

Stability Selection

Nicolai Meinshausen and Peter Bühlmann
University of Oxford and ETH Zürich

May 16, 2009

Abstract

Estimation of structure, such as in variable selection, graphical modelling or cluster analysis is notoriously difficult, especially for high-dimensional data. We introduce 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 regularisation for structure estimation. Variable selection and structure estimation improve markedly for a range of selection methods if stability selection is applied. We prove for randomised Lasso that stability selection will be variable selection consistent even if the necessary conditions needed for consistency of the original Lasso method are violated. We demonstrate stability selection for variable selection and Gaussian graphical modelling, using real and simulated data.

1 Introduction

Estimation of discrete structure, such as graphs or clusters, or variable 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 variable 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 exceeds the number of observations n ; for an early overview see for example van de Geer and van Houwelingen (2004). In a similar spirit, graphical modelling 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 optimisation problems are discrete, as well as determining the right amount of regularisation, 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 optimisation 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; Zhang, 2009). The practical usefulness of such procedures has been demonstrated in various applications. However, the general issue of selecting a proper amount of regularisation (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 regularisation with a very generic subsampling approach (bootstrapping would behave similarly). **We show that subsampling can be used to determine the amount of regularisation** such that a certain familywise type I error rate in 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 regularisation, the subsampling approach yields a new structure estimation or variable selection scheme. For the more specialised case of high-dimensional linear models, we prove what we expect in greater generality: namely that subsampling in conjunction with ℓ_1 -penalised estimation requires much weaker assumptions on the design matrix for asymptotically consistent variable selection than what is needed for the (non-sampled) ℓ_1 -penalty scheme. Furthermore, we show that additional improvements can be achieved by randomising 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 non-zero. 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 $(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})$ with univariate response variable Y and p -dimensional covariates X . We typically assume that $(X^{(i)}, Y^{(i)})$'s are i.i. distributed. The vector β could be the coefficient vector in a linear model

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

where $Y = (Y_1, \dots, Y_n)$, X is the $n \times p$ design matrix and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ 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 penalising 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 penalised optimisation problem becomes quickly unfeasible 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),

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

where $\lambda \in \mathbb{R}^+$ is a regularisation 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 optimisation problem in (2) is convex. Furthermore, the Lasso is able to select variables by shrinking certain estimated coefficients exactly to 0. We can then estimate the set S of non-zero β coefficients by $\hat{S}^\lambda = \{k; \hat{\beta}_k^\lambda \neq 0\}$ which involves convex optimisation only. Substantial understanding has been gained over the last few years about consistency of such Lasso variable 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 regularisation λ for consistent variable 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 modelling. The data is assumed to be

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

The goal is to infer conditional dependencies among the d variables or components in $X = (X_1, \dots, X_d)$. 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 unfeasible 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., 2008):

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

This amounts to an ℓ_1 -penalised estimator of the Gaussian log-likelihood, partially maximised over the mean vector μ , when minimising 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 optimisation only and is computationally feasible for large values of d .

Another potential area of application is clustering. Choosing the correct number of cluster is a notoriously difficult problem. Looking for clusters that are stable under perturbations or subsampling of the data can help to get a better sense of a meaningful number of clusters and to validate results. Indeed, there has been some activity in this area, most notably in the context of *consensus clustering* (Monti et al., 2003). For an early application see Bhattacharjee et al. (2005). Our proposed false discovery control can be applied to consensus clustering, yielding good estimates of the parameters of a suitable base clustering method for consensus clustering.

1.2 Outline

The use of resampling for purposes of validation is certainly not new; we merely try to put it into a more formal framework and to show certain empirical and theoretical advantages of doing so. It seems difficult to give a complete coverage of all previous work in the area, as notions of stability, resampling and perturbations are very natural in the context of structure estimation and variable selection. We reference and compare with previous work throughout the paper.

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 and Gaussian graphical models are presented in Section 2. A detailed asymptotic analysis of Lasso and randomised 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 variable 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. We assume throughout this Section 2 that the data, denoted here by $Z^{(1)}, \dots, Z^{(n)}$, are independent and identically distributed (e.g. $Z^{(i)} = (X^{(i)}, Y^{(i)})$ with covariate $X^{(i)}$ and response $Y^{(i)}$).

For a generic structure estimation or variable selection technique, we have a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines the amount of regularisation. This tuning parameter could be the penalty parameter in ℓ_1 -penalised regression, see (2), or in Gaussian graphical modelling, 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 does not cause conceptual problems. For every value $\lambda \in \Lambda$, we obtain a structure estimate $\hat{S}^\lambda \subseteq \{1, \dots, p\}$. It is then of interest to determine whether there exists a $\lambda \in \Lambda$ such that \hat{S}^λ is identical to S with high probability and how to achieve that right amount of regularisation.

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 regularisation paths. A regularisation path is given by the coefficient value of each variable over all regularisation parameters: $\{\hat{\beta}_k^\lambda; \lambda \in \Lambda, k = 1, \dots, p\}$. Stability paths (defined below) are, in contrast, the *probability* for each variable to be selected when randomly resampling from the data. For any given regularisation parameter $\lambda \in \Lambda$, the selected set \hat{S}^λ is implicitly a function of the samples $I = \{1, \dots, n\}$. 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 $\{1, \dots, n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement. For every set $K \subseteq \{1, \dots, p\}$, the probability of being in the selected set $\hat{S}^\lambda(I)$ is*

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

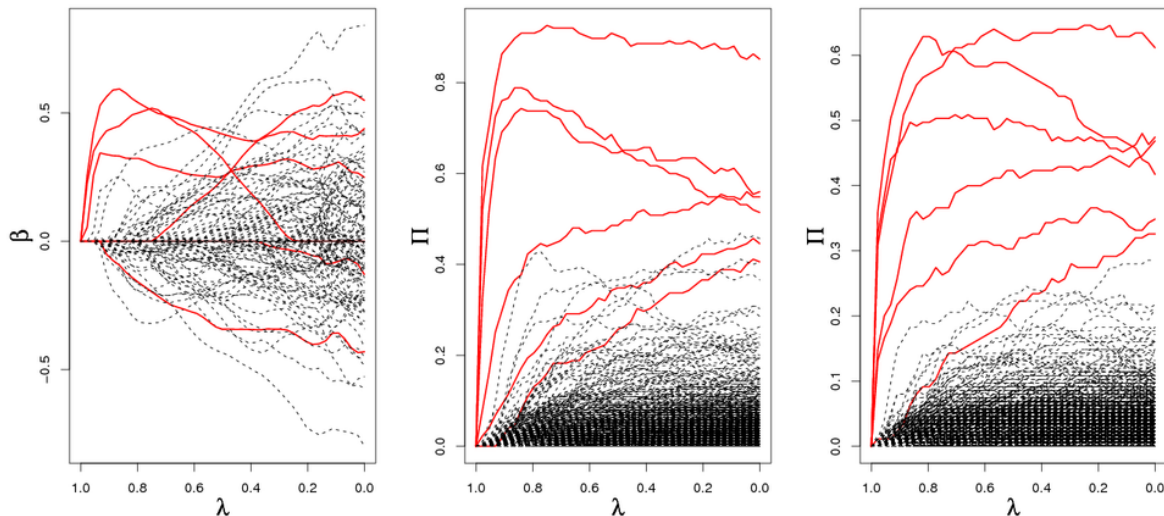


Figure 1: *Left: The Lasso path for the vitamin gene-expression dataset. The paths of the 6 non-permuted genes are plotted as solid, red lines, while the paths of the 4082 permuted genes are shown as broken, black lines. Selecting a model with all 6 unpermuted genes invariably means selecting a large 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 ‘randomised Lasso’ with weakness $\alpha = 0.2$, introduced in Section 3.1. Now all 6 non-permuted variables are chosen before any noise variable enters the model.*

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

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. Subsampling has also been advocated in a related context in Valdar et al. (2009).

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

In the remainder of the manuscript, we look at the selection probabilities of individual variables. The definition above covers also sets of variables. We could monitor the selection probability of a set of functionally related variables, say, by asking how often *at least one* variable in this set is chosen or how often *all* variables in the set are chosen.

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 $S = \{k; \beta_k \neq 0\}$.

