Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 4, *pp.* 417–473

Stability selection

Nicolai Meinshausen

University of Oxford, UK

and Peter Bühlmann

Eidgenössiche Technische Hochschule Zürich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 3rd, 2010*, Professor D. M. Titterington *in the Chair*]

Summary. 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 regularization for structure estimation. Variable selection and structure estimation improve markedly for a range of selection methods if stability selection is applied. We prove for the randomized lasso that stability selection will be variable selection consistent even if the necessary conditions 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.

Keywords: High dimensional data; Resampling; Stability selection; Structure estimation

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 data sets 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 optimization problems are discrete, as well as determining the right amount of regularization, e.g. in an asymptotic sense for consistent structure estimation. Substantial progress has been made over recent 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

Address for correspondence: Nicolai Meinshausen, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
E-mail: meinshausen@stats.ox.ac.uk

© 2010 Royal Statistical Society 1369–7412/10/72417

some discrete optimization problems, e.g. by l_1-penalty approaches (Donoho and Elad, 2003; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; 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 regularization (for the procedures that were mentioned above and for many others) for obtaining a right-sized structure or model has largely remained a problem with unsatisfactory solutions.

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

1.1. Preliminaries and examples

In general, let β be a p-dimensional vector, where β is sparse in the sense that $s < p$ components are 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 $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ with univariate response variable Y and p-dimensional covariates X. We typically assume that $(X^{(i)}, Y^{(i)})$s are independent and identically distributed (IID). The vector β could be the coefficient vector in a linear model

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

where $Y = (Y_1, \ldots, Y_n)$, X is the $n \times p$ design matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the random noise whose components are IID. Thus, inferring the set S from data is the well-studied variable selection problem in linear regression. A main stream of classical methods proceeds to solve this problem by penalizing the negative log-likelihood with the l_0-norm $\|\beta\|_0$ which equals the number of non-zero components of β. The computational task to solve such an l_0-norm penalized optimization problem becomes quickly unfeasible if p is growing 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),

$$\hat{\beta}^{\lambda} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right), \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter and we typically assume that the covariates are on the same scale, i.e. $\|X_k\|_2 = \Sigma_{i=1}^n (X_k^{(i)})^2 = 1$. An attractive feature of the lasso is its computational feasibility for large p since the optimization problem (2) is convex. Furthermore, the lasso can 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 optimization 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, 2009; Yuan and Lin, 2007), and we present the details in Section 3.1. Among the challenges are the issue of choosing a proper amount of regularization λ for consistent 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 are assumed to be

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

The goal is to infer conditional dependences among the d variables or components in $X = (X_1, \ldots, X_d)$. It is well known that X_j and X_k are conditionally dependent given all other components $\{X_{(l)}; l \neq j, k\}$ 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 $\mathcal{G} = \{(j,k); 1 \leqslant j < k \leqslant d\}$ 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 dependences 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 difficult 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):

$$\hat{\Theta}^\lambda = \underset{\Theta \text{ nonneg.def.}}{\arg\min} [-\log\{\det(\Theta)\} + \operatorname{tr}(S\Theta) + \lambda \sum_{j<k} |\Theta_{jk}|]. \tag{4}$$

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

Another potential area of application is clustering. Choosing the correct number of clusters is a notoriously difficult problem. Looking for clusters that are stable under perturbations or subsampling of the data can help to obtain 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 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 the lasso and randomized lasso for high dimensional linear models is given in Section 3 and more numerical results are described in Section 4. After a discussion in Section 5, we collect all the technical proofs in Appendix A.

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 that the data, which are denoted here by $Z^{(1)}, \ldots, Z^{(n)}$, are IID (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 assume that we have a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines the amount of regularization. This tuning parameter could be the penalty parameter in l_1-penalized regression (see estimator (2)) or in Gaussian graphical modelling (see expression (4)), or it may be the number of steps in forward variable selection or orthogonal matching pursuit (OMP) (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 the 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 $\hat{S}^{\lambda} \subseteq \{1, \ldots, p\}$. It is then of interest to determine whether there is a $\lambda \in \Lambda$ such that $\hat{S}^{\lambda}$ is identical to S with high probability and how to achieve that right amount of regularization.

2.1. Stability paths

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

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

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

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

The sample size of $\lfloor n/2 \rfloor$ is chosen as it resembles most closely the bootstrap (Freedman, 1977; Büchlmann and Yu, 2002) while allowing computationally efficient implementation. Note that random subsampling can be viewed as a computational short cut for computing the relative frequency for $K \subseteq \hat{S}^{\lambda}(I_b)$ over all $\binom{n}{m}$ subsets I_b, $b = 1 \ldots, \binom{n}{m}$, of size $m = \lfloor n/2 \rfloor$, which itself is a U-statistic of order $m = \lfloor n/2 \rfloor$. Subsampling has also been advocated in a related context in Valdar *et al.* (2009).

For every variable $k = 1, \ldots, p$, 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 $k = 1, \ldots, p$ as a function of the regularization parameter. It can be seen in Fig. 1 that this simple path plot is potentially very useful for improved variable selection for high dimensional data.

In the remainder of the paper, we look at the selection probabilities of individual variables. The definition above covers sets of variables also. 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 that is defined in equation (2). We work with a gene expression data set for illustration which was 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, i.e. we consider a linear model as in equation (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) were close in spirit to our approach by arguing for 'consensus' gene signatures which assess the stability of selection, whereas Zucknick *et al.* (2008) proposed to measure stability of so-called 'molecular profiles' by the Jaccard index.

To see how the lasso and the related stability path cope with noise variables, we randomly permute all except six of the 4088 gene expressions across the samples, using the same permutation to keep the dependence structure between the permuted gene expressions intact. The set of six 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 Fig. 1(a), as a function of the regularization parameter λ (rescaled so that $\lambda = 1$ is the minimal λ-value for which the null model is selected and $\lambda = 0$ amounts to the basis pursuit solution). Three of the 'relevant' (unpermuted) genes stand out, but all remaining three variables are hidden within the paths of noise (permuted) genes. Fig. 1(b) shows the stability path. At least four relevant variables stand out much clearer now than they did in the regularization path plot. Fig. 1(c) shows the stability plot for the randomized lasso which will be introduced in Section 3.1: now all six unpermuted variables stand above the permuted variables and the separation between (potentially) relevant variables and irrelevant variables is even better.

Choosing the right regularization parameter is very difficult for the original path. The prediction optimal and cross-validated choices 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 that was chosen by cross-validation. Fig. 1 motivates that choosing the right regularization parameter is much less critical for the stability path and that we have a better chance of selecting truly relevant variables.

2.3. Stability selection

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

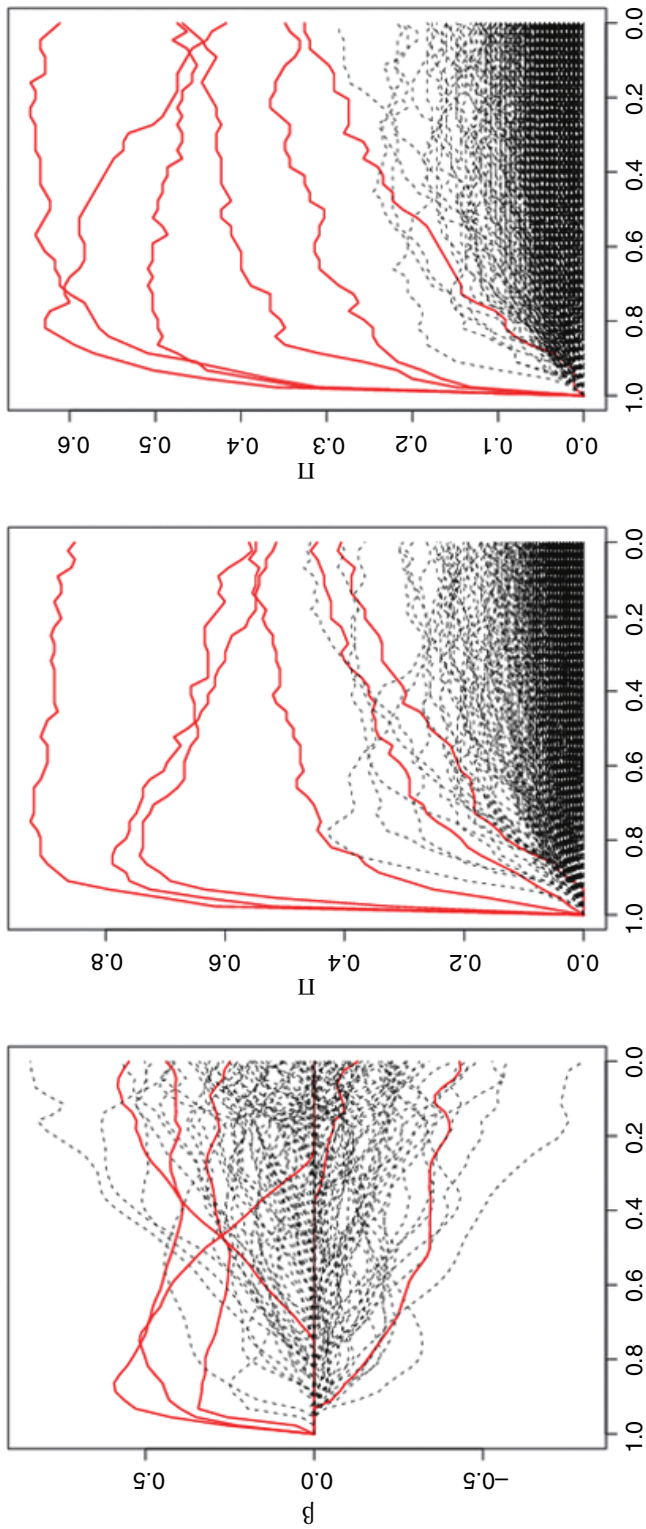


Fig. 1. (a) Lasso path for the vitamin gene expression data set (———, paths of six non-permuted genes; - - - - - -, paths of the 4082 permuted genes; selecting a model with all six unpermuted genes invariably means selecting a large number of irrelevant noise variables), (b) stability path of the lasso (the first four variables chosen with stability selection are truly non-permuted variables) and (c) stability path for the randomized lasso with weakness $\alpha = 0.2$, introduced in Section 3.1 (now all six non-permuted variables are chosen before any noise variable enters the model)

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

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

With stability selection, we do not simply select one model in the list (6). Instead the data are perturbed (e.g. 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 cut-off π_{thr} with $0 < \pi_{\text{thr}} < 1$ and a set of regularization parameters Λ, the set of stable variables is defined as

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

We keep variables with a high selection probability and disregard those with low selection probabilities. The exact cut-off π_{thr} with $0 < \pi_{\text{thr}} < 1$ is a tuning parameter but the results vary surprisingly little for sensible choices in a range of the cut-off. Nor do results depend strongly on the choice of regularization λ or the regularization region Λ. See Fig. 1 for an example.

Before we present some guidance on how to choose the cut-off parameter and the regularization region Λ below, it is worthwhile to point out that there have been related ideas in the literature on Bayesian model selection. Barbieri and Berger (2004) showed certain predictive optimality results for the so-called *median probability model*, consisting of variables which have posterior probability of being in the model of $\frac{1}{2}$ 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 regularization 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 regularization parameter is hence crucial. An advantage of our stability selection is that the choice of the initial set of regularization parameters Λ typically has not a very strong influence on the results, as long as Λ is varied within reason. Another advantage, which we focus on below, is the ability to choose this set of regularization parameters in a way that guarantees, under stronger assumptions, a certain bound on the expected number of false selections.

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

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

In general, it is very difficult 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 $\{\mathbf{1}_{\{k \in \hat{S}^{\lambda}\}}, k \in N\}$ is exchangeable for all $\lambda \in \Lambda$. Also, assume that the original procedure is not worse than random guessing, i.e.

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

The expected number V of falsely selected variables is then bounded for $\pi_{\mathrm{thr}} \in (\frac{1}{2}, 1)$ by

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

We shall 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 or, if divided by p, the per-comparison error rate in multiple testing (Dudoit *et al.*, 2003). Choosing fewer variables (reducing q_{Λ}) or increasing the threshold π_{thr} for selection will, unsurprisingly, reduce the expected number of falsely selected variables, with a minimal achievable non-trivial value of $1/p^2$ (for $\pi_{\mathrm{thr}} = 1$ and $q_{\Lambda} = 1$) for the per-family error rate. This seems sufficiently low for all practical purposes as long as $p > 10$, say.

The exchangeability assumption involved is perhaps stronger than we would wish, but there does not seem to be a way of achieving error control in the same generality without making similar assumptions. In recent independent work, Fan *et al.* (2009) made use of a similar condition for error control in regression. In regression and classification, the exchangeability assumption is fulfilled for all reasonable procedures $\hat{S}$ (whose results do not depend on the ordering of variables) if the design is random and the distribution of (Y, X_S, X_N) is invariant under permutations of variables in N. The simplest example is independence between each X_k, $k \in N$, and all other variables, including Y. To give another example for regression in model (1), the condition is satisfied if the error has a normal distribution and the variable X has a joint normal distribution where it holds true for all pairs $k, k' \in N$ that $\mathrm{cov}(X_k, X_l) = \mathrm{cov}(X_{k'}, X_l)$ for all $l = 1, \ldots, p$. 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 that are 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, $\pi_{\mathrm{thr}} \in (0.6, 0.9)$, results tend to be very similar. Once the threshold has been chosen at some default value, the regularization region Λ is determined by the error control desired. Specifically, for a default cut-off value $\pi_{\mathrm{thr}} = 0.9$, choosing the regularization parameters Λ such that say $q_{\Lambda} = \sqrt{(0.8p)}$ will control $E(V) \leqslant 1$, or choosing Λ such that $q_{\Lambda} = \sqrt{(0.8\alpha p)}$ controls the familywise error rate at level α, i.e. $P(V > 0) \leqslant \alpha$. Of course, we can proceed the other way round by fixing the regularization region Λ and choosing π_{thr} such that $E(V)$ is controlled at the desired level.

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

Without stability selection, the regularization parameter λ invariably must depend on the unknown noise level of the observations. The advantages of stability selection are 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 difficult to estimate the noise level in these settings.

2.4.1. Pointwise control

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

2.5. Example II: graphical modelling

Stability selection is also promising for graphical modelling. Here we focus on Gaussian graphical models as described in Section 1.1 around formulae (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 data set (without the response variable) that was introduced in Section 2.2. We want to infer the set of non-zero entries in the inverse covariance matrix Σ^{-1}. Part of the resulting regularization path of the graphical lasso showing graphs for various values of the regularization parameter λ, i.e. $\{\hat{S}^\lambda; \lambda \in \Lambda\}$ where $\hat{S}^\lambda = \{(j,k); (\hat{\Sigma}^{-1})_{jk}^\lambda \neq 0\}$, are shown in Figs 2(a)–2(f). For reasons of display, variables (genes) are ordered first using hierarchical clustering and are symbolized by nodes arranged in a circle. Stability selection is shown in Figs 2(g)–2(l). We pursue a pointwise control approach. For each value of λ, we select the threshold π_{thr} to guarantee that $E(V) \leqslant 30$, i.e. we expect fewer than 30 wrong edges among the 12 720 possible edges in the graph. The set $\hat{S}^{\text{stable}}$ varies remarkably little for the majority of the path and the choice of q (which is implied by λ) does not seem to be critical, as already observed for variable selection in regression.

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

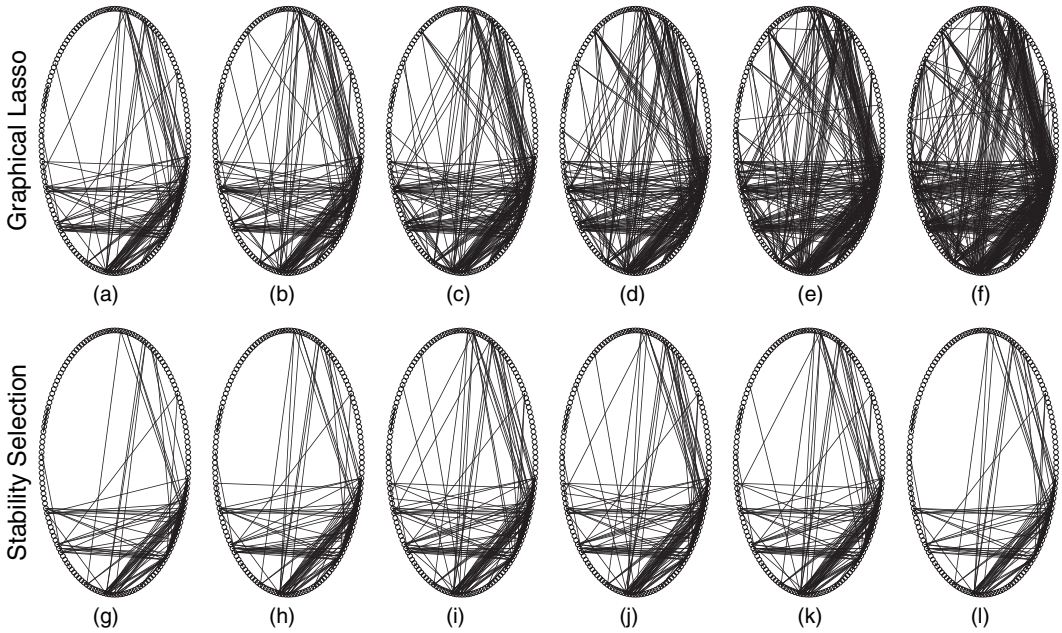


Fig. 2. Vitamin gene expression data set—(a)–(f) regularization path of the graphical lasso and (g)–(l) the corresponding pointwise stability-selected models: (a), (g) $\lambda = 0.46$; (b), (h) $\lambda = 0.448$; (c), (i) $\lambda = 0.436$; (d), (j) $\lambda = 0.424$; (e), (k) $\lambda = 0.412$; (f), (l) $\lambda = 0.4$

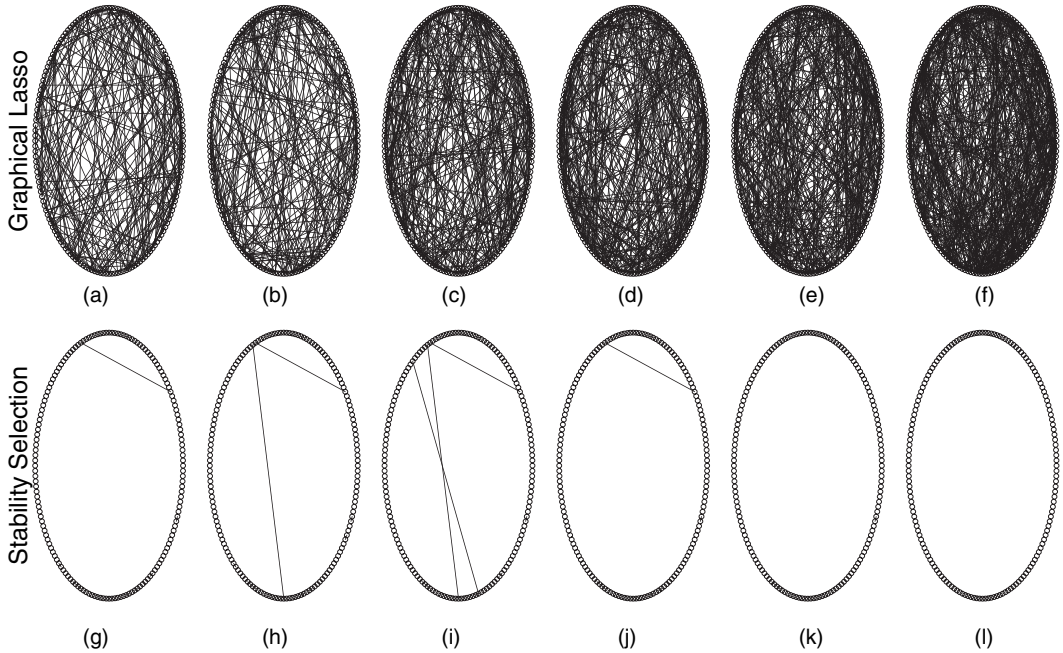


Fig. 3. Same plots as in Fig. 2 but with the variables (expression values of each gene) permuted independently (the empty graph is the true model; with stability selection, only a few errors are made, as guaranteed by the error control made): (a), (g) $\lambda = 0.065$; (b), (h) $\lambda = 0.063$; (c), (i) $\lambda = 0.061$; (d), (j) $\lambda = 0.059$; (e), (k) $\lambda = 0.057$; (f), (l) $\lambda = 0.055$

2.6. Computational requirements

Stability selection demands that we rerun $\{\hat{S}^{\lambda}; \lambda \in \Lambda\}$ multiple times. Evaluating selection probabilities over 100 subsamples seems sufficient in practice. The algorithmic complexity of the lasso in expression (2) or in expression (13) in Section 3.1 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 data set and running 100 simulations is 25 times the cost of running a single fit on the full data set. This cost could be compared with the cost of cross-validation, as this is what we must resort to often in practice to select the regularization parameter. Running tenfold cross-validation uses approximately $10 \times 0.9^2 = 8.1$ as many computational resources as the single fit on the full data set. Stability selection is thus roughly three times more expensive than tenfold cross-validation. This analysis is based on the fact that the computational complexity scales like $O(n^2)$ with the number of observations (assuming that $p > n$). If computational costs would scale linearly with sample size (e.g. for the lasso with $p < n$), this factor would increase to roughly 5.5.

Stability selection with the lasso (using 100 subsamples) for a data set with $p = 1000$ and $n = 100$ takes about 10 s on a 2.2-GHz 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, which is 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, 2009). We consider a linear model as in equation (1) with Gaussian noise,

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

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

Stability selection is attractive for two reasons. First, the choice of a proper regularization parameter for variable selection is crucial and notoriously difficult, especially because the noise level is unknown. With stability selection, results are much less sensitive to the choice of the regularization. Second, we shall 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

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

It is clearly of interest to know under which conditions consistent variable selection can be achieved. In the high dimensional context, this places a restriction on the growth of the number p of variables and sparsity $|S|$, typically of the form $|S|\log(p) = o(n)$ (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009). Although this assumption is often realistic, there are stronger assumptions on the design matrix that need to be satisfied

for consistent variable selection. For the 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 OMP (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 regularization. For brevity, we shall only discuss in detail the case of the lasso whereas the analysis of OMP is just alluded to.

