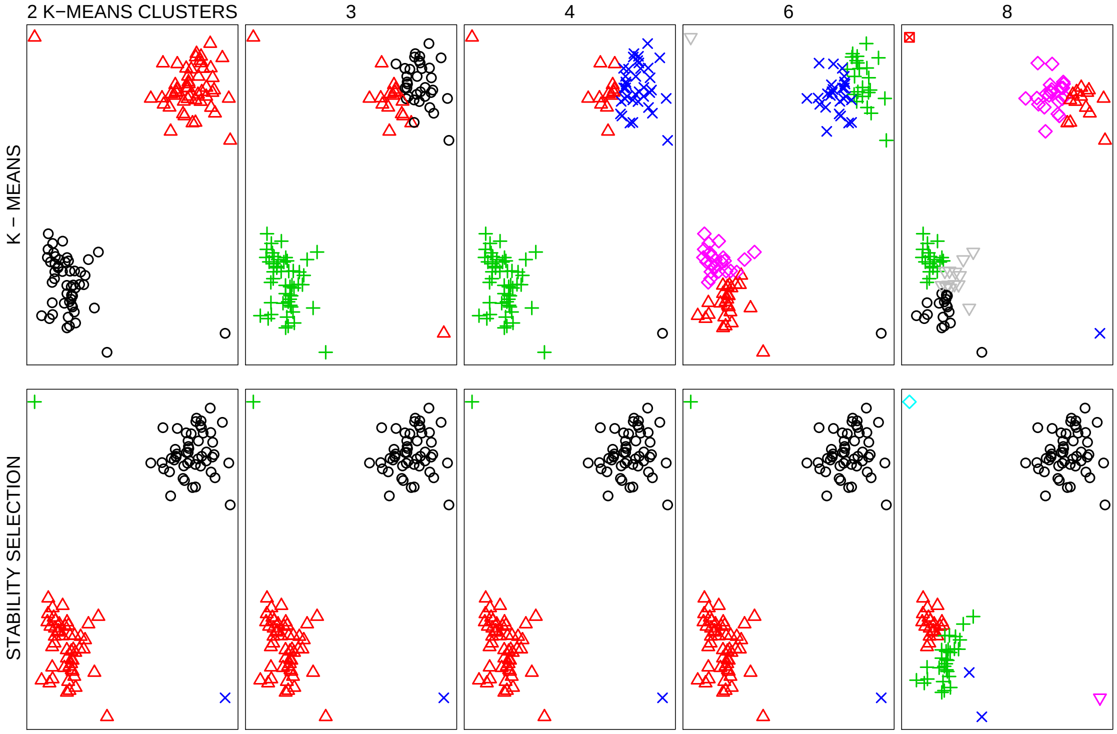

Figure 4: *Stability selection applied to clustering. Top row: K-means results if using 2-8 cluster centers. Points with identical class after K-means clustering are shown in the same color and plotting symbol. For more than 2 centers, the two obvious clusters are split in various ways. The outlying points are often associated with elements of the two clusters. Bottom row: the same result if using stability selection. The outcome is identical for 3-7 cluster centers, putting the two obvious clusters into one cluster each and assigning the two outlying points to yet two other separate clusters.*

For stability selection, we employ a slightly different perturbation as previously described to demonstrate that the idea is more widely applicable. We repeat the clustering repeatedly, each time using only a subset $I \subseteq \{1, \ldots, n\}$ of all samples with $|I| = \lfloor n/2 \rfloor$ of the original n samples. Let k be the number of cluster centers chosen for K-means and let $C_1, \ldots, C_k$ be the partitioning of $\{1, \ldots, n\}$ into the k clusters by K-means. The quantity $\hat{\Pi}^k_{i,j}$ is then defined to be the probability of having samples i and j in the same K-means cluster, *conditional* of having both i and j in the set I,

$$\hat{\Pi}^k_{i,j} \; = \; P^*\big(\text{there exists } 1 \leq k' \leq k : \{i,j\} \in C_{k'} \big| \{i,j\} \in I\big).$$

Let the stability clusters $C_1, \ldots, C_m$ be the m connectivity components of the graph given by putting edges between nodes $\{i,j\}$ if and only if $\hat{\Pi}^k_{i,j} \geq \pi_{thr}$ with $0 < \pi_{thr} < 1$ here chosen as 0.75. Note that the number m of connectivity components can both be larger or smaller than the number k of clusters used in the underlying K-means clustering. Indeed, in Figure 4, the connectivity components are always the desired two clusters for all values $k = 3, \ldots, 7$, while the two outliers comprise an individual cluster each. For $k = 8$, one of the two clusters is broken apart. Stability selection has the same effect for clustering as for regression and graphical model selection. The choice of the regularization parameter (here k) matters much less than in the original procedure. Also, insight can be gained into the underlying structure (i.e. having two main clusters and two outliers) that is not directly accessible in the original procedure. We point out that this approach is very different from Lange et al. (2004) whose aim is to find an optimal number of clusters via subsampling and stability, whereas we obtain with our stability selection a new clustering rule.

3 Consistent variable selection

Stability selection is a general technique, applicable to a wide range of applications, some of which we have discussed above. Here, we want to discuss advantages and properties of stability selection for the specific application of variable selection in regression with high-dimensional data which is a well-studied topic nowadays (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2006). We consider a linear model as in (1) with Gaussian noise,

$$Y = X\beta + \varepsilon, \tag{10}$$

with fixed $n \times p$ design matrix X and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$. The predictor variables are normalized with $\|X_k\|_2 = (\sum_{i=1}^n (X_k^{(i)})^2)^{1/2} = 1$ for all $k \in \{1, \ldots, p\}$.

Stability selection is attractive for two reasons. First, the choice of a proper regularization parameter for variable selection is crucial and notoriously difficult, especially because the noise level is unknown. With stability selection, results are much less sensitive to the choice of the regularization. Second, we will show that stability selection makes variable selection consistent in settings where the original methods fail.

We give general conditions under which consistent variable selection is achieved with stability selection. Consistent variable selection is understood to be equivalent to

$$P(\hat{S}^{stable} = S) \rightarrow 1 \qquad n \rightarrow \infty. \tag{11}$$

It is clearly of interest to know under which conditions consistent variable selection can be achieved. In the high-dimensional context, this places a restriction on the growth of the number p of variables and sparsity $|S|$, typically of the form $|S| \cdot \log p = o(n)$ (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2006). While this assumption is often realistic, there are stronger assumptions on the design matrix that need to be satisfied for consistent variable selection. For Lasso, it amounts to the 'neighborhood stability' condition (Meinshausen and Bühlmann, 2006) which is equivalent to the 'irrepresentable condition' (Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007). For Orthogonal Matching Pursuit (which is essentially forward variable selection), the so-called 'exact recovery criterion' (Tropp, 2004) is sufficient and necessary for consistent variable selection.

Here, we show that these conditions can be circumvented more directly by using stability selection, also giving guidance on the proper amount of regularization. Due to the restricted length of the paper, we will only discuss the case of Lasso whereas the analysis of Orthogonal Matching Pursuit is indicated by Remark 3 below.

An interesting aspect is that stability selection with the original procedures alone yields often very large improvements already. Moreover, when adding some extra sort of randomness in the spirit of Random Forests Breiman (2001) weakens considerably the conditions needed for consistent variables selection as discussed next.

