Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

p-Values for High-Dimensional Regression

Nicolai Meinshausen, Lukas Meier & Peter Bühlmann

To cite this article: Nicolai Meinshausen, Lukas Meier & Peter Bühlmann (2009) *p*-Values for High-Dimensional Regression, Journal of the American Statistical Association, 104:488, 1671-1681, DOI: 10.1198/jasa.2009.tm08647

To link to this article: https://doi.org/10.1198/jasa.2009.tm08647

Published online: 01 Jan 2012.

p-Values for High-Dimensional Regression

Nicolai MEINSHAUSEN, Lukas MEIER, and Peter BÜHLMANN

Assigning significance in high-dimensional regression is challenging. Most computationally efficient selection algorithms cannot guard against inclusion of noise variables. Asymptotically valid *p*-values are not available. An exception is a recent proposal by Wasserman and Roeder that splits the data into two parts. The number of variables is then reduced to a manageable size using the first split, while classical variable selection techniques can be applied to the remaining variables, using the data from the second split. This yields asymptotic error control under minimal conditions. This involves a one-time random split of the data, however. Results are sensitive to this arbitrary choice, which amounts to a "*p*-value lottery" and makes it difficult to reproduce results. Here we show that inference across multiple random splits can be aggregated while maintaining asymptotic control over the inclusion of noise variables. We show that the resulting *p*-values can be used for control of both family-wise error and false discovery rate. In addition, the proposed aggregation is shown to improve power while reducing the number of falsely selected variables substantially.

KEY WORDS: Data splitting; False discovery rate; Family-wise error rate; High-dimensional variable selection; Multiple comparisons.

1. INTRODUCTION

The problem of high-dimensional variable selection has received tremendous attention in the last decade. Sparse estimators like the Lasso (Tibshirani 1996) and extensions thereof (Zou 2006; Meinshausen 2007) have been shown to be very powerful because they are suitable for high-dimensional data sets and because they lead to sparse, interpretable results.

In the usual workflow for high-dimensional variable selection problems, the user sets potential tuning parameters to their prediction optimal values and uses the resulting estimator as the final result. In the classical low-dimensional setup, some error control based on *p*-values is a widely used standard in all areas of sciences. So far, *p*-values are not available in high-dimensional situations, except for the proposal of Wasserman and Roeder (2009). An ad hoc solution for assigning relevance is to use the bootstrap to analyze the stability of the selected predictors and focus on those selected most often (or even always). Bach (2008) and Meinshausen and Bühlmann (2008) showed that for the Lasso, this leads to a consistent model selection procedure under fewer restrictions than for the nonbootstrap case.

More recently, some progress has been made in obtaining error control (Meinshausen and Bühlmann 2008; Wasserman and Roeder 2009). Here we build on the approach of Wasserman and Roeder (2009) and show that an extension of their "screen and clean" algorithm leads to a more powerful variable selection procedure. Moreover, family-wise error rate (FWER) and false discovery rate (FDR) can be controlled, whereas Wasserman and Roeder (2009) focused on variable selection rather than assigning significance via *p*-values. We also extend the methodology to control of the false discovery rate (Benjamini and Hochberg 1995) for high-dimensional data. Although the main application of our procedure is for high-dimensional data, where the number p of variables can greatly exceed sample size n, we show that the method also is quite competitive with more standard error control for $n > p$ settings, indeed often providing better detection power in the presence of highly correlated variables.

This article is organized as follows. We briefly discuss the single-split method of Wasserman and Roeder (2009) in Section 2, noting that the results can depend strongly on the arbitrary choice of a random sample split. We propose a multi-split method, which eliminates this dependence. In Section 3 we prove FWER and FDR control of the multisplit method, and in Section 4 we show numerically that for simulated and real data sets, the method is more powerful than the single-split version while significantly reducing the number of false discoveries. We outline some possible extensions of the proposed methodology in Section 5.

2. SAMPLE SPLITTING AND HIGH–DIMENSIONAL VARIABLE SELECTION

We consider the usual high-dimensional linear regression setup with a response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and an $n \times p$ fixed design matrix $\mathbf{X}$ such that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ is a random error vector with ε_i iid $\mathcal{N}(0, \sigma^2)$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter vector. Extensions to other models are given in Section 5.

