Applying Least Absolute Shrinkage and Selection Operator for Error Control

Farhad Bahadori-Jahromi
Department of Electrical and computer Engineering, Fasa Branch, Islamic Azad University, Fasa, Iran.
bahadori.fr@gmail.com

Abstract:
In the recent past, the development of statistical methods for high-dimensional problems has greatly advanced leading to methods for model selection such as the lasso. However, the question of error control in high-dimensional settings has proven to be difficult. Recently, an approach called stability selection has been proposed to tackle the problem. It combines a method for model selection and sub sampling to deliver a form of error control. In this paper, some variants of stability selection are introduced. It was tested if error control would actually hold up. Furthermore, some conditions were isolated where using these variants might have beneficial effects.

Keywords: Least Absolute Shrinkage and Selection Operator (LASSO), Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), independent and identically distributed (i.i.d.**)**

1. Introduction

This work is about the statistical problem of model selection in linear models.

Given some data $(Y_i, X_i^{(1)}, ..., X_i^P (i = 1,..., n)$, one assumes that there is a linear relationship in the following fashion

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i (i = 1,...n)$$

(1)

where $\varepsilon_i, (i = 1,..., n)$ are random and i.i.d with mean zero. This can be written more compactly as

$$Y = \mu 1 + X\beta + \varepsilon$$

(2)

where Y, ε and 1 are n-dimensional vectors, β is a p-dimensional vector and X is an $n \times p$-matrix with the first column set to one. Additionally, we assume that many β_j are zero (where β_j refers to the jth entry in the vector β). Our goal is to isolate those variables for which $\beta_j \neq 0$. One obvious way to do this is to look at each predictor variable separately. Let the jth column of X be denoted by $X^{(j)}$. Then, for each $X^{(j)}$, we would need to test the model for significance taking into account that this is a multiple testing problem.

$$Y = \mu 1 + X^{(j)}\beta_j + \varepsilon$$

(3)

However, this approach will not take the correlations between the predictors $X^{(j)}, (j = 1,..., p)$ into consideration. An alternative way is to find an estimate of the whole vector of coefficients β that leads to a good fit to the data while at the same time ensuring that our estimated vector $\hat{\beta}$ is sparse in the sense that few entries in β are non zero. This should yield a model with high predictive power that is also interpretable due to the low numbers of coefficients in the model. Since we also assumed that most coefficients β_j are truly zero, we can even hope to uncover the true model. One generic way to accomplish those goals is to find minimize of a function consisting of the negative log-likelihood and an additional term penalizing the size of the model. This approach is also called regularization. There are many methods that fit into this general framework and in Section 2, we will present one in detail, namely the lasso. It was proposed by Tibshirani (1996) and has become very popular due to its good theoretical properties as well as computational feasibility (see for example Buhlmann and van de Geer (2011)). However, there are some difficulties associated with this approach. The first concerns the trade-off between goodness of fit and sparsity of the solution. It is difficult to decide how much one should penalize model size. Secondly, it is not obvious how to establish some form of error control. Meinshausen and Buhlmann (2010) addressed these problems with an ensemble approach called stability selection. This method consists of repeatedly drawing subsamples from the data, then applying a selection method to each subsample and looking for consensus in the collection of proposed solutions. Stability selection is not a new variable selection technique (see for example Sauerbrei and Schumacher (1992)). However, Meinshausen and Buhlmann used it for establishing error control and to resolve the issue of deciding how much regularization is necessary. Section 3 describes this approach in more detail and presents some variants for which we can establish a very similar form of error control. Section 4 contains simulation results comparing the various approaches, and Section 5 a final discussion.

2. The Lasso

In this section, we first present the idea behind regularization and its connection to model selection. We then give some reasons why we can hope to accomplish

both with the lasso. We go on to present some asymptotic results for this estimator and talk about how to decide how much regularization is necessary. This exposition follows Buhlmann and van de Geer (2011), where also many of the following results are treated in much greater detail.

2.1. Regularization

Consider again the linear model

$$Y = \mu 1 + X\beta + \varepsilon \tag{4}$$

where Y, β, ε and 1 are n-dimensional vectors, X is an $n \times p$-matrix and the entries of $\varepsilon, \varepsilon_i, (i = 1,...,n)$, are random, i.i.d and have expectation zero. For simplicity, we assume that the intercept μ is zero and that all covariates $(X^{(1)},...,X^{(p)})$ are centered and on the same scale. This is approximately achieved by subtracting the empirical means $\overline{Y}, \overline{X}^{(j)}$, $j = 1,..., p$, and additionally scaling the covariates with the estimate of the standard deviation $\hat{\sigma}_j = \sqrt{\sum_{i=1}^{n} (X_i^{(j)} - \overline{X}^{(j)})^2 / n}$. The goal is now to estimate the coefficients $\beta_j, (j = 1,..., p)$. The classical method to solve this problem is to minimize the residual sum of squares

$$\hat{\beta} = \arg_\beta \min\left(\|Y - X\beta\|_2^2\right) \tag{5}$$

This is called the ordinary least squares (OLS) estimator. If $X^T X$ is not singular, this minimum is unique and can be calculated as follows

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

One case where this cannot hold is when p > n. But also, when $X^T X$ is almost singular, say because the predictors are highly correlated, the estimate becomes unreliable, because for the OLS-estimate we have

$$Cov\left(\hat{\beta}\right) = \sigma^2 \left(X^T X\right)^{-1} \tag{7}$$

which increases as ($X^T X$) gets closer to singularity. We can decompose the mean squared error into variance and the square of the bias of the estimator. Although the OLS estimate is unbiased, its variance is large if $X^T X$ is close to singularity as we have seen. It follows that our mean squared error will be large as well. One way to potentially improve the quality of the OLS-estimator is to modify it so that the new estimator has a bias but has dramatically reduced variance. This is what regularization does. It typically takes the form

$$\hat{\beta} = \arg_\beta \min\left(\|Y - X\beta\|_2^2 / n + \lambda pen(\beta)\right) \tag{8}$$

where $pen(\beta)$ measures model size and $\lambda > 0$ is a tuning parameter. The most natural choice for complexity penalization is l_0-penalization. Here,

$$\hat{\beta} = \arg_\beta \min\left(\|Y - X\beta\|_2^2 / n + \lambda \|\beta\|_0\right) \tag{9}$$

where

$$\|\beta\|_0 = \sum_{i=1}^{p} 1\{\beta_j \neq 0\} \tag{10}$$

l_0-penalization favors models with few coefficients different from zero. This is a desirable property because of better interpretability. Also, it is often reasonable to assume that most predictors have zero coefficients. For this approach, there are choices for the size of λ which have a theoretical justification, for example the Akaike Information Criterion (AIC), or Bayes Information Criterion (BIC).

However, for large p, the optimization problem becomes infeasible since the penalty function is neither convex nor continuous. One method to overcome this is forward selection. Here, the model size is increased one variable at the time. One selects always the variable which lowers the residual sum of squares of the resulting model the most. Note that the minimum of the residual sum of squares does not increase when an additional variable is added, since the parameter space over which we optimize becomes larger. However, this procedure turns out to be instable, because it corresponds to a very greedy search in the model space. Other related stepwise procedures were proposed, but they all tend to show this instable behavior.

2.2. Lasso and Variable Selection

A popular alternative is the lasso estimator which stands for Least Absolute Shrinkage and Selection Operator. This estimator is defined as

$$\hat{\beta} = \arg_\beta \min\left(\|Y - X\beta\|_2^2 / n + \lambda \|\beta\|_1\right) \tag{11}$$

Because the objective function is convex, computation of this estimator is feasible. Also, the estimator has the property that it sets many coefficients to zero, i.e. $\hat{\beta}_j(\lambda) = 0$ for some j. This makes this estimator a candidate for a variable selection procedure. Additionally, one can establish via lagrangian duality that problem 2.0.2.1 is equivalent to

$$\hat{\beta}_{prim}(R) = \underset{\beta:\|\beta\|_1 \leq R}{\arg\min} \left(\|Y - X\beta\|_2^2 / n\right) \tag{12}$$

and that there is a one-to-one correspondence between R and λ depending on the data. Hence, one can think of the Lasso estimate as an OLS solution constrained to lie in the set $\{\beta : \|\beta\|_1 \leq R\}$One can derive an intuitive understanding why the lasso will set some coefficients to zero and why this doesn't occur in ridge regression,

$$\hat{\beta}_{prim}(R) = \underset{\beta:\|\beta\|_2 \leq R}{\arg\min} \left(\|Y - X\beta\|_2^2 / n\right) \tag{13}$$

The situation is schematically depicted in Figure 1. The contour lines of β-values leading to equal values of residual sum of squares in the picture reach the l_1-ball in a corner where-l_2. This is not the case for the $\beta_1 = 0$ ball.

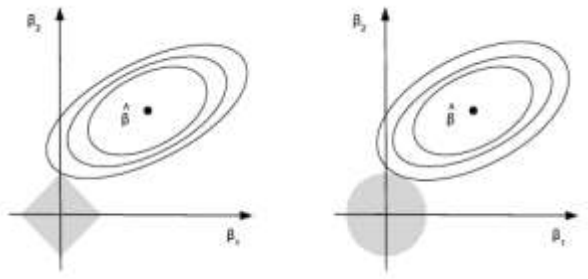

Fig. 1. Contour lines of residual sum of squares.