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

3.1. Lasso and randomized lasso

The lasso (Tibshirani, 1996; Chen *et al.*, 2001) estimator is given in expression (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 some assumptions, the strongest of which is arguably 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):

$$\max_{k \in N} |\mathrm{sgn}(\beta_S)^{\mathrm{T}} (X_S^{\mathrm{T}} X_S)^{-1} X_S^{\mathrm{T}} X_k| < 1. \tag{12}$$

Condition (12) is sufficient and (almost) necessary (the word 'almost' refers to the fact that a necessary relationship uses '$\leqslant$' instead of '$<$'). If this condition is violated, all that we can hope for is recovery of the regression vector β in an l_2-sense of convergence by achieving $\|\hat{\beta}^{\lambda} - \beta\|_2 \to_{\mathrm{p}} 0$ for $n \to \infty$. The main assumption here is 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 randomized lasso as an alternative. Despite its simplicity, it is consistent for variable selection even though the irrepresentable condition (12) is violated.

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

For the randomized lasso with weakness $\alpha \in (0, 1]$, let W_k be IID random variables in $[\alpha, 1]$ for $k = 1, \ldots, p$. The randomized lasso estimator $\hat{\beta}^{\lambda, \mathrm{W}}$ for regularization parameter $\lambda \in \mathbb{R}$ is then

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

A proposal for the distribution of the weights W_k is described below, just before theorem 2. The word 'weakness' is borrowed from the terminology of weak greedy algorithms (Temlyakov, 2000) which are loosely related to our randomized lasso. Implementation of estimator (13) is straightforward by appropriate rescaling of the predictor variables (with scale factor W_k for the kth variable). Using these rescaled variables, the standard lasso is solved, using for example the algorithm LARS (Efron *et al.*, 2004) or fast co-ordinatewise approaches (Meier *et al.*, 2008; Friedman *et al.*, 2007). The perturbation of the penalty weights is reminiscent of the reweighting in the adaptive lasso (Zou, 2006). Here, however, the reweighting 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 we surely cannot expect any improvement from such a random perturbation. If applied only with one random perturbation, the randomized lasso is not very useful. However, applying the randomized lasso many times and looking for variables that are chosen often will turn out to be a very powerful procedure.

3.1.1. Consistency for randomized lasso with stability selection

For stability selection with the randomized 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.*, 2009), which was also called the sparse Riesz condition in Zhang and Huang (2008).

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

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

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

We must constrain sparse eigenvalues to succeed.

Assumption 1 (sparse eigenvalues). There are some $C > 1$ and some $\kappa \geqslant 9$ such that

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

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

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

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

We have not specified the exact form of perturbations that we shall be using for the randomized lasso (13). For what follows we consider the randomized lasso (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 often work just as well in practice.

Theorem 2. Consider model (10). For the randomized 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 $a_n \to \infty$ for $n \to \infty$. Let $\lambda_{\min} = 2\sigma\{\sqrt{(2C)s+1}\}\sqrt{\{\log(p \vee a_n)/n\}}$. Assume that $p > 10$ and $s \geqslant 7$ and that the sparse eigenvalue assumption 1 is satisfied. Then there is some $\delta = \delta_s \in (0, 1)$ such that, for all $\pi_{\text{thr}} \geqslant 1 - \delta$, stability selection with the randomized lasso satisfies on a set Ω_A with $P(\Omega_A) \geqslant 1 - 5/(p \vee a_n)$ that no noise variables are selected,

$$N \cap \hat{S}_{\lambda}^{\text{stable}} = \emptyset, \tag{17}$$

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

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

where $S_{\text{small};\lambda} = \{k : |\beta_k| \leqslant 0.3(Cs)^{3/2}\lambda\}$. 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 $\min_{k \in S} |\beta_k| \geqslant (Cs)^{3/2}(0.3\lambda)$, as a consequence of theorem 2,

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

i.e. consistent variable selection for $p \vee a_n \to \infty$ ($p \to \infty$ or $n \to \infty$) in the sense of expression (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 $\lambda \geqslant \lambda_{\min}$. This is noteworthy as it means that, even if the value of λ is chosen too large (i.e. considerably larger than $\lambda_{\min}$), no noise variables will be selected and expression (17) holds true. 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 randomized 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. Although this achieves good empirical results, it seems more advantageous in combination with subsampling of the data. The results extend directly to this case. Let $\tilde{\Pi}^{\lambda}_{k}$ 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 $\tilde{\Pi}^{\lambda}_{k}$ is above the threshold $\pi_{\text{thr}} \in (0, 1)$ is bounded by a Markov-type inequality from below by

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

having used that $E(\tilde{\Pi}^{\lambda}_{k}) \geqslant 1 - 5/(p \vee a_{n/2})$ as a consequence of theorem 2. If $5/(p \vee a_{n/2})$ is sufficiently small in comparison with $1 - \pi_{\text{thr}}$, 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 trade-off when choosing the weakness α. A negative consequence of a low α is that the design can become closer to singularity and can thus lead to unfavourable conditioning of the weighted design matrix. However, 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 $(0.2, 0.8)$ gives very useful results.

3.1.2. Relation to other work

In related and very interesting work, Bach (2008) has proposed the 'bolasso' (for bootstrapped enhanced lasso) and shown that using a finite number of subsamples of the original lasso procedure and applying basically stability selection with $\pi_{\text{thr}} = 1$ yield consistent variables selection under the condition that the penalty parameter λ vanishes faster than typically assumed, at rate $n^{-1/2}$, and that the model dimension p is fixed. Although the latter condition could possibly be technical only, the first distinguishes it from our results. Applying stability selection to the randomized 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 the randomized lasso, only truly relevant variables are chosen (though a few might be missed). If λ is chosen too large with the bolasso, noise variables might be picked up. Fig. 4 is a good illustration. Picking the regularization in Fig. 4(a) (without extra randomness) to select the correct model is much more difficult than in Fig. 4(c), where extra randomness is added. The same distinction can be made with two-stage procedures like the 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 the randomized lasso does not pick wrong variables if λ is chosen too large.

3.2. Example

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

3.3. Randomized orthogonal matching pursuit

Interesting alternatives to the lasso or greedy forward search in this context are the recently proposed forward–backward search (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 optimization 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 OMP, a greedy forward search in the variable space. The iterative sure independence screening procedure (Fan and Lv, 2008) entails OMP as a special case. We shall examine the effect of stability selection under subsampling and additional randomization. To have a clear definition of randomized OMP, with weakness $0 < \alpha < 1$ and q iterations, we define it as follows.

(a) Set $R_1 = Y$. Set $m = 0$ and $\hat{S}^0 = \emptyset$.
(b) For $m = 1, \ldots, q$:
    (i) find $\rho_{\max} = \max_{1 \leqslant k \leqslant p} |X_k^{\mathrm{T}} R_m|$;

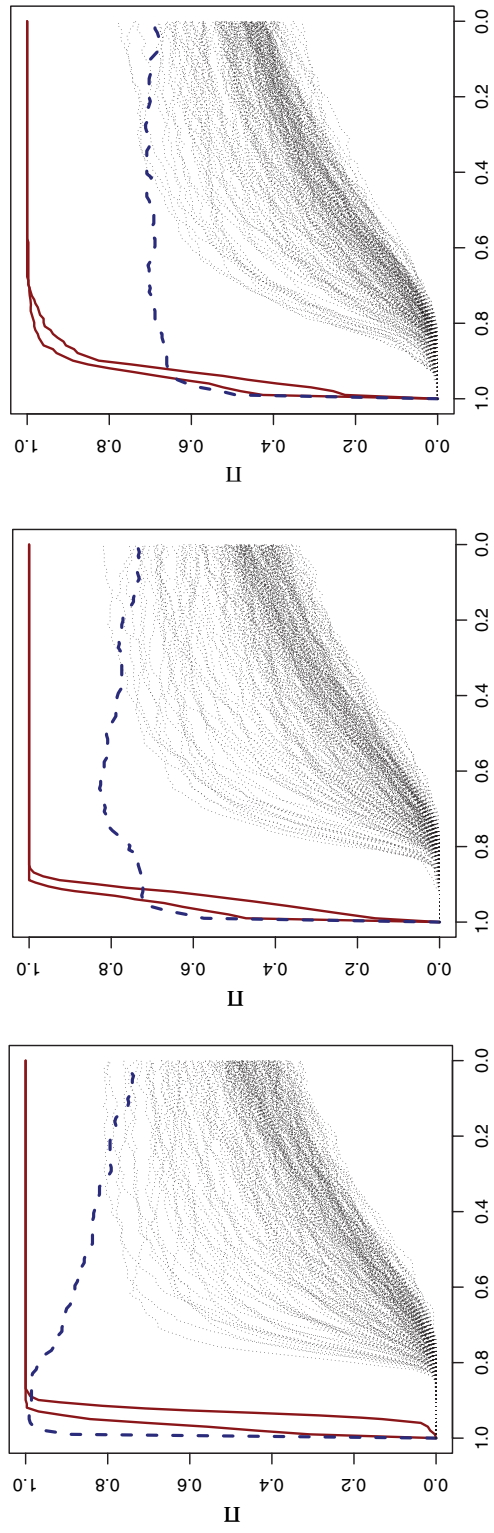


Fig. 4. Stability paths for the randomized lasso with stability selection by using weakness parameters (a) $\alpha = 1$ (identical to the original lasso), (b) $\alpha = 0.5$ and (c) $\alpha = 0.2$ (——, coefficients of the first two (relevant) variables; – – –, coefficient of the third (irrelevant) variable; · · · · · ·, coefficients from all other (irrelevant) variables): introducing the randomized version helps to avoid choosing the third (irrelevant) predictor variable

(ii) define $K = \{k : |X_k^{\mathrm{T}} R_m| \geqslant \alpha \rho_{\max}\}$;
(iii) select randomly a variable k_{sel} in the set K and set $\hat{S}^m = \hat{S}^{m-1} \cup \{k_{\mathrm{sel}}\}$;
(iv) let $R_{m+1} = Y - P_m Y$, where the projection P_m is given by $X_{\hat{S}^m}(X_{\hat{S}^m}^{\mathrm{T}} X_{\hat{S}^m})^{-1} X_{\hat{S}^m}^{\mathrm{T}}$.

(c) Return the selected sets $\hat{S}^1 \subset \hat{S}^2 \subset \ldots \subset \hat{S}^q$.

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

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

This is a sufficient condition for consistent variable selection. If it is not fulfilled, there are 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 quite 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 randomized OMP is asymptotically consistent for variable selection in linear models, even if the right-hand side in condition (19) is not bounded by 1 but instead by a possibly large constant (assuming that the weakness α is sufficiently low). This indicates that stability selection has a more general potential for improved structure estimation, beyond the case for the lasso that was presented in theorem 2. 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 shall give empirical evidence for the usefulness of stability selection under subsampling and additional randomization for OMP 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 the lasso and randomized lasso for both regression and the natural extension to binary classification. The effect on OMP and randomized OMP will also be examined.

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

(a) Independent predictor variables: all $p = 1000$ predictor variables are IID standard normal distributed; sample size $n = 100$ and $n = 1000$.
(b) Block structure with 10 blocks: the $p = 1000$-dimensional predictor variable follows an $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{km} = 0$ for all pairs (k, m) except if $\mathrm{mod}_{10}(k) = \mathrm{mod}_{10}(m)$, for which $\Sigma_{km} = 0.5$; sample size $n = 200$ and $n = 1000$.
(c) Toeplitz design: the $p = 1000$-dimensional predictor variable follows an $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{km} = \rho^{|k-m|}$ and $\rho = 0.99$; sample size $n = 200$ and $n = 1000$.
(d) Factor model with two factors: let ϕ_1 and ϕ_2 be two latent variables following IID standard normal distributions. Each predictor variable X_k, for $k = 1, \ldots, p$, is generated as $X_k = f_{k,1}\phi_1 + f_{k,2}\phi_2 + \eta_k$, where $f_{k,1}$, $f_{k,2}$ and η_k have IID standard normal distributions for all $k = 1, \ldots, p$; sample sizes are $n = 200$ and $n = 1000$, and $p = 1000$.
(e) Data set (e) identical to (d) but with 10 instead of two factors.
(f) Motif regression data set: this is a data set ($p = 660$ 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 data set 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) This data set is the vitamin gene expression data (with $p = 4088$ and $n = 158$) that were described in Section 2.2.

We do not use the response values from the real data sets, however, as we need to know which variables are truly relevant or irrelevant. For this, we create sparse regression vectors by setting $\beta_k = 0$ for all $k = 1, \ldots, p$, except for a randomly chosen set S of coefficients, where 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 data set. For regression, the noise vector $(\varepsilon_1, \ldots, \varepsilon_n)$ is chosen IID $\mathcal{N}(0, \sigma^2/n)$, 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 of 0.5 and 2. For binary classification, we scale the vector β to achieve a given Bayes misclassification rate, either $\frac{1}{8}$ or $\frac{1}{3}$. Each of the 64 scenarios is run 100 times: once using the standard procedure (the lasso or OMP), once using stability selection with subsampling and once using stability selection with subsampling and additional randomization ($\alpha = 0.5$ for the randomized lasso and $\alpha = 0.9$ for randomized OMP). The methods are thus in total evaluated on about 20000 simulations each.

The solution of stability selection cannot be reproduced by simply selecting the right penalty with the lasso, since stability selection provides a fundamentally new solution. To compare the power of both approaches, we look at the probability that γs 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 is a regularization parameter such that at least $\lceil \gamma s \rceil$ variables in S have a non-zero regression coefficient and all variables in $N = \{1, \ldots, p\} \setminus S$ have a zero regression coefficient. For stability selection, recovery without error means that the $\lceil \gamma s \rceil$ variables with highest selection probability $\max_{\lambda \geqslant \lambda_{\min}}(\hat{\beta}_k^{\lambda})$ are all in S. The value $\lambda_{\min}$ is chosen such that at most $\sqrt{(0.8p)}$ variables are selected in the whole path of solutions for $\lambda \geqslant \lambda_{\min}$. Note that this notion neglects the fact that the most advantageous regularization parameter is selected automatically here for the lasso and OMP but not for stability selection.

Results are shown in Fig. 5 for the lasso applied to regression, and in Fig. 6 for OMP applied to regression and the lasso applied to binary classification. In Fig. 5, we also give the median number of variables violating the irrepresentable 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 as or more correct variables than the underlying method itself. In some settings (e.g. in (a) or in (b) and (f) when the number s of relevant variables is very small), stability selection does not improve and yields comparable results with those of the underlying method. That stability selection is not helping 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. However, the gain of stability selection under subsampling is often substantial, irrespective of the sparsity of the signal and the signal-to-noise-ratio. Additional randomization helps in cases where there are many variables violating the irrepresentable condition, e.g. in setting (e). This is in line with our theory.

Instead of giving full receiver operating characteristic curves for each simulation setting, we look at the number of falsely chosen variables when selecting 20% and 80% of all s relevant variables. The mean number of falsely selected variables in each of these two cases is shown in Fig. 7 for the lasso and Fig. 8 for OMP. When using the lasso, stability selection with subsampling can increase the number of falsely selected edges when there is a large number of relevant variables ($s = 50$) and we are looking to identify a large proportion (80%) of these. Yet all methods fail

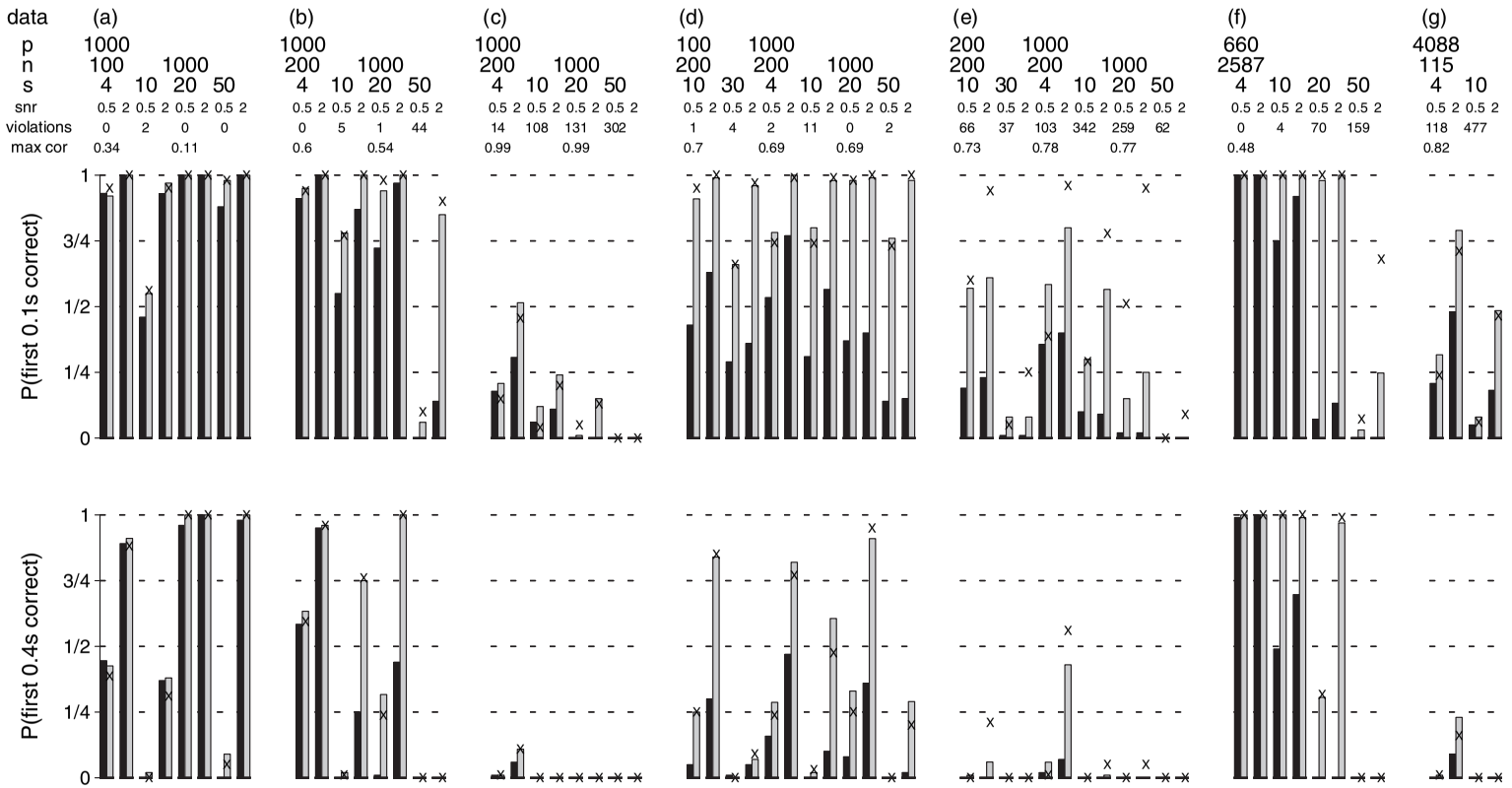


Fig. 5. Probability of selecting $0.1s$ and $0.4s$ important variables without selecting a noise variable with the lasso in the regression setting (▮) and stability selection under subsampling (▮) for the 64 different settings: ×, results for stability selection under additional randomization ($\alpha = 0.5$)

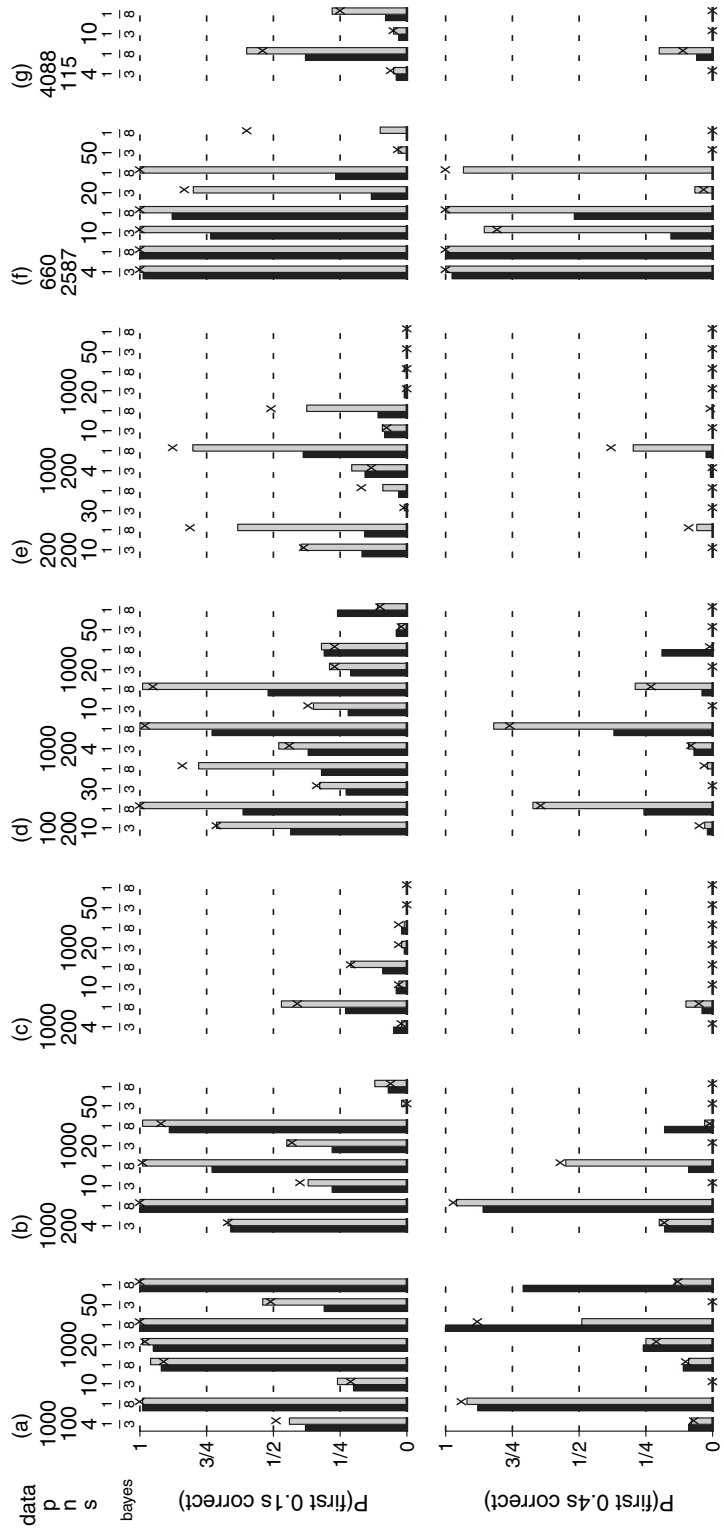


Fig. 6. Equivalent plot to Fig. 5 for the lasso applied to (a) binary classification and (b) OMP applied to regression