3.1 Lasso and randomized Lasso

The Lasso (Tibshirani, 1996; Chen et al., 2001) estimator is given in (2). A natural question to ask is whether the pattern of non-zero estimated coefficients $\hat{S}^\lambda = \{k;\ \hat{\beta}_k^\lambda \neq 0\}$ closely approximates the set S of true non-zero regression coefficients. It turns out that the design needs to satisfy the so-called 'neighborhood stability' condition (Meinshausen and Bühlmann, 2006) which is equivalent to the 'irrepresentable condition' (Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007):

$$\max_{k \in S^c} |\mathrm{sign}(\beta_S)^T (X_S^T X_S)^{-1} X_S^T X_k| < 1. \qquad (12)$$

The condition in (12) is sufficient and (almost) necessary (the word "almost" refers to the fact that a necessary relation is with "$\leq$" instead of "$<$"). If this condition is violated, all one can hope for is recovery of the regression vector β in an ℓ_2-sense of convergence by achieving $\|\hat{\beta}^\lambda - \beta\|_2 \to_p 0$ for $n \to \infty$. The main assumption here are bounds on the sparse eigenvalues as discussed below. This type of ℓ_2-convergence can be used to achieve consistent variable selection in a two-stage procedure by thresholding or, preferably, the adaptive Lasso (Zou, 2006). The disadvantage of such a two-step procedure is the need to choose several tuning parameters without proper guidance on how these parameters can be chosen in practice. We propose the randomized Lasso as an alternative. Despite its simplicity, it is consistent for variable selection even though the 'irrepresentable condition' in (12) is violated.

Randomized Lasso is a new generalization of the Lasso. While the Lasso penalizes the absolute value $|\beta_k|$ of every component with a penalty term proportional to λ, the randomized Lasso changes the penalty λ to a randomly chosen value in the range $[\alpha\lambda, \lambda]$.

Randomized Lasso with weakness $\alpha \in (0, 1]$:

Let W_k be i.i.d. random variables in $[\alpha, 1]$ for $k = 1, \ldots, p$. The randomized Lasso estimator $\hat{\beta}^{\lambda,W}$ for regularization parameter $\lambda \in \mathbb{R}$ is then

$$\hat{\beta}^{\lambda,W} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \ \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k}. \tag{13}$$

A proposal for the distribution of the weights W_k is described below, just before Theorem 2. The word "weakness" is borrowed from the terminology of weak greedy algorithms (Temlyakov, 2000) which are loosely related to our randomized Lasso. Implementation of (13) is straightforward by appropriate re-scaling of the predictor variables (with scale factor W_k for the k-th variable). Using these re-scaled variables, the standard Lasso is solved, using for example the LARS algorithm (Efron et al., 2004) or fast coordinate wise approaches (Friedman et al., 2007; Meier et al., 2008). The perturbation of the penalty weights is reminiscent of the re-weighting in the adaptive Lasso (Zou, 2006). Here, however, the re-weighting is not based on any previous estimate, but is simply chosen at random! As such, it is very simple to implement. However, it seems non-sensical at first sight since one can surely not expect any improvement from such a random perturbation. If applied only with one random perturbation, randomized Lasso is not very useful. However, applying randomized Lasso many times and looking for variables that are chosen often will turn out to be a very powerful procedure.

Consistency for randomized Lasso with stability selection For stability selection with randomized Lasso, we can do without the irrepresentable condition (12) but need only a condition on the sparse eigenvalues of the design (Bickel et al., 2007; Candes and Tao, 2007; Meinshausen and Yu, 2008), also called the sparse Riesz condition in Zhang and Huang (2008).

Definition 4 (Sparse Eigenvalues) *For any $K \subseteq \{1, \ldots, p\}$, let X_K be the restriction of X to columns in K. The minimal sparse eigenvalue $\phi_{\min}$ is then defined for $k \leq p$ as*

$$\phi_{\min}(k) = \inf_{a \in \mathbb{R}^{\lceil k \rceil}, K \subseteq \{1,\ldots,p\} : |K| \leq \lceil k \rceil} \frac{\|X_K a\|_2}{\|a\|_2} \tag{14}$$

We have to constrain sparse eigenvalues to succeed.

Assumption 1 (Sparse eigenvalues) *There exists some $C > 1$ and some $\kappa \geq 10$ such that*

$$\frac{\phi_{\max}(Cs^2)}{\phi_{\min}^{3/2}(Cs^2)} < \sqrt{C}/\kappa, \qquad s = |S|. \tag{15}$$

This assumption (15) is related to the sparse Riesz condition in Zhang and Huang (2008). The equivalent condition there requires the existence of some $\overline{C} > 0$ such that

$$\frac{\phi_{\max}((2 + 4\overline{C})s + 1)}{\phi_{\min}((2 + 4\overline{C})s + 1)} < \overline{C}, \tag{16}$$

compare with Remark 2 in Zhang and Huang (2008). This assumption essentially requires that maximal and minimal eigenvalues, for a selection of order s variables, are bounded away from 0 and ∞ respectively. In comparison, our assumption is significantly stronger than (16), but at the same time much weaker than the standard assumption of the 'irrepresentable condition' typically necessary to get results comparable to ours.

Assumption 2 (Minimal non-zero value) *The minimal non-zero entries of β satisfy, with the same constant C from Assumption 1,*

$$\min_{k:\beta_k \neq 0} |\beta_k| \;\geq\; 0.3\,(Cs)^{3/2}\lambda_{\min},$$

where $s = |S|$ and $\lambda_{\min} = 2\sigma(\sqrt{C}s + 1)\sqrt{\log(p)/n}$.

Assuming that the magnitude of relevant variables vanishes slower than $1/\sqrt{n}$ is standard. Variables whose coefficient vanish faster than $1/\sqrt{n}$ are clearly not detectable. The involved constants in Assumptions 1 and 2 are not very tight. These constants and the reliance on the number s of relevant variables could possibly be improved upon with more elaborate analysis. It is not the aim of this paper to come up with the weakest possible assumptions. We rather aim to show in this section that stability selection offers a new selection scheme which requires much weaker assumptions than the 'irrepresentable condition' in (12) for consistent variable selection with the original Lasso estimator.

We will consider the Lasso solutions for all regularization parameters λ in the region Λ, where $\Lambda = \{\lambda : \lambda \geq \lambda_{\min}\}$, with the minimal value given Assumption 2; we note that this value is bounded from below by the choice of $\lambda_{n,p}$ with $c_0 = a_n = 1$ in Zhang and Huang (2008).

We have not specified the exact form of perturbations we will be using for the randomized Lasso in (13). For the following, we consider the randomized Lasso of (13), where the weights W_k are sampled independently as $W_k = \alpha$ with probability $p_w \in (0,1)$ and $W_k = 1$ otherwise. Other perturbations are certainly possible and work often just as well in practice.

Theorem 2 *Let the weakness α be given by $\alpha^2 = \nu\phi_{\min}(m)/m$, for any $\nu \in (7/\kappa, 1/\sqrt{2})$, and $m = Cs^2$. If Assumptions 1 and 2 are satisfied and $p > 10$ and $s \geq 7$, there exists some $\delta = \delta_s \in (0,1)$ such that for all $\pi_{thr} \geq 1 - \delta$, stability selection with the randomized Lasso satisfies,*

$$P(\hat{S}^{stable} = S) \;\geq\; 1 - 5/p. \tag{17}$$

For $p \to \infty$, we have hence indeed asymptotically consistent variable selection in the sense of (11) even if the irrepresentable condition (12) is violated.

There is an inherent tradeoff when choosing the weakness α. A negative consequence of a low α is that the design can get closer to singularity and can thus lead to unfavourable conditioning of the weighted design matrix. On the other hand, a low value of α makes it less likely that irrelevant variables are selected. This is a surprising result but rests on the fact that irrelevant variables can only be chosen if the corresponding irrepresentable condition (12) is violated. By randomly perturbing the weights with a low α, this condition is bound to fail sometimes, lowering the selection probabilities for such variables. A low value of α will thus help stability selection to avoid selecting noise variables with a violated irrepresentable condition (12). In practice, choosing α in the range of $(0.2, 0.8)$ gives very useful results.