Instability of the selected set of genes has been noted before (Ein-Dor et al., 2005; Michiels et al., 2005), if either using marginal association or variable selection in a regression or classification model. Davis et al. (2006) are close in spirit to our approach by arguing for ‘consensus’ gene signatures which assess the stability of selection, while Zucknick et al. (2008) propose to measure stability of so-called ‘molecular profiles’ by the Jaccard index.

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 set of 6 unpermuted genes has been chosen randomly among the 200 genes with the highest marginal association with the response. The Lasso path $\{\hat{\beta}^\lambda; \lambda \in \Lambda\}$ is shown in the left panel of Figure 1, as a function of the regularisation 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 regularisation path plot. The right panel shows the stability plot for randomised 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 regularisation parameter is very difficult for the original path. The prediction optimal and cross-validated choice include too many variables (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. Figure 1 motivates that choosing the right regularisation 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, variable selection would amount to choosing one element of the set of models

$$\{\hat{S}^\lambda; \lambda \in \Lambda\}, \quad (6)$$

where Λ is again the set of considered regularisation 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 regularisation 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 regularisation parameters Λ , the set of stable variables is defined as*

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

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 regularisation λ or the regularisation region Λ . See Figure 1 for an example.

Before we present some guidance on how to choose the cutoff parameter and the regularisation region Λ below, it is worthwhile pointing out that there have been related ideas in the literature on Bayesian model selection. Barbieri and Berger (2004) show certain predictive optimality results for the so-called *median probability model*, consisting of variables which have posterior probability of being in the model of 1/2 or greater (as opposed to choosing the model with the highest posterior probability). Lee et al. (2003) or Sha et al. (2004) are examples of more applied papers considering Bayesian variable selection in this context.

2.4 Choice of regularisation and error control

When trying to recover the set S , a natural goal is to include as few variables of the set N of noise variables as possible. The choice of the regularisation parameter is hence crucial. An advantage of our stability selection is that the choice of the initial set of regularisation 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 regularisation 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 regularisation λ in the set Λ . Let q_Λ be the average number of selected variables, $q_\Lambda = E(|\hat{S}^\Lambda(I)|)$. Define V to be the number of falsely selected variables with stability selection,

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

In general, it is very hard to control $E(V)$, 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 $\lambda \in \Lambda$,

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

The expected number V of falsely selected variables is then bounded by

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

We will discuss below how to make constructive use of the value q_Λ^2 which is in general an unknown quantity. The expected number of falsely selected variables is sometimes called the per-family error rate (*PFER*) or, if divided by p , the per-comparison error rate (*FCER*) in multiple testing (Dudoit et al., 2003). Choosing less variables (reducing q_Λ) or increasing the threshold π_{thr} for selection will, unsurprisingly, reduce the the expected number of falsely selected variables, with a minimal

achievable non-trivial value of $1/p^2$ (for $\pi_{thr} = 1$ and $q_\Lambda = 1$) for the *PFER*. This seems low enough for all practical purposes as long as $p > 10$, say.

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 for all reasonable procedures \hat{S} 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 for real data.

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 regularisation region Λ is determined by the desired error control. Specifically, for a default cutoff value $\pi_{thr} = 0.9$, choosing the regularisation parameters Λ such that say $q_\Lambda = \sqrt{0.8p}$ 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 regularisation region Λ and choosing π_{thr} such that $E(V)$ is controlled at the desired level.

To do this, we need knowledge about q_Λ . This can be easily achieved by regularisation of the selection procedure $\hat{S} = \hat{S}^q$ in terms of the number of selected variables q . That is, the domain Λ for the regularisation parameter λ determines the number q of selected variables, i.e. $q = q(\Lambda)$. For example, with ℓ_1 -norm penalisation as in (2) or (4), the number q is given by the variables which enter first in the regularisation path when varying from a maximal value λ_{\max} to some minimal value λ_{\min} . Mathematically, λ_{\min} is such that $|\cup_{\lambda_{\max} \geq \lambda \geq \lambda_{\min}} \hat{S}^\lambda| \leq q$.

Without stability selection, the regularisation 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}^λ 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}^λ 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. Results typically do not depend strongly on the utilised regularisation 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 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 $k = 1, \dots, p$.

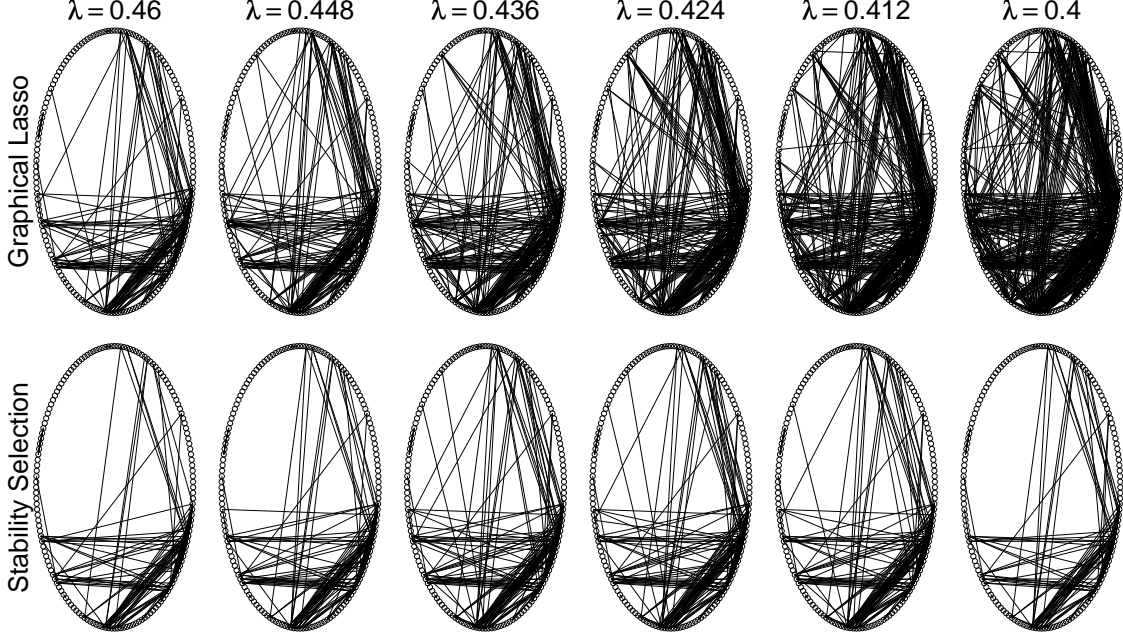


Figure 2: *Vitamin gene-expression dataset. The regularisation path of graphical lasso (top row) and the corresponding point-wise stability selected models (bottom row).*

2.5 Example II: Graphical modelling

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 Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Banerjee and El Ghaoui, 2008; Bickel and Levina, 2008; Rothman et al., 2008). Here we work with the graphical Lasso (Friedman et al., 2008), 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 regularisation path of the graphical Lasso showing graphs for various values of the regularisation parameter λ , i.e. $\{\hat{S}^\lambda; \lambda \in \Lambda\}$ where $\hat{S}^\lambda = \{(j, k); (\hat{\Sigma}^{-1})_{jk}^\lambda \neq 0\}$, are shown in the first row of Figure 2. For reasons of display, variables (genes) are ordered first using hierarchical clustering and are symbolised 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