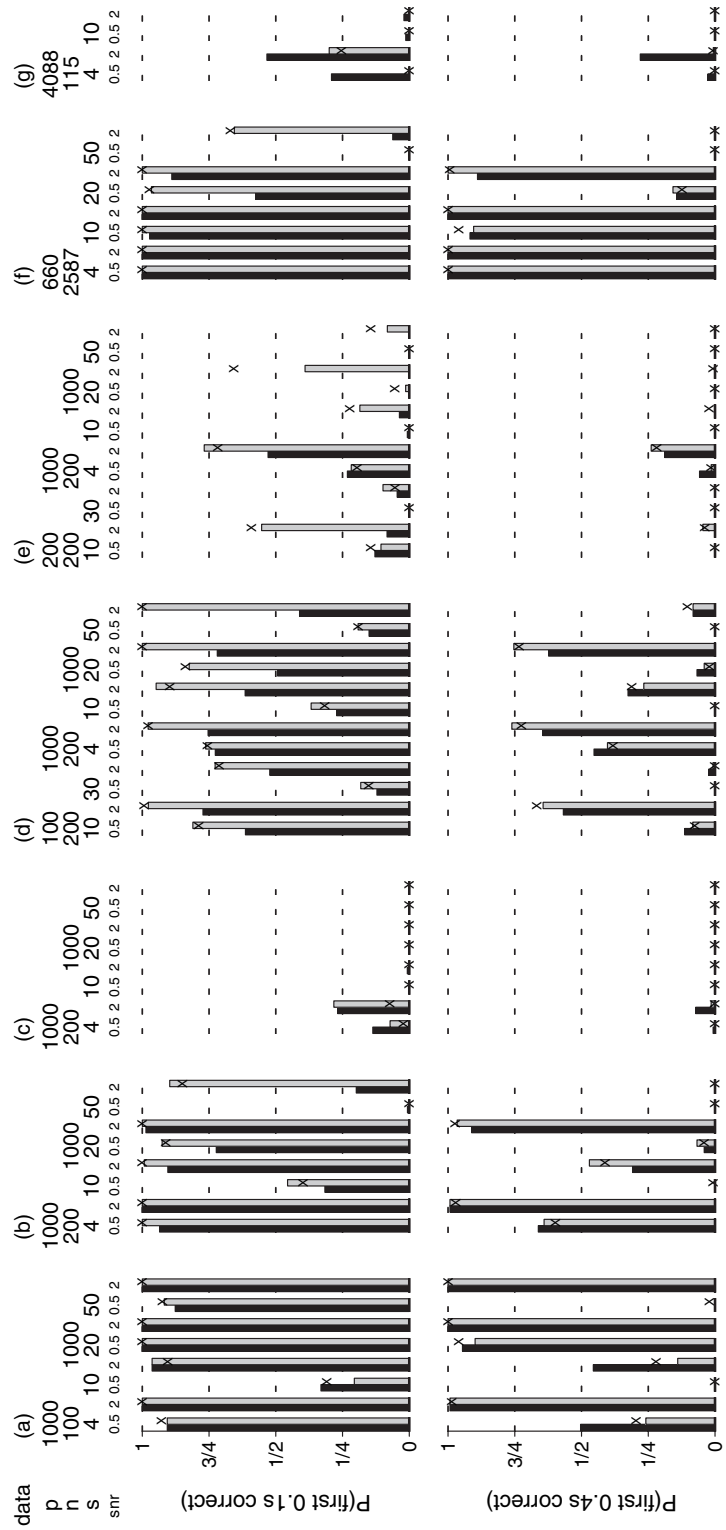


Fig. 6 (*continued*)

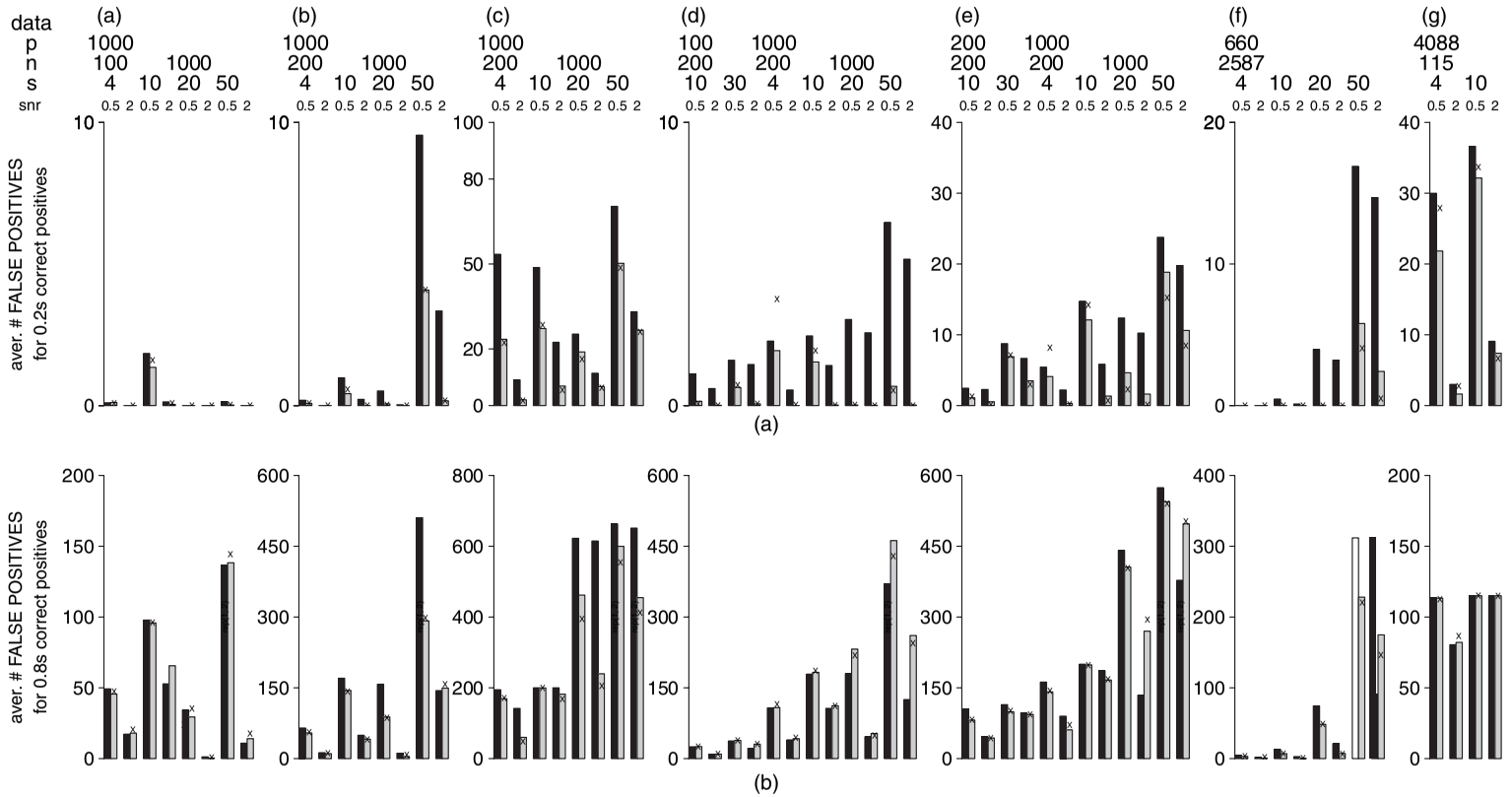


Fig. 7. Average number of falsely chosen variables in the regression setting when selecting (a) 20% or (b) 80% of all s correct variables: ▮, lasso results; ▮, stability selection under subsampling; ×, results for stability selection under additional randomization ($\alpha = 0.5$)

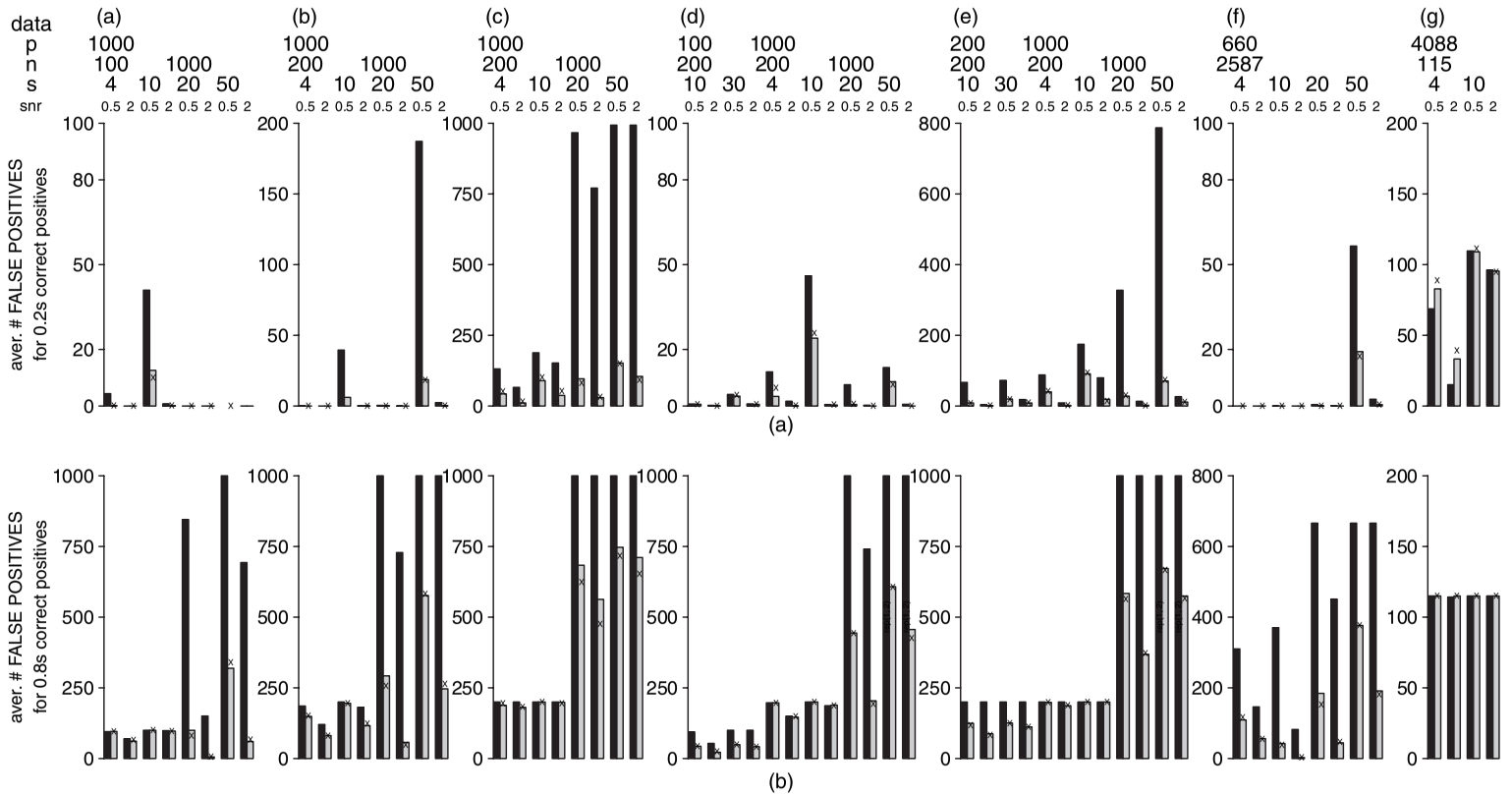


Fig. 8. Equivalent plot to Fig. 7 when using OMP instead of the lasso

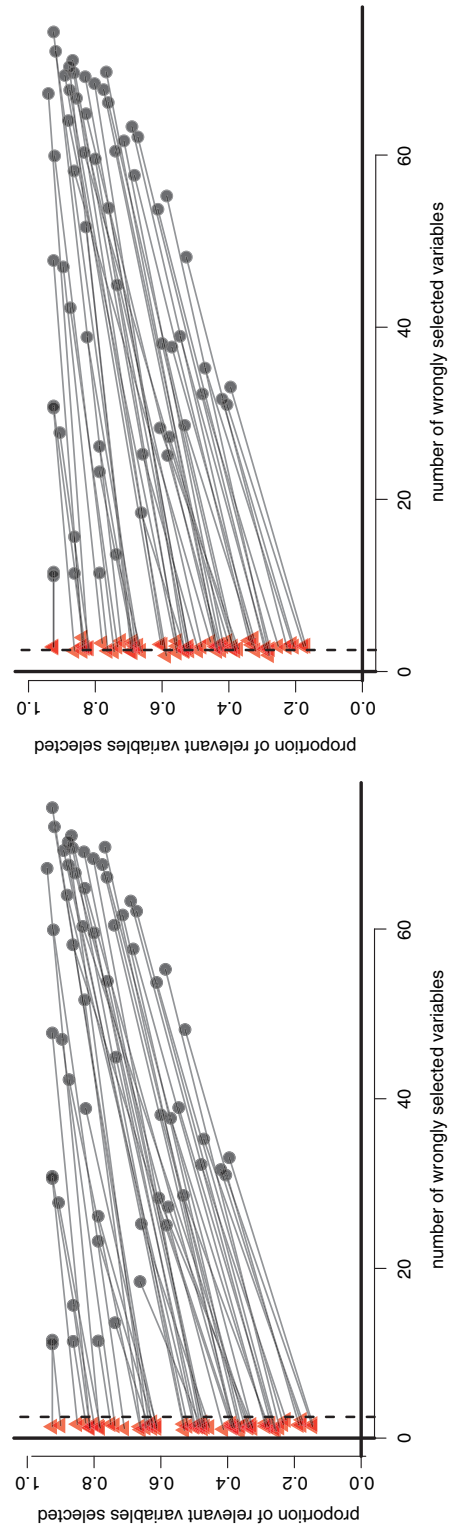


Fig. 9. Comparison of stability selection (for the randomized lasso with (a) $\alpha = 0.5$ and (b) $\alpha = 1$) (▲) with cross-validation (for the standard lasso) (●) for the real data sets (f) and (g), showing the average proportion of correctly identified relevant variables *versus* the average number of falsely selected variables: each ●–▲ pair corresponds to a simulation setting (some specified signal-to-noise ratio and s); ·, value at which the number of wrongly selected variables is controlled, namely $E(V) \leqslant 2.5$; looking at stability selection, the proportion of correctly identified relevant variables is very close to the cross-validation solution, whereas the number of falsely selected variables is reduced dramatically

somehow when trying to recover such a large number of relevant variables, yielding often hundreds of false positive results. In all other settings, stability selection seems advantageous, selecting often substantially fewer variables falsely than the standard lasso. For OMP, the gains are even more pronounced.

Next, we test how well the error control of theorem 1 holds up for these data sets. For the motif regression data set (f) and the vitamin gene expression data set (g), the lasso is applied, with randomization and without. For both data sets, 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 $\pi_{\text{thr}} = 0.6$, corresponding to a control of $E(V) \leqslant 2.5$, 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 probably not fulfilled for the given data set and it is of interest to see how well the bound holds up for real data. Results are shown in Fig. 9. 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 level desired, giving empirical evidence that the error control derived is useful beyond the setting of exchangeability discussed. Stability selection thus helps to select a useful amount of regularization.

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 regularization for structure estimation or variable selection. Furthermore, the solution of stability selection depends surprisingly little on the initial regularization chosen. This is an additional great benefit besides error control.

Another property of stability selection is the improvement over a prespecified selection method. Often computationally efficient algorithms for high dimensional selection are inconsistent, even in rather simple settings. We prove for the randomized lasso that stability selection will be variable selection consistent even if the necessary conditions for consistency of the original method are violated. And, thus, stability selection will asymptotically select the right model in scenarios where the 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 have been demonstrated on simulated and real data.

Acknowledgements

We thank Bin Yu and Peter Bickel for inspiring discussions and the referees for many helpful comments and suggestions which greatly helped to improve the manuscript. NM thanks the Forschungsinstitut für Mathematik at Eidgenössiche Technische Hochschule Zürich for generous support and hospitality.

Appendix A

A.1. Sample splitting

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

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

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

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

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

We work with the selection probabilities that are 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 that are close to 1.

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

$$\hat{\Pi}_K^{\text{simult},\lambda} \geqslant 2\hat{\Pi}_K^\lambda - 1. \tag{21}$$

Proof. Let I_1 and I_2 be the two random subsets in sample splitting of $\{1, \ldots, n\}$ with $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$ and $I_1 \cap I_2 = \emptyset$. Denote by $s_K(\{1,1\})$ the probability $P^*[\{K \subseteq \hat{S}^\lambda(I_1)\} \cap \{K \subseteq \hat{S}^\lambda(I_2)\}]$. Note that the two events are not independent as the probability is only with respect to a random split of the fixed samples $\{1, \ldots, n\}$ into I_1 and I_2. The probabilities $s_K(\{1,0\})$, $s_K(\{0,1\})$ and $s_K(\{0,0\})$ 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 $\hat{\Pi}_K^{\text{simult},\lambda} = s_K(\{1,1\})$ 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 $s_K(\{0,0\}) \geqslant 0$, it also follows that $s_K(\{1,0\}) \leqslant 1 - \hat{\Pi}_K^\lambda$. Hence

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

which completes the proof.

A.2. Proof of theorem 1

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

$$E(V) = \sum_{k \in N} P\{\max_{\lambda \in \Lambda}(\hat{\Pi}_k^{\lambda}) \geqslant \pi_{\text{thr}}\} \leqslant q_{\Lambda}^2 / p(2\pi_{\text{thr}} - 1),$$

which completes the proof.

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

$$P(\hat{\Pi}_K^{\text{simult},\lambda} \geqslant \xi) \leqslant \varepsilon^2 / \xi.$$

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

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

Proof. Let $I_1, I_2 \subseteq \{1, \ldots, n\}$ be, as above, the random split of the samples $\{1, \ldots, n\}$ into two disjoint subsets, where both $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$. Define the binary random variable H_K^{λ} for all subsets $K \subseteq \{1, \ldots, p\}$ as $H_K^{\lambda} := \mathbf{1}\{K \subseteq \{\hat{S}^{\lambda}(I_1) \cap \hat{S}^{\lambda}(I_2)\}\}$. Denote the data (the n samples) by Z. The simultaneous selection probability $\hat{\Pi}_K^{\text{simult},\lambda}$, as defined in expression (20), is then $\hat{\Pi}_K^{\text{simult},\lambda} = E^*(H_K^{\lambda}) = E(H_K^{\lambda}|Z)$, where the expectation E^* is with respect to the random split of the n samples into sets I_1 and I_2 (and additional randomness if $\hat{S}^{\lambda}$ is a randomized algorithm). To prove the first part, the inequality $P\{K \subseteq \hat{S}^{\lambda}\} \leqslant \varepsilon$ (for a sample size $\lfloor n/2 \rfloor$) implies that $P(H_K^{\lambda} = 1) \leqslant P\{K \subseteq \hat{S}^{\lambda}(I_1)\}^2 \leqslant \varepsilon^2$ and hence $E(H_K^{\lambda}) \leqslant \varepsilon^2$. Therefore, $E(H_K^{\lambda}) = E\{E(H_K^{\lambda}|Z)\} = E(\hat{\Pi}_K^{\text{simult},\lambda}) \leqslant \varepsilon^2$. Using a Markov-type inequality, $\xi\, P(\hat{\Pi}_K^{\text{simult},\lambda} \geqslant \xi) \leqslant E(\hat{\Pi}_K^{\text{simult},\lambda}) \leqslant \varepsilon^2$. Thus $P(\hat{\Pi}_K^{\text{simult},\lambda} \geqslant \xi) \leqslant \varepsilon^2 / \xi$, completing the proof of the first claim. The proof of the second part follows analogously.

A.3. Proof of theorem 2

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

Definition 6 (additional notation). For weights W as in expression (13), let X^w be the matrix of rescaled variables, with $X_k^w = X_k W_k$ for each $k = 1, \ldots, p$. Let $\phi_{\max}^w$ and $\phi_{\min}^w$ be the maximal and minimal eigenvalues analogous to expression (14) for X^w instead of X.

The proof rests mainly on the twofold effect that 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 the randomized lasso as it shows that the irrepresentable condition (12) is sometimes fulfilled under randomly sampled weights, even though it is 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 that $s \geqslant 7$. Let W be weights generated randomly in $[\alpha, 1]$, as in expression (13), and let X^w be the corresponding rescaled predictor variables, as in definition 6. For $\alpha^2 \geqslant \nu\, \phi_{\min}(Cs^2)/Cs^2$, with $\nu \in \mathbb{R}^+$, it holds under assumption 1 for all random realizations W that

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

Proof. Using assumption 1,

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

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

$$\frac{\phi_{\max}(Cs^2)}{\phi_{\min}(Cs^2)} \leqslant \frac{3+4\bar{C}}{\kappa} \sqrt{\left\{ \frac{\phi_{\min}(Cs^2)}{Cs^2} \right\}}. \tag{23}$$

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

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

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 $v^{\mathrm{T}} v = \tilde{v}^{\mathrm{T}} \mathcal{W}^{-1,\mathrm{T}} \mathcal{W}^{-1} \tilde{v}$ and thus $\tilde{v}^{\mathrm{T}} \tilde{v} \leqslant v^{\mathrm{T}} v \leqslant \alpha^{-2} \tilde{v}^{\mathrm{T}} \tilde{v}$ for all $\mathcal{W}$ with diagonal entries in $[\alpha, 1]$. The corresponding argument for $\phi_{\min}(m)$ yields the bound $\phi^w_{\min}(m) \geqslant \alpha\, \phi_{\min}(m)$ for all $m \in \mathbb{N}$. Claim (22) follows by observing that $\bar{C} \geqslant 1$ for $s \geqslant 7$, since $C \geqslant 1$ by assumption 1 and hence $3 + 4\bar{C} \leqslant 7\bar{C}$.