Remark 3 *In the spirit of Theorem 2, we have also a proof that stability selection for a randomized version of Orthogonal Matching Pursuit is asymptotically consistent for variable selection in linear models, assuming a much weaker condition than the necessary and sufficient exact recovery condition (Tropp, 2004) for (non-randomized) Orthogonal Matching Pursuit. This indicates that stability selection has a more general potential for improved structure estimation, beyond the case for the Lasso presented in Theorem 2.*

Relation to other work In related and very interesting work, Bach (2008) has proposed 'Bolasso' (for bootstrapped enhanced Lasso) and shown that using a finite number of subsamples of the original Lasso procedure and applying basically stability selection with $\pi_{thr} = 1$ yields consistent variables selection under the condition that the penalty parameter λ vanishes faster than typically assumed, at rate $n^{-1/2}$, and that the model dimension p is fixed. While the latter condition could possibly be technical only, the first distinguishes it from our results. Applying stability selection to randomized Lasso, no false variable is selected for all sufficiently large values of λ. In other words, if λ is chosen 'too large' with randomized Lasso, only truly relevant variable are chosen (though a few might be missed). If λ is chosen too large with Bolasso, noise variables might be picked up. Figure 5 is a good illustration. Picking the regularization in the left plot (without extra randomness) to select the correct model is much harder than in the right plot, where extra randomness is added. The same distinction can be made with two-stage procedures like adaptive Lasso (Zou, 2006) or hard-thresholding (Meinshausen and Yu, 2008; Candes and Tao, 2007), where variables are thresholded. Picking λ too large (and λ is notoriously difficult to pick), false variables will invariably enter the model. In contrast, stability selection with randomized Lasso is not picking wrong variables if λ is chosen too large.

3.2 Example

We illustrate the results on randomized Lasso with a small simulation example: $p = n = 200$ and the predictor variables are sampled from a $\mathcal{N}(0, \Sigma)$ distribution, where Σ is the identity matrix, except for the entries $\Sigma_{13} = \Sigma_{23} = \rho$ and their symmetrical counterparts. We use a regression vector $\beta = (1, 1, 0, 0, \ldots, 0)$. The response Y is obtained from the linear model $Y = X\beta + \varepsilon$ in (1), where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, 1/4)$. For $\rho > 0.5$, the irrepresentable condition in (12) is violated and Lasso is not able to correctly identify the first two variables as the truly important ones, as it always includes the third variable superfluously as well. Using the randomized version for Lasso, the two relevant variables are still chosen with probability close to 1, while the irrelevant third variable is only chosen with much lower probability; the corresponding probabilities are shown for randomized Lasso in Figure 5. This allows to separate relevant and irrelevant variables. And indeed, the randomized Lasso is consistent under stability selection.

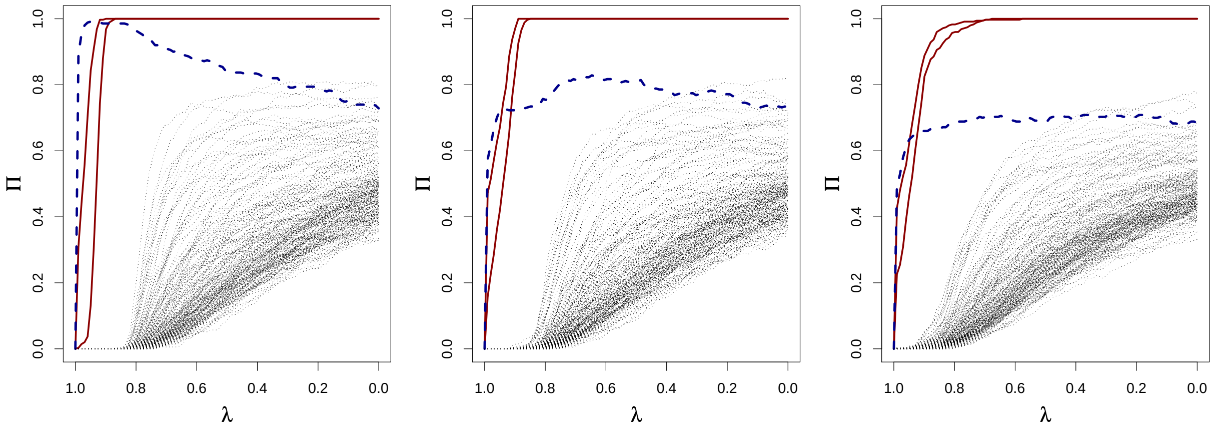

Figure 5: *The stability paths for randomized Lasso with weakness parameters* $\alpha = 1$ *(left panel identical to the original Lasso) and* $\alpha = 0.5$ *(middle) and* $\alpha = 0.2$ *(right). Red solid lines are the coefficients of the first two (relevant variables). The blue broken line is the coefficient of the third (irrelevant) variable and the dotted lines are the coefficients from all other (irrelevant) variables. Introducing the randomized version helps avoid choosing the third (irrelevant) predictor variable.*

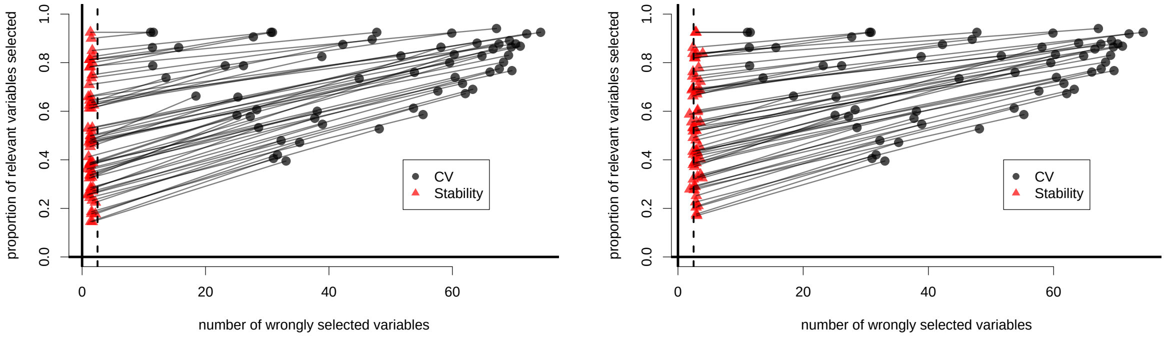

Figure 6: *Comparison of stability selection with cross-validation for motif regression. For each simulation setting, the cross-validated solution (for standard Lasso) is indicated by a dot and the corresponding stability selection (for randomized Lasso,* $\alpha = 0.5$ *on the left and* $\alpha = 1$ *on the right) by a red triangle, showing the average proportion of correctly identified relevant variables versus the average number of falsely selected variables. The cross-validated solution and the stability selection solution of a single setting are joined by a line. The broken vertical line indicates the value at which the number of wrongly selected variables is controlled, namely* $E(V) \leq 2.5$*. Looking at stability selection, the proportion of correctly identified relevant variables is very close to the CV-solution, while the number of falsely selected variables is reduced dramatically.*