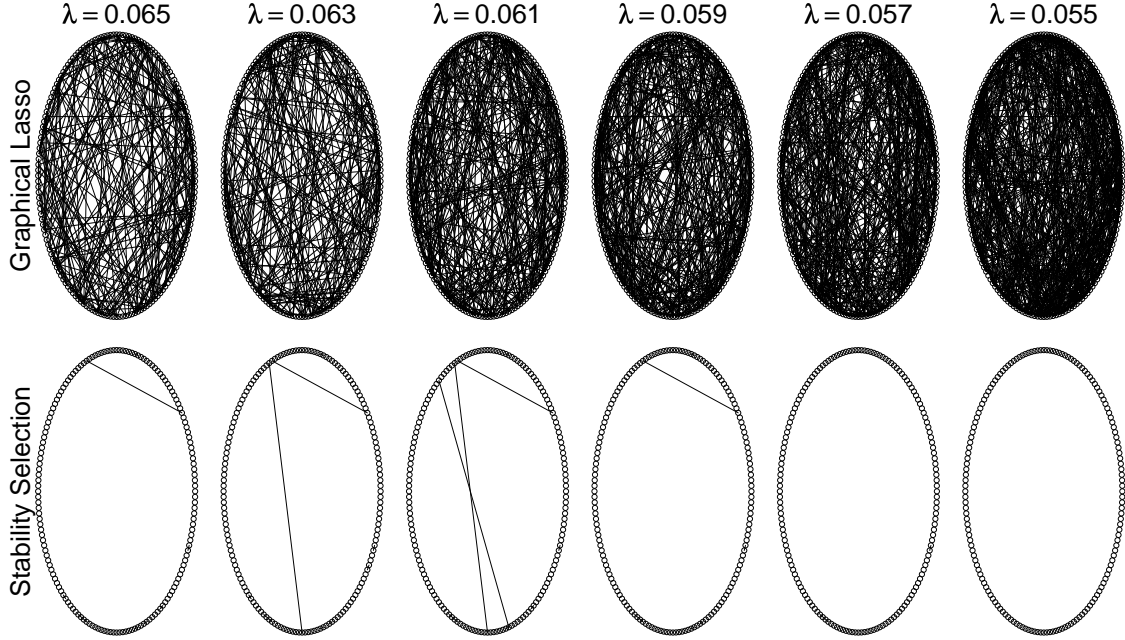


Figure 3: *The same plot as in Figure 2 but with the variables (expression values of each gene) permuted independently. The empty graph is the true model. With stability selection, only a few errors are made, as guaranteed by the made error control.*

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 Computational requirements

Stability selection demands to re-run $\{\hat{S}^\lambda; \lambda \in \Lambda\}$ multiple times. Evaluating selection probabilities over 100 subsamples seems sufficient in practice. The algorithmic complexity of Lasso in (2) or in (13) below is of the order $O(np \min\{n, p\})$, see Efron et al. (2004). In the $p > n$ regime, running the full Lasso path on subsamples of size $n/2$ is hence a quarter of the cost of running the algorithm on the full dataset and running 100 simulations is 25 times the cost of running a single fit on the full dataset. This cost could be compared with the cost of cross-validation, as this is what one has to resort to often in practice to select the regularisation parameter. Running 10-fold cross-validation uses approximately $10 \cdot 0.9^2 = 8.1$ as many computational resources as the single fit on the full dataset. Stability selection is thus roughly three times more expensive than 10-fold CV. This analysis is based on the fact that the computational complexity scales like $O(n^2)$ with the number of observations (assuming $p > n$). If computational costs would scale linearly with sample size (e.g. for Lasso with $p < n$), this factor would increase to roughly 5.5.

Stability selection with the Lasso (using 100 subsamples) for a dataset with $p = 1000$ and $n = 100$ takes about 10 seconds on a 2.2GHz processor, using the implementation of Friedman et al. (2007). Computational costs of this order would often seem worthwhile, given the potential benefits.

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,

$$Y = X\beta + \varepsilon, \quad (10)$$

with fixed $n \times p$ design matrix X and $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$. The predictor variables are normalised with $\|X_k\|_2 = (\sum_{i=1}^n (X_k^{(i)})^2)^{1/2} = 1$ for all $k \in \{1, \dots, p\}$. We allow for high-dimensional settings where $p \gg n$.

Stability selection is attractive for two reasons. First, the choice of a proper regularisation 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 regularisation. 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 for a procedure \hat{S} is understood to be equivalent to

$$P(\hat{S} = S) \rightarrow 1 \quad n \rightarrow \infty. \quad (11)$$

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 ‘neighbourhood 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; Zhang, 2009) 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 regularisation. Due to the restricted length of the paper, we will only discuss in detail the case of Lasso whereas the analysis of Orthogonal Matching Pursuit is just indicated.

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 randomised Lasso

The Lasso (Tibshirani, 1996; Chen et al., 2001) estimator is given in (2). For consistent variable selection using $\hat{S}^\lambda = \{k; \hat{\beta}_k^\lambda \neq 0\}$, it turns out that the design needs to satisfy the so-called

‘neighbourhood stability’ condition (Meinshausen and Bühlmann, 2006) which is equivalent to the ‘irrepresentable condition’ (Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007):

$$\max_{k \in N} |\text{sign}(\beta_S)^T (X_S^T X_S)^{-1} X_S^T X_k| < 1. \quad (12)$$

The condition in (12) is sufficient and (almost) necessary (the word ‘almost’ refers to the fact that a necessary relation is using ‘ \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 $\|\hat{\beta}^\lambda - \beta\|_2 \rightarrow_p 0$ for $n \rightarrow \infty$. 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; Huang et al., 2008). 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 randomised Lasso as an alternative. Despite its simplicity, it is consistent for variable selection even though the ‘irrepresentable condition’ in (12) is violated.

Randomised Lasso is a new generalisation of the Lasso. While the Lasso penalises the absolute value $|\beta_k|$ of every component with a penalty term proportional to λ , the randomised Lasso changes the penalty λ to a randomly chosen value in the range $[\lambda, \lambda/\alpha]$.

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

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

$$\hat{\beta}^{\lambda, W} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^p \frac{|\beta_k|}{W_k}. \quad (13)$$

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 randomised 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 (Meier et al., 2008; Friedman et al., 2007). 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 nonsensical at first sight since one can surely not expect any improvement from such a random perturbation. If applied only with one random perturbation, randomised Lasso is not very useful. However, applying randomised Lasso many times and looking for variables that are chosen often will turn out to be a very powerful procedure.

Consistency for randomised Lasso with stability selection. For stability selection with randomised Lasso, we can do without the irrepresentable condition (12) but need only a condition on the sparse eigenvalues of the design (Candes and Tao, 2007; van de Geer, 2008; Meinshausen and Yu, 2009; Bickel et al., 2007), also called the sparse Riesz condition in Zhang and Huang (2008).

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

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

and analogously for the maximal sparse eigenvalue ϕ_{\max} .

We have to constrain sparse eigenvalues to succeed.

Assumption 1 (Sparse eigenvalues) There exists some $C > 1$ and some $\kappa \geq 9$ such that

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

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

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

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 typically much weaker than the standard assumption of the ‘irrepresentable condition’ necessary to get results comparable to ours.

We have not specified the exact form of perturbations we will be using for the randomised Lasso in (13). For the following, we consider the randomised 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 Consider the model in (10). For randomised Lasso, let the weakness α be given by $\alpha^2 = \nu \phi_{\min}(m)/m$, for any $\nu \in ((7/\kappa)^2, 1/\sqrt{2})$, and $m = Cs^2$. Let a_n be a sequence with $a_n \rightarrow \infty$ for $n \rightarrow \infty$. Let $\lambda_{\min} = 2\sigma(\sqrt{2Cs} + 1)\sqrt{\log(p \vee a_n)/n}$. Assume that $p > 10$ and $s \geq 7$ and that the sparse eigenvalue Assumption 1 is satisfied. Then there exists some $\delta = \delta_s \in (0, 1)$ such that for all $\pi_{thr} \geq 1 - \delta$, stability selection with the randomised Lasso satisfies on a set Ω_A with $P(\Omega_A) \geq 1 - 5/(p \vee a_n)$ that no noise variables are selected,

$$N \cap \hat{S}_{\lambda}^{stable} = \emptyset, \quad (17)$$

where $\hat{S}_{\lambda}^{stable} = \{k : \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}$ with $\lambda \geq \lambda_{\min}$. On the same set Ω_A ,

$$(S \setminus S_{small;\lambda}) \subseteq \hat{S}_{\lambda}^{stable} \quad (18)$$

where $S_{small;\lambda} = \{k : |\beta_k| \leq 0.3(Cs)^{3/2}\lambda\}$. This implies that all variables with sufficiently large regression coefficient are selected.

Remark 2 Under the condition that the minimal non-zero regression coefficient is bounded from below by $\min_{k \in S} |\beta_k| \geq (Cs)^{3/2}(0.3\lambda)$, as a consequence of Theorem 2,

$$P(S = \hat{S}_\lambda^{\text{stable}}) \geq 1 - 1/(p \vee a_n),$$

i.e. consistent variable selection for $p \vee a_n \rightarrow \infty$ ($p \rightarrow \infty$ or $n \rightarrow \infty$) in the sense of (11) even if the irrepresentable condition (12) is violated. If no such lower bound holds, the set of selected variables might miss variables with too small regression coefficients, which are, by definition, in the set $S_{\text{small};\lambda}$.