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

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

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

Proof. The proof follows mostly from theorem 1 in Zhang and Huang (2008). For this, set $c_0 = 0$ in their notation. We also have $Cs^2 \leqslant (2+4\bar{C})s+1$, 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 $\phi^w_{\max}(Cs^2)/\phi^w_{\min}(Cs^2)$. It is bounded for all random realizations of $W = w$, as long as $\alpha^2 \geqslant (7/\kappa)^2\, \phi_{\min}(Cs^2)/Cs^2$, using lemma 3, by

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

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

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

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

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

having used for the first inequality that, in the notation of Zhang and Huang (2008), $1/c^* c_* \leqslant c^*/c_*$. The n^{-2}-factor has been omitted to account for our different normalization. For the second inequality, we used $4\bar{C} \leqslant Cs$. 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.

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

$$P_w[\|\{(X^w_{K(w)})^{\mathrm{T}} X^w_{K(w)}\}^{-1} (X^w_{K(w)})^{\mathrm{T}} X^w_k\|_1 \leqslant 2^{-1/4}] \geqslant p_w(1-p_w)^m, \tag{27}$$

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 realization of W for which $\tilde{w}_k = \alpha$ and $\tilde{w}_j = 1$ for all other $j \in \{1, \ldots, p\} \setminus k$. The probability of $W = \tilde{w}$ is clearly $p_w(1 - p_w)^{p-1}$ under the sampling scheme that is used 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 $w_j \leqslant \tilde{w}_j$ for all $j \in \{1, \ldots, p\}$ with equality for $j \in A$. For all weights $w \in \mathcal{W}$, it follows moreover that

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

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

$$\sup_{A:|A| \leqslant m} \sup_{k \notin A} \left\{ \|(X_A^{\mathrm{T}} X_A)^{-1} X_A^{\mathrm{T}} X_k\|_1^2 \right\} \leqslant m / \phi_{\min}(m). \tag{28}$$

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

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

As $X_A \gamma$ is the projection of X_k into the space that is spanned by X_A and $\|X_k\|_2^2 = 1$, it holds that $\|X_A \gamma\|_2^2 \leqslant 1$. Using $\|X_S \gamma\|_2^2 = \gamma^{\mathrm{T}} (X_A^{\mathrm{T}} X_A) \gamma \geqslant \phi_{\min}(|A|) \|\gamma\|_2^2$, it follows that $\|\gamma\|_2^2 \leqslant 1 / \phi_{\min}(|A|)$, which shows result (28) and thus completes the proof.

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

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

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

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

It is now sufficient to show that $P(\Omega_1'') \geqslant 1 - 1/(p \vee a_n)$, showing that this bound is related to a bound in Zhang and Huang (2008), and we repeat a similar argument. Each term $n^{1/2} \|P_A \varepsilon\|_2 / \sigma$ 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),

$$P[n \|P_A \varepsilon\|_2^2 / \sigma^2 \geqslant |A| \{1 + 4 \log(p \vee a_n)\}] \leqslant [(p \vee a_n)^{-4} \{1 + 4 \log(p \vee a_n)\}]^{|A|/2} \leqslant (p \vee a_n)^{-3|A|/2},$$

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

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

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

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 is a set Ω_A with $P(\Omega_A) \geqslant 1 - 5/(p \vee a_n)$ such that, for all $\omega \in \Omega_A$ and $\lambda \geqslant \lambda_{\min}$,

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

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

where $S_{\mathrm{small};\lambda}$ is defined as in theorem 2.

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

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

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

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

where $\hat{\beta}^{\lambda, W, -k}$ is the solution to expression (13) with the constraint that $\hat{\beta}_k^{\lambda, W, -k} = 0$, which is comparable with 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 shall use $\hat{B}$ as a shorthand notation for $\hat{B}^{\lambda, W, -k}$. Let $P_{\hat{B}}^w$ be the projection operator into the space that is spanned by all variables in the set $\hat{B}$. For all $W = w$, this is identical to

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

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

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

it holds for the right-hand term in expression (34) that

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

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

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

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

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

We now apply lemma 5 to the rightmost term. The set $\hat{B}$ is a function of the weight vector and satisfies for every realization of the observations $Y \in \Omega_0$ the conditions in lemma 5 on the set $K(w)$. First, $|\hat{B}| \leqslant Cs^2$. Second, by definition of $\hat{B}$ above, $k \notin \hat{B}$ for all weights w. Third, it follows by the Karush–Kuhn–Tucker conditions for the 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 as $v_j = w_j$ for all $j \in \hat{A}^{\lambda, W, -k}$ and $v_j \leqslant w_j$ for all $j \notin \hat{A}^{\lambda, W, -k}$ (increasing the penalty on zero coefficients will leave them zero, if the penalty for non-zero coefficients is kept constant). Hence there is a set Ω_w in the sample space of W with $P_w(\Omega_w) \geqslant 1 - \delta_w$ such that $\|\{(X_{\hat{B}}^w)^{\mathrm{T}} X_{\hat{B}}^w\}^{-1} (X_{\hat{B}}^w)^{\mathrm{T}} X_k^w\|_1 \leqslant 2^{-1/4}$. Moreover, for the same set Ω_w, we have $\|X_k^w\|_2 / \|X_k\|_2 = \alpha \leqslant 1/s \leqslant 1/7$. Hence, for all $\omega \in \Omega_0 \cap \Omega_1$ and, for all $\omega \in \Omega_w$, the left-hand side of condition (33) is bounded from above by $\lambda_{\min}/7 + 2^{-1/4}\lambda < \lambda$ 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 $S_{\mathrm{small};\lambda} \subseteq S$, defined in theorem 2. For all $\omega \in \Omega_0$, however, it follows directly from lemma 4 that $(S \backslash S_{\mathrm{small};\lambda}) \subseteq \hat{A}^{\lambda, W}$. Hence, for all $k \in S \backslash S_{\mathrm{small};\lambda}$, the selection probability satisfies $\hat{\Pi}_k^{\lambda} = 1$ for all $Y \in \Omega_0$, which completes the proof.

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

References

Bach, F. (2008) Bolasso: model consistent lasso estimation through the bootstrap. In *Proc. 25th Int. Conf. Machine Learning*, pp. 33–40. New York: Association for Computing Machinery.

Banerjee, O. and El Ghaoui, L. (2008) Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Barbieri, M. and Berger, J. (2004) Optimal predictive model selection. *Ann. Statist.*, **32**, 870–897.

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D. and Meyerson, M. (2005) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Bioinformatics*, **21**, 3301–3307.

Bickel, P. and Levina, E. (2008) Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Breiman, L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

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

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

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

Conlon, E., Liu, X., Lieb, J. and Liu, J. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natn. Acad. Sci. USA*, **100**, 3339–3344.

Davis, C., Gerick, F., Hintermair, V., Friedel, C., Fundel, K., Kuffner, R. and Zimmer, R. (2006) Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, **22**, 2356–2363.

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

Dudoit, S., Shaffer, J. and Boldrick, J. (2003) Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, **18**, 71–103.

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

Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional variable selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 1989–2014.

Freedman, D. (1977) A remark on the difference between sampling with and without replacement. *J. Am. Statist. Ass.*, **72**, 681.

Freund, Y. and Schapire, R. (1996) Experiments with a new boosting algorithm. In *Proc. 13th Int. Conf. Machine Learning*, pp. 148–156. San Francisco: Morgan Kaufmann.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

van de Geer, S. (2008) High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36**, 614–645.

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

Huang, J., Ma, S. and Zhang, C.-H. (2008) Adaptive lasso for sparse high-dimensional regression models. *Statist. Sin.*, **18**, 1603–1618.

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

Lee, K., Sha, N., Dougherty, E., Vannucci, M. and Mallick, B. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

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

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

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc.* B, **70**, 53–71.

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

Meinshausen, N. and Yu, B. (2009) Lasso-type recovery of sparse representations from high-dimensional data. *Ann. Statist.*, **37**, 246–270.

Michiels, S., Koscielny, S. and Hill, C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.
Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, **2**, 494–515.
Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C. and Falciani, F. (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. *Adv. Computnl Math.*, **12**, 213–227.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tropp, J. (2004) Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, **50**, 2231–2242.
Valdar, W., Holmes, C., Mott, R. and Flint, J. (2009) Mapping in structured populations by resample model averaging. *Genetics*, **182**, 1263–1277.
Wainwright, M. (2009) Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Trans. Inform. Theor.*, **55**, 2183–2202.
Yuan, M. and Lin, Y. (2007) Model selection and estimation in the gaussian graphical model. *Biometrika*, **94**, 19–35.
Zhang, C.-H. (2007) Penalized linear unbiased selection. *Technical Report 2007-003*. Department of Statistics, Rutgers University, Piscataway.
Zhang, T. (2008) Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proc. Neural Information Processing Systems*. Boston: MIT Press .
Zhang, T. (2009) On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, **10**, 555–568.
Zhang, C. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zucknick, M., Richardson, S. and Stronach, E. A. (2008) Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statist. Appl. Genet. Molec. Biol.*, **7**, article 7.

Discussion on the paper by Meinshausen and Bühlmann

Sylvia Richardson *(Imperial College London)*

This stimulating paper on combining resampling with l_1-selection algorithms makes important contributions for the analysis of high dimensional data. What I found particularly appealing in this paper is that it puts on a firm footing the idea of using the stability under resampling to select a set of variables, by

(a) estimating for each variable X_k its inclusion probability $\hat{\Pi}_k^{\lambda}$ in resampled subsets and
(b) formulating a selection rule based on the maximum of these over a regularization domain Λ.

Using inclusion probabilities in resampled subsets had been discussed in an informal way for a considerable time in applied work, in particular in genomics. Early work on extracting prognostic signatures from gene expression data was soon questioned as it was noticed that such signatures had little reproducibility. The idea of intersecting or combining resampled signatures followed. For example Zucknick *et al.* (2008) investigated the stability of gene expression signature in ovarian cancer derived by different supervised learning procedures including the lasso, the elastic net and random forests by computing their inclusion frequencies under resampling for profiles of different sizes (see Fig. 2 of Zucknick *et al.* (2008)).

I shall focus my discussion on the variable selection aspect rather than the graphical modelling, and specifically I shall comment on two aspects:

(a) trying to understand better the applicability of the bound in theorem 1 and the performance of this method beyond the reported simulations;
(b) putting 'stability' into a broader context and relating or comparing it with other approaches.

The focus of theorem 1 is on the control of the familywise error rate, control which depends on two quantities: the threshold π_{thr} and the average number of selected variables over the regularization domain q_Λ. It is informative to work out the bounds that are obtained as successive variables are selected for particular thresholds. On the vitamin data set, using the recommended $q_\Lambda = p\sqrt{0.8} = 57$, $\max_{\lambda\in\Lambda}(\hat{\Pi}_k^{\lambda})$ reaches 0.9 for only one variable, with a bound $E(V) \leqslant 1$. Lowering π_{thr} to 0.6 selects three variables with a small $q_\Lambda = 3.9$,

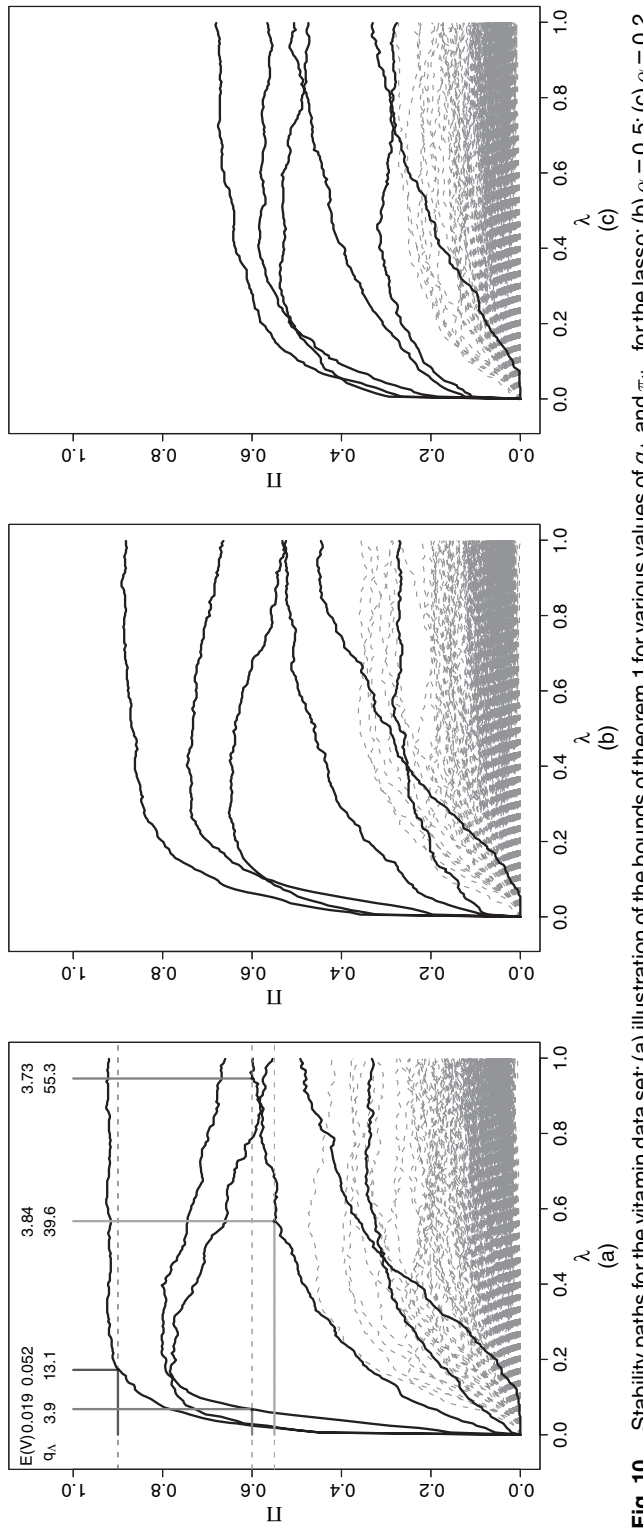


Fig. 10. Stability paths for the vitamin data set: (a) illustration of the bounds of theorem 1 for various values of q_Λ and π_{thr} for the lasso; (b) $\alpha = 0.5$; (c) $\alpha = 0.2$

and hence a useful bound $E(V) \leqslant 0.02$. But the bound on $E(V)$ reaches 3.8 if the domain is extended till a fourth variable is included. Hence, in this example, the bounds of theorem 1 would restrict the selection to three variables (Fig. 10(a)). If the practical use of the stability plots is extended to a ranking of the features according to the values of $\max_{\lambda \in \Lambda}(\hat{\Pi}_k^{\lambda})$ as suggested in the simulations, the bounds of theorem 1 thus appear to be quite conservative for deriving a cut-off.

With respect to the randomized lasso, the relevant quantities in the consistency theorem 2 are the threshold δ and the bound on the $\beta_k s$. Unfortunately, these quantities do not seem amenable to explicit computations. The authors seem to rely instead on a semiqualitative interpretation of the plots described in terms such as 'variables standing out', 'better separated', ... without giving quantitative guidelines on how to judge such a separation. However, the values of $\hat{\Pi}_k^{\lambda}$ are clearly influenced by the choice of weakness α (see Figs 10(b) and 10(c)), indicating that the thresholds for stability selection should be adapted with respect to α. Besides the elegant theoretical results of theorem 2, it is thus not entirely clear how to use the randomized stability paths in practice.

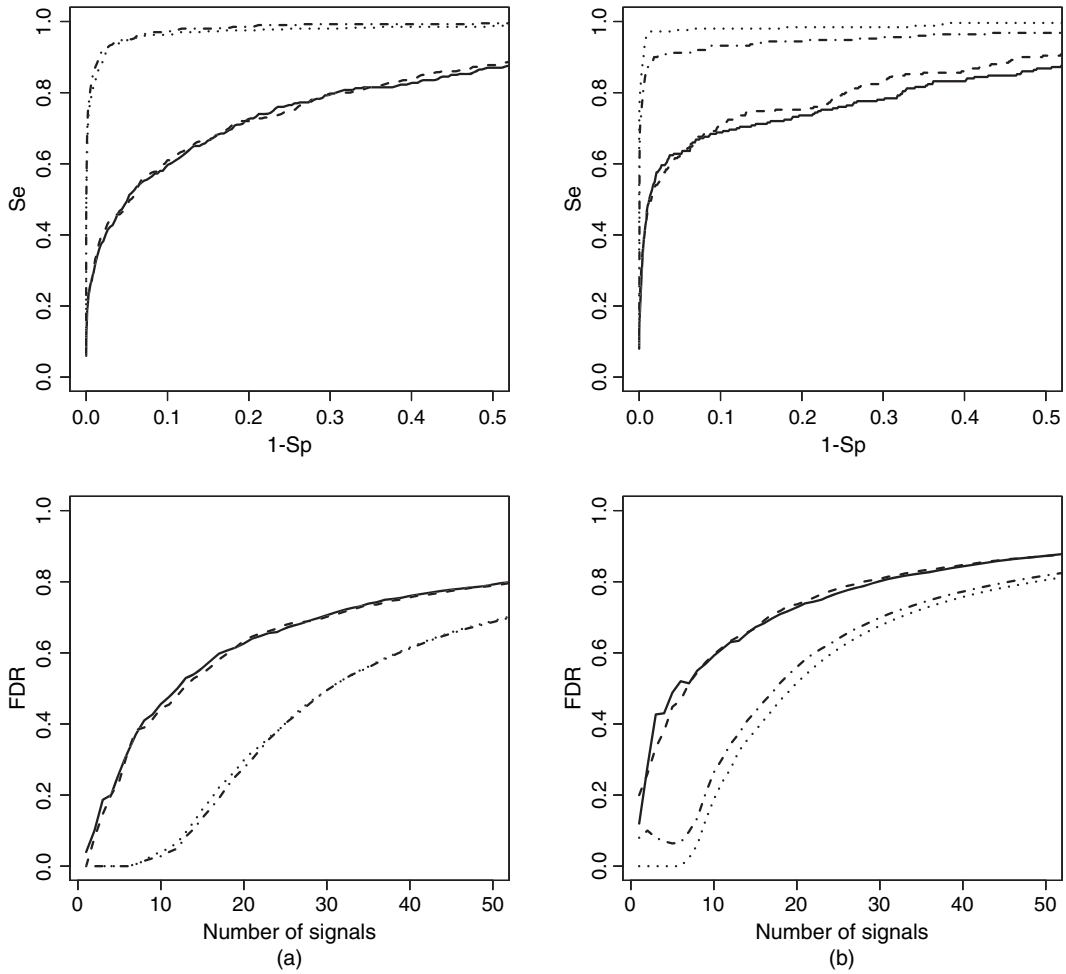


Fig. 11. Receiver operating characteristic curves comparing stability selection (with (– – –, $\alpha = 0.5$) and without (———) randomization) and Bayesian variable selection using either stochastic shotgun search (· - - - - -) or evolutionary stochastic search (·····, $E(p) = 5$) (variables are ranked by marginal probabilities of inclusion (Bayesian methods) or by $\max_{\lambda \in \Lambda}(\hat{\pi}_k^{\lambda})$; results are averaged over 25 simulated data sets and details of the simulation set-up can be found in Bottolo and Richardson (2010)): (a) $p = 300$, $n = 120$, $s = 16$ and average maximum correlation 0.68; (b) HapMap data example, $p = 775$, $n = 120$, $s = 10$ and average maximum correlation 0.88

Broadly speaking, stability selection and machine learning methods can both be viewed as 'ensemble learning' procedures following Hastie *et al.* (2009). Counting the number of times that a variable is selected in each of the resampled $n/2$ subsets for particular values of λ is just one way of combining the information of a collection of lasso learners. In this respect, it is a little surprising that the authors have not opened up the discussion on connections between their approach and ensemble methods such as 'bagging' or 'stacking'. Exploiting this connection could potentially lead to revisiting some of the choices made in their procedure, such as the set of learners that are combined (e.g. involving learners with more complex penalties such as in the elastic net) and the size of the subsamples, and to investigate the performance of combination rules that would exploit more than the marginal information, e.g. the order, or the stability of subsets.

Linking stability selection to Bayesian approaches provides further intriguing questions. It is well known that the penalty λ can be viewed as a parameter in a Laplace prior on the regressions coefficients β. To 'stabilize' inference, the authors take the maximum of $\hat{\Pi}_k^\lambda$ over a domain Λ. From a Bayesian perspective, the choice of using the maximum rather than some form of *integration over* Λ is questionable. Have the authors considered alternative choices to the maximum and would some of their results carry over?

This naturally leads me to discuss the connection with the Bayesian variable selection (BVS) context, where stability and predictive performance are achieved, not by resampling the data but by allowing parameter and model uncertainty. In this light, model averaging for BVS could be viewed as an ensemble method. There are several strategies for BVS, differing in their prior model of the regression coefficients and the model search strategy. One way (but by no means the only way) to exploit the output of the BVS search is to compute marginal probabilities of inclusion for each variable, averaging over the space of models that are visited. In the large p, small n paradigm, ranking the posterior probabilities of inclusion to select relevant variables is commonly done. Of course, when the covariates are dependent, joint rather than marginal thresholding should be also considered.