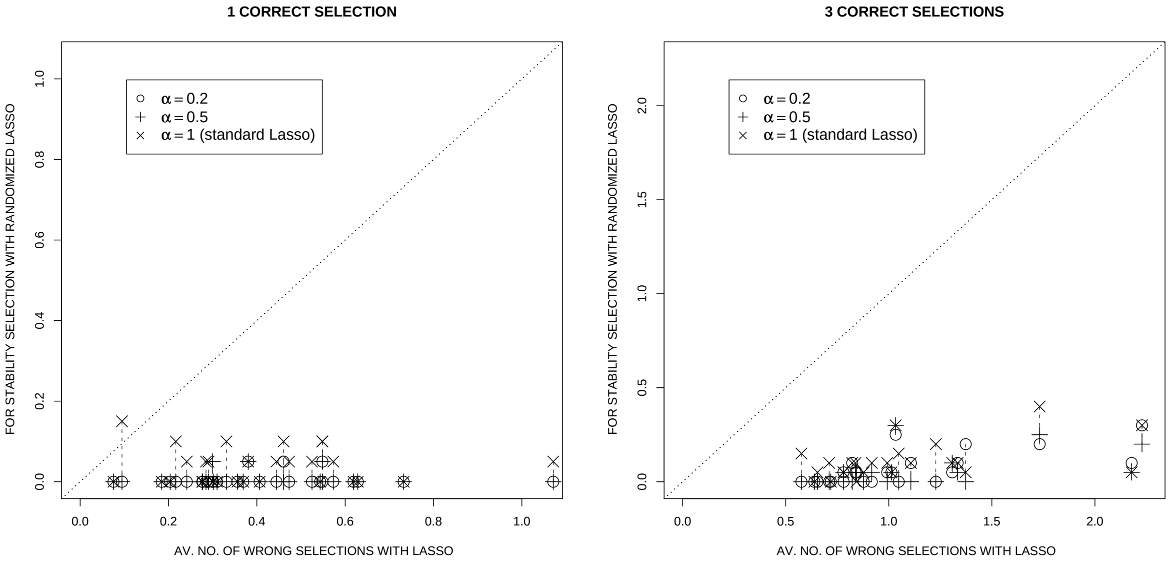

Figure 7: *The average number of falsely selected variables of stability selection (with randomized Lasso and $\alpha \in \{0.2, 0.5, 1\}$) as a function of the average number of falsely selected variables for standard Lasso. The tuning parameters are chosen in each case to guarantee that 1 relevant variable is selected (left) or that 3 relevant variables are selected (right). Solutions for the same setup and different weaknesses α are joined by a vertical lines. Stability selection with randomized Lasso selects far fewer variables falsely than standard Lasso, across the whole range of signal-to-noise ratios and sparsities.*

4 Numerical Results

To investigate further the effects of stability selection numerically, we focus here on the application of stability selection to Lasso and randomized Lasso. We use two real datasets as basis for our simulations. The first is the already mentioned vitamin gene expression data (with $p = 4088$ and $n = 158$) described in Section 2.2. The second dataset ($p = 660$ and $n = 750$) is about motif regression for finding transcription factor binding sites (motifs) in DNA sequences. The response variable consists of gene expression values (with genes being the samples) and the real-valued predictor variables are abundance scores for p candidate motifs (for each of the genes). Our dataset is from a heat-shock experiment with yeast. For a general description and motivation about motif regression we refer to Conlon et al. (2003).

Each dataset is once used with all n available samples and once with sample size reduced to $\lfloor 0.6n \rfloor$. We do not use the response values from these datasets, however, as we need to know which variables are truly relevant or irrelevant. To this end, we create sparse regression vectors by setting $\beta_k = 0$ for all $k = 1, \ldots, p$, except for a randomly chosen set S of coefficients, where $\beta_k = 1$ for all $k \in S$. The response Y is then simulated as $Y = X\beta + \varepsilon$, where ε_i i.i.d. $\mathcal{N}(0, \sigma^2/n)$ for all $i = 1, \ldots, n$, where the rescaling of the variance with n is due to the rescaling of the predictor variables to unit norm, i.e. $\|X^{(k)}\|_2 = 1$. The size $s = |S|$ of the active set is increased in steps of 4 from 4 to 40. The noise level σ^2 is chosen to effectively have signal-to-noise ratios (SNR) of $0.25, 1$ and 4.

For each of these settings, Lasso is applied, with randomization and without, and simulations are run 20 times. We are firstly interested in the cross-validated solution, compared with stability selection. For stability selection, we chose $q_\Lambda = \sqrt{0.8p}$ and thresholds of $\pi_{thr} = 0.6$, corresponding to a control of $E(V) \leq 2.5$, where V is the number of wrongly selected variables. The control is mathematically derived under the assumption of exchangeability for the distribution of noise variables, see Theorem 1. This assumption is most likely not fulfilled for the given dataset and it is of interest to see how well the bound holds up for real data.

Results are shown in Figure 6 for the motif regression dataset. Stability selection reduces the number of falsely selected variables dramatically, while maintaining almost the same power to detect relevant variables. The number of falsely chosen variables is remarkably well controlled at the desired level, giving empirical evidence that the derived error control is useful beyond the discussed setting of exchangeability. Stability selection thus helps to select a useful amount of regularization.

The solution of stability selection cannot be reproduced by simply selecting the right penalty with Lasso, since stability selection provides a fundamentally new solution. To compare the power of both approaches, we look at the ranking of variables produced by both approaches. For Lasso, a variable k is ranked higher than variable k' if and only if there exists some λ such that $\hat{\beta}_k^\lambda \neq 0$ and $\hat{\beta}_{k'}^{\lambda'} = 0$ for all $\lambda' \geq \lambda$, i.e. according to the first appearance in the regularization path. For stability selection, we simply rank variables according to their selection probability $\hat{Pi}_k^\lambda$, where λ is chosen such that $\sqrt{0.8p}$ variables are selected. For a given number $u \geq 1$ of desired 'discoveries', we look in both lists how many irrelevant variables need to be selected to include u relevant variables. Results are shown in Figure 7. Stability selection identifies as many or more correct variables than the underlying method itself. Often the gain is substantial, irrespective of the sparsity of the signal and the signal-to-noise-ratio. The weakness of the underlying randomized Lasso seems to be of secondary importance, although having $\alpha < 1$ generally helps slightly compared to the non randomized version with $\alpha = 1$, as one would expect from the theory.

5 Discussion

Stability selection addresses the notoriously difficult problem of structure estimation or model selection, especially for high-dimensional problems. For example, cross-validation fails often for high-dimensional data, sometimes spectacularly. Stability selection is based on subsampling in combination with (high-dimensional) selection algorithms. The method is extremely general and we demonstrate its applicability for variable selection in regression, for Gaussian graphical modeling and for clustering.

Stability selection provides finite sample familywise multiple testing error control (or control of other error rates of false discoveries) and hence a transparent principle to choose a proper amount of regularization for structure estimation or model selection. Furthermore, the solution of stability selection depends surprisingly little on the chosen regularization which is an additional great benefit besides error control.

Another property of stability selection is the improvement over a pre-specified selection method. It is often the case that computationally efficient algorithms for high-dimensional selection are inconsistent, even in very simple settings. We prove for randomized Lasso, and also for randomized Orthogonal Matching Pursuit, that stability selection will be model selection consistent even if the necessary conditions needed for consistency of the original method are violated. And thus, stability selection will asymptotically select the right model in scenarios where Lasso or Orthogonal Matching Pursuit fail.

In short, stability selection is the marriage of subsampling and high-dimensional selection algorithms, yielding finite sample familywise error control and markedly improved structure estimation or selection methods. Both of these main properties are demonstrated on simulated and real data.