Remark 3 Theorem 2 is valid for all $\lambda \geq \lambda_{\min}$. This is noteworthy as it means that even if the value of λ is chosen too large (i.e. considerably larger than λ_{\min}), no noise variables will be selected (formula (17)). Only some important variables might be missed. This effect has been seen in the empirical examples as stability selection is very insensitive to the choice of λ . In contrast, a hard-thresholded solution of the Lasso with a value of λ too large will lead to the inclusion of noise variables. Thus, stability selection with the randomised Lasso exhibits an important property of being conservative and guarding against false positive selections.

Remark 4 Theorem 2 is derived under random perturbations of the weights. While this achieves good empirical results, it seems more advantageous in combination with subsampling of the data. The results extend directly to this case. Let $\tilde{\Pi}_k^\lambda$ be the selection probability of variable $k \in S \setminus S_{\text{small};\lambda}$, while doing both random weight perturbations and subsampling $n/2$ out of n observations. The probability that $\tilde{\Pi}_k^\lambda$ is above the threshold $\pi_{\text{thr}} \in (0, 1)$ is bounded by a Markov-type inequality from below by

$$P(\tilde{\Pi}_k^\lambda \geq \pi_{\text{thr}}) \geq \frac{E(\tilde{\Pi}_k^\lambda) - \pi_{\text{thr}}}{1 - \pi_{\text{thr}}} \geq 1 - \frac{5}{(p \vee a_{n/2})(1 - \pi_{\text{thr}})},$$

having used that $E(\tilde{\Pi}_k^\lambda) \geq 1 - 5/(p \vee a_{n/2})$ as a consequence of Theorem 2. If $5/(p \vee a_{n/2})$ is sufficiently small in comparison to $1 - \pi_{\text{thr}}$, this elementary inequality implies that important variables in $S \setminus S_{\text{small};\lambda}$ are still chosen by stability selection (subsampling and random weights perturbation) with very high probability. A similar argument shows that noise variables are also still not chosen with very high probability. Empirically, combining random weight perturbations with subsampling yields very competitive results and this is what we recommend to use.

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.

Relation to other work. In related and very interesting work, Bach (2008) has proposed ‘Bo-lasso’ (for bootstrapped enhanced Lasso) and shown that using a finite number of subsamples of the original Lasso procedure and applying basically stability selection with $\pi_{\text{thr}} = 1$ yields consistent

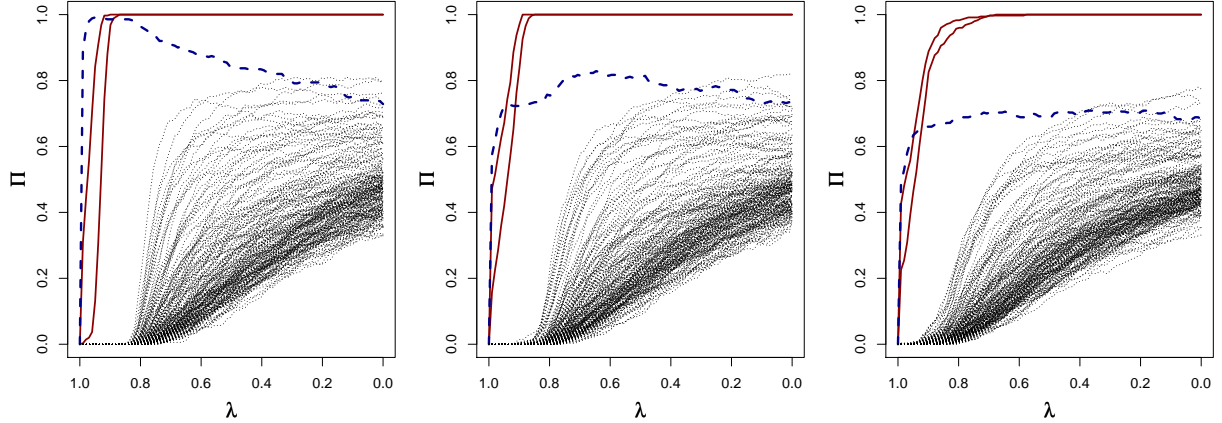


Figure 4: The stability paths for randomised Lasso with stability selection using weakness parameters $\alpha = 1$ (left panel identical to the original Lasso) and $\alpha = 0.5$ (middle) and $\alpha = 0.2$ (right). Red solid lines are the coefficients of the first two (relevant) variables). The blue broken line is the coefficient of the third (irrelevant) variable and the dotted lines are the coefficients from all other (irrelevant) variables. Introducing the randomised version helps avoid choosing the third (irrelevant) predictor variable.

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 randomised Lasso, no false variable is selected for all sufficiently large values of λ , see Remark 3. In other words, if λ is chosen ‘too large’ with randomised 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 4 is a good illustration. Picking the regularisation 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 (Candes and Tao, 2007; Meinshausen and Yu, 2009), where variables are thresholded. Picking λ too large (and λ is notoriously difficult to choose), false variables will invariably enter the model. In contrast, stability selection with randomised Lasso is not picking wrong variables if λ is chosen too large.

3.2 Example

We illustrate the results on randomised 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 $\beta = (1, 1, 0, 0, \dots, 0)$. The response Y is obtained from the linear model $Y = X\beta + \varepsilon$ in (1), where $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, 1/4)$. For $\rho > 0.5$, the irrerepresentable condition in (12) is violated and Lasso is not able to correctly identify the first two variables as the truly important ones, since it always includes the third variable superfluously as well. Using the randomised 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 randomised Lasso in Figure 4. This allows to separate relevant and irrelevant variables. And indeed, the randomised Lasso is consistent under stability selection.

3.3 Randomised orthogonal Matching Pursuit

An interesting alternative to Lasso or greedy forward search in this context are the recently proposed forward-backward search FOBA (Zhang, 2008) and the MC+ algorithm (Zhang, 2007), which both provably lead to consistent variable selection under weak conditions on sparse eigenvalues, despite being greedy solutions to non-convex optimisation problems. It will be very interesting to explore the effect of stability selection on these algorithms, but this is beyond the scope of this paper.

Here, we look instead at orthogonal matching pursuit, a greedy forward search in the variable space. The iterative SIS procedure (Fan and Lv, 2008), entails orthogonal matching pursuit as a special case. We will examine the effect of stability selection under subsampling and additional randomisation. To have a clear definition of randomised orthogonal matching pursuit (ROMP), we define it as follows.

Randomised orthogonal matching pursuit with weakness $0 < \alpha < 1$ and q iterations.

1. Set $R_1 = Y$. Set $m = 0$ and $\hat{S}^0 = \emptyset$.
2. For $m = 1, \dots, q$:
 - (a) Find $\rho_{\max} = \max_{1 \leq k \leq p} |X_k^T R_m|$
 - (b) Define $K = \{k : |X_k^T R| \geq \alpha \rho_{\max}\}$.
 - (c) Select randomly a variable k_{sel} in the set K and set $\hat{S}^m = \hat{S}^{m-1} \cup \{k_{sel}\}$.
 - (d) Let $R_{m+1} = Y - P_m Y$, where the projection P_m is given by $X_{\hat{S}^m} (X_{\hat{S}^m}^T X_{\hat{S}^m})^{-1} X_{\hat{S}^m}^T$.
3. Return the selected sets $\hat{S}^1 \subset \hat{S}^2 \subset \dots \subset \hat{S}^q$.

A drawback of OMP is clearly that conditions for consistent variable selection are quite strong. Following Tropp (2004), the exact recovery condition for OMP is defined as

$$\max_{k \in N} \|(X_S^T X_S)^{-1} X_S^T X_k\|_1 < 1. \quad (19)$$

This is a sufficient condition for consistent variable selection. If it is not fulfilled, there exist regression coefficients that cause OMP or its weak variant to fail in recovery of the exact set S of relevant variables. Surprisingly, this condition is rather similar to the irrepresentable (Zhao and Yu, 2006) or neighbourhood stability condition (Meinshausen and Bühlmann, 2006).