To understand better the power and sensitivity of stability selection, and to investigate further the claim that is made by the authors of empirical evidence of good performance even when the 'irrepresentable condition' is violated, we have implemented their procedure on a set of simulated examples under two scenarios of large p, small n, the first inspired by classical test cases for variable selection that were devised by George and McCullogh (1997) and the second based on phased genotype data from the HapMap project. In both cases, a few of the regressors have strong correlation with the noise variables. In parallel, we have run two Bayesian stochastic search algorithms, shotgun stochastic search (Hans *et al.*, 2007) and evolutionary stochastic search (Bottolo and Richardson, 2010), on the same data sets. Receiver operating characteristic curves and false discovery rate curves averaged over 25 replicates are presented in Fig. 11. It is clear from the plots that, in these two cases of large p, small n, good power for stability selection is only achieved at the expense of selecting a large number of false positive discoveries, a fact that can also be clearly seen in Fig. 7 of the paper. The Bayesian stochastic algorithms outperform stability selection procedures in the two scenarios. By their capacity to explore efficiently important parts of the parameter and model space and to perform averaging according to the support of each model, here Bayesian learners have an enhanced performance.

As can be surmised from my comments, I have found this paper enjoyable, thought provoking and rich for future research directions, and I heartily congratulate the authors.

John Shawe-Taylor and Shiliang Sun (*University College London*)

We congratulate the authors on a paper with an exciting mix of novel theoretical insights and practical experimental testing and verification of the ideas. We provide a personal view of the developments that were introduced by the paper, mentioning some areas where further work might be usefully undertaken, before presenting some results assessing the generalization performance of stability selection on a medical data set.

The paper introduces a general method for assessing the reliability of including component features in a model. They independently follow a similar line to that proposed by Bach (2008), in which the author proposed to run the lasso algorithm using boostrap samples and only included features that occur in all the models thus created. Meinshausen and Bühlmann refine this idea by assessing the probability that a feature is included in models created with random subsets of $\lfloor n/2 \rfloor$ training examples. Features are included if this probability exceeds a threshold π_{thr}.

Theorem 1 provides a theoretical bound on the expected number of falsely selected variables in terms of π_{thr} and q_Λ, the expected number of features to be included in the models for a fixed subset of the training data, but a range of values of the regularization parameter $\lambda \in \Lambda$. The theorem is quite general, but makes one non-trivial assumption: that the distribution over the inclusion of false variables is exchangeable.

In their evaluation of this bound on a range of real world training sets, albeit with artificial regression functions, they demonstrate a remarkable agreement between the bound value (chosen to equal 2.5) and the true number of falsely included variables.

We would have liked to have seen further assessment of the reliability of the bound in different regimes, i.e. bound values as fixed by different q_Λ and π_{thr}. The experimental results indicate that in the data sets that were considered the exchangeability assumption either holds or, if it fails to hold, does not adversely affect the quality of the bound. We believe that it would have been useful to explore in greater detail which of these explanations is more probable.

One relatively minor misfit between the theory and practical experiments was the fact that the theoretical results are in terms of the expected value of the quantities over random subsets, whereas in practice a small sample is used to estimate the features to include as well as quantities such as q_Λ. Perhaps finite sample methods for estimating fit with the assumption of exchangeability could also be considered. This might lead to an on-line approach where samples are generated until the required accuracy is achieved.

Theorem 2 provides a more refined analysis in that it also provides guarantees that relevant examples are included provided that they play a significant part in the true model, which is something that theorem 1 does not address. Though stability selection as defined refers to the use of random subsampling and all the experiments make use of this strategy, theorem 2 analyses the effect of a 'randomized lasso' algorithm that randomly rescales the features before training on the full set. Furthermore, the proof of theorem 2 does not make it easy for the reader to gain an intuitive understanding of the key ideas behind the result.

Our final suggestion for further elucidation of the usefulness of the ideas that are presented in the paper is to look at the effects of stability selection on the generalization performance of the resulting models. As an example we have applied the approach to a data set that is concerned with predicting the level of cholesterol of subjects on the basis of risk factors and single-nucleotide polymorphism genotype features.

The data set includes 1842 subjects or examples. The feature set (input) includes six risk factors (age, smo, bmi, apob, apoa, hdl) and 787 genotypes. Each genotype takes a value in $\{1, 2, 3\}$. As preprocessing, each risk factor is normalized to have mean 0 and variance 1. For each example, its output is the averaged level of cholesterol over five successive years. The whole data were divided into a training set of 1200 examples and a test set of the remaining 642 examples. We shall report the test performance averaged across 10 different random divisions of training and test sets. The performance is evaluated through the root-mean-square error. In addition to standard 'stability selection' we report performance for a variant in which complementary pairs of subsets are used.

We report results for four methods:

(a) ridge regression with the original features (method M1);
(b) the lasso with the original features (method M2);
(c) ridge regression with the features identified by stability selection (method M3);
(d) the lasso with the features identified by stability selection (method M4).

The variants of M3 and M4 based on complementary pairs of subsets are denoted M3c and M4c. The performances of the first two methods are independent of π_{thr} and provide a baseline given in Table 1.

For the two methods involving stability selection we experiment with values of π_{thr} from the set $\{0.2, 0, 25, 0.3, 0.35, 0.4, 0.45, 0.5\}$. The results for various values of π_{thr} for methods M3 and M4 using standard subsampling and the ranadomized lasso are given in Table 2, whereas using the complementary sampling gives the results of Table 3.

The results suggest that stability selection has not improved the generalization ability of the resulting regressors, though clearly the lasso methods outperform ridge regression. The performance is remarkably

Table 1. Mean (and standard deviations in parentheses) of the test performance and number of retained features for methods M1 and M2

	Results for method M1	*Results for method M2*
Root-mean-square error	0.752 (0.017)	0.707 (0.017)
Number of retained features	792 (0.66)	109 (5.22)

Table 2. Mean (and standard deviations in parentheses) of the test performance and number of retained features for methods M3 and M4

π_{thr}	*Number of features*	*Results for method M3*	*Results for method M4*
0.20	117.4 (6.2)	0.722 (0.017)	0.716 (0.017)
0.25	86.8 (5.2)	0.720 (0.016)	0.715 (0.016)
0.30	64.7 (4.1)	0.719 (0.017)	0.715 (0.017)
0.35	45.3 (4.1)	0.716 (0.016)	0.715 (0.017)
0.40	27.3 (3.8)	0.714 (0.016)	0.713 (0.016)
0.45	17.7 (1.9)	0.712 (0.016)	0.710 (0.016)
0.50	11.4 (1.6)	0.714 (0.019)	0.713 (0.019)

Table 3. Mean (and standard deviations in parentheses) of the test performance and number of retained features for methods M3c and M4c

π_{thr}	*Number of features*	*Results for method M3c*	*Results for method M4c*
0.20	116.5 (4.4)	0.721 (0.017)	0.715 (0.017)
0.25	83.8 (3.0)	0.720 (0.017)	0.715 (0.017)
0.30	62.4 (3.6)	0.718 (0.017)	0.714 (0.016)
0.35	44.2 (3.2)	0.717 (0.015)	0.716 (0.016)
0.40	27.4 (3.4)	0.714 (0.015)	0.713 (0.015)
0.45	18.2 (1.7)	0.714 (0.012)	0.710 (0.013)
0.50	11.8 (1.8)	0.715 (0.014)	0.713 (0.014)

stable across different values of π_{thr} despite the number of stable variables undergoing an order of magnitude reduction.

The vote of thanks was passed by acclamation.

Tso-Jung Yen (*Academia Sinica, Taipei*) **and Yu-Min Yen** (*London School of Economics and Political Science*)

We congratulate the authors for tackling a challenging statistical problem with an effective and easily implementable method. Our comments and interest in the paper are as follows. First, the authors claim that, under the method, tuning parameter λ is insensitive to the final result. However, we have found that it may still be affected by its range, particularly in the $p \leqslant m$ situation, where m is the subsampling size. In this situation, when $\lambda \to 0$, the subsampling estimation results of the lasso will approach those of ordinary least squares. Consequently, $\lambda_{\min} \to 0$ will lead to $\max_{\lambda \in \Lambda}(\hat{\Pi}_k^{\lambda}) \to 1$ for all $\{1, \ldots, p\} \in K$ with high probability.

The authors suggest the use of $m = n/2$. Although $p \leqslant n/2$ is not a common case in current genomic and genetic studies, it is often seen in other research disciplines. By using the diabetes data that were presented in Efron *et al.* (2004), we then demonstrate that unsuitable $\lambda_{\min}$ limits the use of equation (9) when $p \leqslant m$. We use

$$\hat{q}_{\Lambda} = \frac{1}{B} \sum_{i=1}^{B} \hat{S}^{\Lambda}(I_i)$$

to estimate q_{Λ}, where I_i is the index of the ith subsample. We fix $\lambda_{\max} = 100$ and $B = 200$ with $m = n/2$. Fig. 12 shows the estimation results of $\hat{q}_{\Lambda}$ with the lasso when $\lambda_{\min}$ varies. Suppose that we require $E(V) \leqslant 2$; then, with $\lambda_{\min} = 10.11$, the corresponding $\hat{q}_{\Lambda} \approx 4.57$. Calibrating the value into equation (9) we obtain an unfeasible value $\pi_{\text{thr}} = 1.022 > 1$.

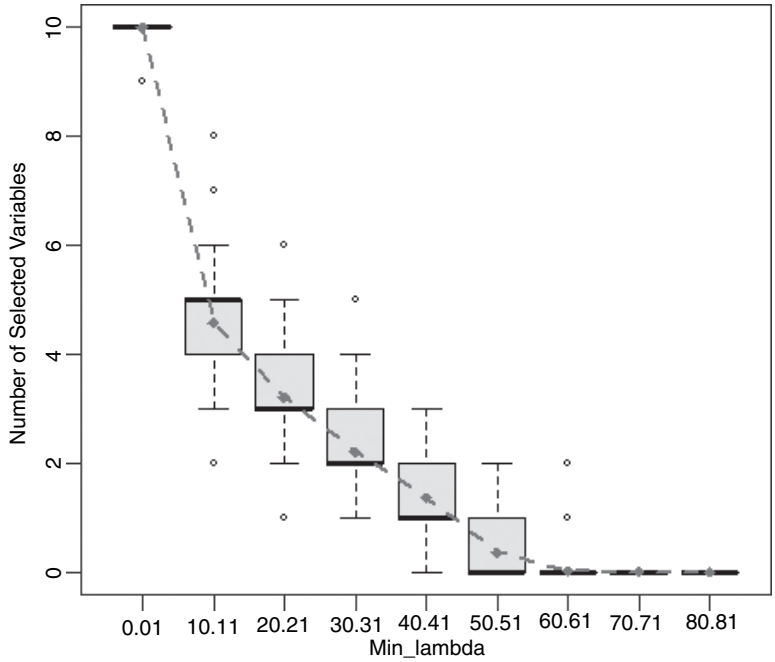


Fig. 12. Boxplots for the estimated number of selected variables (— — —, $\hat{q}_\Lambda$, average values of 200 subsampling estimations of $\hat{S}^\Lambda$): the data set used is the diabetes data presented in Efron *et al.* (2004), with $p = 10$ and $n = 442$; the method used is the lasso and $\lambda_{\min}$ is varied at nine different levels

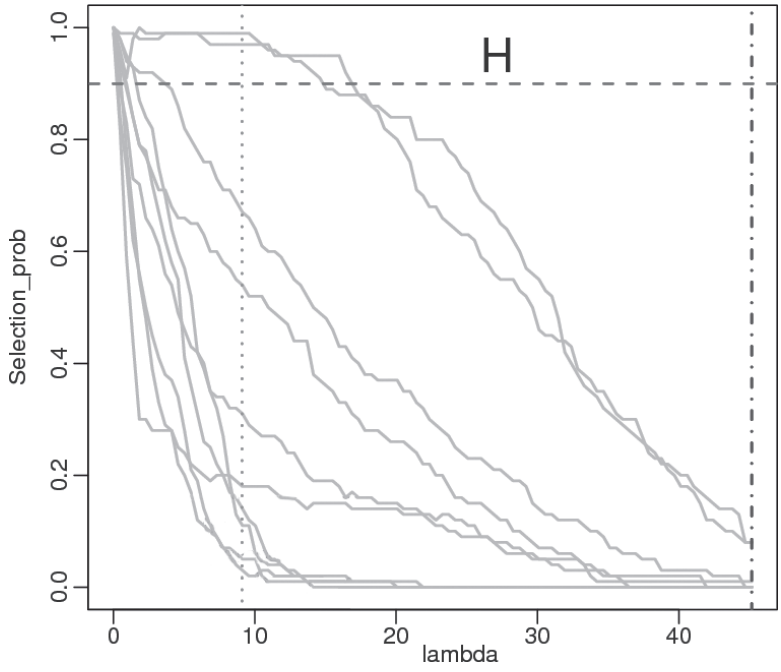


Fig. 13. Result of stability selection for the diabetes data with $\pi_{\text{thr}} = 0.9$ (– – –), $p = 10$, $q^* = 2.828$, $\hat{\lambda}_{\max} = 45.160$ (⋮) and $\hat{\lambda}_{\min} = 9.090$ (⋮); the method used is the randomized lasso with $W_k \in [0.5,1]$; the plot indicates that stability selection will only select variables with paths falling in region H

We propose a solution to this problem by directly estimating the regularization region $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ by $\hat{\lambda}_{\max} = \max_j |n^{-1} x_j^{\mathrm{T}} y|$ and

$$\hat{\lambda}_{\min} = \arg\max_{\lambda} \{ |\hat{\lambda}_{\max} - \lambda| : 0 \leqslant \lambda \leqslant \hat{\lambda}_{\max}, \hat{q}_{[\lambda, \hat{\lambda}_{\max}]} = q^* \}.$$

Given $E(V) \leqslant 2$, $\pi_{\mathrm{thr}} = 0.9$ and $p = 10$, we have $q^* = 4$, $\hat{\lambda}_{\max} = 45.160$ and $\hat{\lambda}_{\min} = 9.090$. The estimation results with the randomized lasso are shown in Fig. 13, which indicates that only paths falling in region H (two variables) are selected.

Secondly, in addition to $E(V)/p$, we may be also interested in controlling the false discovery rate $E(V/|\hat{S}^{\mathrm{stable}}|)$. Conventionally, the quantity may be approximated by $E(V)/E(|\hat{S}^{\mathrm{stable}}|)$, but it is unknown whether such an approximation works well in regression-based variable selection.

Finally, we are interested in what the relationship between q_Λ and the degrees of freedom of the lasso ($\mathrm{df}_{\mathrm{lasso}}$) is. As indicated in Efron *et al.* (2004) and Zou *et al.* (2007), $E(|\hat{S}^\lambda|) = \mathrm{df}_{\mathrm{lasso}}$. $E(|\hat{S}^\lambda(I)|) = q_\Lambda$ by definition. They are different: $|\hat{S}^\lambda|$ relies on a single λ and the whole sample, but $|\hat{S}^\lambda(I)|$ depends on Λ and the subsample set I. We are wondering whether they will have some common features, and this may be useful to link the method with other traditional methods such as C_p, Akaike's information criterion and the Bayes information criterion.

Rajen Shah and Richard Samworth (*University of Cambridge*)

We congratulate the authors for their innovative and thought-provoking paper. Here we propose a minor variant of the subsampling algorithm that is the basis of stability selection. Instead of drawing individual subsamples at random, we advocate drawing disjoint pairs of subsamples at random. This variant appears to have favourable properties.

Below, we use the same notation as the paper. Our method of subsampling involves splitting $\{1, \ldots, n\}$ into two halves at random and picking a subset of size $\lfloor n/2 \rfloor$ in each half. Repeating this M times, we obtain a sequence of subsets $I_1, \ldots, I_{2M}$ with $I_{2i} \cap I_{2i-1} = \emptyset$, $i = 1, \ldots, M$. For $k \in \{1, \ldots, p\}$, define

$$\tilde{\Pi}^\lambda_{k,M} = \frac{1}{2M} \sum_{i=1}^{2M} \mathbb{1}\{k \in \hat{S}^\lambda(I_i)\}.$$

Similarly to the stability selection algorithm, we select variable k when $\max_{\lambda \in \Lambda}(\tilde{\Pi}^\lambda_{k,M}) \geqslant \pi_{\mathrm{thr}}$.

(a) Letting V_M be the number of falsely selected variables $\mathbb{E}(V_M)$ satisfies the same upper bound as in theorem 1 of the paper. Briefly, defining

$$\tilde{\Pi}^{\mathrm{simult},\lambda}_{k,M} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{k \in \hat{S}^\lambda(I_{2i-1}) \cap \hat{S}^\lambda(I_{2i})\},$$

the result corresponding to lemma 1 of the paper is

$$0 \leqslant \frac{1}{M} \sum_{i=1}^{M} [1 - \mathbb{1}\{k \in \hat{S}^\lambda(I_{2i-1})\}][1 - \mathbb{1}\{k \in \hat{S}^\lambda(I_{2i})\}] = 1 - 2\tilde{\Pi}^\lambda_{k,M} + \tilde{\Pi}^{\mathrm{simult},\lambda}_{k,M}.$$

The arguments of lemma 2 and theorem 1 follow through since $\mathbb{E}(\tilde{\Pi}^{\mathrm{simult},\lambda}_{k,M}) = \mathbb{E}(\hat{\Pi}^{\mathrm{simult},\lambda}_{k})$. Thus we have the same error control as in the paper even for finite M, as well as the infinite subsampling case.

(b) Simulations suggest that we obtain a slight decrease in the Monte Carlo variance. A heuristic explanation is that, when n is even, each observation is contained in the same number of subsamples. This minimizes the sum of the pairwise intersection sizes of our subsamples.

(c) With essentially no extra computational cost, we obtain estimates of simultaneous selection probabilities, which can also be useful for variable selection; see Fan *et al.* (2009).

(d) If, in addition to the assumptions of theorem 1, we also assume that the distribution of $\max_{\lambda \in \Lambda}(\hat{\Pi}^{\mathrm{simult},\lambda}_{k})$ is unimodal, we obtain improved bounds:

$$\mathbb{E}(V_M) \leqslant \begin{cases} \dfrac{1}{2(2\pi_{\mathrm{thr}} - 1 - 1/2M)} \dfrac{q_\Lambda^2}{p} & \text{if } \pi_{\mathrm{thr}} \in (q_\Lambda^2/p^2 + \tfrac{1}{2}, \tfrac{3}{4}]; \\ \dfrac{4(1 - \pi_{\mathrm{thr}} + 1/2M)}{1 + 1/M} \dfrac{q_\Lambda^2}{p} & \text{if } \pi_{\mathrm{thr}} \in (\tfrac{3}{4}, 1]. \end{cases}$$

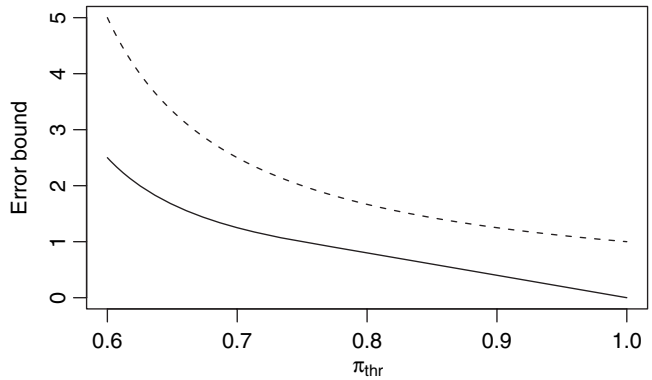


Fig. 14. Factor multiplying q_Λ^2/p against π_{thr} for each of the bounds: the bound of theorem 1 (-------) and the new bound with $M = \infty$ (———)

For a visual comparison between this bound and that of theorem 1, see Fig. 14. The improvement suggests that using sample splitting with this bound can lead to more accurate error control than using standard stability selection.

(e) This new bound gives guidance about the choice of M. For instance, when $\pi_{thr} = 0.6$ choosing $M > 52$ ensures that the bound on $\mathbb{E}(V_M)$ is within 5% of its limit as $M \to \infty$. When $\pi_{thr} = 0.9$, choosing $M > 78$ has the same effect.

Christian Hennig (*University College London*)

Stability selection seems to be a fruitful idea.

As usually done with variable selection, the authors present it as a mathematical problem in which the task is to pick a few variables with truly non-zero coefficients out of many variables with true $\beta_k = 0$. However, in reality we do not believe model assumptions to be precisely fulfilled, and in most cases we believe that the (in some sense) closest linear regression approximation to reality does not have any regression coefficients precisely equal to zero.

It is of course fine to have theory about the idealized situation with many zero coefficients, but in more realistic situations the quality of a variable selection method cannot be determined by considering models and data alone. It would be necessary to specify 'costs' for including or excluding variables with 'small' true β_k, which may depend on whether we would rather optimize the predictive quality or rather favour models with small numbers of variables enabling simple interpretations. We may even be interested in stability of the selection in its own right. Accepting the dependence of the choice of a method on the aim of data analysis, it would be very useful for promising methods such as stability selection to have a 'profile' of potential aims for which this is particularly suited, or rather not preferable.

Considering the author's remark at the end of Section 1.1, in Hennig (2010) it is illustrated in which sense the problem of finding the correct number of clusters s cannot be decided on the basis of models and data alone, and also in some simulation set-ups given there it turns out that s is not necessarily estimated most stable if it is chosen by a subsampling method looking for stable clusterings *given* s (based on 'prediction strength'; Tibshirani and Walther (2005)).

Paul D. W. Kirk, Alexandra M. Lewin and Michael P. H. Stumpf (*Imperial College London*)

We consider stability selection when several of the *relevant* variables are correlated with one another. Like the authors, we are interested in variable relevance, rather than prediction; hence we wish to select all relevant variables.

To illustrate, we use a simulated example, which is similar to that of the authors, in which $p = 500$, $n = 50$, the predictors are sampled from an $\mathcal{N}(0, \Sigma)$ distribution and the response is given by $Y = \Sigma_{i=1,\ldots,8} X_i + \varepsilon$, where ε is a zero-centred Gaussian noise term with variance 0.1. Here Σ is the identity matrix except for the elements $\Sigma_{1,2} = \Sigma_{3,4} = \Sigma_{4,5} = \Sigma_{3,5} = 0.8$ and their symmetrical counterparts. Thus two sets of predictors are correlated: $\{X_1, X_2\}$ and $\{X_3, X_4, X_5\}$.