Restriction to the l_1-ball for the lasso solution and the l_2-ball for the ridge solution is shown on the left and right respectively.

Therefore, it is possible to use the lasso as a variable screening method. Just like forward selection regression, lasso favors sparse models where many coefficients are set to zero. One can show that for each λ, lasso selects at most min (n, p) covariates (Efron, Hastie, Johnstone, and Tibshirani (2004)). Additionally, the number of different sub models that are selected for various values of λ is typically O (min (n, p)) (Rosset and Zhu (2007)). So, the Lasso can be used to derive sparse models. Obviously, one wants to know if the model proposed has good properties such as good prediction and if it can detect the correct model assuming that the true underlying model is sparse. For a good estimator of our regression function, we want the empirical squared prediction error

$$\left\| X_n\left(\hat{\beta}_n(\lambda_n) - \beta_n^0\right)\right\|_2^2 / n \tag{14}$$

to be small. Here, β^0 denotes the true coefficient vector. This is also the quantity of interest when we want to predict a new response. Another question of interest is if the lasso is able to estimate β^0 well. Also, since we are interested in variable screening we want to know if the lasso can uncover the set $S_0 = \left\{j : \beta_j^0 \neq 0\right\}$

2.3. Consistency of the Lasso

To derive the favorable properties of the OLS estimator, one resorts to asymptotic (see for example Sen and Srivastava (1990)). Assuming that the number of observations tends to infinity and p is fixed, one can, under some conditions on the design $X^T X$, show that

$$\left\| X\left(\hat{\beta} - \beta\right)\right\|_2^2 \xrightarrow{d} \sigma^2 \chi_p^2 \tag{15}$$

and

$$E\left[\left\| X\left(\hat{\beta} - \beta\right)\right\|_2^2\right] \xrightarrow{d} \sigma^2 p \tag{16}$$

Via Chebyshev's inequality, we get that

$$\left\| X\left(\hat{\beta} - \beta\right)\right\|_2^2 = Op(1) \tag{17}$$

This result is derived under the assumption that n goes to infinity while p is fixed. However, the situations where we want to use the lasso are not modeled well by this assumption since we want to allow the case $p >> n$. One way to capture this situation is by triangular arrays of observations

$$Y_{n,i} = \sum_{j=1}^{p_n} X_{n.i}^{(j)} \beta_{n.j}^0 + \varepsilon_{n,i},\ i = 1,...,n; \qquad n = 1,2,.... \tag{18}$$

Here, $p_n > n$ is allowed. Note that we also allow the true coefficient vector β_n^0 to depend on n. One can show consistency of the lasso estimator if $\left\|\beta_n^0\right\|_1$ doesn't grow too fast, i.e. for a suitable choice of λ, it holds that

$$\left\| X_n\left(\hat{\beta}_n(\lambda_n) - \beta_n^0\right)\right\|_2^2 / n = Op\left(\left\|\beta_n^0\right\|_1 \sqrt{\log(p_n)/n}\right) \tag{19}$$

where the asymptotic is with respect to 2.0.3.4.

Therefore, if $\left(\left\|\beta_n^0\right\|_1 \sqrt{\log(p_n)/n}\right)$ converges to zero, consistency holds. Note that if we assume the classical setting with p fixed as $n \to \infty$, this rate is slower than what equation 2.0.3.3 suggests for the OLS solution. With an additional assumption on the design X called compatibility (or restricted Eigenvalue) condition, one can prove the following faster error rate

$$\left\| X\left(\hat{\beta} - \beta\right)\right\|_2^2 / n = Op\left(\frac{|S_0| \log(p)}{n\phi^2}\right) \tag{20}$$

where ϕ^2 is the so-called compatibility constant, which depends on the design X as well as the active set S_0. If this constant is bounded from below for all n, we see that the squared prediction error of the lasso solution has the same rate of convergence as the OLS-solution up to a factor log (p). Therefore, the rate of convergence is almost as fast as if one knew beforehand which coefficients truly is nonzero and applied OLS. This is called an oracle property. Under exactly the same conditions, it can be shown that

$$\left\|\hat{\beta}(\lambda)-\beta^0\right\|_1 = Op\left(\frac{|S_0|\sqrt{\log(p)}}{\sqrt{n}\phi^2}\right) \quad (21)$$

Therefore, if ϕ is bounded from below
$\left(|S_0|\sqrt{\log(p)}/(\sqrt{n}\phi^2)\right)\to 0$

we have
$$\left\|\hat{\beta}(\lambda)-\beta^0\right\|_1 \xrightarrow{p} 0 \quad (22)$$

(Also, a similar result can be proven for the l_2-norm under more stringent compatibility conditions.) If we predefine, for some cutoff C > 0, the set $S_0^{relevant}(C)=\left\{j:\left|\beta_j^0\right|\geq C\right\}$ result 2.0.3.8 ensures that, asymptotically, we will find all covariates that are in $S_0^{relevant}(C)$, i.e.

$$P\left[\hat{S}(\lambda)\supset S_0^{relevant}[C]\right]\longrightarrow 1,(n\to\infty) \quad (23)$$

since

$$p\left[\hat{S}(\lambda)\supset S_0^{relevant}(C)\right]\leq P\left[\left\|\hat{\beta}(\lambda)-\beta^0\right\|_1\geq C\right]\to 0 \quad (24)$$

This property is called the screening property since it ensures that relevant variables are retained. Because of this and the fact that the lasso will at most select min (n; p) variables if $p >> n$, we can use the lasso to get rid of most noise covariates while retaining the relevant variables with high probability.

2.4. Choice of the Tuning Parameter

Typically, the tuning parameter is chosen via cross validation. However, cross validation estimates the prediction error and tries to find an optimal amount of regularization for prediction. But the optimal amount of regularization for prediction and the optimal amount for uncovering the set of active variables S0 do not need to coincide. Choosing the tuning parameter via cross validation often leads to inclusion of too many covariates. As an illustration, Figure 2 shows a so-called regularization path and the choice made via 10-fold cross validation.

$$t=\frac{R}{\left\|\hat{\beta}^{OLS}\right\|_1} \quad (25)$$

where $\left\|\hat{\beta}^{OLS}\right\|_1$ refers to the smallest possible l_1-norm of an OLS solution. This parameterization has the advantage that roughly represents the fraction of coefficients that are in the model compared to the full model. Note how the 10-fold cross validated solution finds almost all truly active variables but also a large number of inactive variables.

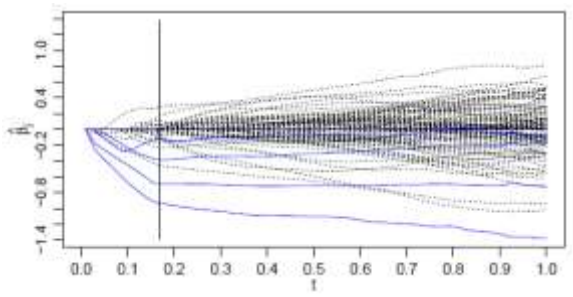

Fig. 2. Plot of estimates $\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_p$ versus $t=\left\|\hat{\beta}\right\|_1/\left\|\hat{\beta}^{OLS}\right\|_1$ for the lasso.

Solid blue lines correspond to active covariates, dashed black lines to inactive covariates. Data was generated according to design b) in section 4.0.3 with n = 100 and p = 200. The number of active variables was 5 and the signal-to-noise ratio 2.

This is not so surprising since the screening property ensures that, asymptotically, our relevant variables will be included in the solution, but does not say if irrelevant ones will be excluded for sure. This is referred to as consistent variable selection which we will discuss next.

2.5. Consistent Variable Selection

We have seen that under so-called compatibility conditions $\left\|\beta^0-\hat{\beta}\right\|\xrightarrow{p}0$, but this does not imply that $P\left[\hat{S}(\lambda)=S_0\right]\to 1 \quad (n\to\infty)$ (26)

It can be shown that, for this to hold, the so called irrepresentable is sufficient if all active coefficients are bounded away from zero. Designate X_{S0} as the restriction of the design matrix X to the columns corresponding to S0. Then we say that the irrepresentable condition holds if for every $j\in S_0^C$

$$\left|\tau_{S0}\left(X_{S0}^T X_{S0}\right)^{-1}X_{S0}^T X^{(j)}\right|\leq\theta \quad (27)$$

for some $0<\theta<1$, where τ_{S0} is the restriction of the vector $(sign(\beta_1^0),...,sign(\beta_p^0))^T$ to the active coefficients. Also, for consistent variable selection to be at all possible, it is necessary that for every $j\in S_0^C$

$$\left|\tau_{S0}\left(X_{S0}^T X_{S0}\right)^{-1}X_{S0}^T X^{(j)}\right|\leq 1 \quad (28)$$

One can show that the irrepresentable condition is stronger than the compatibility condition.