In the spirit of Theorem 2, we have also a proof that stability selection for randomised Orthogonal Matching Pursuit (ROMP) is asymptotically consistent for variable selection in linear models, even if the right hand side in (19) is not bounded by 1 but instead by a possibly large constant (assuming the weakness α is low enough). This indicates that stability selection has a more general potential for improved structure estimation, beyond the case for the Lasso presented in Theorem 2.

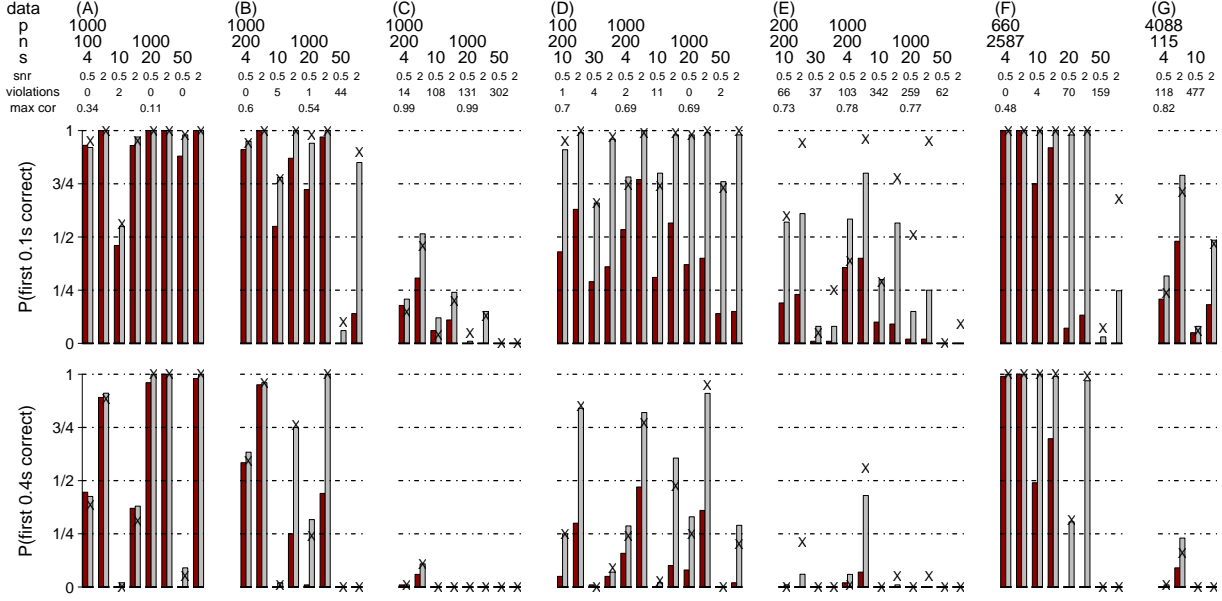


Figure 5: *Probability to select 0.1s and 0.4s important variables without selecting a noise variable with the Lasso in the regression setting (dark red bar) and stability selection under subsampling (light grey bar) for the 64 different settings. Black crosses mark the result for stability selection with additional randomisation ($\alpha = 0.5$).*

It is noteworthy that our proof involves artificial adding of noise covariates. In practice, this seems to help often but a more involved discussion is beyond the scope of this paper. We will give empirical evidence for the usefulness of stability selection under subsampling and additional randomisation for orthogonal matching pursuit in the numerical examples below.

4 Numerical Results

To investigate further the effects of stability selection, we focus here on the application of stability selection to Lasso and randomised Lasso for both regression and the natural extension to classification. The effect on OMP and randomised OMP will also be examined.

For regression (Lasso and OMP), we generate observations by $Y = X\beta + \varepsilon$. For classification, we use the logistic linear model under the binomial family. To generate the design matrices X , we use two real and five simulated datasets,

- (A) Independent predictor variables. All $p = 1000$ predictor variables are i.i.d. standard normal distributed. Sample size $n = 100$ and $n = 1000$.
- (B) Block structure with 10 blocks. The $p = 1000$ -dimensional predictor variable follows a $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{km} = 0$ for all pairs (k, m) except if $\text{mod}_{10}k = \text{mod}_{10}m$, for which $\Sigma_{km} = 0.5$. Sample size $n = 200$ and $n = 1000$.
- (C) Toeplitz design. The $p = 1000$ -dimensional predictor variable follows a $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{km} = \rho^{|k-m|}$ and $\rho = 0.99$. Sample size $n = 200$ and $n = 1000$.

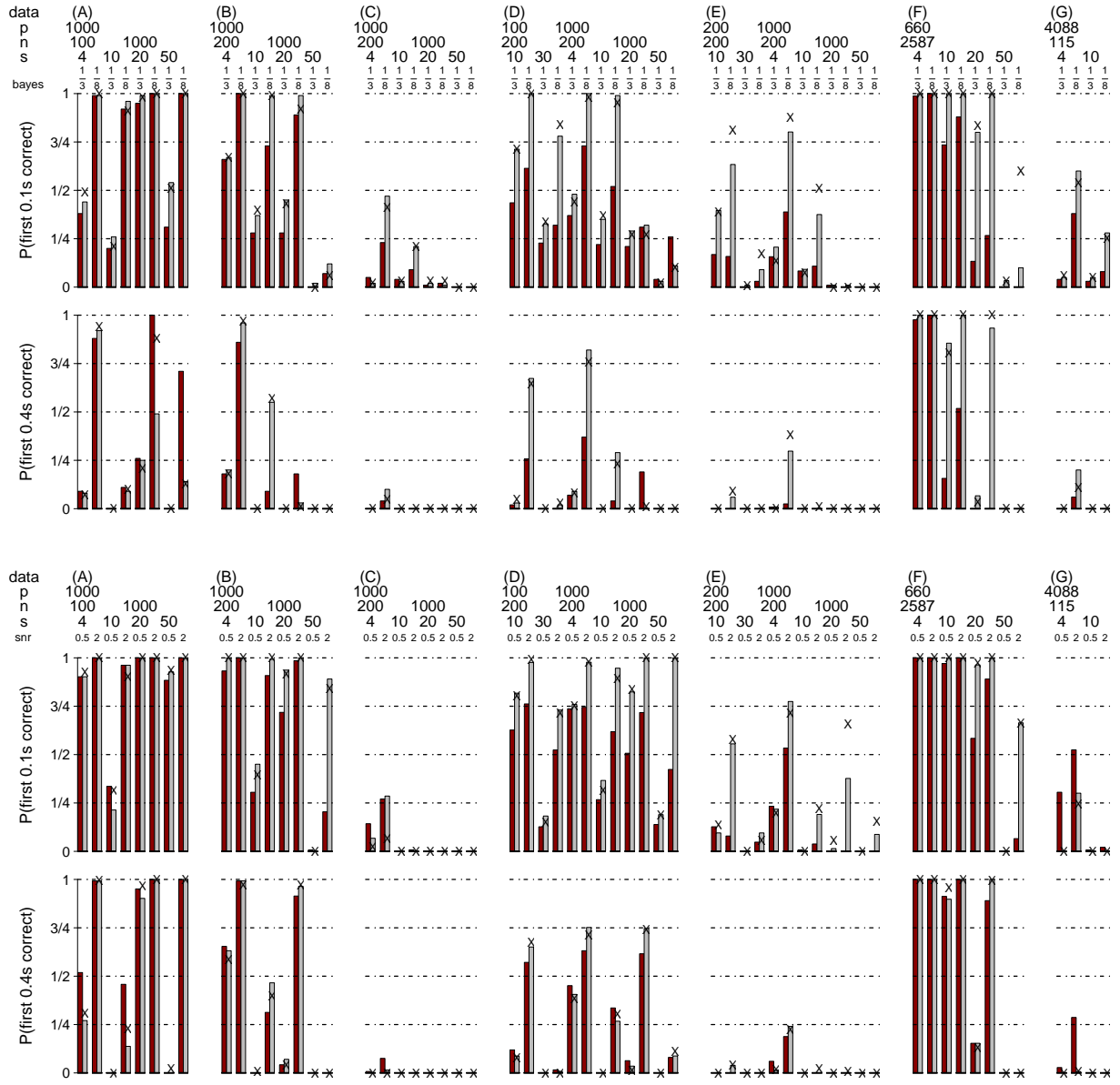


Figure 6: The equivalent plot to Fig. 2 for Lasso applied to classification (top two rows) and OMP applied to regression (bottom two rows).