Denote by

$$S = \{j; \beta_j \neq 0\}$$

the set of active predictors, and similarly by $N = S^c = \{j; \beta_j = 0\}$ the set of noise variables. Our goal is to assign *p*-values for the null hypotheses $H_{0,j} : \beta_j = 0$ versus $H_{A,j} : \beta_j \neq 0$ and to infer the set S from a set of n observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$. We allow for potentially high-dimensional designs, that is, $p \gg n$. This makes statistical inference very challenging. An approach proposed by Wasserman and Roeder (2009) is to split the data into two parts, reducing the dimensionality of predictors on one part to a manageable number of predictors (keeping the important variables with high probability), and then assign *p*-values and make a final selection on the second part of the data, using classical least squares estimation.

Nicolai Meinshausen is University Lecturer, Department of Statistics, University of Oxford, Oxford OX1 3TG, U.K. (E-mail: *meinshausen@stats.ox.ac.uk*). Lukas Meier is Ph.D. Student, Peter Bühlmann is Professor, Seminar für Statistik, ETH Zurich, 8092 Zurich, Switzerland. Nicolai Meinshausen acknowledges the generous support and hospitality shown during his stay at Forschungsinstitut für Mathematik at ETH Zürich.

Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2009.tm08647

2.1 Family-Wise Error Rate Control With the Single-Split Method

The procedure of Wasserman and Roeder (2009) attempts to control the family-wise error rate (FWER), defined as the probability of making at least one false rejection. The method relies on sample splitting, performing variable selection and dimensionality reduction on one part of the data and classical significance testing on the other part. The data are split randomly into two disjoint groups, $D_{in} = (\mathbf{X}_{in}, \mathbf{Y}_{in})$ and $D_{out} = (\mathbf{X}_{out}, \mathbf{Y}_{out})$, of equal size. Let $\tilde{S}$ be a variable selection or screening procedure that estimates the set of active predictors. Abusing notation slightly, we also denote by $\tilde{S}$ the set of selected predictors. Then variable selection and dimensionality reduction is based on D_{in}; that is, we apply $\tilde{S}$ only on D_{in}. This includes the selection of potential tuning parameters involved in $\tilde{S}$. The idea is to break down the large number, p, of potential predictor variables to a smaller number, $k \ll p$, with k at most a fraction of n, while keeping all relevant variables. The regression coefficients and the corresponding p-values, $\tilde{P}_1, \ldots, \tilde{P}_p$, of the selected predictors are determined based on D_{out} using ordinary least squares estimation on the set $\tilde{S}$ and setting $\tilde{P}_j = 1$ for all $j \notin \tilde{S}$. If the selected model $\tilde{S}$ contains the true model S (i.e., $\tilde{S} \supseteq S$), then the p-values based on D_{out} are unbiased. Finally, each p-value, $\tilde{P}_j$, is adjusted by a factor $|\tilde{S}|$ to correct for the multiplicity of the testing problem.

The selected model is given by all variables in $\tilde{S}$ for which the adjusted p-value is below a cutoff, $\alpha \in (0, 1)$,

$$\hat{S}_{single} = \{j \in \tilde{S} : \tilde{P}_j |\tilde{S}| \le \alpha\}.$$

Under suitable assumptions (discussed later), this yields asymptotic control against inclusion of variables in N (false positives) in the sense that

$$\limsup_{n\to\infty} \mathbb{P}[|N \cap \hat{S}_{single}| \ge 1] \le \alpha,$$

that is, control of the FWER. The method is easy to implement and yields the asymptotic control under weak assumptions. The single-split method relies on an arbitrary split into D_{in} and D_{out}, however, and the results can change drastically if this split is chosen differently. This in itself is unsatisfactory, because then the results are not reproducible.