The problem

For variables that are correlated with each other, different realizations of the simulation example above result in different stability paths; for example some realizations will stably select X_1 with high probability but not X_2, some will stably select X_2 but not X_1 (as in Fig. 15(a)) and others will select both variables with lower probability, and hence may not select either with sufficiently high probability to be chosen in the final analysis. In fact there is a clear relationship between the marginal selection probabilities for X_1 and X_2, as shown in Fig. 15(b), which shows these probabilities for 1000 realizations.

Solution 1

One approach is to use the lasso as before, but to calculate selection probabilities for sets of correlated predictors. Fig. 15(c) shows the stability paths for grouped predictors for the same realization as in Fig. 15(a),

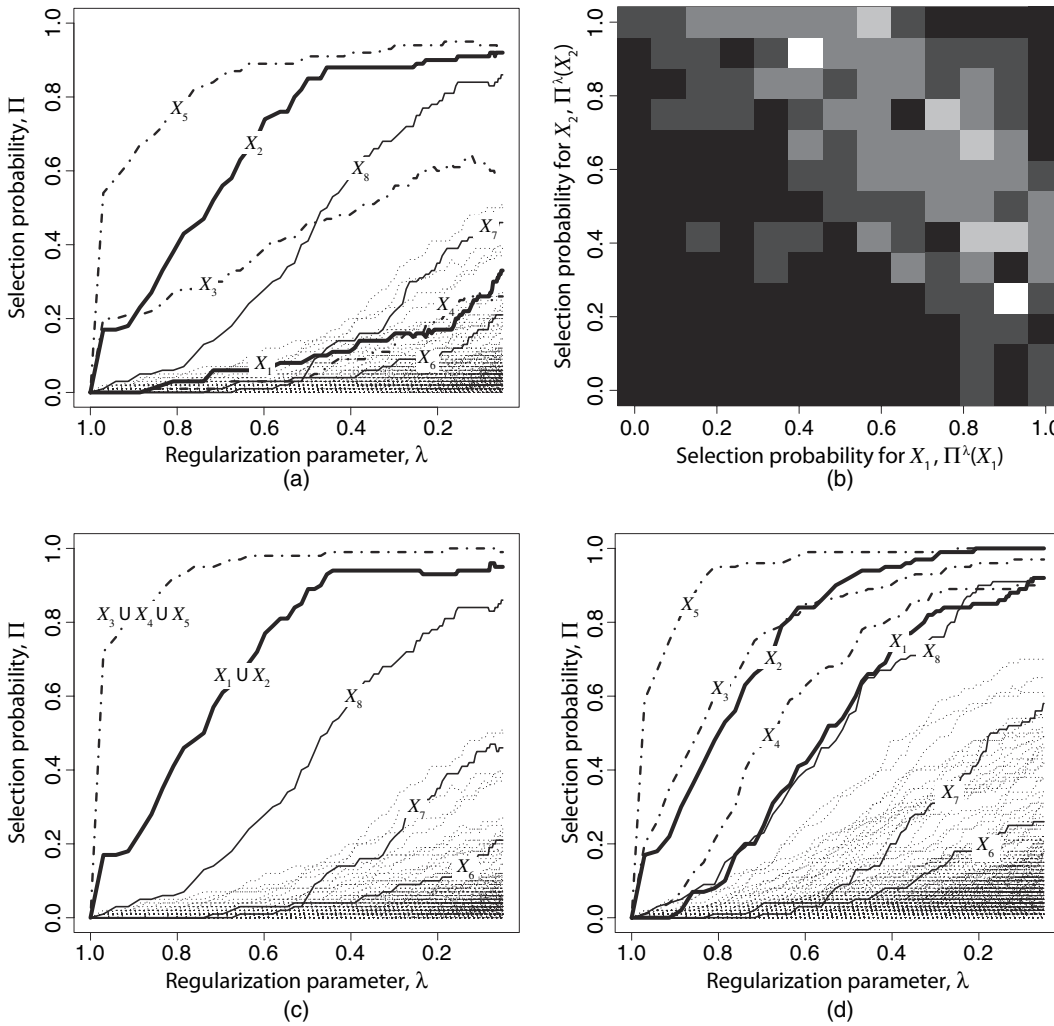


Fig. 15. (a) Stability path for a particular realization of the simulation example (———, X_1 and X_2; · · · · · ·, X_3, X_4 and X_5; ———, X_6, X_7 and X_8; · · · · · · · ·, irrelevant variables); (b) for 1000 realizations, selection probabilities for X_1 and X_2 at $\lambda = 0.25$ estimated by using the authors' subsampling method (the plot illustrates the density of these points in the 0–1 square (with lighter squares indicating higher density), showing a clear negative relationship); (c) using the same realization as in (a), the stability path when correlated variables are grouped together; (d) again using the same realization, a stability path by using the elastic net (with mixing parameter set to 0.2)

in which only one member of each correlated set would have been selected with high probability. Grouping them enables us to select the groups as required.

Solution 2

The obvious drawback to selection probabilities for groups is that the groups must be defined from the outset. We propose to use the elastic net (Zou and Hastie, 2005), which uses a linear combination of the lasso l_1-penalty and the ridge l_2-penalty. The l_2-penalty lets the algorithm include groups of correlated variables, whereas the l_1-penalty ensures that most regression coefficients are set to 0. We find that using marginal selection probabilities with the elastic net can give us all members of the correlated groups without defining them in advance, as shown in Fig. 15(d).

J. T. Kent (*University of Leeds*)

This has been a fascinating paper dealing, in particular, with the important problem of variable selection in regression. I have two simple questions about the methodology in this setting.

First, if we are willing to assume joint normality of the Y, X data, then all the information in the data will be captured by the sufficient statistics, namely the first two sample moments, together with the sample size n. Presumably there is no need to resample from the data in this situation; in principle, inferences could be made analytically from the set of sample correlations between the variables, though in practice a version of the parametric bootstrap might be used. More generally, the use of resampling methods seems to carry with it an implicit assumption or accommodation of non-normality and leads to the question how the methodology of the paper will be affected by different types of non-normality.

Second, I am not entirely clear what happens under approximate collinearity between the explanatory variables. In the conventional forward search algorithm in regression analysis, we are often faced with the situation where two variables x_1 and x_2 have similar explanatory power. If x_1 is in the model, then there is no need to include x_2; conversely, if x_2 is in the model there is no need to include x_1. If I understand your procedure correctly, you will tend to include x_1 half the time and x_2 half the time, leading to stability probabilities of about 50% each. If so, you might falsely conclude that neither variable is needed.

Axel Gandy (*Imperial College London*)

I congratulate the authors on their stimulating paper. The following comments concern the practical implementation of selecting stable variables.

The paper defines the set of stable variables in expression (7) as those k for which $\max_{\lambda\in\Lambda}(\hat{\Pi}_k^\lambda) \geqslant \pi_{\mathrm{thr}}$ for a fixed $0 < \pi_{\mathrm{thr}} < 1$. In practice, $\hat{\Pi}_k^\lambda$, and therefore also the set of stable variables, cannot be evaluated explicitly.

One way to evaluate $\hat{\Pi}_k^\lambda$ via Monte Carlo simulation is to generate J independent subsamples I_j of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ each and to approximate $\hat{\Pi}_k^\lambda$ by

$$\Pi_k^{*,\lambda} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\{k \in \hat{S}^\lambda(I_j)\}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Owing to the simulation, $\hat{\Pi}_k^\lambda$ and $\hat{\Pi}_k^{*,\lambda}$ can be on different sides of the threshold π_{thr}, leading potentially to a different set of stable variables.

To tackle this problem, the sequential algorithm in Gandy (2009) can be used. This sequential algorithm was originally suggested for the implementation of Monte Carlo tests, but it can also be used in the current situation for computing $\hat{\Pi}_k^\lambda$. Applied to the situation of the present paper, the algorithm will produce an estimate $\tilde{\Pi}_k^\lambda$ of $\hat{\Pi}_k^\lambda$ with a bound on the probability of $\hat{\Pi}_k^\lambda$ being on a different side of π_{thr} from $\hat{\Pi}_k^\lambda$. More precisely, for an (arbitrarily small) $\varepsilon > 0$,

$$\tilde{P}\{\mathbf{1}(\hat{\Pi}_k^\lambda \geqslant \pi_{\mathrm{thr}}) \neq \mathbf{1}(\tilde{\Pi}_k^\lambda \geqslant \pi_{\mathrm{thr}})\} \leqslant \varepsilon,$$

where $\tilde{P}$ denotes the probability distribution of the simulation conditionally on the observed data.

Besides this guaranteed performance, the algorithm in Gandy (2009) is a sequential algorithm and will come to a decision based on only a small number of samples if $\hat{\Pi}_k^\lambda$ is far from the threshold π_{thr}.

If Λ is a finite set with $\#\Lambda$ elements then also the following simultaneous bound on the sampling error for all variables can be obtained:

$$\tilde{P}\{\exists \lambda \in \Lambda, k \in \{1, \ldots, p\} : \mathbf{1}(\hat{\Pi}_k^\lambda \geqslant \pi_{\mathrm{thr}}) \neq \mathbf{1}(\tilde{\Pi}_k^\lambda \geqslant \pi_{\mathrm{thr}})\} \leqslant \varepsilon.$$

This can be accomplished by running the algorithm of Gandy (2009) for each λ and k with the Bonferroni corrected threshold $\varepsilon/(p\#\Lambda)$. These can be run in parallel using the same subsamples I_j. The Bonferroni

correction would be conservative. Devising a less conservative correction could be a topic for further research.

Howell Tong (*London School of Economics and Political Science*)
I join the others in congratulating the authors on a thought-provoking paper. It may be constructive to look beyond the independent and identically distributed data case and the exchangeable case. I would welcome the authors' reaction to what follows, some of which I have alluded to elsewhere (Tong, 2010). There are many examples of ill-posed problems with dependent data. I only need to mention the well-known seasonal adjustment, which is as old as time series itself. Akaike (1980) considered the classic decomposition of a time series $y_i, i=1,\ldots,n$, into $y_i = T_i + S_i + I_i$ where T_i, S_i and I_i are respectively the trend, the seasonal and the irregular component. He treated the problem as one of smoothing. By incorporating the prior belief of some underlying structural stability, e.g. a smooth trend or gradual change of the seasonal component, he minimized

$$\sum_{i=1}^{n} \{(y_i - T_i - S_i)^2 + d^2(A_i^2 + r^2 B_i^2 + z^2 C_i^2)\},$$

where d, r and z are regularization parameters, $A_i = T_i - 2T_{i-1} + T_{i-2}$, $B_i = S_i - S_{i-12}$ and $C_i = S_i + S_{i-i} + \ldots + S_{i-11}$. He estimated the regularization parameters by adopting a Bayesian approach based on what he called the 'ABIC', which is minus 2 times the log-likelihood of a Bayesian model. Akaike's approach has three important aspects:

(a) treating the problem as one of smoothing;
(b) focusing on structural stability rather than variable selection;
(c) treating the regularization parameters as some hyperparameters in a Bayesian framework.

Finally, I have a minor question. Have the authors tried to use their stability selection on model selection in time series modelling?

Chris Holmes (*Oxford University*)
The authors are to be congratulated on a ground breaking paper. The following comments are made from the perspective of a casual Bayesian observer.

In applied statistics it is quite rare to encounter problems of hard or crisp variable selection whereby the statistician is required to select a single subset of variables on the basis of the output of a computer program. When such circumstances arise, there is almost by definition the notion of an action to be performed using the variable set and hence a loss function $l\{a(\hat{S}_k), S_{k'}\}$ occurred in taking actions, $a(\cdot)$, using $\hat{S}_k$ when nature is really in $S_{k'}$. An optimal way to proceed then is according to the principle of maximum expected utility, i.e. to choose $\hat{S}_k$ so as to minimize your expected loss,

$$\hat{S}_k = \arg\min_{\tilde{S}} \Big[\sum_{S} l\{a(\tilde{S}_k), S\}\, \pi(S|\mathcal{D}) \Big]$$

where $\pi(\mathcal{S}|\mathcal{D})$ is the posterior mass assigned to subset S which characterizes all the information about the unknown state of nature. Such an approach provides provably coherent decision making in the face of uncertainty and is prescriptive in how to take actions by using variable selection; see Lindley (1968) and Brown *et al.* (1999) for instance. It feels to me that crisp variable selection without incorporation of utility is a little like breaking the eggs without making the omelette. The job is only half done—with apologies to Savage (1954) for the analogy.

Crisp variable subset selection is rare but much more common is for the statistician to work with the owner of the data to determine the relevance of the measured variables and to understand better the dependence structures within the data, i.e. as part of a dialogue whereby statistical evidence is combined with expert judgement. On the one hand the Bayesian works with the posterior distribution $\hat{\Pi}(\mathcal{S}|\mathcal{D}) = \mathcal{F}[\{y, x\}, \pi(\mathcal{S})]$, which is a function of the data and a prior distribution which captures any rich domain knowledge that may exist; on the other hand stability selection reports $\hat{\Pi}(\mathcal{S}) = \mathcal{F}_{\mathcal{A}}(\{y, x\}, \mathcal{A})$, which is a function of the data and the algorithm. It seems to the casual Bayesian observer that the former is more objective while providing a formal mechanism to incorporate any domain knowledge which might exist about the problem.

The following contributions were received in writing after the meeting.

Ismaïl Ahmed and Sylvia Richardson (*Imperial College London*)

The object of this contribution is to discuss further the vitamin example that was provided by the authors. This example is given to 'see how the lasso and the related stability path cope with noise variables'. It shows that, on the basis of a graphical analysis of the stability path, we can select five of the six permuted genes whereas, with the lasso path, the situation seems to be much less clear.

Thanks to the authors, we had the opportunity to reanalyse the vitamin data set that was used in the paper. The first thing that we would like to remark is that by performing a simple univariate analysis, i.e. by using each of the 4088 genes one at a time and then adjusting the corresponding p-values for multiplicities at a 5% level for the false discovery rate, we also pick up five of the six unpermuted covariates. The results are illustrated by Fig. 16, which also shows that there is an important discrepancy between the first five q-values and the remaining values.

Furthermore, we also performed a standard multivariate regression analysis restricted to the six unpermuted covariates, removing thus all the noise variables. The results, which are illustrated in Table 4, show that only one unpermuted gene is associated with a p-value that is less than 0.05 and that three unpermuted genes have a p-value that is less than 0.10. Consequently, it seems unclear whether any multivariate selection method could or should pick up more than these three variables. And indeed, when applying the shotgun stochastic search algorithm of Hans *et al.* (2007) on the whole data set with 20 000 iterations, no more than these three variables could possibly be selected with regard to their posterior importance measure (as defined in equation (2) of Hans *et al.* (2007)) over the 100 000 top visited models.

It thus seems to us puzzling that, on this example, stability selection behaves more like a univariate approach rather than a multivariate approach.

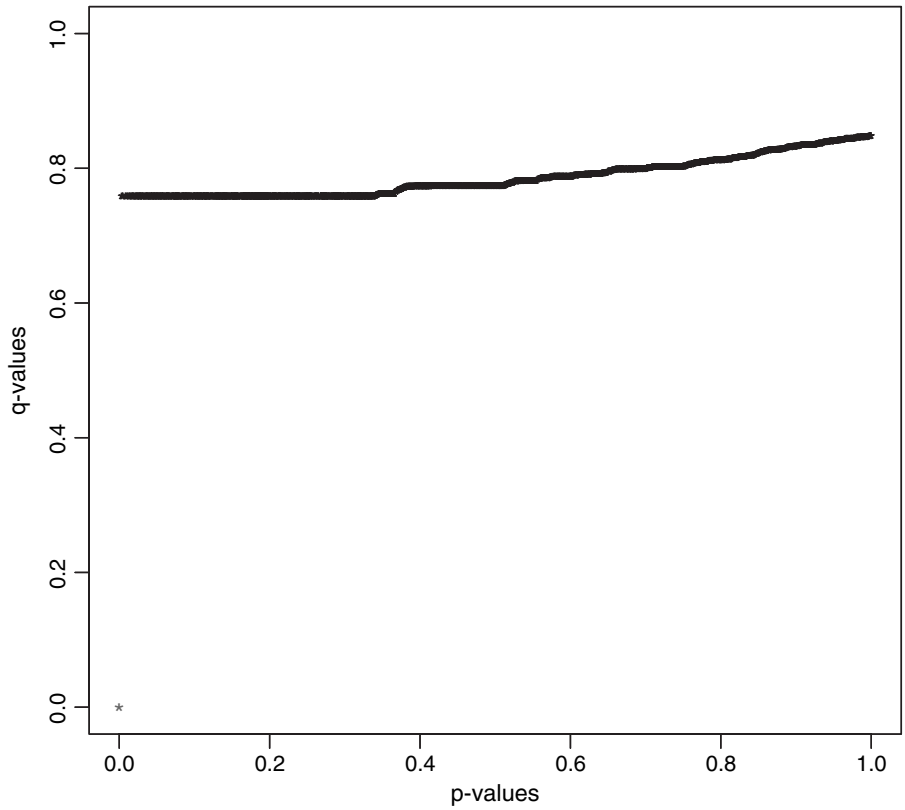


Fig. 16. Estimated q-values according to the p-values resulting from the 4088 univariate analyses: the estimated q-values are obtained with the location-based estimator (Dalmasso *et al.*, 2005); for a false discovery rate of 0.05 or lower, five null hypotheses are rejected (∗)

Table 4. Results of a multivariate regression restricted to the six unpermuted covariates

Parameter	*Estimate*	*Standard error*	*t-value*	*Pr(>|t|)*
(Intercept)	−7.7107	0.1782	−43.27	0.0000
X1407	−0.1221	0.1912	−0.64	0.5246
X1885	0.6665	0.3888	1.71	0.0894
X3228	−0.1094	0.2716	−0.40	0.6880
X3301	0.4697	0.2750	1.71	0.0905
X3496	0.6183	0.3077	2.01	0.0470
X3803	−0.1729	0.3271	−0.53	0.5982

Phil Brown and Jim Griffin (*University of Kent, Canterbury*)

We comment on the use of the randomized lasso (equation (13)). This, with subsampling, attempts to remedy the seductive appeal of convex penalization, which is a property of the lasso. Demanding a single solution when there is inherent uncertainty and interchangeability of predictors leads to the present paper's suggestion of subsampling for inference. In the Bayesian modal analysis of Griffin and Brown (2007) it is the multiplicity from a non-convex penalization which allows posterior exploration of alternative models without the need for external randomization. Our generalization of the lasso, a hierarchical scale mixture-of-normals prior, is the flexible normal–exponential–gamma distribution. The first two stages generate a double exponential, which is the equivalent of L_1-penalization. The third stage puts a gamma$(\alpha, 1/\lambda^2)$ distribution on the natural parameter of the exponential second-stage mixing. Thus the penalization can be written as

$$\lambda \sum_{k=1}^{p} \sqrt{a_k}|\beta_k|,$$

where $a_1, a_2, \ldots, a_p$ are independently realized χ^2 random variables with 2α degrees of freedom weighting each β_k in each simulation. This third-stage gamma distribution is somewhat different from the authors' advocacy of an inverse truncated uniform distribution, whose implied prior distribution for β_k is less natural, and we feel needs more justification. Combining all three stages, the $\beta_1, \beta_2, \ldots, \beta_p$ are independent and identically distributed *a priori* where β_k follows a normal–exponential–gamma distribution which can be explicitly written in terms of a parabolic cylinder function, and is a unimodal spiked distribution with tails whose heaviness depends on the shape parameter α. When $\alpha = 0.5$, it is the quasi-Cauchy distribution of Johnstone and Silverman (2005), and the robustness prior of Berger (1985), section 4.7.10. The third-stage stochastic generation of the gamma (i.e. χ^2-) distribution gives a *stochastic lasso* allowing fast algorithms such as LARS, and we thank the authors for that suggestion. We would ask though whether the other form of randomization, subsampling of observations, is necessary with such rich stochastic weighting possibilities. It is better to generate prior data than to throw away real data.

We have also given a full Bayesian analysis in Griffin and Brown (2010) illustrating the limitations of straight lasso penalization using another robustness prior, the variance gamma prior.

David Draper (*University of California, Santa Cruz*)

I have two comments on this interesting and useful paper.

(a) I am interested in pursuing connections with Bayesian ideas beyond the authors' mention of the Barbieri and Berger (2004) results. I reinterpret three of the concepts in the present paper in Bayesian language.
   (i) Frequentist penalization of the log-likelihood to regularize the problem is often equivalent to Bayesian choice of a prior distribution (for example, think of the l_1-norm penalty term in the lasso in the paper's equation (2) as a log-prior for β; might there be an even better prior to achieve the goals of this paper?).
   (ii) Under the assumption that the rows $Z^{(i)}$ of the data matrix are sampled independently from the underlying data-generating distribution, resampling the existing data is like sampling from the posterior predictive distribution (given the data seen so far) for future rows in the data matrix

(and of course, if we already had enough such rows, it would no longer be true that $p \gg n$ and the good predictors would be more apparent).

(iii) When estimated via resampling, stability paths are just Monte Carlo approximations to expectations of the indicator function, for inclusion of a given variable, with respect to the posterior distribution for the unknowns in the model.

I bring this up because it has often proved useful in the past to figure out what is the Bayesian model (both prior and likelihood) for which an algorithmic procedure, like that described in this paper, is approximately a good posterior summary of the quantities of interest, because even better procedures can then be found; a good example is the reverse engineering of neural networks from a Bayesian viewpoint by Lee (2004). Can the authors suggest the next steps in this algorithm-to-Bayes research agenda applied to their procedure?

(b) The authors have cast their problem in inferential language, but the real goal of structure discovery is often decision making (for instance, a drug company trying to maximize production of riboflavin in the example in Section 2.2 will want to decide in which genes to encourage mutation; money is lost both by failing to pursue good genes and by pursuing bad ones); moreover, when the problem is posed inferentially there is no straightforward way to see how to trade off false positive against false negative selections, whereas this is an inescapable part of a decision-making approach. What does the authors' procedure look like in a real world problem when it is optimized for making decisions, via maximization of expected utility?