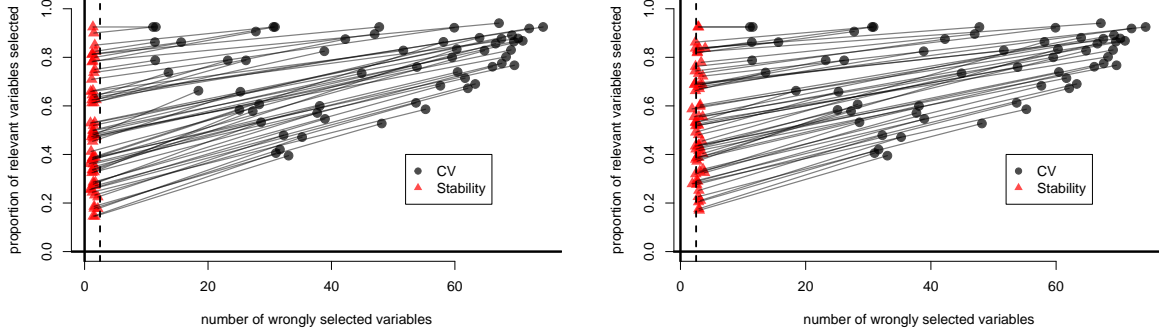


Figure 7: Comparison of stability selection with cross-validation for the real datasets (F) and (G). The cross-validated solution (for standard Lasso) is indicated by a dot and the corresponding stability selection (for randomised Lasso, $\alpha = 0.5$ on the left and $\alpha = 1$ on the right) by a red triangle, showing the average proportion of correctly identified relevant variables versus the average number of falsely selected variables. Each pair consisting of a dot and triangle corresponds to a simulation setting (some specified SNR and s). 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.

- (D) Factor model with 2 factors. Let ϕ_1, ϕ_2 be two latent variables following i.i.d. standard normal distributions. Each predictor variable X_k , for $k = 1, \dots, p$, is generated as $X_k = f_{k,1}\phi_1 + f_{k,2}\phi_2 + \eta_k$, where $f_{k,1}, f_{k,2}, \eta_k$ have i.i.d. standard normal distributions for all $k = 1, \dots, p$. Sample sizes are $n = 200$ and $n = 1000$, while $p = 1000$.
- (E) Identical to (D) but with 10 instead of 2 factors.
- (F) Motif regression dataset. A dataset ($p = 660$ and $n = 2587$) about finding transcription factor binding sites (motifs) in DNA sequences. 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).
- (G) The already mentioned vitamin gene expression data (with $p = 4088$ and $n = 158$) described in Section 2.2.

We do not use the response values from the real 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 $k = 1, \dots, p$, except for a randomly chosen set S of coefficients, where β_k is chosen independently and uniformly in $[0, 1]$ for all $k \in S$. The size $s = |S|$ of the active set is varied between 4 and 50, depending on the dataset. For regression, the noise vector $(\varepsilon_1, \dots, \varepsilon_n)$ is chosen i.i.d. $\mathcal{N}(0, \sigma^2/n)$, 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 noise level σ^2 is chosen to achieve signal-to-noise ratios (SNR) of 0.5 and 2. For classification, we scale the vector β to achieve a given Bayes misclassification rate, either $1/8$ or $1/3$. Each of the 64 scenarios is run 100 times,

once using the standard procedure (Lasso or OMP), once using stability selection with subsampling and once using stability selection with subsampling and additional randomisation ($\alpha = 0.5$ for the randomised Lasso and $\alpha = 0.9$ for randomised OMP). The methods are thus in total evaluated on about 20.000 simulations each.

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 probability that $\gamma \cdot s$ of the s relevant variables can be recovered without error, where $\gamma \in \{0.1, 0.4\}$. A set of γs variables is said to be recovered successfully for the Lasso or OMP selection, if there exists a regularisation parameter such that at least $\lceil \gamma s \rceil$ variables in S have a non-zero regression coefficient and all variables in $N = \{1, \dots, p\} \setminus S$ have a zero regression coefficient. For stability selection, recovery without error means that the $\lceil \gamma s \rceil$ variables with highest selection probability $\max_{\lambda \geq \lambda_{\min}} \hat{\beta}_k^\lambda$ are all in S . The value λ_{\min} is chosen such that at most $\sqrt{0.8p}$ variables are selected in the whole path of solutions for $\lambda \geq \lambda_{\min}$. Note that this notion neglects the fact that the most advantageous regularisation parameter is selected here automatically for Lasso and OMP but not for stability selection.

Results are shown in Figure 5 for Lasso applied to regression, and in Figure 6 for Lasso applied to classification and OMP applied to regression again. In Figure 5, we also give the median number of variables violating the irrerepresentable condition (denoted by ‘violations’) and the average of the maximal correlation between a randomly chosen variable and all other variables (‘max cor’) as two measures of the difficulty of the problem.

Stability selection identifies as many or more correct variables than the underlying method itself in all cases except for scenario (A), where it is about equivalent. That stability selection is not advantageous for scenario (A) is to be expected as the design is nearly orthogonal (very weak empirical correlations between variables), thus almost decomposing into p univariate decisions and we would not expect stability selection to help in a univariate framework.

Often the gain of stability selection under subsampling is substantial, irrespective of the sparsity of the signal and the signal-to-noise-ratio. Additional randomisation helps in cases where there are many variables violating the irrerepresentable condition; for example in setting (E). This is in line with our theory.

Next, we test how well the error control of Theorem 1 holds up for these datasets. For the motif regression dataset (F) and the vitamin gene expression dataset (G), Lasso is applied, with randomisation and without. For both datasets, the signal-to-noise ratio is varied between 0.5, 1 and 2. The number of non-zero coefficients s is varied in steps of 1 between 1 and 12, with a standard normal distribution for the randomly chosen non-zero coefficients. Each of the 72 settings is run 20 times. We are interested in the comparison between the cross-validated solution and 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 7. 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 regularisation.

5 Discussion

Stability selection addresses the notoriously difficult problem of structure estimation or variable selection, especially for high-dimensional problems. 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 and Gaussian graphical modelling.

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 regularisation for structure estimation or variable selection. Furthermore, the solution of stability selection depends surprisingly little on the chosen initial regularisation. This 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 rather simple settings. We prove for randomised Lasso that stability selection will be variable 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 fails.

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. 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 $\{1, \dots, n\}$ 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 $\hat{S}^\lambda(I_2)$,

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

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

$$\hat{\Pi}_K^{simult, \lambda} = P^*(K \subseteq \hat{S}^{simult, \lambda}), \quad (20)$$

where the probability P^* is with respect to the random sample splitting (and any additional randomness if \hat{S}^λ is a randomised 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 latter is used for the proof of Theorem 1. The bound is rather tight for selection probabilities close to 1.

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

$$\hat{\Pi}_K^{simult, \lambda} \geq 2\hat{\Pi}_K^\lambda - 1. \quad (21)$$

Proof. Let I_1 and I_2 be the two random subsets in sample splitting of $\{1, \dots, n\}$ 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 $\{1, \dots, n\}$ into I_1 and I_2 . The probabilities $s_K(\{1, 0\})$, $s_K(\{0, 1\})$, $s_K(\{0, 0\})$ are defined equivalently by $P^*(\{K \subseteq \hat{S}^\lambda(I_1)\} \cap \{K \not\subseteq \hat{S}^\lambda(I_2)\})$, $P^*(\{K \not\subseteq \hat{S}^\lambda(I_1)\} \cap \{K \subseteq \hat{S}^\lambda(I_2)\})$, and $P^*(\{K \not\subseteq \hat{S}^\lambda(I_1)\} \cap \{K \not\subseteq \hat{S}^\lambda(I_2)\})$. 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. \square

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}^Λ 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}^λ , $\lambda \in \Lambda$. Now using Lemma 2 below, it follows that $P(\max_{\lambda \in \Lambda} \hat{\Pi}_k^{simult, q} \geq \xi) \leq (q_\Lambda/p)^2/\xi$ for all $0 < \xi < 1$ and $k \in N$. Using Lemma 1, it follows that $P(\max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda \geq \pi_{thr}) \leq P((\max_{\lambda \in \Lambda} \hat{\Pi}_k^{simult, \lambda} + 1)/2 \geq \pi_{thr}) \leq (q_\Lambda/p)^2/(2\pi_{thr} - 1)$. 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. \square