2.2 Family-Wise Error Rate Control With the New Multisplit Method

An obvious alternative to a single arbitrary split is to divide the sample repeatedly. For each split, we end up with a set of p-values. How to combine and aggregate the results is not obvious, however. Here we describe a possible approach. For each hypothesis, a distribution of p-values is obtained for random sample splitting. We propose that error control can be based on the quantiles of this distribution. We show empirically that, possibly unsurprisingly, the resulting procedure is more powerful than the single-split method. The multisplit method also makes the results reproducible, at least approximately if the number of random splits is chosen to be very large.

The multisplit method uses the following procedure:

For $b = 1, \ldots, B$:

1. Randomly split the original data into two disjoint groups, $D_{in}^{(b)}$ and $D_{out}^{(b)}$, of equal size.
2. Using only $D_{in}^{(b)}$, estimate the set of active predictors, $\tilde{S}^{(b)}$.
3. (a) Using only $D_{out}^{(b)}$, fit the selected variables in $\tilde{S}^{(b)}$ with ordinary least squares and calculate the corresponding p-values, $\tilde{P}_j^{(b)}$, for $j \in \tilde{S}^{(b)}$.
   (b) Set the remaining p-values to 1, that is,
   $$\tilde{P}_j^{(b)} = 1, \qquad j \notin \tilde{S}^{(b)}.$$
4. Define the adjusted (nonaggregated) p-values as
   $$P_j^{(b)} = \min\bigl(\tilde{P}_j^{(b)} |\tilde{S}^{(b)}|, 1\bigr), \qquad j = 1, \ldots, p. \tag{2.1}$$

Finally, aggregate over the B p-values $P_j^{(b)}$, as discussed later.

This procedure leads to a total of B p-values for each predictor $j = 1, \ldots, p$. It will turn out that suitable summary statistics are quantiles. For $\gamma \in (0, 1)$ define

$$Q_j(\gamma) = \min\bigl\{1, q_\gamma\bigl(\{P_j^{(b)}/\gamma; b = 1, \ldots, B\}\bigr)\bigr\}, \tag{2.2}$$

where $q_\gamma(\cdot)$ is the (empirical) γ-quantile function.

A p-value for each predictor $j = 1, \ldots, p$ is then given by $Q_j(\gamma)$, for any fixed $0 < \gamma < 1$. In Section 3 we show that this is an asymptotically correct p-value, adjusted for multiplicity.

Properly selecting γ may be difficult. Error control is not guaranteed if we search for the best value of γ. We propose to instead use an adaptive version that selects a suitable value of the quantile based on the data. Let $\gamma_{\min} \in (0, 1)$ be a lower bound for γ, typically 0.05, and define

$$P_j = \min\Bigl\{1, (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)\Bigr\}. \tag{2.3}$$

The extra correction factor, $1 - \log \gamma_{\min}$, ensures that the FWER remains controlled at level α despite of the adaptive search for the best quantile (see Sec. 3). For the recommended choice of $\gamma_{\min} = 0.05$, this factor is upper-bounded by 4; in fact, $1 - \log(0.05) \approx 3.996$.

We comment briefly on the relation between the proposed adjustment to the FDR (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001) or FWER (Holm 1979) controlling procedures. While we provide a family-wise error control and as such use union-bound corrections as done by Holm (1979), the definition of the adjusted p-value (2.3) and its graphical representation in Figure 1 are vaguely reminiscent of the FDR procedure, rejecting hypotheses if and only if the empirical distribution of p-values crosses a certain linear bound. The empirical distribution in (2.3) is taken for only one predictor variable, though, which is either in S or N. This corresponds to a multiple-testing situation in which we are testing a single hypothesis with multiple statistics. Figure 1 shows an example. Panel (a) presents a histogram of the adjusted p-values, $P_j^{(b)}$, for $b = 1, \ldots, B$, of the selected variable in the real data example in Section 4.3. The single-split method is equivalent to picking one of these p-values randomly and selecting the variable if this randomly chosen p-value is sufficiently small. To avoid this "p-value lottery," the multisplit method computes the empirical distribution of all p-values, $P_j^{(b)}$, for $b = 1, \ldots, B$, and rejects the null hypothesis $H_0 : \beta_j = 0$ (thus selecting variable j and including it into the model) if the empirical distribution crosses the broken line in Figure 1(b). A short derivation of the latter is as follows. Variable j is selected if and only if $P_j \le \alpha$, which occurs if and only if there exists some $\gamma \in (0.05, 1)$ such that