6 Appendix

6.1 Sample splitting

An alternative to subsampling is sample splitting. Instead of observing if a given variable is selected for a random subsample, one can look at a random split of the data into two non-overlapping samples of equal size $\lfloor n/2 \rfloor$ and see if the variable is chosen in both sets simultaneously. Let I_1 and I_2 be two random subsets of $\{1, \ldots, n\}$ with $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$ and $I_1 \cap I_2 = \emptyset$. Define the simultaneously selected set as the intersection of $\hat{S}^\lambda(I_1)$ and $\hat{S}^\lambda(I_2)$,

$$\hat{S}^{simult,\lambda} \quad = \quad \hat{S}^\lambda(I_1) \cap \hat{S}^\lambda(I_2).$$

Definition 5 (Simultaneous selection probability) *Define the simultaneous selection probabilities $\hat{\Pi}$ for any set $K \subseteq \{1, \ldots, p\}$ as*

$$\hat{\Pi}_K^{simult,\lambda} \quad = \quad P^*(K \subseteq \hat{S}^{simult,\lambda}), \tag{18}$$

where the probability P^* is with respect to the random sample splitting (and any additional randomness if $\hat{S}^\lambda$ is a randomized algorithm).

We work with the selection probabilities based on subsampling but the following lemma lets us convert these probabilities easily into simultaneous selection probabilities based on sample splitting. The bound is rather tight for selection probabilities close to 1.

Lemma 1 (Lower bound for simultaneous selection probabilities) *For any set $K \subseteq \{1, \ldots, p\}$, a lower bound for the simultaneous selection probabilities is given by, for every $\omega \in \Omega$, by*

$$\hat{\Pi}_K^{simult,\lambda} \;\geq\; 2\hat{\Pi}_K^\lambda - 1 \tag{19}$$

Proof. Let I_1 and I_2 be the two random subsets in sample splitting of $\{1, \ldots, n\}$ with $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$ and $I_1 \cap I_2 = \emptyset$. Denote by $s_K(\{1,1\})$ the probability $P^*(\{K \subseteq \hat{S}^\lambda(I_1)\} \cap \{K \subseteq \hat{S}^\lambda(I_2)\})$.

Note that the two events are not independent as the probability is only with respect to a random split of the fixed samples $\{1,\ldots,n\}$ into I_1 and I_2. The probabilities $s_K(\{1,0\}), s_K(\{0,1\}), s_K(\{0,0\})$ are defined equivalently by $P^*(\{K \subseteq \hat{S}^\lambda(I_1)\} \cap \{K \nsubseteq \hat{S}^\lambda(I_2)\})$, $P^*(\{K \nsubseteq \hat{S}^\lambda(I_1)\} \cap \{K \subseteq \hat{S}^\lambda(I_2)\})$, and $P^*(\{K \nsubseteq \hat{S}^\lambda(I_1)\} \cap \{K \nsubseteq \hat{S}^\lambda(I_2)\})$. Note that $\hat{\Pi}_K^{simult,\lambda} = s_K(\{1,1\})$ and

$$\begin{aligned}\hat{\Pi}_K^\lambda &= s_K(\{1,0\}) + s_K(\{1,1\}) = s_K(\{0,1\}) + s_K(\{1,1\}) \\ 1 - \hat{\Pi}_K^\lambda &= s_K(\{0,1\}) + s_K(\{0,0\}) = s_K(\{1,0\}) + s_K(\{0,0\})\end{aligned}$$

It is obvious that $s_K(\{1,0\}) = s_K(\{0,1\})$. As $s_K(\{0,0\}) \geq 0$, it also follows that $s_K(\{1,0\}) \leq 1 - \hat{\Pi}_K^\lambda$. Hence

$$\hat{\Pi}_K^{simult,\lambda} = s_K(\{1,1\}) = \hat{\Pi}_K^\lambda - s_K(\{1,0\}) \geq 2\hat{\Pi}_K^\lambda - 1,$$

which completes the proof. □

6.2 Proof of Theorem 1

The proof uses mainly Lemma 2. We first show that $P(k \in \hat{S}^\Lambda) \leq q_\Lambda/p$ for all $k \in N$, using the made definitions $\hat{S}^\Lambda = \cup_{\lambda\in\Lambda}\hat{S}^\lambda$ and $q_\Lambda = E(|\hat{S}^\Lambda|)$. Define furthermore $N_\Lambda = N \cap \hat{S}^\Lambda$ to be the set of noise variables (in N) which appear in $\hat{S}^\Lambda$ and analogously $U_\Lambda = S \cap \hat{S}^\Lambda$. The expected number of falsely selected variables can be written as $E(|N_\Lambda|) = E(|\hat{S}^\Lambda|) - E(|U_\Lambda|) = q_\Lambda - E(|U_\Lambda|)$. Using the assumption (8) (which asserts that the method is not worse than random guessing), it follows that $E(|U_\Lambda|) \geq E(|N_\Lambda|)|S|/|N|$. Putting together, $(1 + |S|/|N|)E(|N_\Lambda|) \leq q_\Lambda$ and hence $|N|^{-1}E(|N_\Lambda|) \leq q_\Lambda/p$. Using the exchangeability assumption, we have $P(k \in \hat{S}^\Lambda) = E(|N_\Lambda|)/|N|$ for all $k \in N$ and hence, for $k \in N$, it holds that $P(k \in \hat{S}^\Lambda) \leq q_\Lambda/p$, as desired. Note that this result is independent of the sample size used in the construction of $\hat{S}^\lambda$, $\lambda \in \Lambda$. Now using Lemma 2 below, it follows that $P(\max_{\lambda\in\Lambda}\hat{\Pi}_k^{simult,q} \geq \xi) \leq (q/p)^2/\xi$ for all $0 < \xi < 1$ and $k \in N$. Using Lemma 1, it follows that $P(\max_{\lambda\in\Lambda}\hat{\Pi}_k^\lambda \geq \pi_{thr}) \leq P((\max_{\lambda\in\Lambda}\hat{\Pi}^{simult,\lambda} + 1)/2 \geq \pi_{thr}) \leq (q_\Lambda/p)^2/(2\pi_{thr} - 1)$. Hence $E(V) = \sum_{k\in N} P(\max_{\lambda\in\Lambda}\hat{\Pi}_k^\lambda \geq \pi_{thr}) \leq q_\Lambda^2/(p(2\pi_{thr} - 1))$, which completes the proof. □

Lemma 2 *Let $K \subset \{1,\ldots,p\}$ and $\hat{S}^\lambda$ the set of selected variables based on a sample size of $\lfloor n/2 \rfloor$. If $P(K \subseteq \hat{S}^\lambda) \leq \varepsilon$, then*

$$P(\hat{\Pi}_K^{simult,\lambda} \geq \xi) \leq \varepsilon^2/\xi.$$

If $P(K \subseteq \cup_{\lambda\in\Lambda}\hat{S}^\lambda) \leq \varepsilon$ for some $\Lambda \subseteq \mathbb{R}^+$, then

$$P(\max_{\lambda\in\Lambda}\hat{\Pi}_K^{simult,\lambda} \geq \xi) \leq \varepsilon^2/\xi.$$

Proof. Let $I_1, I_2 \subseteq \{1,\ldots,n\}$ be, as above, the random split of the samples $\{1,\ldots,n\}$ into two disjoint subsets, where both $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$. Define the binary random variable H_K^λ for all subsets $K \subseteq \{1,\ldots,p\}$ as $H_K^\lambda := \mathbf{1}\{K \subseteq \{\hat{S}^\lambda(I_1) \cap \hat{S}^\lambda(I_2)\}\}$. Denote the data (the n samples) by Z. The simultaneous selection probability $\hat{\Pi}_K^{simult,\lambda}$, as defined in (18), is then $\hat{\Pi}_K^{simult,\lambda} = E^*(H_K^\lambda) = E(H_K^\lambda|Z)$, where the expectation E^* is with respect to the random split of the n samples into sets I_1 and I_2 (and additional randomness if $\hat{S}^\lambda$ is a randomized algorithm). To prove the first part, the inequality $P(K \subseteq \hat{S}^\lambda) \leq \varepsilon$ (for a sample size $\lfloor n/2 \rfloor$), implies that $P(H_K^\lambda = 1) \leq P(K \subseteq \hat{S}^\lambda(I_1))^2 \leq$