Zhou Fang (*Oxford University*)

The authors suggest an interesting and novel way of enhancing variable selection methods and should be congratulated for their contribution. However, recalculation with subsamples can be computationally costly. Here, a heuristic is suggested that may allow some benefits of stability selection with the lasso without resampling.

Consider the weighted lasso, for observation weights W. Differentiation gives then, for $\tilde{X}$, $\tilde{\beta}$ being the design matrix and coefficient estimates for the selected variables $\hat{S}^\lambda$,

$$\tilde{\beta} = (\tilde{X}^{\mathrm{T}} W \tilde{X})^{-1}(\tilde{X}^{\mathrm{T}} W Y - \lambda).$$

Differentiating by W gives, after substitution,

$$A := \left.\frac{\mathrm{d}\tilde{\beta}}{\mathrm{d}W}\right|_{W=1} = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\,\tilde{X}^{\mathrm{T}}\mathrm{diag}(Y - \tilde{X}\tilde{\beta}).$$

$A\delta$ is a linear approximation to changes in the lasso non-zero estimates under a small perturbation of weights δ. This is directly analogous to influence calculations in normal linear regression. (See for example Belsley *et al.* (1980).) Although this approximation may be inaccurate, the components of A will all already be calculated as part of LARS or similar algorithms, reducing the computational burden.

Consider a simulation. Using $n = 50$ and $p = 100$, generate $(Z^{(1)}, \ldots, Z^{(p)}, V)$ independently standard normal and set $X^{(k)} = Z^{(k)} + V$. We set $Y = X\beta + \varepsilon$, with $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$ and ε independent $N(0, \frac{1}{4})$ noise. Fig. 17 shows results from various variable selection techniques.

In Fig. 17(a), V creates correlation in the predictors, hence generating false positive selections. For high regularization, these spurious fits rival the true fits in magnitude. As suggested in the paper, stability selection is more effective at controlling false positive selections.

For perturbation methods based on the original single lasso path, we may examine the relative norms of the rows $A_1, \ldots, A_{|\hat{S}^\lambda|}$ of A, as defined by

$$\|A_k\|^2_{\mathrm{REL}} = \sum_{i=1}^{n}\left(\frac{A_{ki}}{\tilde{\beta}_k}\right)^2.$$

In our experiments, we see that calculating this also distinguishes the true covariates from the irrelevant covariates—with most λ, the true positive selections vary relative to their estimates much less under perturbation. For Fig. 17(d), we naively convert the relative norms into an approximate upper bound on selection probability under weight perturbations of variance 1, using the Cantelli inequality:

$$P(\tilde{\beta}_k = 0) \approx P(-A_k\delta > \tilde{\beta}_k) \leqslant 1 - \frac{1}{\tilde{\beta}_k^2/\mathrm{var}(A_k\delta) + 1} = 1 - \frac{1}{1/\|A_k\|^2_{\mathrm{REL}} + 1}.$$

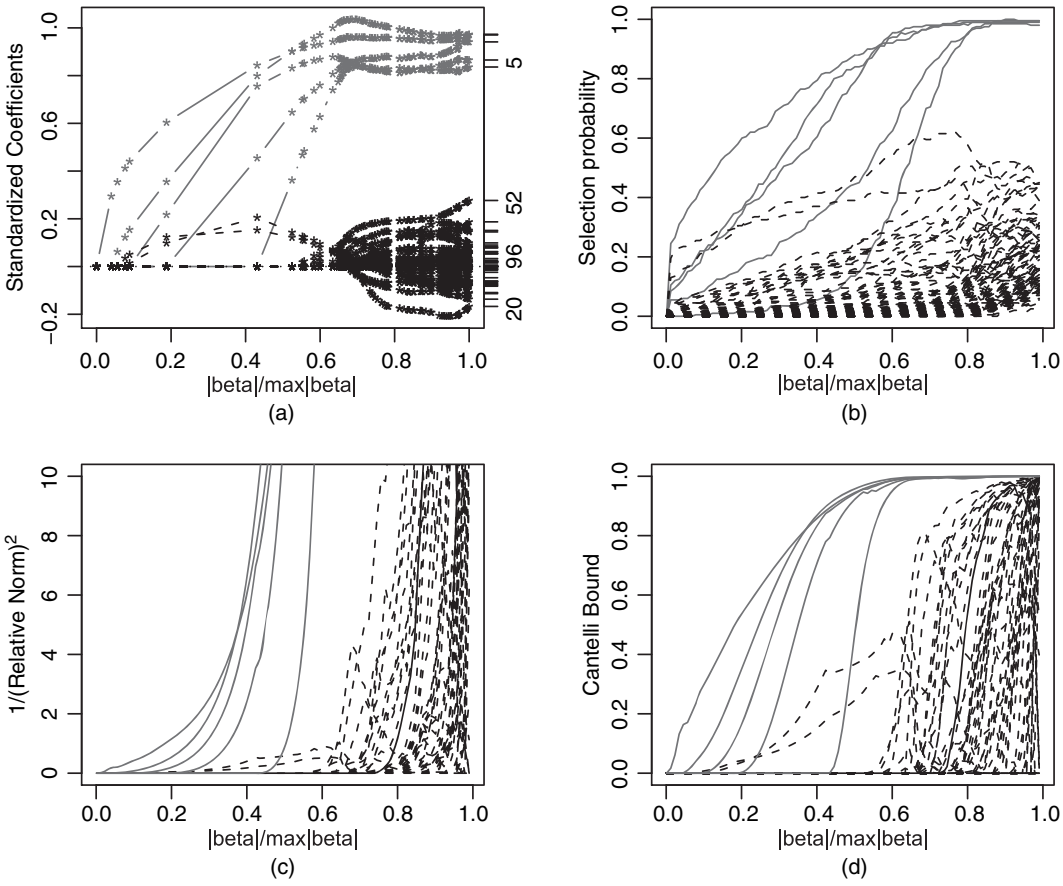


Fig. 17. (a) Lasso, (b) stability selection, (c) perturbation and (d) perturbation probability paths on a simulated data set: ———, true covariates; -------, irrelevant covariates

This approximates the stability selection path, except for low regularization, where failure to consider cases of unselected variables being selected under a perturbation means that we overestimate stability. However, this calculation, even inefficiently implemented, took 0.5 s to compute, compared with 14 s for 200 resamples of stability selection.

Torsten Hothorn (*Ludwig-Maximilians-Universität München*)

Stability selection brings together statistical error control and model-based variable selection. The method, which controls the probability of selecting—and thus potentially interpreting—a model containing at least one non-influential variable, will increase the confidence in scientific findings obtained from high dimensional or otherwise complex models.

The idea of studying the stability of the variable selection procedure applied to a specific problem by means of resampling is simple and easy to implement. And the authors point out that this straightforward approach has actually been used much earlier. The first reference that I could find is a paper on model selection in Cox regression by Sauerbrei and Schumacher (l992). Today, multiple-testing procedures utilizing the joint distribution of the estimated parameters can be applied in such low dimensional models for variable and structure selection under control of the familywise error rate (Haufe *et al.* (2010) present a nice application to multivariate time series). With their theorem 1, Nicolai Meinshausen and Peter Bühlmann now provide the means for proper error control also in much more complex models.

Two issues seem worth further attention to me: the exchangeability assumption that is made in theorem 1 and the prediction error of models fitted by using only the selected variables. One popular approach for

variable selection in higher dimensions is based on the permutation variable importance measure that is used in random forests. Interestingly, it was found by Strobl *et al.* (2008) that correlated predictor variables receive a higher variable importance than is justified by the data-generating process. The reason is that exchangeability is (implicitly) assumed by the permutation scheme that is applied to derive these variable importances. The problem can be addressed by applying a conditional permutation scheme and I wonder whether a more elaborate resampling technique taking covariate information into account might allow for a less strong assumption for stability selection as well.

Concerning my second point, the simulation results show that stability selection controls the number of falsely selected variables. I wonder how the performance (measured by the out-of-sample prediction error) of a model that is fitted to only the selected variables compares with the performance of the underlying standard procedure (including a cross-validated choice of hyperparameters). If the probability that an important variable is missed by stability selection is low, there should not be much difference. However, if stability selection is too restrictive, I would expect the prediction error of the underlying standard model to be better. This would be another hint that interpretable models and high prediction accuracy might not be achievable at the same time.

Chenlei Leng and David J. Nott (*National University of Singapore*)

We congratulate Meinshausen and Bühlmann on an elegant piece of work which shows the usefulness of introducing additional elements of randomness into the lasso and other feature selection procedures through subsampling and other mechanisms. It is now well understood that certain restrictive assumptions (Zhao and Yu, 2006; Wainwright, 2009) must be imposed on the design matrix for the lasso to be a consistent model selector although adaptive.versions of the lasso can circumvent the problem (Zou, 2006). However, as convincingly pointed out by Meinshausen and Bühlmann, by considering multiple sparse models obtained from perturbations of the original feature selection problem the performance of the original lasso, which uses just a single fit, can be improved.

We believe that a Bayesian perspective has much to offer when thinking about randomized versions of the lasso. We offer two alternative approaches, where the randomness comes from an appropriate posterior distribution.

(a) Our first approach puts a prior on the parameters in the full model. Given a draw of the parameters, say β^* from the posterior distribution, we consider projecting the model that is formed by this realization onto subspaces defined via some form of l_1-constraint on the parameters. Defining the loss function as the expected Kullback–Leibler divergence between this model and its projection, we use any of the following constraints on the subspace

$$S(\lambda) = \{\beta : \sum_{k=1}^{p} |\beta_k| \leqslant \lambda\},$$

$$S(\beta^*, \lambda) = \{\beta : \sum_{k=1}^{p} |\beta_k|/|\beta_k^*| \leqslant \lambda\}$$

inspired by the lasso and adaptive lasso penalty respectively. Owing to the l_1-penalty, in the posterior distribution of the projection there is positive probability that some parameters are exactly 0 and the posterior distribution on the model space that is induced by the projection allows exploration of model uncertainty. This idea is discussed in Nott and Leng (2010) and extends a Bayesian variable selection approach of Dupuis and Robert (2003) which considers projections onto subspaces that are defined by sets of active covariates.

(b) In on-going work, we consider the following adaptive lasso (Zou, 2006):

$$\|Y - X\beta\|^2 + \sum_{k=1}^{p} \lambda_k |\beta_k|. \tag{35}$$

In comparison with the usual methods which determine a single estimate of $\{\lambda_k\}_{k=1}^p$ (Zou, 2006; Wang *et al.*, 2007),we generalize the Bayesian lasso method in Park and Casella (2008) to produce a posterior sample of $\{\lambda_k\}_{k=1}^p$, which is denoted as $\{\lambda_k^{*b}\}_{k=1}^p, b = 1, \ldots, B$. For each b, we plug $\{\lambda_k^{*b}\}_{k=1}^p$ into expression (35), which gives a sparse estimate β^{*b} of β. The estimated parameters $\{\beta^{*b}\}_{b=1}^B$ can then be used for prediction and assessing model uncertainty. This is very much like the randomized lasso of Meinshausen and Bühlmann, but the randomness enters very naturally through a posterior distribution on hyperparameters. Our preliminary results show that this approach works competitively in prediction and model selection compared with the lasso and adaptive lasso.

Rebecca Nugent, Alessandro Rinaldo, Aarti Singh and Larry Wasserman (*Carnegie Mellon University, Pittsburgh*)

Meinshausen and Bühlmann argue for using stability-based methods. We suspect that the methods that are introduced in the current paper will generate much interest.

Stability methods have gained popularity lately. See Lange *et al.* (2004) and Ben-Hur *et al.* (2002) for example. There are cases where stability can lead to poor answers (Ben-David *et al.*, 2006). Some caution is needed.

General view of stability

Let $\{\hat{\theta}_h : h \in H\}$ be some class of procedures indexed by a tuning parameter h. We think of larger h as corresponding to larger bias. Our view of the stability approach is to use the least biased procedure subject to having an acceptable variability. This has a Neyman–Pearson flavour to it since we optimize what we cannot control subject to bounds on what we can control. The advantage is that variance is estimable whereas bias, generally, is not. There is no notion of approximating the 'truth' so it is not required that the model be correct. In contrast, Meinshausen and Bühlmann seem to be more focused on finding the 'true structure'.

Rinaldo and Wasserman (2010) applied this idea to finding stable density clusters as follows. Randomly split the data into three groups $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$. Construct a kernel density estimator $\hat{p}_h$ from X (with bandwidth h) and construct a kernel density estimator $\hat{q}_h$ from Y. Define the instability by

$$\Xi(h) = \hat{P}_Z(\{\hat{p}_h > \lambda\} \Delta \{\hat{q}_h > \lambda\})$$

where $\hat{P}_z$ is the empirical distribution based on Z. Under certain conditions, Rinaldo and Wasserman (2010) showed the following theorem.

Theorem 3. Let h_* be the diameter of $\{p > \lambda\}$ and let d be the dimension of the support of X_i. Then:

(a) $\Xi(0) = 0$ and $\Xi(h) = 0$, for all $h \geqslant h_*$;
(b) $\sup_{0<h<h_*}[\mathbb{E}\{\Xi(h)\}] < \frac{1}{2}$;
(c) As $h \to 0$, $\mathbb{E}\{\Xi(h)\} \asymp h^d$;
(d) for each $h \in (0, h_*)$,

$$c_1^n (h_* - h)^{d(n+1)} \leqslant \mathbb{E}\{\Xi(h)\} \leqslant 2c_2^n (h_* - h)^{n+1}$$

for constants c_1 and c_2.

We suggest using

$$\hat{h} = \inf[h : \sup_{t>h} \{\Xi(t)\} \leqslant \alpha], \tag{36}$$

where $\Xi(h)$ measures the variability and α is a user-defined acceptable amount of variability. Currently, we are generalizing the results to hold under weaker conditions and to hold uniformly over cluster trees rather than a single level set. The same ideas can be applied to graphs.

True structure?

The authors spend time discussing the search for true structure. In general, we feel that there is too much emphasis on finding true structure. Consider the linear model. It is a virtual certainty that the model is wrong. Nevertheless, we all use the linear model because it often leads to good predictions. The search for good predictors is much different from the search for true structure. The latter is not even well defined when the model is wrong, which it always is.

Adam J. Rothman, Elizaveta Levina and Ji Zhu (*University of Michigan, Ann Arbor*)

We congratulate the authors on developing a clever and practical method for improving high dimensional variable selection, and establishing an impressive array of theoretical performance guarantees. We are particularly interested in stability selection in graphical models, which is illustrated with one brief example in the paper. To investigate the performance of stability selection combined with the graphical lasso a little further, we performed the following simple simulation. The data are generated from the $N_p(0, \Omega^{-1})$ distribution, where $\Omega_{ii} = 1$, $\Omega_{i,i-1} = \Omega_{i-1,i} = 0.3$ and the rest are 0. We selected $p = 30$ and $n = 100$, and performed 50 replications. Stability selection with pointwise control was implemented with bootstrap samples of size $n/2$ drawn 100 times.

We selected four different values of the tuning parameter λ for the graphical lasso, which correspond

to the marked points along the receiver operating characteristic (ROC) curves for the graphical lasso in Fig. 18. The ROC curve showing false positive and true positive rates of detecting 0s in Ω for the graphical lasso was obtained by varying the tuning parameter λ and averaging over replications. For each fixed λ, we applied stability selection varying π_{thr} within the recommended range of 0.6–0.9, which resulted in an ROC curve for stability selection. The ROC curves show that stability selection reduces the false positive rate, as it should, and shifts the graphical lasso result down along the ROC curve; essentially, it is equivalent to the graphical lasso with a larger λ. Figs 18(a) and 18(b) have λs which are too small, and stability selection mostly improves on the graphical lasso result, but it does appear somewhat sensitive to the exact value of λ: if λ is very small (Fig. 18(a)), stability selection only improves on the graphical lasso for large values of π_{thr}. In Figs 18(c) and 18(d), λ is just right or too large, and then applying stability selection makes the overall result worse. This example confirms that stability selection is a useful computational tool to improve on the false positive rate of the graphical lasso when tuning over the

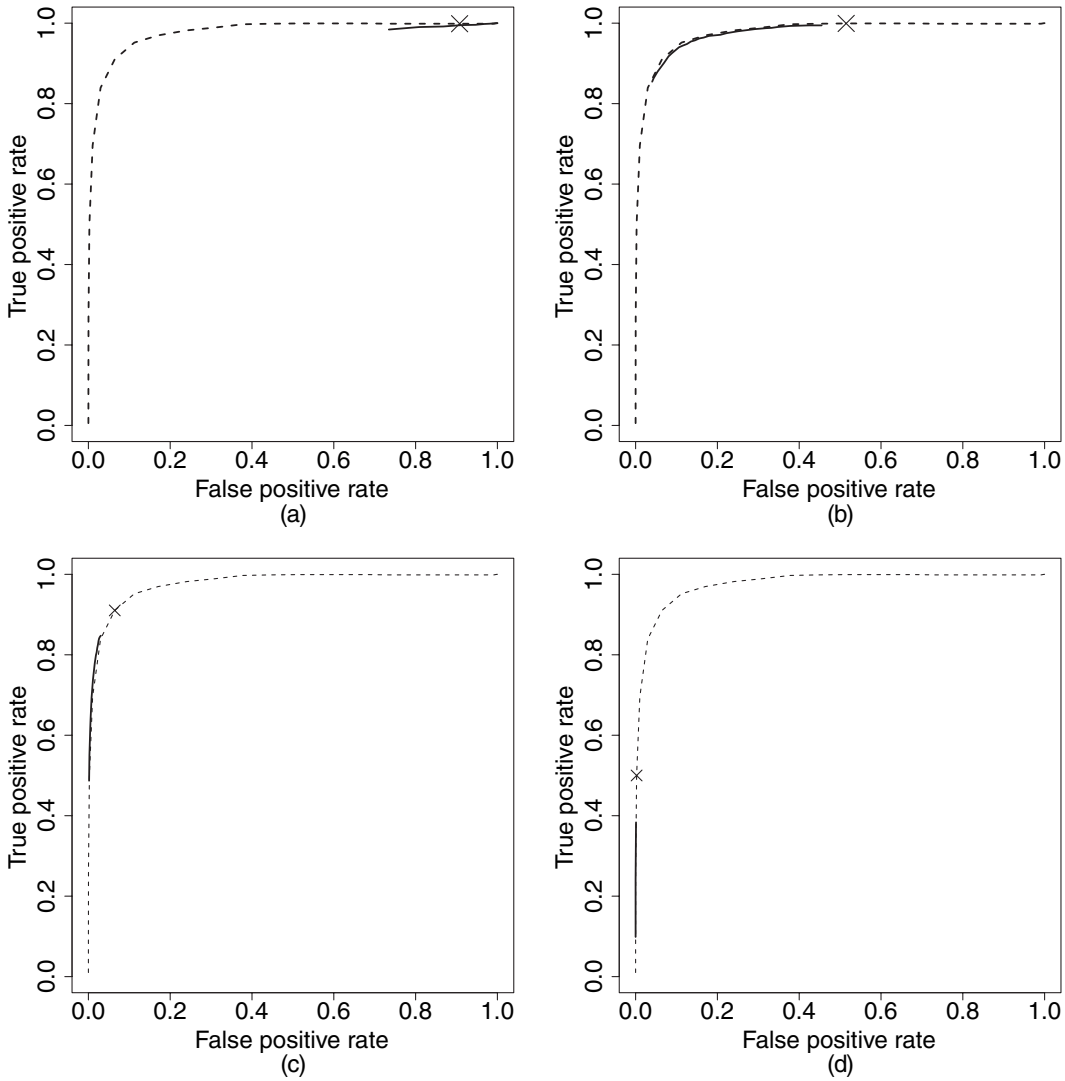


Fig. 18. Graphical lasso ROC curve (-------) and four different stability selection ROC curves (———) obtained by varying π_{thr} from 0.6 to 0.9 for fixed values of λ of (a) 0.01, (b) 0.06, (c) 0.23 and (d) 0.40: × marks the point on the graphical lasso ROC curve corresponding to the fixed λ

full range of λ is more expensive than doing bootstrap replications. However, since it does seem somewhat sensitive to the choice of a suitable small λ, it seems that combining it with some kind of initial crude cross-validation could result in even better performance. It would be interesting to consider whether there are particular types of the inverse covariance matrix that benefit from stability selection more than others, and whether any theoretical results can be obtained specifically for such structures; in particular, it would be interesting to know whether stability selection can perform better than the graphical lasso with oracle λ.

A. B. Tsybakov (*Centre de Recherche en Economie et Statistique, Université Paris 6 and Ecole Polytechnique, Paris*)

I congratulate the authors on a thought-provoking paper, which pioneers many interesting ideas. My question is about the comparison with other selection methods, such as the adaptive lasso or thresholded lasso (TL). In the theory these methods have better selection properties than those stated in theorem 2. For example, consider the TL $\tilde{\beta}_k = \hat{\beta}_k \, I\{|\hat{\beta}_k| > c\tau\sqrt{\|\hat{\beta}\|_0}\}$ where $\hat{\beta}$ is the lasso estimator with λ as in Bickel *et al.* (2009), $\tau = \sqrt{\{\log(p)/n\}}$ and $c > 0$ is such that $\|\hat{\beta} - \beta\|_2 \leqslant cs^{1/2}\tau$ with high probability under the restricted eigenvalue condition of Bickel *et al.* (2009). Then a two-line proof using expression (7.9) in Bickel *et al.* (2009) shows that, with the same probability, under the RE condition $\tilde{\beta}$ selects S correctly whenever $\min_{k \in S} |\beta_k| > Cs^{1/2}\tau$ for some $C > 0$ depending only on σ^2 and the eigenvalues of $X'X/n$. Since also c depends only on X and σ^2 (see Bickel *et al.* (2009)), c can be evaluated from the data. The restricted eigenvalue condition is substantially weaker than assumption 1 of theorem 2 and $\min_{k \in S} |\beta_k|$ need not be as large as greater than $C's^{3/2}\tau$, as required in theorem 2. We may interpret it as the fact that stability selection is successful if the relevant β_k are very large and the Gram matrix is very nice, whereas for smaller β_k and less diagonal Gram matrices it is safer to use the TL. Of course, here we compare only the 'upper bound', but it is not clear why stability selection does not achieve at least similar behaviour to that of the TL. Is it only technical or is there an intrinsic reason?