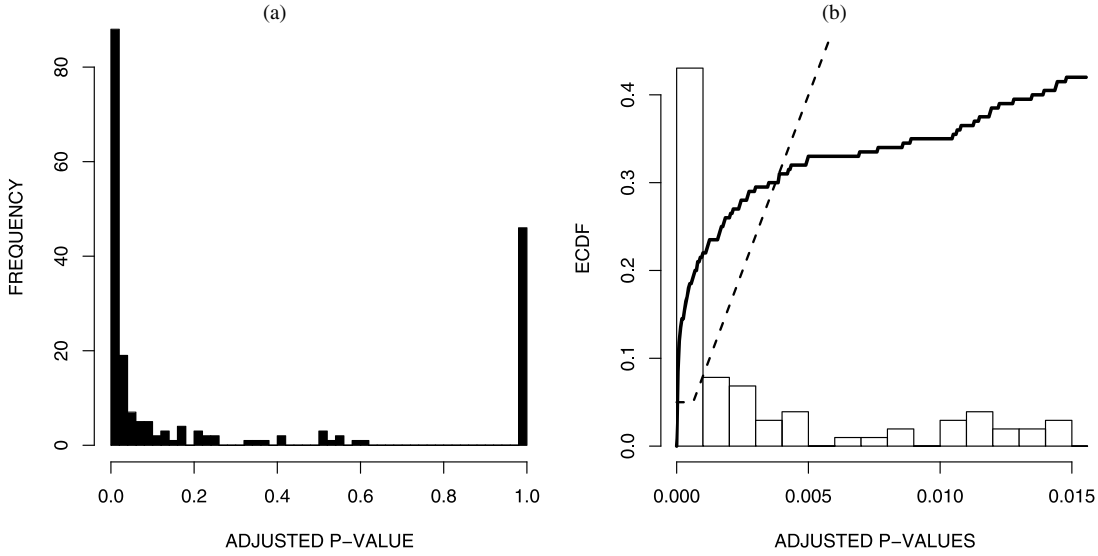

Figure 1. (a) A histogram of adjusted p-values, $P_j^{(b)}$, for the selected variable in the motif regression data example of Section 4.3. The single-split method randomly picks one of these p-values (a "p-value lottery") and rejects if it is below α. For the multisplit method, we reject if and only if the empirical distribution function of the adjusted p-values crosses the broken line [which is $f(p) = \max\{0.05, (3.996/\alpha)p\}$] for some $p \in (0, 1)$. This bound is shown as a broken line for $\alpha = 0.05$ in (b). For this example, the bound is indeed exceeded, and the variable is thus selected.

$Q_j(\gamma) \le \alpha/(1 - \log 0.05) \approx \alpha/3.996$. Equivalently, using definition (2.2), the γ-quantile of the adjusted p-values, $q_\gamma(P_j^{(b)})$, must be smaller than or equal to $\alpha\gamma/3.996$. This in turn is equivalent to the situation where the empirical distribution of the adjusted p-values, $P_j^{(b)}$, for $b = 1, \ldots, B$, is crossing above the bound $f(p) = \max\{0.05, (3.996/\alpha)p\}$ for some $p \in (0, 1)$. This bound is shown as a broken line in Figure 1(b).

The resulting adjusted p-values, $P_j, j = 1, \ldots, p$, can then be used for both FWER and FDR control. For FWER control at level $\alpha \in (0, 1)$, simply all p-values below α are rejected, and the selected subset is

$$\hat{S}_{multi} = \{j : P_j \le \alpha\}. \tag{2.4}$$

In Section 3.2 we show that indeed, asymptotically, $\mathbb{P}(V > 0) \le \alpha$, where $V = |\hat{S}_{multi} \cap N|$ is the number of falsely selected variables under the proposed selection (2.4). Besides better reproducibility and asymptotic family-wise error control, the multisplit version is, maybe unsurprisingly, more powerful than the single-split selection method.