Lemma 2 Let $K \subset \{1, \dots, p\}$ and \hat{S}^λ 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 $I_1, I_2 \subseteq \{1, \dots, n\}$ be, as above, the random split of the samples $\{1, \dots, n\}$ 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 $K \subseteq \{1, \dots, p\}$ 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 (20), 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}^λ is a randomised 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 $P(H_K^\lambda = 1) \leq P(K \subseteq \hat{S}^\lambda(I_1))^2 \leq \varepsilon^2$ and hence $E(H_K^\lambda) \leq \varepsilon^2$. Therefore, $E(H_K^\lambda) = E(E(H_K^\lambda|Z)) = E(\hat{\Pi}_K^{simult,\lambda}) \leq \varepsilon^2$. Using a Markov-type inequality, $\xi P(\hat{\Pi}_K^{simult,\lambda} \geq \xi) \leq E(\hat{\Pi}_K^{simult,\lambda}) \leq \varepsilon^2$. Thus $P(\hat{\Pi}_K^{simult,\lambda} \geq \xi) \leq \varepsilon^2/\xi$, completing the proof of the first claim. The proof of the second part follows analogously. \square

6.3 Proof of Theorem 2

Instead of working directly with form (13) of the randomised 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 $X_k^w = X_k \cdot W_k$ for each $k = 1, \dots, p$. Let ϕ_{\max}^w and ϕ_{\min}^w 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 randomised 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 \bar{C} by $(2 + 4\bar{C})s + 1 = Cs^2$ and assume $s \geq 7$. 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 $\alpha^2 \geq \nu \phi_{\min}(Cs^2)/(Cs^2)$, with $\nu \in \mathbb{R}^+$, it holds under Assumption 1 for all random realisations W that

$$\frac{\phi_{\max}^w(Cs^2)}{\phi_{\min}^w(Cs^2)} \leq \frac{7\bar{C}}{\kappa\sqrt{\nu}}. \quad (22)$$

Proof. Using Assumption 1,

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

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

$$\frac{\phi_{\max}(Cs^2)}{\phi_{\min}(Cs^2)} \leq \frac{3 + 4\bar{C}}{\kappa} \sqrt{\frac{\phi_{\min}(Cs^2)}{Cs^2}}. \quad (23)$$

Now, let \mathcal{W} be again the $p \times p$ -diagonal matrix with diagonal entries $\mathcal{W}_{kk} = W_k$ for all $k = 1, \dots, p$ and 0 on the non-diagonal elements. Then $X^w = X\mathcal{W}$ and, taking suprema over all \mathcal{W} with diagonal entries in $[\alpha, 1]$,

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

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 $\phi_{\min}^w(m)$ yields the bound $\phi_{\min}^w(m) \geq \alpha \phi_{\min}(m)$ for all $m \in \mathbb{N}$. The claim (22) follows by observing that $\bar{C} \geq 1$ for $s \geq 7$, since $C \geq 1$ by Assumption 1 and hence $3 + 4\bar{C} \leq 7\bar{C}$. \square

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

$$|\hat{A}^{\lambda, w} \cup S| \leq Cs^2 \text{ and } (S \setminus S_{\text{small}; \lambda}) \subseteq \hat{A}^{\lambda, w}, \quad (24)$$

where $S_{\text{small}; \lambda}$ is defined as in Theorem 2.

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

$$\frac{\phi_{\max}^w((2 + 4\bar{C})s + 1)}{\phi_{\min}^w((2 + 4\bar{C})s + 1)} \leq \bar{C}.$$

Hence all assumptions of Theorem 1 in Zhang and Huang (2008) are fulfilled, with $\eta_1 = 0$, for any random realisation $W = w$. 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 $P(Y \in \Omega_0) \geq 2 - \exp(2/(p \vee a_n)) - 2/(p \vee a_n)^2 \geq 1 - 3/(p \vee a_n)$ for all $p \geq 5$, such that if $Y \in \Omega_0$, from (2.21) in Zhang and Huang (2008),

$$|\hat{A}^{\lambda, w} \cup S| \leq (2 + 4\bar{C})s \leq Cs^2, \quad (25)$$

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

$$\sum_{k \in S} |\beta_k|^2 1\{k \notin \hat{A}^{\lambda, w}\} \leq \left(\frac{2}{3}\bar{C} + \frac{28}{9}\bar{C}^2 + \frac{16}{9}\bar{C}^3\right)s\lambda^2 \leq 5.6\bar{C}^3 s^3 \lambda^2 \leq (0.3(Cs)^{3/2}\lambda)^2, \quad (26)$$

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 normalisation. For the second inequality, we used $4\bar{C} \leq Cs$. The last inequality implies, by definition of $S_{\text{small}; \lambda}$ in Theorem 2, that $S \setminus S_{\text{small}; \lambda} \subseteq \hat{A}^{\lambda, w}$, which completes the proof. \square

Lemma 5 Set $m = Cs^2$. Let $k \in \{1, \dots, p\}$ and let $K(w) \subseteq \{1, \dots, p\}$ be a set which can depend on the random weight vector W . Suppose that $K(w)$ satisfies $|K(w)| \leq m$ and $k \notin K(w)$ for all realisations $W = w$. Suppose furthermore that $K(w) = A$ for some $A \subseteq \{1, \dots, p\}$ implies that $K(v) = A$ for all pairs $w, v \in \mathbb{R}^p$ of weights that fulfill $v_j \leq w_j$ for all $j \in \{1, \dots, p\}$, with equality for all $j \in A$. Then, for $\alpha^2 \leq \phi_{\min}(m)/(\sqrt{2}m)$,

$$P_w(\|(X_{K(w)}^w)^T X_{K(w)}^w)^{-1} (X_{K(w)}^w)^T X_k^w\|_1 \leq 2^{-1/4}) \geq p_w(1 - p_w)^m. \quad (27)$$

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 \tilde{w} be the realisation of W for which $\tilde{w}_k = \alpha$ and $\tilde{w}_j = 1$ for all other $j \in \{1, \dots, p\} \setminus k$. The probability of $W = \tilde{w}$ is clearly $p_w(1 - p_w)^{p-1}$ under the used sampling scheme for the weights. Let $A := K(\tilde{w})$ be the selected set of variables under these weights. Let now $\mathcal{W} \subseteq \{1, \alpha\}^p$ be the set of all weights for which $w_k = \alpha$ and $w_j = 1$ for all $j \in A$, and arbitrary values in $\{\alpha, 1\}$ for all w_j with $j \notin A \cup k$. The probability for a random weight being in this set is $P_w(w \in \mathcal{W}) = p_w(1 - p_w)^{|A|}$. By the assumption on K , it holds that $K(w) = A$ for all $w \in \mathcal{W}$, since $w_j \leq \tilde{w}_j$ for all $j \in \{1, \dots, p\}$ with equality for $j \in A$. For all weights $w \in \mathcal{W}$, it follows moreover that

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

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

$$\sup_{A: |A| \leq m} \sup_{k \notin A} \|(X_A^T X_A)^{-1} X_A^T X_k\|_1^2 \leq m/\phi_{\min}(m). \quad (28)$$

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

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

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 $\|X_A \gamma\|_2^2 \leq 1$. Using $\|X_S \gamma\|_2^2 = \gamma^T (X_A^T X_A) \gamma \geq \phi_{\min}(|A|)\|\gamma\|_2^2$, it follows that $\|\gamma\|_2^2 \leq 1/\phi_{\min}(|A|)$, which shows (28) and thus completes the proof. \square

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 $P(\Omega_1) \geq 1 - 2/(p \vee a_n)$, such that for all $\omega \in \Omega_1$,

$$\sup_{A: |A| \leq m} \sup_{k \notin A} |X_k^T (1 - P_A) \varepsilon| < 2\sigma(\sqrt{2m} + 1)\sqrt{\log(p \vee a_n)/n}. \quad (29)$$