3. Stability Selection

As we have seen, even though the lasso is powerful tools for model selection in cases were $p >> n$, there are some problems. One issue is how to select the correct amount of regularization to achieve a certain goal such as variable selection. We have seen that cross validation often leads to the inclusion of too many variables. However, it is often desirable to control the number of false positives, even at the expense of missing some active variables, for example, if costly and time-consuming follow-up experiments is performed to validate the findings. It is not clear how to achieve this simply by choosing the correct regularization parameter. One approach is stability selection proposed by Meinshausen and Buhlmann (2010). As mentioned before, it is an ensemble approach. It relies on repeatedly drawing subsamples from the data, applying a variable selection method, such as the lasso, and look for consensus in the ensemble of solutions. Let the situation be as stated in the introduction. We are trying to estimate the set of coefficients that are active in the linear model

$$Y = X\beta^0 + \varepsilon \tag{29}$$

i.e. find set S where $S = \left\{ j : \beta_j^0 \right\} \neq 0$. We additionally assume that the data

$$Z = (x_1, y_1), ..., (x_n, y_n) \tag{30}$$

are independent and identically distributed. In particular, we do not think of our design as fixed. Assume that I is some uniform random subsample of size $\lfloor n/2 \rfloor$ of the index set $\{1, ..., n\}$ and use this index set to subsample from the data yielding Z (I). For this subset of the data and a given regularization parameter $\lambda \in \Lambda$, we can apply the lasso procedure yielding

$$\left\{ \hat{\beta}_k^\lambda (I) : \lambda \in \Lambda, k = 1, ..., p \right\}$$

and

$$\hat{S}^\lambda (I) = \left\{ k : \hat{\beta}_k^\lambda (I) \neq 0, k = 1, ..., p \right\}.$$

We can now define the conditional selection probability of covariate k as

$$\hat{\Pi}_\kappa^\lambda = P\left[k \in \hat{S}^\lambda (I) \middle| Z \right] \tag{31}$$

This quantity can be estimated with arbitrary accuracy by repeated sub sampling and application of the above procedure. Analogously to the regularization paths explained above, we are in a position to plot stability paths where the quantities $\hat{\Pi}_\kappa^\lambda$ are plotted over the whole set of potential regularization parameters. Figure 3 displays such a stability path for the same data as Figure 2. As we see, almost all variables that display high stability are truly active. Furthermore, the stability for very stable variables changes comparatively little over a large portion of the stability path. This stands in stark contrast to the lasso trace, where the solution varies a lot over the whole regularization path.

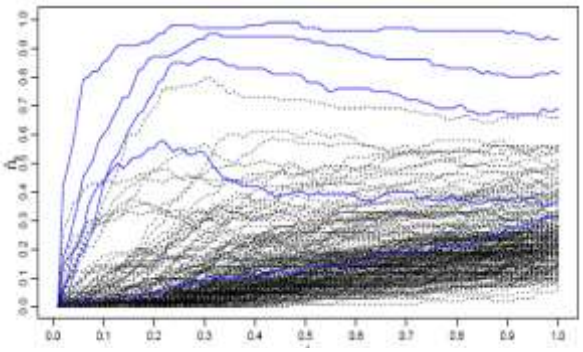

Fig. 3. Example of a lasso stability path: Plot of estimates $\hat{\Pi}_\kappa$ versus $t = \left\| \hat{\beta} \right\|_1 / \left\| \hat{\beta}^{OLS} \right\|_1$ on the same data as in Figure 2.

Solid blue lines correspond to active covariates, dashed black lines to inactive covariates. 100 subsamples were drawn.

For our estimate of S, we choose those covariates whose selection probability is above a certain predefined threshold τ_{thr} for some $\lambda \in \Lambda$, i.e.

$$\hat{S}^{stable} = \left\{ k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda, \geq \tau_{thr} \right\} \tag{32}$$

The new tuning parameters of the method are the regularization region λ and the threshold τ_{thr}, where $0 < \tau_{thr} < 1$.

3.1. Error Control

One advantage of the described method is that it allows for a certain form of error control under some additional conditions. For a given regularization region, one can derive a bound on the expected number of false positives depending on the cut-off τ_{thr}. In order to define the additional conditions, we introduce some notation. We set $\hat{S}^\Lambda = \cup_{\lambda \in \Lambda} \hat{S}^\lambda$ and let $q_\Lambda = E\left[\left| \hat{S}^\Lambda (I) \right| \right]$ to be the mean number of selected variables. We define N as the set of noise covariates, i.e. $N = \left\{ j : \beta_j^0 = 0 \right\}$. Also, we define V as the number of falsely selected variables with stability selection,

$$V = \left| N \cap \hat{S}^{stable} \right| \tag{33}$$

We are then in a position to give a bound-on E [V], also called the per-family error rate.

Theorem: Assume that the expectation of the random indicator variables $1_{\{k\in\hat{S}^{\Lambda}\}}$ is equal for all $k \in N$. Also assume that the procedure with respect to the whole regularization region is not worse than random guessing, i.e.

$$\frac{E\left[\left|S \cap \hat{S}^{\Lambda}\right|\right]}{E\left[\left|N \cap \hat{S}^{\Lambda}\right|\right]} \geq \frac{|S|}{|N|} \tag{34}$$

Then E [V] is bounded by

$$E[V] \leq \frac{1}{2\tau_{thr} - 1} \frac{q_{\Lambda}^2}{p} \tag{35}$$

The bound depends on the size of the regularization region via $q_{\Lambda} = E\left[\left|\hat{S}^{\Lambda}(I)\right|\right]$ which can be approximated via $\hat{q}_{\Lambda} = E\left[\left|\hat{S}^{\Lambda}(I)\right| \,|Z\right]$.

Again, this quantity can be approximated arbitrarily well by sub sampling. For a given regularization region Λ, one can choose an adequate threshold τ_{thr} such that E[V] is bounded as desired. Alternatively, one can fix the threshold and choose the regularization region q_{Λ} adequately. In classical regularization, the optimal amount of regularization depends on the noise level, which is hard to estimate, and the solution is quite sensitive to the choice of the regularization parameter. Stability selection, on the other hand, seems to be much less sensitive to the choice of the regularization region. At the same time, it offers a way to achieve exact error control without any estimate of the noise level. Note that in Meinshausen and Buhlmann (2010), the assumption of equal selection probability is replaced by the stronger assumption of exchangeability which demands that the joint probability distribution of the random variables $1_{\{k\in\hat{S}^{\Lambda}\}}$,. Nr permutations in is invariant unde $k \in N$
We used the weaker assumption to highlight that our modifications of stability selection need somewhat stronger assumptions albeit weaker than exchangeability.

The condition in itself is quite strong and, presumably, often not fulfilled for real data. However, it seems that an assumption of this kind has to be made to guarantee error control in this generality. Intuitively, the proof relies on the assumption that there is competition among noise variables for selection. This lowers the chance of any specific noise variable to be stably selected.

Via above theorem, we are not only able to control the per-family error rate E [V] but also the so-called family-wise error rate P[V > 0] (Dudoit, Shaffer, and Boldrick (2003)). This holds because of the markov-type inequality,

$$kP[V \geq k] \leq E[V] \tag{36}$$

Using k = 1 leads to the desired bound for $P[V > 0]$It seems natural to look at $\Lambda = [\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max}$ corresponds to the beginning of the regularization path where no variable is included and $\lambda_{\min}$ is chosen such that q_{Λ} has adequate size. To guarantee low error rate via inequality (35), Λ is typically chosen such that q_{Λ} is of order $\sqrt{p}$. If $p >> n$, then Λ might be a substantial part of the total regularization path. If however $p \approx n$, one has to restrict Λ to the beginning of the stability path. As an example, assume we set $\lambda_{\min}$ such that $q_{\Lambda} = \sqrt{p}$ and we choose $\tau_{thr} = 0.75$. Then the right side of inequality (35) is two. Assume that n = 100. If p = 4900, then we want to choose Λ such that $E\left[\left|\hat{S}^{\Lambda}(I)\right|\right] = 70$. At the end of the regularization path we know that 50 variables are included since $n/2 = 50$. If Λ is set to cover the whole regularization path, $\left|\hat{S}^{\Lambda}(I)\right|$ will be somewhat larger than 50 because it also contains all variables that were present at some point in the regularization path but are missing at its end. Still, we might be able to chose Λ as the whole regularization path and still achieve error control. If, however, p = 100, we would need to restrict our attention to a region Λ where about 10 variables have been selected. In general, one should not choose q_{Λ} too small because even variables that show high overall stability might not have fully stabilized yet. An example of this behavior can be seen in Figure. 3.

Obviously, one can choose Λ to contain only one value λ. This is called point wise control. To avoid the need to estimate q_{Λ}, one can use a base procedure that just selects a given number of variables. Say, one wants to fix q_{Λ} at 20 to achieve a certain amount of error control. Then one would define the base method to select the first 20 variables that enter in the stability path. With this base method, one then proceeds as before.

3.2. Consistent Variable Selection

We have seen that for the lasso it is impossible to have consistent variable selection in the sense that $P[\hat{S}(\lambda) = S] \to 1 \ (n \to \infty)$ if the design does not satisfy the irrepresentable condition. Meinshausen and Buhlmann (2010) introduced the randomized lasso which makes use of additional randomness. They showed that for this modification, consistent variable selection is possible for cases where the irrepresentable condition is not satisfied. The randomized lasso works via randomly perturbing the weighting for each variable. Specifically, let $\alpha \in (0,1]$ and $W_k, (k = 1,..., p)$ be i.i.d. random

variables taking either α or 1 as values where the two states have probability p_w and $(1-p_w)$ respectively. The randomized lasso estimator is then defined as

$$\hat{\beta}^{\lambda,W} = \arg\min_{\beta\in R} \|Y - X\beta\|_2^2 / n + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \quad (37)$$

This estimator is calculated repeatedly for different realizations of $W_k, (k = 1,..., p)$ and one looks for consensus in the collection of solutions. Consistency for this procedure was shown, granted that α and p_w are chosen sensibly. Combined with sub sampling of the data, the consistency result still holds.