Cun-Hui Zhang (*Rutgers University, Piscataway*)

I congratulate the authors for their correct call for attention to the utility of randomized variable selection and great effort in studying its effectiveness.

In variable selection, a false variable may have a significant observed association with the response variable by representing a part of the realized noise through luck or by correlating with the true variables. A fundamental challenge in such structure estimation problems with high dimensional data is to deal with the competition of many such false variables for the attention of a statistical learning algorithm.

The solution proposed here is to simulate the selection probabilities of each variable with a randomized learning algorithm and to estimate the structure by choosing the variables with high simulated selection probabilities. The success of the proposed method in the numerical experiments is very impressive, especially in some cases at a level of difficulty that has rarely been touched on earlier. I applaud the authors for raising the bar for future numerical experiments in the field.

On the theoretical side, the paper considers two assumptions to guarantee the success of the method proposed:

(a) many false variables compete among themselves at random so each false variable has only a small chance of catching the attention of the randomized learning algorithm;
(b) the original randomized learning algorithm is not worse than random guessing.

The first assumption controls false discoveries whereas the second ensures a certain statistical power of detecting the true structure. Under these two assumptions, theorem 1 asserts in a broad context the validity of an upper bound for the total number of false discoveries. This result has the potential for an enormous influence, especially in biology, text mining and other areas that are overwhelmed with poorly understood large data.

Because of the potential for great influence of such a mathematical inequality in the practice of statistics, possibly by many non-statisticians, we must proceed with equally great caution. In this spirit, I comment on the two assumptions as follows.

Assumption (a) is the exchangeability condition in theorem 1. As mentioned in the paper, it is a consequence of the exchangeability of X_N given X_S in linear regression. The stronger condition implies a correlation structure for the design as

$$\begin{pmatrix} \Sigma_S & \mathrm{diag}(\rho_S)\mathbf{1} \\ \mathbf{1}\,\mathrm{diag}(\rho_S) & (1-\rho_N)I_N + \rho_N \mathbf{1}\mathbf{1}' \end{pmatrix},$$

where $\rho_S \in \mathbb{R}^S$, $\rho_N \in \mathbb{R}$, I_N is the identity and **1** denotes matrices of proper dimensions with 1 for all entries. I wonder whether such an assumption could be tested.

Assumption (b) may not always hold for the lasso. For $q_\Lambda < |S|$, a counterexample seems to exist with $X_N = \rho_S' X_S \mathbf{1} + \sqrt{(1 - \|\rho_S\|^2)} Z_N$, where X_S and Z_N are independent standard normal vectors and the components of ρ_S are of the same sign as those of β.

Hui Zou (*University of Minnesota, Minneapolis*)

I congratulate Dr Meinshausen and Professor Bühlmann on developing stability selection for addressing the difficult problem of variable selection with high dimensional data. Stability selection is intuitively appealing, general and supported by finite sample theory.

Regularization parameter selection in sparse learning is often guided by some model comparison criteria such as the Akaike information criterion and Bayes information criterion in which prediction accuracy measurement is a crucial component. It is quite intriguing to see that stability selection directly targets variable selection without using any prediction measurement. The advantage of stability selection is well demonstrated by theorem 1 in which inequality (9) controls the number of false selections. In the context of variable selection, inequality (9) is very useful when the number of missed true variables is small. In an ideal situation we wish to have $\hat{S}^{\text{stable}} \supseteq S$ while controlling the number of false selections. An interesting theoretical problem is whether a non-trivial lower bound could be established for $E(|S \cap \hat{S}^{\text{stable}}|)$.

To see how well stability selection identifies relevant variables, we did some simulations where stability selection was applied to sure independence screening (SIS) (Fan and Lv, 2008). In the linear regression case, SIS picks the top $d \ll p$ variables that have the highest correlation with the response variable. Denote by $\hat{S}^d$ the SIS selector with reduced dimension d. Then $\hat{\Pi}_k^d = P^*(k \in \hat{S}^d)$ and $\hat{S}_d^{\text{stable}} = \{k : \hat{\Pi}_k^d \geqslant \pi_{\text{thr}}\}$. We need to consider only pointwise control because $\hat{\Pi}_k^d$ is monotonically increasing with increasing d. The simulation model is $y = \sum_{j=1}^p x_j \beta_j + N(0, 1)$ with $(x_1, \ldots, x_p)$ independent and identically distributed $N(0, 1)$ and $\beta_1 = \beta_2 = \ldots = \beta_{10} = 1$ and $\beta_j = 0$ for $j > 10$. Following Meinshausen and Bühlmann, we let $d = \lfloor \sqrt{\{(2\pi_{\text{thr}} - 1)\}} p \rfloor$ to guarantee $E(V) \leqslant 1$. Moreover, in the $p \gg n$ setting $\hat{S}^d$ discovers all relevant variables with very high probability (Fan and Lv, 2008). For the simulation study we considered an asymptotic high dimension setting ($p = 20\,000$; $n = 800$) and a more realistic high dimension setting ($p = 4000$; $n = 200$).

Table 5 summarizes the simulation results. First of all, in all four cases the number of false selections by stability selection is much smaller than 1, the nominal upper bound. For the case of $p = 20\,000$ and $n = 800$, both SIS and stability selection select all true variables. In particular, stability selection using $\pi_{\text{thr}} = 0.9$ achieves the perfect variable selection in all 100 replications. When $p = 4000$ and $n = 200$, SIS still has a reasonably low missing rate (less than 5%), but stability selection using $\pi_{\text{thr}} = 0.6$ and $\pi_{\text{thr}} = 0.9$ selects about six and three relevant variables respectively. The performance is not very satisfactory. From this example we also see that with finite samples the choice of π_{thr} can have a significant effect on the missing rate of stability selection, although its effect on the false discovery rate is almost ignorable.

Table 5. Simulation for SIS plus stability selection based on 100 replications

π_{thr}	d	$E(\lvert S \cap \hat{S}^d \rvert)$, $\lvert S \rvert = 10$	$E(\lvert S \cap \hat{S}_d^{\text{stable}} \rvert)$, $p = 20000$	$E(\lvert N \cap \hat{S}_d^{\text{stable}} \rvert)$, $n = 800$
0.6	63	10	10	0.22
0.9	126	10	10	0
			$p = 4000$	$n = 200$
0.6	28	9.52	5.91	0.09
0.9	56	9.68	3.23	0.01

The **authors** replied later, in writing, as follows.

We are very grateful to all the discussants for their many insightful and inspiring comments. Although we cannot respond in a brief rejoinder to every issue that has been raised, we present some additional thoughts relating to the stimulating contributions.

Connections to Bayesian approaches
Richardson, Brown and Griffin, Draper, and Leng and Nott discuss interesting possible connections between stability selection (or other randomized selection procedures) and Bayesian approaches with appropriately chosen priors. The randomized lasso has the most immediate relation, as pointed out by Brown and Griffin and connecting with their interesting paper (Griffin and Brown, 2007). They also raise the question whether subsampling is then still necessary. Although we do not have a theoretical answer here, it seems that subsampling improves in practice a randomized procedure (or the equivalent Bayesian counterpart). We are also not 'throwing away real data' with subsampling since the final selection probabilities over subsampled data are U-statistics of order $\lfloor n/2 \rfloor$ and are using all n samples, not just a random subset. Stability selection is closely related to bagging (Breiman, 1996), as pointed out by Richardson. Stability selection is aggregating selection outcomes rather than predictions and assigning an error rate via our theorem 1. Nott and Leng (2010) seems to be very interesting in the context of Bayesian variable selection.

Bayesian decision theoretic framework
Draper and Holmes point out that the Bayesian framework is natural for a decision theoretic framework. And, indeed, this is one of the advantages of Bayesian statistics. In our examples of biomarker discovery and, more generally for variable selection (and also for example graphical modelling), the workflow consists of two steps. The first aim is to obtain

(a) a good ranked list of potentially interesting markers or variables which we
(b) then need to cut off at some position in the list.

Although a decision theoretic analysis is mostly helpful in step (b), stability selection is potentially improving both (a) and (b). The issue where to cut in the list in step (b) involves in the frequentist set-up a choice of an acceptable type I error rate. The choice of a type I error rate is maybe not as satisfying as a full decision theoretic treatment but it is often useful in practice. Each 'discovery' needs to be validated by further experiments which are often very costly and the chosen framework aims to optimize the number of true discoveries under a given budget that can be spent on falsely chosen variables or hypotheses.

Exchangeability assumption
Shawe-Taylor and Sun, and Zhang raise, very legitimately, the question to what extent the exchangeability assumption in theorem 1 is too stringent. We wrote in the paper that results do seem to hold up very well for real data sets where the assumption is likely to be violated (and theorem 2 is not making use of the strong exchangeability assumption). It is maybe also worthwhile mentioning that the assumptions can be weakened considerably for specific applications. For the special case of high dimensional linear models, we worked out a related solution in follow-up work (Meinshausen *et al.*, 2009).

Tightness of bounds and two-step procedures
Tsybakov correctly points out that sharper bounds on the l_2-distance are available for the standard lasso and these could be exploited for variable selection by using hard thresholding of coefficients or the adaptive lasso. The reasons for having looser results for the randomized lasso are technical in our view, not intrinsic. It is much more difficult to analyse the stability selection algorithm which involves subsampling and randomization of covariates, which opens up maybe interesting areas for further mathematical investigations. We thought that it was interesting, nevertheless, that the irrepresentable condition can be considerably weakened by using randomization of the covariates instead of using two-step procedures such as hard thresholding or the adaptive lasso.

Power and false positive selections
Zou, Richardson, Shah and Samworth, and Rothman, Levina and Zhu examined the power of the method to detect important variables and compared it with alternative approaches for some examples. Although it is obviously true that no method will be universally 'optimal', stability selection places a strong emphasis on avoiding false positive selections. This is in contrast with say, sure independence screening used by Zou, which is a screening method (by name!) and is sitting at the opposite end of the spectrum by placing a large emphasis on a large power while accepting many false positive selections. For the simulation results of Zou, we suspect that sure independence screening would have a much larger false positive rate for $p = 4000$ but we could not see it being reported. Rothman, Levina and Zhu compare the receiver operating characteristic curve for the example of graphical modelling. It does not come entirely unexpected from our point of view that the gain of stability selection is very small or, indeed, non-existent since the simulation takes place in

a Toeplitz design case which is very close to complete independence between all variables. For regression, it was shown already in the paper that stability selection cannot be expected to improve performance for independent or very weakly correlated variables. And our theorem 2 showed that we can expect major improvements only if the irrepresentable condition is violated, which has analogies in Gaussian graphical modelling (Meinshausen, 2008; Ravikumar *et al.*, 2008).

Generalization performance and sparsity

Richardson, Shawe-Taylor and Sun, Tsybakov and Hothorn discuss the connection between generalization performance and sparsity of the selected set of variables. Hothorn mentions that achieving both optimal predictive accuracy and consistent variable selection might be very difficult, as manifested also in the Akaike information criterion–Bayes information criterion dilemma for lower dimensional problems. Shawe-Taylor and Sun illustrate that stability selection will in general produce rather sparse models, which is in agreement with the discussion on false positive selections above. Their example demonstrates though also impressively that the predictive performance is sometimes compromised only very marginally when using much sparser models than those produced by the lasso under cross-validation. In general, stability selection will yield much sparser models than the lasso with cross-validation. How much predictive performance one is willing to sacrifice for higher sparsity of the model, if any, should be application driven. If the answer is 'none', stability selection might not be appropriate.

Approximating the true model

Hennig and Nugent, Rinaldo, Singh and Wasserman rightly question the assumed existence of a true linear model and whether coefficients are ever exactly vanishing. Firstly, sometimes it *is* true that β_ks are exactly zero for some $k \in \{1, \ldots, p\}$, namely if observed variables contain truly just noise (in astronomy and physics this is often so—in biology not so much). Secondly, any study of variable importance or variable relevance will necessarily be model based in some form or another, be it in a low or high dimensional linear model or a random-forest framework, to name two examples. In general, no low or high dimensional parametric or non-parametric model is ever correct in practice. And yet it is legitimate in our view to be interested in assessing variable importance or relevance. In this context, it is maybe worthwhile to look instead at some sparse approximation of the data-generating distribution (which always exists) and to treat the question of variable importance and variable selection in this light. For the lasso, this has been worked out in Bunea *et al.* (2007) and Bickel *et al.* (2009) for estimation of the approximating regression parameters whereas for example van de Geer *et al.* (2010) deal explicitly with the problem of variable selection when the linear model is a sparse linear approximation for a true possibly non-linear regression function. Our theorem 1 can be extended to such settings where the 'true structure' S is defined via a sparse approximation: we need to replace the true set S by an approximation set S_{approx}. For example, in a linear approximating model for a general regression function $f(\cdot) = E(Y|X = \cdot)$, we can define

$$S_{\text{approx}} = \operatorname*{arg\,min}_{M \subseteq \{1,\ldots,p\}} (\|f_M - f\|_2^2/n + C^2 + |M|),$$
$$f_M = \operatorname*{arg\,min}_{f^* = X\beta_M^*} (\|f^* - f\|_2^2/n) \tag{37}$$

and β_M^* has non-zero components only in the set $M \subseteq \{1, \ldots, p\}$. Here, C^2 is a suitable positive number, typically depending on n, and we denote by f, f^* and f_M the $n \times 1$ vectors evaluated at the observed covariates. Clearly, if the true model is linear and sparse with many regression coefficients equal to 0 and where the few non-zero regression coefficients are all sufficiently large, then the set S_{approx} in expression (37) equals the set S of the true active variables. Theorem 1 will remain valid under an appropriate exchangeability assumption for selection of variables in the complement of S_{approx} which might or might not be realistic. The mathematical arguments for extending theorem 2 to such a setting seem to be more involved.

Correlated predictor variables

Kirk, Lewin and Stumpf, and Kent raise the issue of correlated predictor variables and examine the behaviour of stability selection for highly correlated designs. This is a very important discussion point. As mentioned already above, stability selection puts a large emphasis on avoiding false positive selections and, as a consequence, might miss important variables if they are highly correlated with irrelevant variables. This is similar to the behaviour of a classical test for the regression coefficient $p \ll n$ situations. For situations where we are more interested in whether there are interesting variables in a certain group of variables, the proposal of Kirk, Lewin and Stumpf on testing stability of sets of variables (and finding those sets possibly by the elastic net) seems very interesting and useful.

Numerical example of vitamin gene expression data
Ahmed and Richardson analyse our gene expression data set with several competing methods and come to the conclusion that at most three genes should be selected. They raise the question whether stability selection is selecting too many variables. However, as shown in the initial contribution to the discussion by Richardson, stability selection is in fact also selecting only three genes under reasonable type I error control. The methods seem to be in agreement here.

Choice of regularization
Yen and Yen mention that the number q of selected variables can grow very large for small regularization parameters λ and propose an interesting way to choose a suitable region for the regularization parameter. Yet, instead of restricting λ to larger values, a useful alternative in practice is to select only the first q variables that appear when lowering the regularization. And q can be chosen *a priori* to yield non-trivial bounds in theorem 1.

Computational improvements
Gandy and Fang both propose interesting extensions that help to alleviate the computational challenge of having to fit a model on many subsamples of the data. An interesting alternative to the procedure that was proposed by Gandy is the improved bounds suggested by Shah and Samworth.

Dependent data
Tong notes that stability selection makes an inherent assumption of independence between observations. We have not yet tried to apply the method to dependent data such as time series. The standard subsampling scheme will not be suitable in cases of dependence. A block-based approach with independent subsampling of blocks (and where dependence is captured within blocks, at least approximately) along the lines of Künsch (1989) might be an interesting alternative to explore in this context.

Connections to clustering and density cluster estimation
Nugent, Rinaldo, Singh and Wasserman, and Hennig provide fascinating connections to related ideas in clustering and density cluster estimation. As described in the paper, Monti *et al.* (2003) is another interesting connection to consensus clustering.

Related ideas
Richardson and Hothorn mention numerous related previous references. We tried to point out many connections to previous work but have missed important ones. It is maybe worthwhile emphasizing again the similarity, at the crude level, of the work of Bach (2008) on bolasso which has been developed independently and simultaneously.

We reiterate and thank all the contributors again for their many interesting and thoughtful comments which have already opened up and will open up new research in this area. We would like to convey special thanks to Rajen Shah and Richard Samworth, who spotted a mistake in the definition of the assumption 'not worse than random guessing' in an earlier version of the manuscript. Their improved bounds will also make stability selection less conservative and address John Shawe-Taylor's comment regarding the finite amount of random subsampling in practice *versus* our theoretical arguments corresponding to all possible subsamples. Finally, we thank the Royal Statistical Society and the journal for hosting this discussion.

References in the discussion

Akaike, H. (1980) Seasonal adjustment by a Bayesian modeling. *J. Time Ser. Anal.*, **1**, 1–13.
Bach, F. (2008) Bolasso: model consistent Lasso estimation through the bootstrap. In *Proc. 25th Int. Conf. Machine Learning*, pp. 33–40. New York: Association for Computing Machinery.
Barbieri, M. and Berger, J. (2004) Optimal predictive model section. *Ann. Statist.*, **32**, 870–897.
Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*, ch. 2. New York: Wiley.
Ben-David, S., von Luxburg, U. and Pall, D. (2006) A sober look at clustering stability. *Learn. Theor.*, no. 4005, 5–19.
Ben-Hur, A., Elisseef, A. and Guyon, I. (2002) A stability based method for discovering structure in clustered data. *Pacific Symp. Biocomputing*.
Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer.
Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bottolo, L. and Richardson, S. (2010) Evolutionary Stochastic Search for Bayesian model exploration. *Preprint*. (Available from `http://arxiv.org/abs/1002.2706`.)
Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.
Brown, P. J., Fearn, T. and Vannucci, M. (1999) The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, **60**, 627–641.
Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, **1**, 169–194.
Dalmasso, C., Broët, P. and Moreau, T. (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.
Dupuis, J. A. and Robert, C. P. (2003) Variable selection in qualitative models via an entropic explanatory power. *J. Statist. Planng Inf.*, **111**, 77–94.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.
Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh diemensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.
Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.
Gandy, A. (2009) Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *J. Am. Statist. Ass.*, **104**, 1504–1511.
van de Geer, S., Zhou, S. and Bühlmann, P. (2010) Prediction and variable selection with the adaptive Lasso. *Preprint arXiv:1001.5176v1*.
George, E. I. and McCullogh, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.
Griffin, J. E. and Brown, P. J. (2007) Bayesian adaptive lassos with non-convex penalisation. Institute of Mathematics and Statistics, University of Kent, Canterbury. (Available from `http://www.kent.ac.uk/ims/personal/jeg28/`.)
Griffin, J. E. and Brown, P. J. (2010) Inference with normal-gamma prior distributions in regression problems. *Bayesn Anal.*, **5**, 171–181.
Hans, C., Dobra, A. and West, M. (2007) Shotgun Stochastic Search for large p regression. *J. Am. Statist. Ass.*, **102**, 507–517.
Hastie, T., Tibshirani, R. and Friedman, J. (2008) *The Elements of Statistical Learning*, 2nd edn. New York: Springer.
Haufe, S., Müller, K.-R., Nolte, G. and Krämer, N. (2010) Sparse casual discovery in multivariate time series. In *Journal of Machine Learning Research Wrkshp Conf. Proc.*, vol. 6, *Causality: Objectives and Assessment*, pp. 97–106. (Availabe from `http://www.JMLR.org`.)
Hennig, C. (2010) Methods for merging Gaussian mixture components. In *Adv. Data Anal. Classificn*, **4**, 3–34.
Johnstone, I. M. and Silverman, B. W. (2005) Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, **33**, 1700–1752.
Künsch, H.-R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.
Lange, T., Roth, V., Braun, M. and Buhmann, J. (2004) Stability-based validation of clustering solutions. *Neur. Computn*, **16**, 1299–1323.
Lee, H. K. H. (2004) *Bayesian Nonparametrics via Neural Networks*. Philadelphia: Society for Industrial and Applied Mathematics.
Lindley, D. V. (1968) The choice of variables in multiple regression (with discussion). *J. R. Statist. Soc.* B, **30**, 31–66.
Meinshausen, N. (2008) A note on the Lasso for graphical Gaussian model selection. *Statist. Probab. Lett.*, **78**, 880–884.
Meinshausen, N., Meier, L. and Bühlmann, P. (2009) P-values for high-dimensional regression. *J. Am. Statist. Ass.*, **104**, 1671–1681.
Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.
Nott, D. J. and Leng, C. (2010) Bayesian projection approaches to variable selection in generalized linear models. *Computnl Statist. Data Anal.*, to be published.
Park, T. and Casella, G. (2008) The Bayesian Lasso. *J. Am. Statist. Ass.*, **103**, 681–686.
Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2008) High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Preprint arXiv:0811.3628*.
Rinaldo, A. and Wasserman, L. (2010) Generalized density clustering. *Ann. Statist.*, to be published. (Available from `http://arxiv.org/abs/0907.3454`.)
Sauerbrei, W. and Schumacher, M. (1992) A bootstrap resampling procedure for model-building—application to the Cox regression-model. *Statist. Med.*, **11**, 2093–2109.
Savage, L. J. (1954) *The Foundations of Statistics*. New York: Wiley.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional variable importance for random forests. *BMC Bioinform.*, **9**.

Tibshirani, R. and Walther, G. (2005) Cluster validation by prediction strength. *J. Computnl Graph. Statist.*, **14**, 511–528.
Tong, H. (2010) Obituary of Hirotugu Akaike. *J. R. Statist. Soc.* A, **173**, 451–454.
Wainwright, M. (2009) Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Trans. Inform. Theor.*, **55**, 2183–2202.
Wang, H., Li, G. and Tsai, C. L. (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **69**, 63–78.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H., Hastie, T. and Tibshirani, R. (2007) On the degrees of freedom of the LASSO. *Ann. Statist.*, **35**, 2173–2192.
Zucknick, M., Richardson, S. and Stronach, E. A. (2008) Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statist. Applic. Genet. Molec. Biol.*, **7**, no. 1, article 7.