Proof. Let Ω'_1 be the event that $\max_{k \in \{1, \dots, p\}} |X_k^T \varepsilon| \leq 2\sigma\sqrt{\log(p \vee a_n)/n}$. As entries in ε are i.i. $\mathcal{N}(0, \sigma^2)$ distributed, $P(\Omega'_1) \geq 1 - 1/(p \vee a_n)$ for all $\delta \in (0, 1)$. Note that, for all $A \subset \{1, \dots, p\}$ and $k \notin A$, $|X_k^T P_A \varepsilon| \leq \|P_A \varepsilon\|_2$. Define Ω''_1 as

$$\sup_{|A| \leq m} \|P_A \varepsilon\|_2 \leq 2\sigma\sqrt{2m \log(p \vee a_n)/n}. \quad (30)$$

It is now sufficient to show that $P(\Omega''_1) \geq 1 - 1/(p \vee a_n)$. Showing this bound 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),

$$P\left(n\|P_A\varepsilon\|_2^2/\sigma^2 \geq |A|(1+4\log(p \vee a_n))\right) \leq ((p \vee a_n)^{-4}(1+4\log(p \vee a_n)))^{|A|/2} \leq (p \vee a_n)^{-3|A|/2},$$

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

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

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

Lemma 7 *Let $\delta_w = p_w(1 - p_w)^{Cs^2}$ and $\hat{\Pi}_k^\lambda = P_w(k \in \hat{A}^{\lambda,W})$ be again 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 $p > 10$, there exists a set Ω_A with $P(\Omega_A) \geq 1 - 5/(p \vee a_n)$ such that for all $\omega \in \Omega_A$ and $\lambda \geq \lambda_{\min}$,*

$$\max_{k \in N} \hat{\Pi}_k^\lambda < 1 - \delta_w \quad (31)$$

$$\min_{k \in S \setminus S_{\text{small};\lambda}} \hat{\Pi}_k^\lambda \geq 1 - \delta_w, \quad (32)$$

where $S_{\text{small};\lambda}$ is defined as in Theorem 2.

Proof. We let $\Omega_A = \Omega_0 \cap \Omega_1$, where Ω_0 is the event defined in Lemma 4 and event Ω_1 is defined in Lemma 6. Since, using these two lemmas,

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

it is sufficient to show (31) and (32) for all $\omega \in \Omega_0 \cap \Omega_1$. We begin with (31). A variable $k \notin S$ is in the selected set $\hat{A}^{\lambda,W}$ only if

$$|(X_k^w)^T(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k})| \geq \lambda, \quad (33)$$

where $\hat{\beta}^{\lambda,W,-k}$ is the solution to (13) with the constraint that $\hat{\beta}_k^{\lambda,W,-k} = 0$, comparable to the analysis in Meinshausen and Bühlmann (2006). Let $\hat{A}^{\lambda,W,-k} := \{j : \hat{\beta}_j^{\lambda,W,-k} \neq 0\}$ be the set of non-zero coefficients and $\hat{B}^{\lambda,W,-k} := \hat{A}^{\lambda,W,-k} \cup S$ be the set of regression coefficients which are either truly non-zero or estimated as non-zero (or both). We will use \hat{B} as a short-hand notation for $\hat{B}^{\lambda,W,-k}$. Let $P_{\hat{B}}^w$ be the projection operator into the space spanned by all variables in the set \hat{B} . For all $W = w$, this is identical to

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

Then, splitting the term $(X_k^w)^T(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k})$ in (33) into the two terms

$$(X_k^w)^T(1 - P_{\hat{B}}^w)(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k}) + (X_k^w)^T P_{\hat{B}}^w(Y - X_{-k}^w \hat{\beta}^{\lambda,W,-k}), \quad (34)$$

it holds for the right term in (34) that

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

Looking at the left term in (34), since $Y \in \Omega_0$, we know by Lemma 4 that $|\hat{B}| \leq Cs^2$ and, by definition of \hat{B} above, $S \subseteq \hat{B}$. Thus the left term in (34) is bounded from above by

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

having used Lemma 6 in the last step and $\lambda_{\min} = 2\sigma(\sqrt{2Cs} + 1)\sqrt{\log(p \vee a_n)/n}$. Putting together, the two terms in (34) are bounded, for all $\omega \in \Omega_0 \cap \Omega_1$, by

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

We now apply Lemma 5 to the rightmost term. The set \hat{B} is a function of the weight vector and satisfies for every realisation of the observations $Y \in \Omega_0$ the conditions in Lemma 5 on the set $K(w)$. First, $|\hat{B}| \leq Cs^2$. Second, by definition of \hat{B} above, $k \notin \hat{B}$ for all weights w . Third, it follows by the KKT conditions for Lasso that the set of non-zero coefficients of $\hat{\beta}^{\lambda, w, -k}$ and $\hat{\beta}^{\lambda, v, -k}$ is identical for two weight vectors w and v , as long $v_j = w_j$ for all $j \in \hat{A}^{\lambda, W, -k}$ and $v_j \leq w_j$ for all $j \notin \hat{A}^{\lambda, W, -k}$ (increasing the penalty on zero coefficients will leave them at zero, if the penalty for non-zero coefficients is kept constant). Hence there exists a set Ω_w in the sample space of W with $P_w(\Omega_w) \geq 1 - \delta_w$ such that $\|((X_{\hat{B}}^w)^T X_{\hat{B}}^w)^{-1} (X_{\hat{B}}^w)^T X_k^w\|_1 \leq 2^{-1/4}$. 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 (33) 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 (31) of the proof.

For the second part (32), we need to show that, for all $\omega \in \Omega_0 \cap \Omega_1$, all variables k in S are chosen with probability at least $1 - \delta_w$ (with respect to random sampling of the weights W), except possibly for variables in $S_{\text{small}; \lambda} \subseteq S$, defined in Theorem 2. For all $\omega \in \Omega_0$, however, it follows directly from Lemma 4 that $(S \setminus S_{\text{small}; \lambda}) \subseteq \hat{A}^{\lambda, W}$. Hence, for all $k \in S \setminus S_{\text{small}; \lambda}$, the selection probability satisfies $\hat{\Pi}_k^\lambda = 1$ for all $Y \in \Omega_0$, which completes the proof. \square

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

Acknowledgments

Both authors would like to thank anonymous referees for many helpful comments and suggestions which greatly helped to improve the manuscript. 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.
- Barbieri, M. and J. Berger (2004). Optimal predictive model selection. *Annals of Statistics* 32, 870–897.
- Bhattacharjee, A., W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al. (2005). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* 102, 3301–3307.
- Bickel, P. and E. Levina (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* 36, 199–227.
- 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. *Annals of Statistics* 30, 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 Sciences* 100, 3339 – 3344.
- Davis, C., F. Gerick, V. Hintermair, C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer (2006). Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics* 22, 2356–2363.
- Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 -minimization. *Proceedings of the National Academy of Sciences* 100, 2197–2202.
- Dudoit, S., J. Shaffer, and J. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* 18, 71–103.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics* 32, 407–451.
- Ein-Dor, L., I. Kela, G. Getz, D. Givol, and E. Domany (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21, 171–178.
- Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B (with discussion)* 70, 849–911.
- 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 (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.
- Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 18, 1603–1618.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Lee, K., N. Sha, E. Dougherty, M. Vannucci, and B. Mallick (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* 19, 90–97.
- 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. *IEEE Transactions on Signal Processing* 41, 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 (2009). Lasso-type recovery of sparse representations from high-dimensional data. *Annals of Statistics* 37, 246–270.
- Michiels, S., S. Koscielny, and C. Hill (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet* 365, 488–492.
- Monti, S., P. Tamayo, J. Mesirov, and T. Golub (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 52, 91–118.
- Rothman, A., P. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.
- Sha, N., M. Vannucci, M. Tadesse, P. Brown, I. Dragoni, N. Davies, T. Roberts, A. Contestabile, M. Salmon, C. Buckley, and F. Falciani (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* 60, 812–819.
- 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. *IEEE Transactions on Information Theory* 50, 2231–2242.
- Valdar, W., C. Holmes, R. Mott, and J. Flint (2009). Mapping in structured populations by resample-based model averaging. *Genetics*, to appear.
- van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics* 36, 614–645.
- 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, 1567–1594.
- Zhang, C.-H. (2007). Penalized linear unbiased selection. *Technical Report No. 2007-003, Department of Statistics, Rutgers University*.
- Zhang, T. (2008). Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models. In *Proceedings of Neural Information Processing Systems*.
- Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research* 10, 555–568.
- 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.
- Zucknick, M., S. Richardson, and E. Stronach (2008). Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical Applications in Genetics and Molecular Biology* 7.