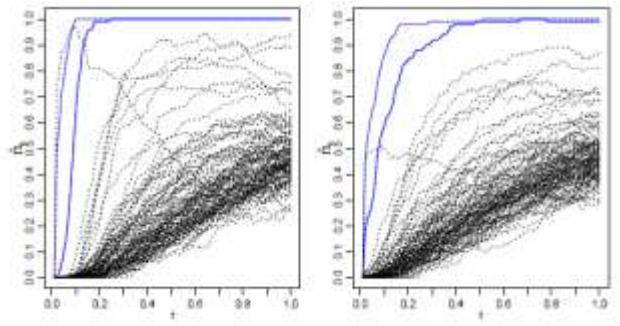

Fig. 4. Example stability paths for the toy model described in the main text using standard lasso on the right and randomized lasso on the left.

Solid blue lines correspond to active covariates, dashed black lines to inactive covariates. 100 subsamples were drawn.

To illustrate this, we reproduce a toy example presented in Meinshausen and Buhlmann (2010). The predictor variables were drawn from $N(0,\Sigma)$ where Σ was the identity up to entries Σ_{23}, Σ_{13} and their symmetric counterparts which were set to 0.6. Only the first two covariates were active having coefficient values one. The signal-to-noise ratio was set to 2. In this setting, the irrepresentable condition is violated for the third covariate. Figure 4 shows the results for regular stability selection and the randomized lasso procedure. One clearly sees that the randomized lasso is protected against picking the inactive third variable.

3.3. Modifications of Stability Selection

In Figure 4, we see that the strong bias for selection of the variable which violates the irrepresentable condition is weaker when the amount of regularization is smaller. The truly active variables on the other hand are still stably selected. If this were a general phenomenon then a test statistic favoring variables showing high stability over the whole regularization path might have better properties in this scenario. We will in the following introduce two test statistics that try to exploit this behavior and formulate their corresponding error bounds in spirit of 3.0.6.1.

3.3.1 Integral-Modification of Stability Selection

As an alternative to the test statistic from Meinshausen and Buhlmann (2010), this is

$$\hat{S}^{stable} = \left\{ k : \max_{\lambda\in\Lambda} \hat{\Pi}_k^{\lambda} \geq \tau_{thr} \right\} \quad (38)$$

we can replace the maximum by a scaled integral

$$\hat{S}_{\text{int}}^{stable} = \left\{ k : \int_{\Lambda} \hat{\Pi}_k^{\lambda} \, d\mu(\lambda) \geq \tau_{thr} \right\} \quad (39)$$

where

$$\int d\mu(\lambda) = 1 \quad (40)$$

This will select variables that show high stability on the whole regularization region. Under similar assumptions as above, an analogous kind of error control can be achieved. Again, we introduce some appropriate notation. Set

$$W_k(I) = \int_{\Lambda} 1_{\{k\in\hat{S}^{\lambda}(I)\}} \, d\mu(\lambda) \quad (41)$$

$$V_{\text{int}} = \left| N \cap \hat{S}_{\text{int}}^{stable} \right| \quad (42)$$

and for some b

$$q_{\Lambda}^{b} = E\left[\sum_{k\in T} 1_{\{W_k(I)\geq b\}} \right] \quad (43)$$

where T refers to the set of all covariates, i.e. $S \cup N = T$.

Theorem: Assume for any c, it holds that

$$\frac{E\left[\sum_{k\in S} 1_{\{W_k(1)\geq c\}}\right]}{E\left[\sum_{k\in N} 1_{\{W_k(1)\geq c\}}\right]} \geq \frac{|S|}{|N|} \quad (44)$$

Additionally, assume that $W_k(I)$ has the same distribution function for all $k \in N$. Then the expected number of falsely selected variables is bounded in the following way

$$E\left[|V_{\text{int}}|\right] \leq \frac{1-b}{(2\tau_{thr}-1-b)p} \left(q_{\Lambda}^{b}\right)^2, \quad (45)$$

where b needs to be chosen such that $2\tau_{thr} - 1 \geq b$.

Admittedly, this error bound yields a method that tends to have less power than the original method because it is too

conservative. However, we know that, roughly speaking, it works via showing that another test statistic, namely

$$P\left[W_k(I) \geq b \middle| Z\right] \tag{46}$$

has a lower bound depending on Z

$$\int_{\Lambda} \hat{\Pi}_k^{\lambda} \, d\mu(\lambda) \tag{47}$$

Then the error bound is proven for (46). As an alternative strategy, we could directly use this statistic. We will do this in the following.

3.3.2 Regularization-Stable Modification of Stability Selection

We define the statistic

$$T_{k,b} = P\left[W_k(I) \geq b \middle| Z\right] \tag{48}$$

Now we may define our stable set as

$$\hat{S}_{reg}^{stable} = \left\{k : T_{k,b} \geq \tau_{thr}\right\} \tag{49}$$

If we define

$$V_{reg} = \left| N \cap \hat{S}_{reg}^{stable} \right| \tag{50}$$

we can prove the following error bound theorem.

Theorem: Assume for any c, it holds that

$$\frac{E\left[\sum_{k \in S} 1_{\{W_k(1) \geq c\}}\right]}{E\left[\sum_{k \in N} 1_{\{W_k(1) \geq c\}}\right]} \geq \frac{|S|}{|N|} \tag{51}$$

Additionally, assume that $W_k(I)$ has the same distribution function for all $k \in N$. Then

the expected number of falsely selected variables is bounded in the following way

$$\mathrm{E}\left[\left|V_{reg}\right|\right] \leq \frac{1}{(2\tau_{thr} - 1)p} \left(q_{\Lambda}^{b}\right)^2 \tag{52}$$

Intuitively, the regularization-stable procedure will only include a variable k into the stable set $\hat{S}_{reg}^{stable}$ if, for most subsamples, it is present in a large portion of the regularization path for the considered regularization region. In the simulations presented below, not a weighting over Λ was used but rather an uniform weighting over the l 1-norm of the solution vector from zero to the size of the full OLS-solution. One just has to regard the function that maps the data to the set $\left\{k : W_k \geq b\right\}$ as the base selection method. For this method, we then use point wise stability selection. Note that this variant will yield the same results as standard stability selection if the selected set can only increase along the regularization path. However, this is not the case when using the lasso as base procedure.

4. Numerical Results

In this section, we explore the properties of the different variants of stability selection via simulation experiments. Two main questions need answering. Firstly, since the error bounds are proven under restrictive conditions which cannot be tested in practice, it is of interest to see if the error bound holds up for the various methods. Secondly, we want to compare the power of the various methods. However, since we are mainly interested in preventing false positives, we will look at a different power measure than the standard ROC-curves. Instead, we will ask how likely it is to uncover some fraction of the model without committing an error. We will focus on the comparison between standard stability selection and regularization-stable stability selection and neglect integral stability selection because it cannot compete with regularization-stable stability selection. For regularization stable stability selection, we will use the whole regularization path as Λ in q_{Λ}^{b} and use changes in b to achieve the desired error control. All calculations were performed in R. Lasso traces were calculated by the LARS algorithm (Efron et al. (2004)) implemented in the lars-package.

4.1. Data Sets

To test our method on realistic data, two real data sets were used. However, since we need to know the true underlying model to evaluate the quality of our different approaches, we discard the response variable and replace it by simulated values for some known linear model. Models were constructed by choosing sets of covariates of predefined size *s* at random. For these, the corresponding β-values were drawn from a uniform distribution on the $[-1,1]$-interval. A normal error $\varepsilon \sim N(0, \sigma^2)$ was added where the standard deviation was chosen such that the signal-to-noise ratio reached a certain predefined value S/N, i.e.

$$S/N = \frac{\left\|\mathrm{X}\beta\right\|_2^2}{n \operatorname{var}(\varepsilon)} \tag{53}$$

Data sets derived from real experiments were also presented in Buhlmann and van de Geer (2011).

4.2. Gene-Expression Data

This data set from DSM Nutritional Products (Switzerland) stems from a gene expression experiment. The goal was to relate gene expression to the production rate of riboflavin in B. Subtilis. To this end, samples from various fermentation processes were taken and the

expression levels of 4088 genes were measured via microarray technology. At the same time, production rate of riboavine was measured. This data set can be written as a 4088×111 design matrix $X \in R^{4088 \times 111}$ and a response variable $Y \in R^{111}$.

4.3. Motif Regression Data

This data set stems from a CHIP-chip study. The goal was to identify short DNA segments, called motifs, to which the transcription factor HIF1α binds. To this end, binding strengths of HIF1α to relatively long stretches of DNA was determined. Then, the abundance of motifs within the longer segments was determined via a computational biology algorithm (Liu, Brutlag, and Liu (2002)). This data set can be written as a 660×2587 Design matrix $X \in R^{660 \times 2587}$ and a response variable $Y \in R^{2587}$. However, to speed up calculations, only the first two thirds of all the covariates and observations were used.