ε^2 and hence $E(H_K^\lambda) \le \varepsilon^2$. Therefore, $E(H_K^\lambda) = E(E(H_K^\lambda|Z)) = E(\hat{\Pi}_K^{simult,\lambda}) \le \varepsilon^2$ Using a Markov-type inequality, $\xi P(\hat{\Pi}_K^{simult,\lambda} \ge \xi) \le E(\hat{\Pi}_K^{simult,\lambda}) \le \varepsilon^2$. Thus $P(\hat{\Pi}_K^{simult,\lambda} \ge \xi) \le \varepsilon^2/\xi$, completing the proof of the first claim. The proof of the second part follows analogously. □

6.3 Proof of Theorem 2

Instead of working directly with form (13) of the randomized Lasso estimator, we consider the equivalent formulation of the standard Lasso estimator, where all variables have initially unit norm and are then rescaled by their random weights W.

Definition 6 (Additional notation) *For weights W as in (13), let X^w be the matrix of re-scaled variables, with $X^w_k = X_k \cdot W_k$ for each $k = 1, \dots, p$. Let $\phi^w_{\max}$ and $\phi^w_{\min}$ be the maximal and minimal eigenvalues analogous to (14) for X^w instead of X.*

The proof rests mainly on the two-fold effect a weakness $\alpha < 1$ has on the selection properties of the Lasso. The first effect is that the singular values of the design can be distorted if working with the reweighted variables X^w instead of X itself. A bound on the ratio between largest and smallest eigenvalue is derived in Lemma 3, effectively yielding a lower bound for useful values of α. The following Lemma 4 then asserts for such values of α that the relevant variables in S are chosen with high probability under any random sampling of the weights. The next Lemma 5 establishes the key advantage of randomized Lasso as it shows that the 'irrepresentable condition' (12) is sometimes fulfilled under randomly sampled weights (even though its not fulfilled for the original data). Variables which are wrongly chosen because condition (12) is not satisfied for the original unweighted data will thus not be selected by stability selection. The final result is established in Lemma 7 after a bound on the noise contribution in Lemma 6.

Lemma 3 *Define $\overline{C}$ by $(2 + 4\overline{C})s + 1 = Cs^2$ and assume $s \ge 7$. Let W be weights generated randomly in $(\alpha, 1]$, as in (13), and let X^w be the corresponding rescaled predictor variables, as in Definition 6. For $\alpha^2 = \nu\phi_{\min}(Cs^2)/(Cs^2)$, with $\nu \in \mathbb{R}^+$, it holds under Assumption 1 for all random realizations W that*

$$\frac{\phi^w_{\max}(Cs^2)}{\phi^w_{\min}(Cs^2)} \le \frac{7\overline{C}}{\kappa\sqrt{\nu}}. \tag{20}$$

Proof. As $\phi_{\max}$ and $1/\phi_{\min}$ are monotonically increasing in their respective arguments, it hence follows that

$$\frac{\phi_{\max}(Cs^2)}{\phi^{3/2}_{\min}(Cs^2)} < \frac{\sqrt{C}}{\kappa} = (Cs^2)^{-1/2}\frac{((2+4\overline{C})s+1)/s}{\kappa} \le (Cs)^{-1/2}(3+4\overline{C})/\kappa,$$

where the first inequality follows by Assumption 1, the equality by $(2 + 4\overline{C})s + 1 = Cs^2$ and the second inequality by $s \ge 1$. It follows that

$$\frac{\phi_{\max}(Cs^2)}{\phi_{\min}(Cs^2)} \le \frac{3+4\overline{C}}{\kappa}\sqrt{\frac{\phi_{\min}(Cs^2)}{Cs^2}}. \tag{21}$$

Now, let $\mathcal{W}$ be again the $p \times p$-diagonal matrix with diagonal entries $\mathcal{W}_{kk} = W_k$ for all $k = 1, \dots, p$. Then $X^w = X\mathcal{W}$ and, taking suprema over all $\mathcal{W}$ with diagonal entries in $(\alpha, 1]$,

$$\begin{aligned}(\phi^w_{\max}(m))^2 &\leq \sup_{\mathcal{W}} \sup_{v \in \mathbb{R}^p : \|v\|_0 \leq m} (\|X^w v\|_2 / \|v\|_2)^2 \\ &= \sup_{\mathcal{W}} \sup_{v \in \mathbb{R}^p : \|v\|_0 \leq m} (v^T \mathcal{W}^T X^T X \mathcal{W} v)/v^T v \leq (\phi_{\max}(m))^2,\end{aligned}$$

where the last step follows by a change of variable transform $\tilde{v} = \mathcal{W}v$ and the fact that $\|v\|_0 = \|\mathcal{W}v\|_0$ as well as $v^T v = \tilde{v}^T \mathcal{W}^{-1,T} \mathcal{W}^{-1} \tilde{v}$ and thus $\tilde{v}^T \tilde{v} \leq v^T v \leq \alpha^{-2} \tilde{v}^T \tilde{v}$ for all $\mathcal{W}$ with diagonal entries in $(\alpha, 1]$. The corresponding argument for $\phi_{\min}(m)$ yields the bound $\phi^w_{\min}(m) \geq \alpha \phi_{\min}(m)$. The claim (20) follows by observing that $\overline{C} \geq 1$ for $s \geq 7$, since $C \geq 1$ by Assumption 1 and hence $3 + 4\overline{C} \leq 7\overline{C}$. □

Lemma 4 *Let $\hat{A}^{\lambda,W}$ be the set $\{k : \hat{\beta}^{\lambda,W} \neq 0\}$ of selected variables of the randomized Lasso with weakness $\alpha \in (0,1]$ and randomly sampled weights W. Suppose that the weakness $\alpha^2 \geq (7/\kappa)\phi_{\min}(Cs^2)/(Cs^2)$ and that $\min_{k:\beta_k \neq 0} |\beta_k| \geq 0.3\,(Cs)^{3/2}\lambda$. Under the assumptions of Theorem 2, there exists a set Ω_0 in the sample space of Y with $P(Y \in \Omega_0) \geq 1 - 3/p$, such that for all realizations $W = w$, for $p \geq 5$, if $Y \in \Omega_0$,*

$$|\hat{A}^{\lambda,w}| \leq Cs^2 \text{ and } S \subseteq \hat{A}^{\lambda_{\min},w}. \tag{22}$$

Proof. Follows mostly from Theorem 1 in Zhang and Huang (2008). To this end, set either $a_n = p$ and $c_0 = 0$ or, equivalently, $a_n = c_0 = 1$ in their notation. Setting also $q^* = Cs^2$, we have $q^* \leq (2 + 4\overline{C})s + 1$, as, by definition, $(2 + 4\overline{C})s + 1 = Cs^2$, as in Lemma 3. The quantity $C = c^*/c_*$ in Zhang and Huang (2008) is identical to our notation $\phi^w_{\max}(Cs^2)/\phi^w_{\min}(Cs^2)$. It is bounded for all random realizations of $W = w$, as long as $\alpha^2 \geq (7/\kappa)\phi_{\min}(Cs^2)/(Cs^2)$, using Lemma 3, by

$$\frac{\phi^w_{\max}((2 + 4\overline{C})s + 1)}{\phi^w_{\min}((2 + 4\overline{C})s + 1)} \leq \overline{C}.$$