4.4. Simulated Designs

Additionally, we generated designs by simulation. Sample size n was fixed at 200 for all simulation runs.

a) Factor model with 10 factors: The design X satisfies the following conditions

$$\underset{n \times p}{X} = \underset{n \times 10}{\Lambda} \cdot \underset{10 \times 1}{f} \times \varepsilon \tag{54}$$

where each entry in Λ, f and ε is derived from N (0, 1).

b) Block-wise structure: Each observation is drawn from $N_p(0, \sum)$, where $\sum$ has block diagonal structure. Each block has size 20×20 and contains only entry values 0.9 except on the diagonal, where it contains ones.

4.5. Checking Error Bound

We performed stability selection for simulated data, where the design was either derived from the real data sets described above, or alternatively, also simulated. Denote the estimated upper bound for E [V] with E_{bound}, i.e.

$$E_{bound} = \frac{1}{2\tau_{thr} - 1} \frac{\left(\hat{q}\hat{\Lambda}\right)^2}{p} \tag{55}$$

(or analogously defined with $\hat{q}_{\Lambda}^{b}$ for the regularization-stable variant of stability selection). Figures 4.1 and 4.2 shows the estimated values of E [V] versus Ebound for various values of τ_{thr}, where model size s and signal-to-noise ratio were fixed. To achieve the desired values for Ebound, either $\lambda_{\min}$ in the case of standard stability selection or b in the case of regularization-stable stability selection were adjusted. Figure 5 shows the results for the riboflavin data set and the motif regression data set. We see that error control is usually conservative, for high values of Ebound. For the riboflavin data set, we see that, for the standard method, the false positive rate seems to be quite insensitive to changes in the demanded error control for fixed thresholds τ_{thr}. This corresponds to an insensitivity of the method to changes in the regularization region Λ for this data set. In the case of very strict error control and low threshold values, this behavior seems to lead to problems in one setting for the riboflavin data set.

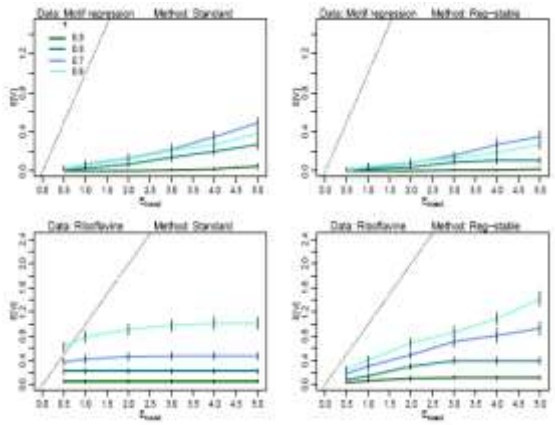

Fig. 5. E [V] versus Ebound for the truncated motif regression data set and the riboflavin data sets (top row and bottom row respectively).

The two methods employed were standard and regularization-stable stability selection (first and second column respectively). Simulation settings where S=N = 2 and s = 20 for motif regression data and s = 16 for riboflavin data. 100 simulation runs were performed. The displayed standard deviations were estimated via the bootstrap.

4.2 show the same for simulated designs of type a) and b). For design b), both methods work fine whereas design a) leads to violation of the estimated error bound for standard stability selection. Regularization-stable stability selection seems better able to control the error rate in this setting. Another way to check if the error bound holds up is to fix τ_{thr} and Ebound and adapt $\hat{q}\Lambda$ or $\hat{q}_{\Lambda}^{b}$ adequately. We ran simulations for various models. Figure 7 and 8 show the results of such a simulation experiment for the riboflavin data set and the truncated motif regression data set respectively. For the riboflavin data set, we again see a conservative error control in most cases. However, for very low values of τ_{thr} and Ebound, we see that the error bound is not respected in all settings for standard stability selection. Regularization stable stability selection seems more sensitive to values of Ebound but seems to stay always well within the bound set by Ebound. This respect of the error bound is paid with loss of power in settings, where tight error control is demanded. In the truncated motif regression data set, the

situation is reversed. Here, the regularization-stable version of stability selection shows less sensitivity to the choice of Ebound. However, error control is much more conservative overall, and nowhere is the bound violated.

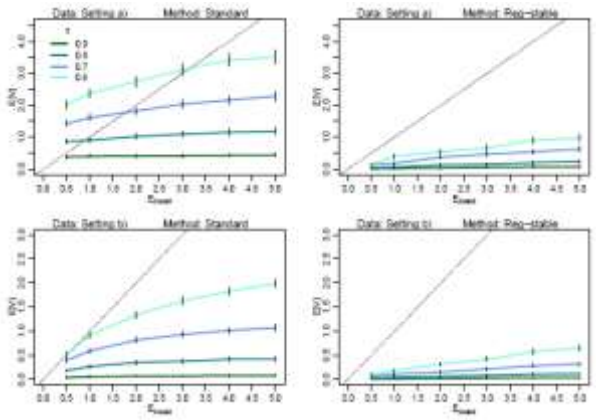

Fig. 6. Analogous plot as figure 5 for the for simulated designs of type a) and b) (top row and bottom row respectively).

Simulation settings where S=N = 2, p = 1000 and s = 14 for design a) and s = 8 for design b).

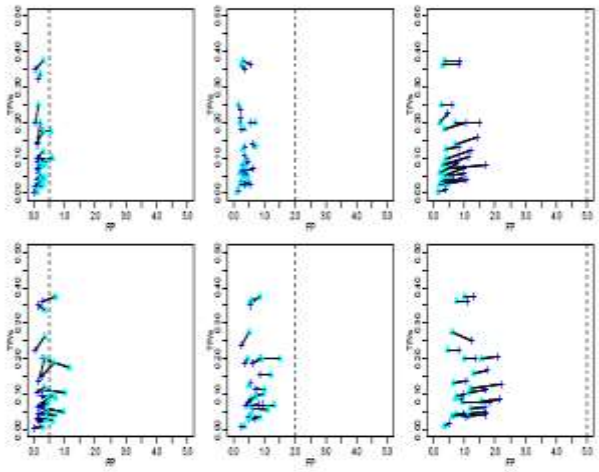

Fig. 7. Average proportion of relevant variables selected versus average of false positives.

The design was taken from the riboflavin data set. For each displayed point, an average was taken over 30 runs. τ_{thr} was set to (0.7, 0.6) for the upper and lower row respectively. Ebound was set to (0.5, 2.5) from left to right. Triangles denote results for standard stability selection and crosses for regularization-stable stability selection. Simulation settings were all combinations of s = (2, 4, 6, 8, 10) and S/N = (4, 3, 2, 1).

4.6. Power Comparison

To compare the power between various approaches, we performed standard stability selection, stability selection employing randomized lasso and regularization-stable stability Selection.

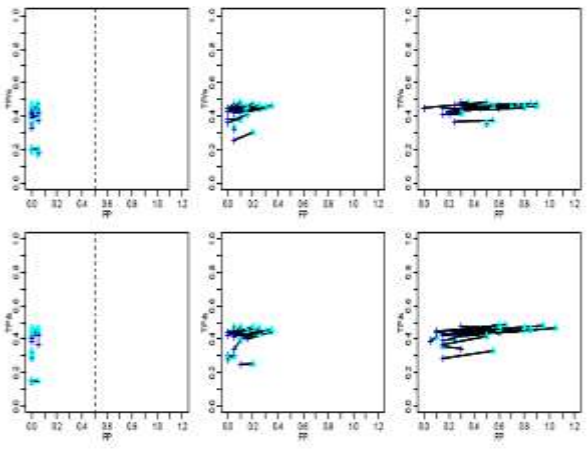

Fig. 8. Average proportion of relevant variables selected versus average of false positives.

The design was taken from the truncated motif regression data sets. For each displayed point, an average was taken over 30 runs. The settings for Ebound and τ_{thr} were the same as in Figure 7. Simulation settings were all combinations of s = (3, 4, 5, 10, 20) and S/N = (4, 3, 2, 1).

We chose Λ such that $q_\Lambda = \sqrt{0.8p}$. Analogously, for regularization-stable stability selection, b was chosen such that $q_\Lambda^b = \sqrt{0.8p}$. This setting is also used in the numerical simulations of Meinshausen and Buhlmann (2010). For randomized lasso, both weakness α and p_w were set to 0.5.

Figure 4.5 shows the estimated probability to select 10% of the relevant variables without selecting any false positives. For comparison to standard lasso, we checked if at any position in the regularization path 10% of relevant variables were chosen without including any false positives.

Standard stability selection and regularization-stable stability selection give comparable results for the real data sets. For simulation setting a), regularization-stable stability selection outperforms all other methods substantially. Standard lasso is weaker in almost all settings than either standard stability selection or regularization-stable stability selection. Randomized lasso performs better than standard stability selection in setting a) but is still outperformed by regularization-stable stability selection. Figure 4.6 shows, for the same data, the estimated probability to select 40% of relevant variables without including any false positives. To illustrate how the advantage in setting a) of the regularization-stable variant over the standard approach and even the randomized lasso approach comes about, Figure 11 displays sample stability paths for regular lasso and randomized lasso respectively. One sees that the demand for strict error control leads to a suboptimal

regularization region. For standard stability selection, two inactive variables show high stability initially, but this

stability disappears further down the regularization path. Therefore, regularization-stable stability selection will not select them. Except for the setting a), results for the standard and the regularization-stable method look very similar. The only other setting where regularization-stable stability selection seems beneficial is in setting b) when noise is low, and the numbers of truly active variables is high. To explore this parameter setting, we generated truncated versions of the riboflavin data set.