Hence all assumptions of Theorem 1 in Zhang and Huang (2008) are fulfilled, with $\eta_1 = 0$, for any random realization $W = w$. Using (2.20)-(2.24) in Zhang and Huang (2008), it follows that there exists a set Ω_0 in the sample space of Y with $P(Y \in \Omega_0) \geq 2 - \exp(2/p) - 2/p^2 \geq 1 - 3/p$ for all $p \geq 5$, such that if $Y \in \Omega_0$, from (2.21) in Zhang and Huang (2008),

$$|\hat{A}^{\lambda,w} \cup S| \leq (2 + 4\overline{C})s \leq Cs^2, \tag{23}$$

and, from (2.23) in Zhang and Huang (2008),

$$\sum_{k \in S} |\beta_k|^2 1\{k \notin \hat{A}^{\lambda,w}\} \leq (\frac{2}{3}\overline{C} + \frac{28}{9}\overline{C}^2 + \frac{16}{9}\overline{C}^3)s\lambda^2 \leq 5.6\overline{C}^3 s^3 \lambda^2 \leq (0.3\,(Cs)^{3/2}\lambda)^2, \tag{24}$$

having used, for the first inequality, that, in the notation of Zhang and Huang (2008), $1/(c^* c_*) \leq c^*/c_*$. The n^{-2} factor was omitted to account for our different normalization. For the second inequality, we used $4\overline{C} \leq Cs$. Using the assumption about the minimal absolute non-zero value of β, the last equation implies $S \subseteq \hat{A}^{\lambda,w}$, which completes the proof. □

Lemma 5 *Let $A \subseteq \{1, \dots, p\}$ and $k \in \{1, \dots, p\} \setminus A$. For $\alpha^2 \le \phi_{min}(m)/(\sqrt{2}m)$, with $m = Cs^2$, there exists a set Ω_w of the sample space of W with $P_w(\Omega_w) \ge p_w(1-p_w)^m$ such that for all $w \in \Omega_w$,*

$$\sup_{A:|A|\le m} \sup_{k \notin A} \|((X_A^w)^T X_A^w)^{-1}(X_A^w)^T X_k^w\|_1 \le 2^{-1/4}. \tag{25}$$

where the probability P_w is with respect to random sampling of the weights W and p_w is, as above, the probability of choosing weight α for each variable and $1-p_w$ the probability of choosing weight 1.

Proof. Let w be a realization of W such that $w_k = \alpha$ and $w_j = 1$ for all $j \in A$. The probability of $W = w$ is clearly $p_w(1-p_w)^{|A|}$ under the used sampling scheme for the weights. For these weights,

$$((X_A^w)^T X_A^w)^{-1}(X_A^w)^T X_k^w = \alpha(X_A^T X_A)^{-1} X_A^T X_k$$

Using the bound on α, it hence only remains to be shown that, if $\|X_l\|_2 = 1$ for all $l \in \{1, \dots, p\}$, for all A and $k \notin A$ and $m \in \mathbb{N}$,

$$\sup_{A:|A|\le m} \sup_{k \notin A} \|(X_A^T X_A)^{-1} X_A^T X_k\|_1^2 \le m/\phi_{\min}(m). \tag{26}$$

Since $\|\gamma\|_1 \le \sqrt{|A|}\|\gamma\|_2$ for any vector $\gamma \in \mathbb{R}^{|A|}$, it is sufficient to show, for $\gamma := (X_A^T X_A)^{-1} X_A^T X_k$,

$$\sup_{A:|A|\le m} \sup_{k \notin A} \|\gamma\|_2^2 \le 1/\phi_{\min}(m).$$

As $X_A\gamma$ is the projection of X_k into the space spanned by X_A and $\|X_k\|_2^2 = 1$, it holds that $\|X_A\gamma\|_2^2 \le 1$. Using $\|X_A\gamma\|_2^2 = \gamma^T(X_A^T X_A)\gamma \ge \phi_{\min}(|A|)\|\gamma\|_2^2$, it follows that $\|\gamma\|_2^2 \le 1/\phi_{\min}(|A|)$, which completes the proof. □

Lemma 6 *Let $P_A = X_A(X_A^T X_A)^{-1} X_A^T$ be the projection into the space spanned by all variables in subset $A \subseteq \{1, \dots, p\}$. Suppose $p > 10$. Then there exists a set Ω_1 with $P(\Omega_1) \ge 1 - 2/p$, such that for all $\omega \in \Omega_1$,*

$$\sup_{A:|A|\le m} \sup_{k \notin A} |X_k^T(1 - P_A)\varepsilon| < 2\sigma(\sqrt{m} + 1)\sqrt{\log(p)/n}. \tag{27}$$

Proof. Let Ω_1' be the event that $\max_{k\in\{1,\dots,p\}} |X_k^T\varepsilon| \le \sigma\sqrt{2\log(p^2)/n}$. As entries in ε are i.i. $\mathcal{N}(0,\sigma^2)$ distributed, $P(\Omega_1') \ge 1 - 1/p$ for all $\delta \in (0,1)$. Note that, for all $A \subset \{1, \dots, p\}$ and $k \notin A$, $|X_k^T P_A\varepsilon| \le \|P_A\varepsilon\|_2$. Define Ω_1'' as

$$\sup_{|A|\le m} \|P_A\varepsilon\|_2 \le 2\sigma\sqrt{m\log(p)/n}. \tag{28}$$

It is now sufficient to show that $P(\Omega_1'') \ge 1 - 1/p$. Showing the bound $P(\Omega_1'') \ge 1 - 1/p$ for the set (28) is related to a bound in Zhang and Huang (2008) and we repeat a similar argument. Each term $\sqrt{n}\|P_A\varepsilon\|_2/\sigma$ has a $\chi^2_{|A|}$ distribution as long as X_A is of full rank $|A|$. Hence, using the same standard tail bound as in the proof of Theorem 3 of Zhang and Huang (2008),

$$P\Big(n\|P_A\varepsilon\|_2^2/\sigma^2 \ge |A|(1 + 4\log p)\Big) \le (p^{-4}(1 + 4\log p))^{|A|/2} \le p^{-3|A|/2},$$

having used $1 + 4\log p \leq p$ for all $p > 10$ in the last step and thus, using $\binom{p}{|A|} \leq p^{|A|}/|A|!$,

$$P(\Omega_1'') \geq 1 - \sum_{|A|=2}^{m} \binom{p}{|A|} p^{-3|A|/2} \geq 1 - \sum_{|A|=2}^{m} p^{-|A|/2}/(|A|)! \geq 1 - 1/p,$$

which completes the proof if setting $\Omega_1 = \Omega_1' \cap \Omega_1''$ and concluding that $P(\Omega_1) \geq 1 - 2/p$ for all $p > 10$. □

Lemma 7 *Let $Z = (X^{(i)}, Y^{(i)})$ be the observed data with $i = 1, \ldots, n$. Let $\delta_w = p_w(1 - p_w)^{Cs^2}$. Let again $\hat{\Pi}_k^\lambda = P_w(k \in \hat{A}^{\lambda,W})$ be the probability for variable k of being in the selected subset, with respect to random sampling of the weights W. Then, under the assumptions of Theorem 2, for all $k \notin S$ and $p > 10$,*

$$P\Big(\max_{\lambda\in\Lambda} \hat{\Pi}_k^\lambda \geq 1 - \delta_w \text{ if and only if } k \in S\Big) \geq 1 - 5/p. \tag{29}$$