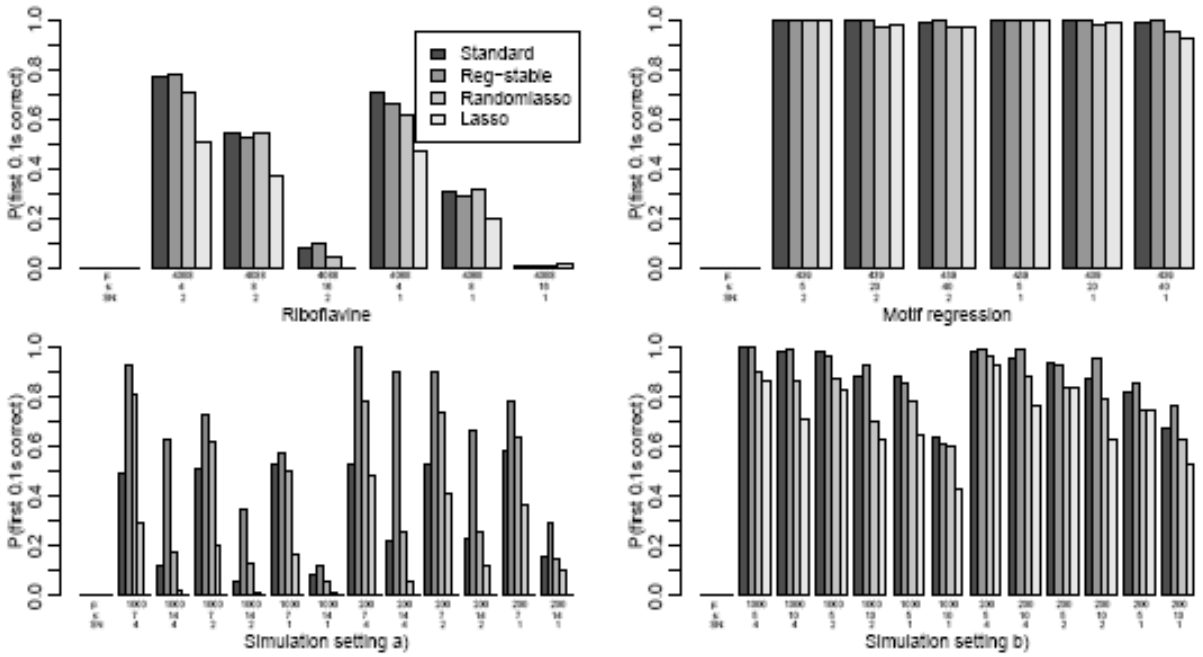

Fig. 9. Probability of selecting 0.1s active variables without selection any noise variables, where $\hat{q\Lambda} = \hat{q}_{\Lambda}^{b} = \sqrt{0.8p}$
'Standard' refers to the procedure proposed by Meinshausen et al, without random weights and 'Random lasso' to the procedure with random weights. 'Reg-stable' refers to the regularization-stable version and 'Lasso' to the lasso trace procedure described in the main text. 100 simulation runs were performed.

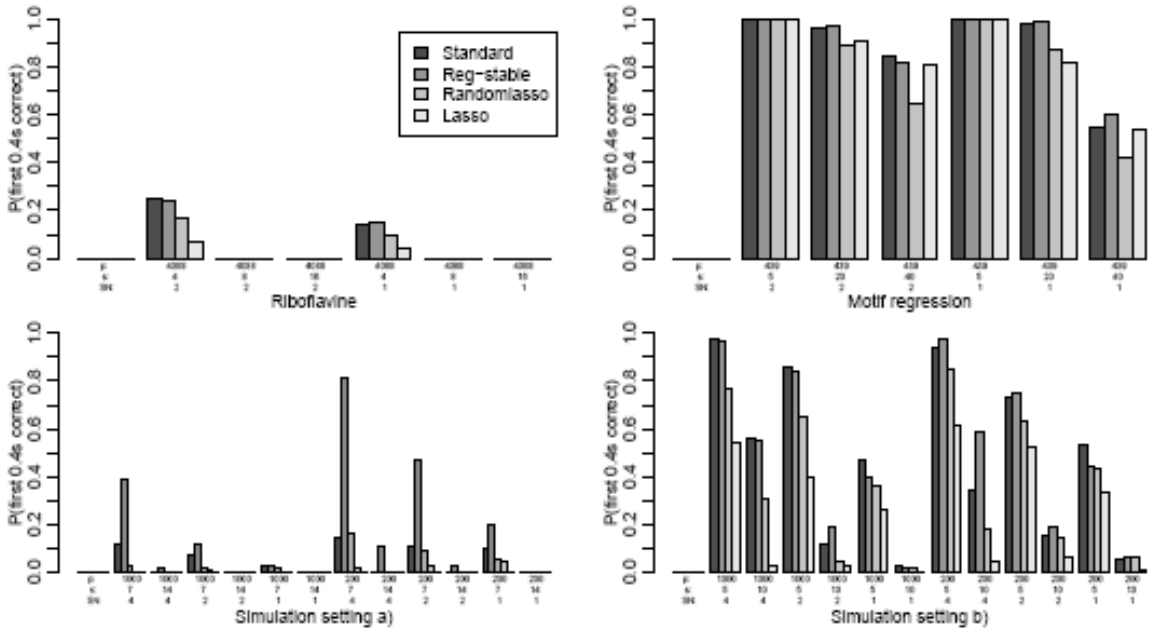

Fig. 10. Probability of selecting 0.4s active variables without selection any noise variables, where $\hat{q\Lambda} = \hat{q}_{\Lambda}^{b} = \sqrt{0.8p}$.
100 simulation runs were performed.

In particular, we took the p_{new} columns of the original design leading to a data matrix of size $p_{new} \times 111$. The resulting designs still tend to contain some high correlations, which is shown in Figure 12. This is the case because the genes were ordered alphabetically, and genes of the same gene family often have the same names up to numerals and also tend to be co expressed.

Figure 11 explores the setting where $p = p_{new}$ is quite small and $s \approx \sqrt{p}$. Again, we choose Λ such that $\hat{q}\Lambda = \hat{q}_{\Lambda}^{b} = \sqrt{0.8p}$. Regularization stable stability selection seems advantageous in low to medium noise settings when the model is quite complex such that $\hat{q}\Lambda < s$. A similar scenario is depicted in Figure 14. Again, $s \approx \sqrt{p}$ but p is chosen larger.

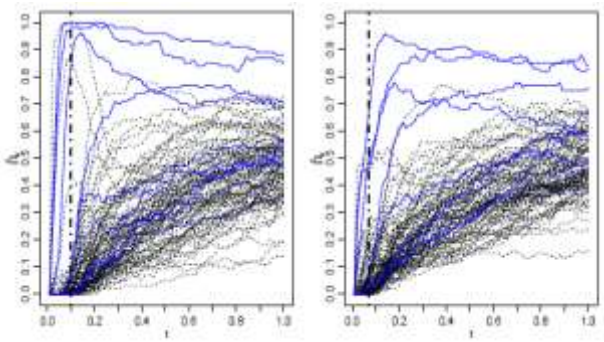

Fig. 11. Sample stability paths for regular lasso (left panel) and randomized lasso (right panel). Simulation setting b) was used. Model parameters were s=10, n=100, p=100, S/N=4. Also, the regularization region for $\hat{q}\Lambda = \sqrt{0.8p}$ is marked.

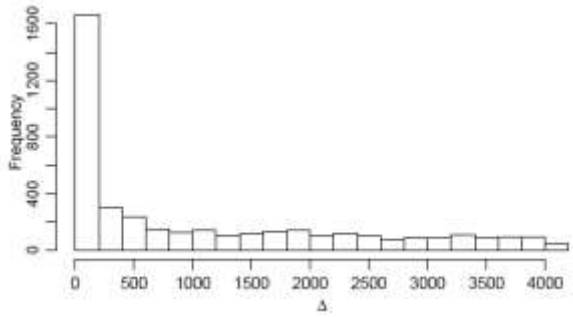

Fig. 12. Histogram of distances for the riboflavin data. Only genes having a maximal absolute correlation to some other gene of 0.8 or higher were considered. For each gene, the distance (in number of columns) to the gene which had the highest absolute correlation with it was measured.

Additionally, to the previous plots, we added results of stability selection where only the very end of the regularization path is used. Due to the complexity of the model compared to the numbers of observations, we only demanded that one truly active variable be recovered without including any inactive variables. In the low noise settings, we see an advantage of regularization stable stability selection over standard stability selection, particularly where $\hat{q}\Lambda \approx s$ or even $\hat{q}\Lambda < s$ but also a slight advantage for $\hat{q}\Lambda > s$. For higher noise settings, the advantage gradually disappears. This seems to happen more quickly for small values of s. End-of-path stability selection performs very well in low noise settings but clearly loses its advantage for higher noise settings. Figure 15 shows similar results for even larger models. Here, the overall chance of recovery of any active variable without including an inactive one is quite small. Nevertheless, regularization stable stability selection seems to have beneficial effects. Also note that the advantage of stability selection over regular lasso remains intact even though n is not much larger than s. To check if the regularization-stable stability selection improves even for high noise settings if $\hat{q}\Lambda$ is substantially smaller than s, we replotted results shown in Figure 13 and 14 for $\hat{q}\Lambda = \hat{q}_{\Lambda}^{b} = \sqrt{0.1p}$ instead of $\sqrt{0.8p}$. Now, regularization-stable stability selection seems advantageous even for high noise settings. The advantage of standard stability selection over regular lasso seems to have completely vanished for p = 100.

From these simulations, it seems that the regularization-stable method has greater power if $\hat{q}\Lambda < s$. However, when we actually want to use our error control approach, it is also necessary, that variables show higher stability than our cut-o_ _thr. We explore this in Figure 18. We chose $\tau_{thr} = 0.6$ and Ebound = 0.5 so that $\hat{q}\Lambda = \hat{q}_{\Lambda}^{b} = \sqrt{0.1p}$. We see that the error rate for regularization-stable stability selection is smaller while the corresponding number of true positives is similar. Also, error control in terms of bound 3.0.6.3 is not well achieved in standard stability selection.

4.7. Discussion of the Simulation Results

Meinshausen and Buhlmann (2010) prove that randomized lasso consistently selects truly active variables even in cases when regular lasso fails because the irrepresentable condition does not hold. As an empirical validation of this finding, they show marked improvement in simulations using the 10-factor model. However, randomized lasso has the drawback that additional parameters need to be set in the procedure. Also, the additional perturbation tends to decrease stability values for variables showing high stability. When using error bound 3.0.6.3, one needs high selection probabilities for truly active variables, so randomized

lasso selects fewer active variables. It is therefore interesting to note that regularization-stable stability selection also leads to a marked improvement for the 10-factor model.

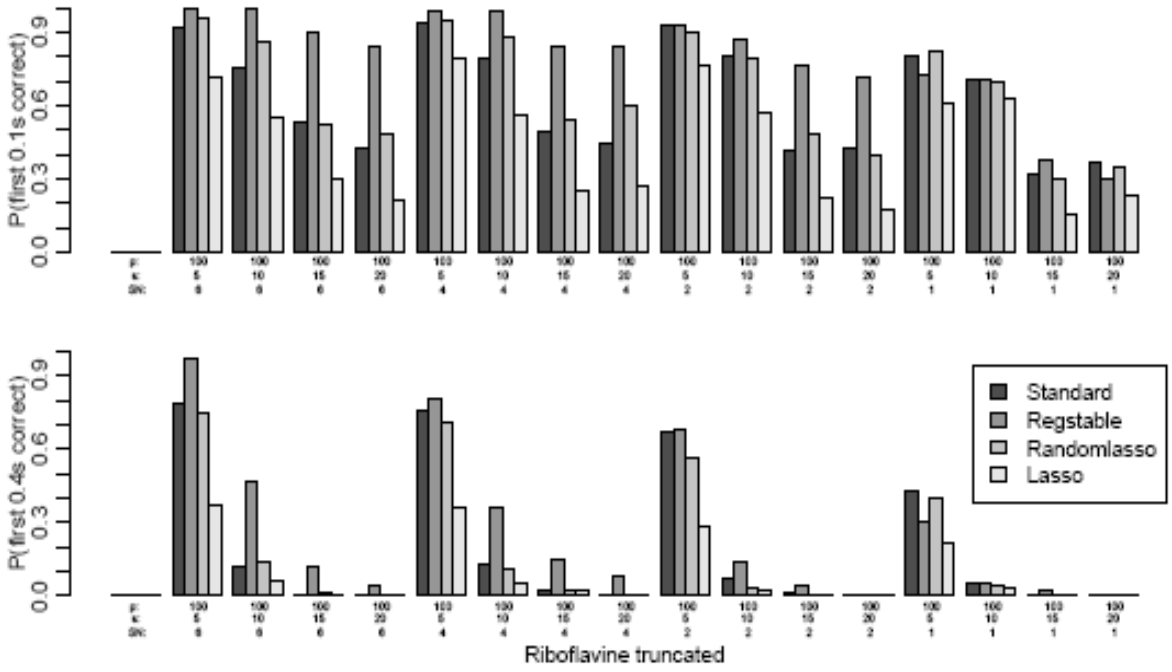

Fig. 13. Probability of selection 0:1s (top row) and 0:4s (bottom row) of active variables without selection any noise variables, where $\hat{q\Lambda} = \hat{q}_{\Lambda}^{b} = \sqrt{0.8p}$. 100 simulation runs were performed.

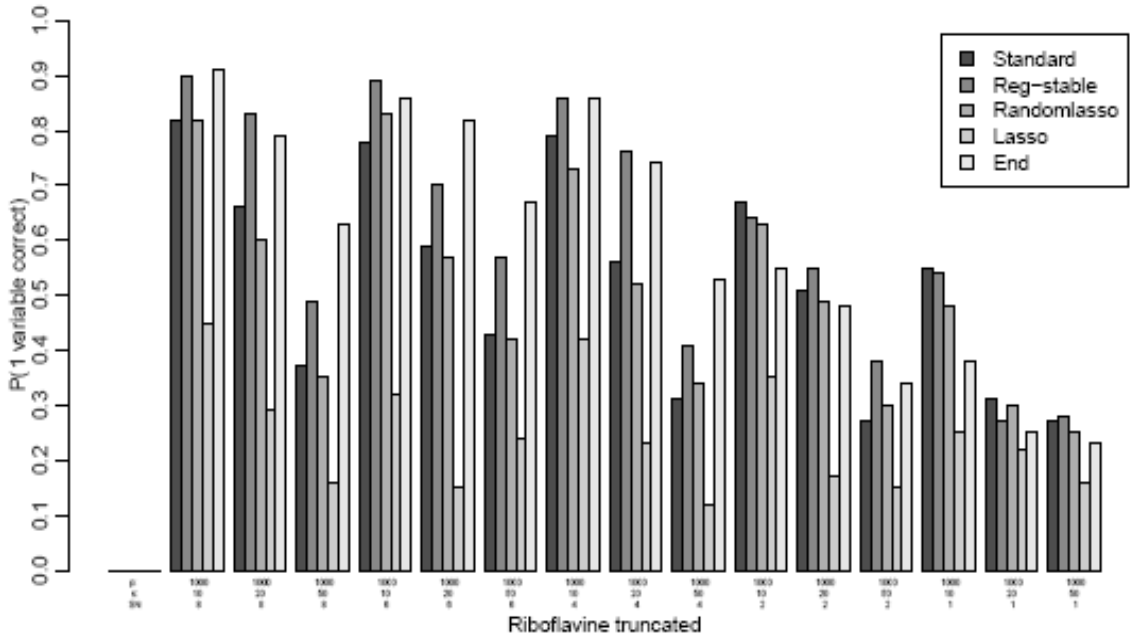

Fig. 14. Probability of selecting one active variable without selection of any noise variables. Additional to the power plots above, we also included results of stability selection via regular lasso, where only the final fit in the lasso trace was used. Again, $\hat{q\Lambda} = \hat{q}_{\Lambda}^{b} = \sqrt{0.8p}$ and 100 simulation runs were performed.

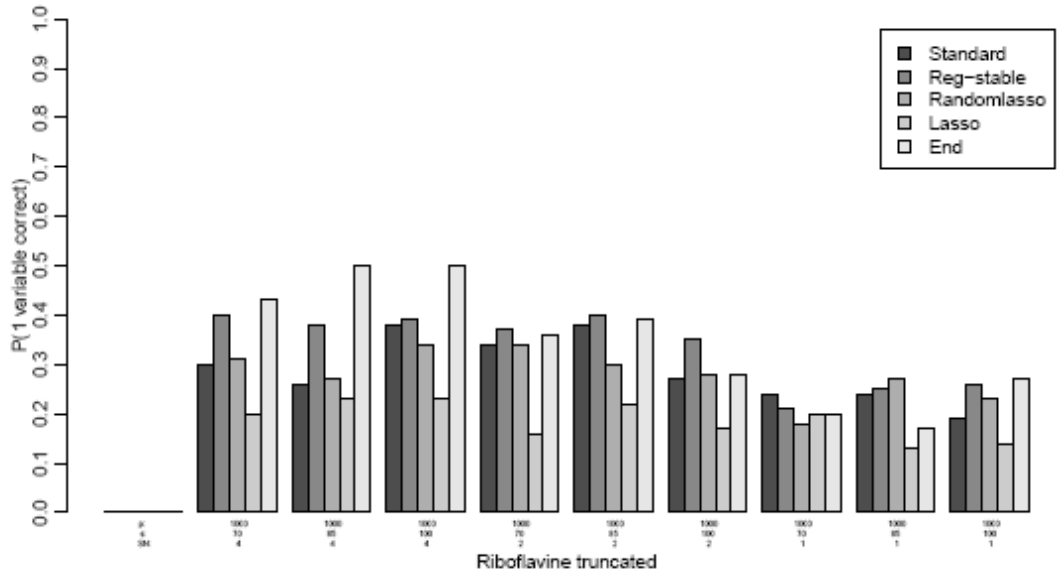

Fig. 15. Probability of selecting one active variable without selection of any noise variables.