Proof. We use event Ω_0 of Lemma 4 and event Ω_1 of Lemma 6. Since, using these two lemmas,

$$P(\Omega_0 \cap \Omega_1) \geq 1 - P(\Omega_0^c) - P(\Omega_1^c) \geq 1 - 3/p - 2/p = 1 - 5/p,$$

it is sufficient to show that, for all $\omega \in \Omega_0 \cap \Omega_1$,

$$\max_{\lambda\in\Lambda} \hat{\Pi}_k^\lambda \geq 1 - \delta_w \text{ if and only if } k \in S \tag{30}$$

We begin with the "only if" part in (30). A variable $k \notin S$ is in the selected set $\hat{A}^{\lambda,W}$ only if

$$|(X_k^w)^T(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k})| \geq \lambda, \tag{31}$$

where $\hat{\beta}^{\lambda,W,-k}$ is the solution to (13) with the constraint that $\hat{\beta}_k^{\lambda,W} = 0$, comparable to the analysis in Meinshausen and Bühlmann (2006). Let $\hat{A} = \{k : \hat{\beta}^{\lambda,W,-k} \neq 0\}$ be the set of non-zero components. Let $P_{\hat{A}}^w$ be the projection operator into the space spanned by all variables in the set $\hat{A}$. For all $W = w$, this is identical to

$$P_{\hat{A}}^w = X_{\hat{A}}^w((X_{\hat{A}}^w)^T X_{\hat{A}}^w)^{-1} X_{\hat{A}}^w = X_{\hat{A}}(X_{\hat{A}}^T X_{\hat{A}})^{-1} X_{\hat{A}} = P_{\hat{A}}.$$

Then splitting the term $(X_k^w)^T(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k})$ in (31) into the two terms

$$(X_k^w)^T(1 - P_{\hat{A}}^w)(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k}) \;+\; (X_k^w)^T P_{\hat{A}}^w (Y - X_{-k}^w \hat{\beta}_{-k}^{\lambda,W}), \tag{32}$$

it holds for the right term in (32) that

$$\begin{aligned}(X_k^w)^T P_{\hat{A}}^w (Y - X_{-k}^w \hat{\beta}_{-k}^{\lambda,W}) &\leq (X_k^w)^T X_{\hat{A}}^w((X_{\hat{A}}^w)^T X_{\hat{A}}^w)^{-1} \text{sign}(\hat{\beta}^{\lambda,W,-k})\lambda \\ &\leq \|((X_{\hat{A}}^w)^T X_{\hat{A}}^w)^{-1}(X_{\hat{A}}^w)^T X_k^w\|_1 \lambda.\end{aligned}$$

Looking at the left term in (32), since $Y \in \Omega_0$, we know by Lemma 4 that $|\hat{A}| \leq Cs^2$ and $S \subseteq \hat{A}$. Thus the left term in (32) is bounded from above by

$$\begin{aligned}(X_k^w)^T(1 - P_{\hat{A}}^w)\varepsilon &\leq \sup_{A:|A|\leq Cs^2} \sup_{k\notin A} |(X_k)^T(1 - P_{\hat{A}})\varepsilon| \cdot \|X_k^w\|_2/\|X_k\|_2 \\ &< \lambda_{\min}\|X_k^w\|_2/\|X_k\|_2,\end{aligned}$$

having used Lemma 6 in the last step with $\lambda_{\min}$. Putting together, the two terms in (32) are bounded, for all $\omega \in \Omega_0 \cap \Omega_1$, by

$$\lambda_{\min}\|X_k^w\|_2/\|X_k\|_2 + \|((X_{\hat{A}}^w)^T X_{\hat{A}}^w)^{-1}(X_{\hat{A}}^w)^T X_k^w\|_1\lambda.$$

Using Lemma 5, there exists a set Ω_w in the sample space of W with $P_w(\Omega_w) \geq 1 - \delta_w$ such that $\|((X_{\hat{A}}^w)^T X_{\hat{A}}^w)^{-1}(X_{\hat{A}}^w)^T X_k^w\|_1 \leq 2^{-1/4}$. Moreover, for the same set Ω_w, we have $\|X_k^w\|_2/\|X_k\|_2 = \alpha \leq 1/s \leq 1/7$. Hence, for all $\omega \in \Omega_0 \cap \Omega_1$ and, for all $\omega \in \Omega_w$, the lhs of (31) is bounded from above by $\lambda_{\min}/7 + \lambda 2^{-1/4} < \lambda$ and variable $k \notin S$ is hence not part of the set $\hat{A}^{\lambda,W}$. It follows that $\max_{\lambda\in\Lambda} \hat{\Pi}_k^\lambda < 1 - \delta_w$ with $\delta_w = p_w(1-p_w)^{Cs^2}$ for all $k \notin S$. This completes the first part of the proof.

For the second part of the claim (the "if" part), we need to show that, for all $\omega in \Omega_0 \cap \Omega_1$, all variables k are chosen with probability at least $1 - \delta_w$ (with respect to random sampling of the weights W). For all $\omega \in \Omega_0$, however, it follows directly from Lemma 4 that $S \subseteq \hat{A}^{\lambda_{\min},W}$. Hence, for all $k \in S$, $\max_{\lambda\in\Lambda} \hat{\Pi}_k^\lambda \geq \hat{\Pi}_k^{\lambda_{\min}} = 1$, which completes the proof. □

Since the statement in Lemma 7 is a reformulation of the assertion of Theorem 2, the proof of the latter is complete.

Acknowledgements

N.M. would like to thank FIM (Forschungsinstitut für Mathematik) at ETH Zürich for support and hospitality.

References

Bach, F. (2008). Bolasso: Model consistent lasso estimation through the bootstrap. *Arxiv preprint arxiv:0804.1302*.

Banerjee, O. and L. El Ghaoui (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research 9*, 485–516.

Bickel, P. and E. Levina (2008). Regularized estimation of large covariance matrices. *Annals of Statistics 36*(1), 199.

Bickel, P., Y. Ritov, and A. Tsybakov (2007). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics, to appear*.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *The Annals of Statistics 30*(4), 927–961.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*, 2312–2351.

Chen, S., S. Donoho, and M. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM Review 43*, 129–159.

Conlon, E., X. Liu, J. Lieb, and J. Liu (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Science 100*, 3339 – 3344.

Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1-minimization. *PNAS 100*, 2197–2202.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–451.

Freedman, D. (1977). A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association 72*, 681–681.

Freund, Y. and R. Schapire (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference 148*, 156.

Friedman, J., T. Hastie, H. Hoefling, and R. Tibshirani (2007). Pathwise coordinate optimization. *Annals of Applied Statistics 1*, 302–332.

Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*.

Hartigan, J. (1975). *Clustering algorithms*. Wiley.

Lange, T., V. Roth, M. Braun, and J. Buhmann (2004). Stability-based validation of clustering solutions. *Neural Computation 16*, 1299 – 1323.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica 16*, 1273–1284.

Mallat, S. and Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on 41*(12), 3397–3415.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B 70*, 53–71.

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

Meinshausen, N. and B. Yu (2008). Lasso-type recovery of sparse representations from high-dimensional data. *Annals of Statistics, to appear*.

Rothman, A., P. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Temlyakov, V. (2000). Weak greedy algorithms. *Advances Computational Mathematics 12*, 213–227.

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

Tropp, J. (2004). Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on 50*(10), 2231–2242.

van de Geer, S. and H. van Houwelingen (2004). High-dimensional data: $p \gg n$ in mathematical statistics and bio-medical applications. *Bernoulli 10*, 939–943.

Wainwright, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Arxiv preprint math.ST/0605740*.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*, 19–35.

Zhang, C. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics 36*(4), 1567–1594.

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

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