Additional to the power plots above, we also included results of stability selection of regular lasso where only the final fit in the lasso trace was used. $\hat{q}_\Lambda = \hat{q}_\Lambda^b = \sqrt{0.8p}$ and 100 simulation runs were performed.

When using the error-bound 3.0.6.3, the choice of the size of the regularization region is made independently of the noise in the data and the model size. In general, results achieved with standard stability selection depend little on the size of the regularization region. However, it seems intuitive that variable selection cannot be done optimally when the regularization region is chosen such that model size of the true underlying model is bigger than $\hat{q}_\Lambda$ given that enough data is available to get a 'decent' estimate. Simulations seem to imply that regularization-stable stability selection can be beneficial in this setting. One reason could be that information from further down the regularization path is included where model size is more adequate for the problem. If one wants to control the

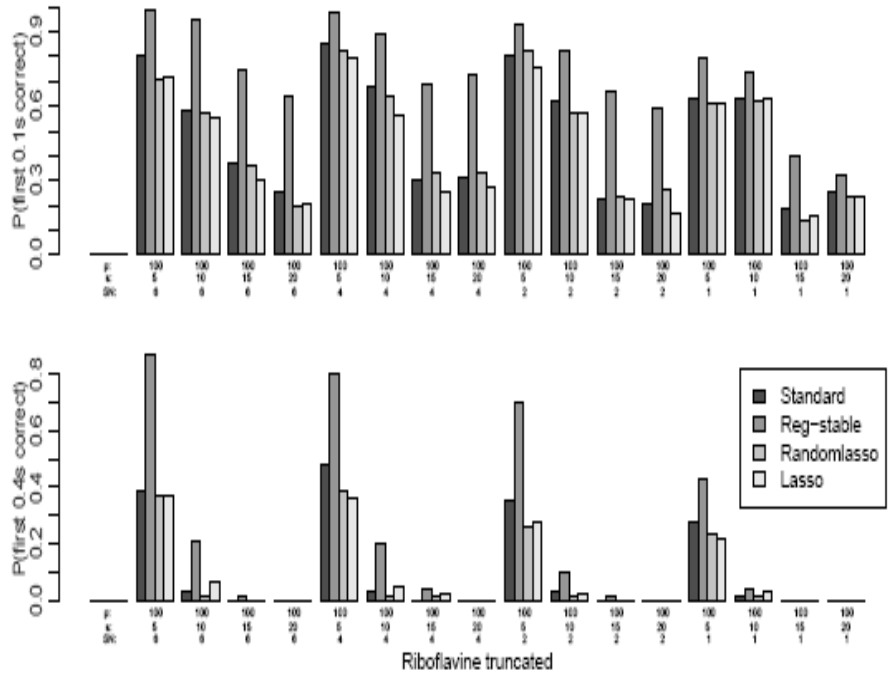

Fig. 16. Probability of selecting 0:1s (top row) and 0:4s (bottom row) of active variables without selection of any noise variables. $\hat{q}_\Lambda = \hat{q}_\Lambda^b = \sqrt{0.1p}$. 100 simulation runs were performed.

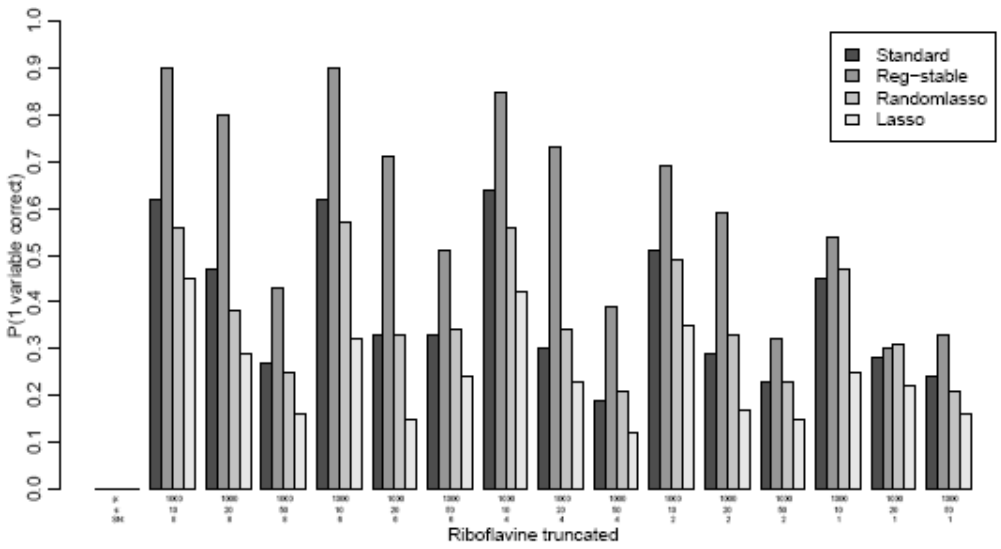

Fig. 17. Probability of selecting one active variable without selection of any noise variables. $\hat{q\Lambda} = \hat{q}_{\Lambda}^{b} = \sqrt{0.1p}$. 100 simulation runs were performed.

Family wise error rate, that is P [V > 0], a setting was $\hat{q\Lambda} < s$ might be quite common. For example, note that in Figure 18 the aim would be to control the family-wise error rate at 0:5. It is also noteworthy that standard stability selection exhibits problems to control error rates at very low values in the example of the riboflavin data set making it infeasible for controlling of the family-wise error rate in this example.

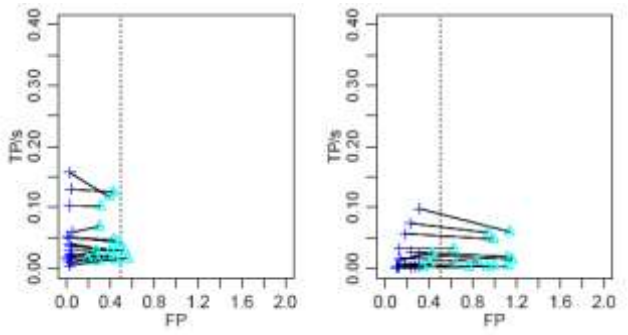

Fig. 18. Average proportion of relevant variables selected versus average of false positives. The design was taken from the riboflavin data set, but only the p first variables were looked at where p was set to 100 for the left and 1000 for the right panel. The settings for Ebound and τ_{thr} r was 0.5 and 0.6 respectively. Simulation settings corresponded to the ones used in Figure 13, and 14 respectively.

5. Discussion

In this paper, some variants of stability selection, proposed by Meinshausen and Buhlmann (2010) were devised. Stability selection combines selection algorithms for high dimensional problems with sub sampling. It provides a principle of error control which guides the amount of regularization. In most cases, the solutions do not depend strongly on the initial regularization. Of the devised methods, one, dubbed the regularization stable version, provided in many cases similar results as the original method. However, it seemed superior in some settings. One was the setting of the 10-factor model, which is a very hard problem for the basic lasso. Improvements for this setting seemed comparable to the randomized lasso. However, the method does not rely on additional randomization to achieve this gain. Since this additional randomization seems to decrease the highest stabilities achieved, this can be seen as an advantage. Additionally, it seemed that the method outperformed the standard procedure if true model size s was larger than the number of covariates that the base procedure would select on the regularization region, (i.e. $q\Lambda < s$) and the noise was not too large. This is more likely to occur when $\sqrt{p}$ is small compared to s and when tight error control is demanded, for example when one tries to control the family-wise error rate. One practical situation where this could be relevant are rescreening, where a prior screen has produced a candidate list and a second screen is performed on this shorter list. One would hope that s is still as large as in the original screen, but p is much smaller. Perhaps even noise would be somewhat reduced because smaller screens are easier to handle. Also, the ratio of s to n is cheaper to decrease. Furthermore, it is plausible that for this second screen, tight error control is desirable, since the next step in the pipeline might be validation through time consuming alternative methods. The seen advantage in the above settings might stem from the fact, that it allows incorporation of information from further down the regularization path, without inflating the number of selected variables for the base method grossly.

One can think of other procedures that also have this property. for example, one could use some truncated version of the lasso, which would have a much sparser solution than the original lasso procedure for the same amount of regularization taking place.

REFRENCES

[1] Buhlmann, P. and S. van de Geer (2011). Statistics for High-Dimensional Data. Statistics. New York: Springer.
[2] Dudoit, S., J. Shaffer, and J. Boldrick (2003). Multiple hypothesis testing in microarray experiments. Statistical Science, 71-103.
[3] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. The Annals of statistics 32 (2), 407-499.
[4] Liu, X., D. Brutlag, and J. Liu (2002). An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature biotechnology 20 (8), 835-839.
[5] Meinshausen, N. and P. Buhlmann (2010). Stability selection. Journal of the Royal Statistical Society. Series B (Methodological) 72, 417-473.
[6] Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. The Annals of Statistics 35 (3), 1012-1030.
[7] Sauerbrei, W. and M. Schumacher (1992). A bootstrap resampling procedure for model building: application to the cox regression model. Statistics in Medicine 11 (16), 2093- 2109.
[8] Sen, A. and M. Srivastava (1990). Regression Analysis; Theory, Methods, and Applications. Springer-Verlag.
[9] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.