2.3 False Discovery Rate Control With the Multisplit Method

Control of the FWER often is considered too conservative. If many rejections are made, Benjamini and Hochberg (1995) proposed instead controlling the expected proportion of false rejections—the FDR. Let $V = |\hat{S} \cap N|$ be the number of false rejections for a selection method $\hat{S}$ and let $R = |\hat{S}|$ be the total number of rejections. The FDR is defined as the expected proportion of false rejections,

$$\mathbb{E}(Q), \quad \text{with } Q = V/\max\{1, R\}. \tag{2.5}$$

For no rejections, $R = 0$, the denominator ensures that the false discovery proportion, Q, is 0, conforming with the definition of Benjamini and Hochberg (1995).

The original FDR controlling procedure of Benjamini and Hochberg (1995) first orders the observed p-values as $P_{(1)} \le P_{(2)} \le \cdots \le P_{(p)}$ and defines

$$k = \max\left\{i : P_{(i)} \le \frac{i}{p} q\right\}. \tag{2.6}$$

It then rejects all variables or hypotheses with the smallest k values, with no rejection made if the set in (2.6) is empty. FDR is controlled in this way at level q under the condition that all p-values are independent. Benjamini and Yekutieli (2001) showed that this procedure is conservative under a wider range of dependencies between p-values (see Blanchard and Roquain 2008 for related work). A great leap of faith would be required to assume any such assumption for our setting of high-dimensional regression, however. For general dependencies, Benjamini and Yekutieli (2001) showed that control is guaranteed at level $q \sum_{i=1}^{p} i^{-1} \approx q(1/2 + \log(p))$.

The standard FDR procedure is to work with the raw p-values, which are assumed to be uniformly distributed on $[0, 1]$ for true null hypotheses. The division by p in (2.6) is an effective correction for multiplicity. But the proposed multisplit method produces already adjusted p-values, as in (2.3). Because we are already working with multiplicity-corrected p-values, the division by p in (2.6) turns out to be superfluous. Instead, we can order the corrected p-values, $P_j, j = 1, \ldots, p$, in increasing order, $P_{(1)} \le P_{(2)} \le \cdots \le P_{(p)}$, and select the h variables with the smallest p-values, where

$$h = \max\{i : P_{(i)} \le iq\}. \tag{2.7}$$

The set of variables selected is denoted, with the value of h given in (2.7), by

$$\hat{S}_{multi;FDR} = \{j : P_j \le P_{(h)}\}, \tag{2.8}$$

with no rejections, $\hat{S}_{multi;FDR} = \emptyset$, if $P_{(i)} > iq$ for all $i = 1, \ldots, p$.

The procedure (2.8) will achieve FDR control at level $q\sum_{i=1}^{p} i^{-1} \approx q(1/2 + \log p)$. To get FDR control at level q, we replace q in (2.7) by $q/(\sum_{i=1}^{p} i^{-1})$, completely analogous to the standard FDR procedure under arbitrary dependence of the p-values of Benjamini and Yekutieli (2001). In the next section, we prove error control. Later, we empirically demonstrate the advantages of the proposed multisplit version over both the single-split and standard FDR controlling procedures, providing numerical results.

3. ERROR CONTROL AND CONSISTENCY

3.1 Assumptions

To achieve asymptotic error control, Wasserman and Roeder (2009) made a few assumptions about the crucial requirements for the variable selection procedure $\tilde{S}$:

(A1) *Screening property*: $\lim_{n\to\infty} \mathbb{P}[\tilde{S} \supseteq S] = 1$.
(A2) *Sparsity property*: $|\tilde{S}| < n/2$.

The *screening property* (A1) ensures that all relevant variables are retained. Irrelevant noise variables are allowed to be selected as well, as long as there are not too many, as required by the *sparsity property* (A2). A violation of the sparsity property would make it impossible to apply classical tests on the retained variables.

The Lasso (Tibshirani 1996) is an important example that satisfies (A1) and (A2) under appropriate conditions discussed by Meinshausen and Bühlmann (2006), Zhao and Yu (2006), van de Geer (2008), Meinshausen and Yu (2009), and Bickel, Ritov, and Tsybakov (2009). The adaptive Lasso (Zou 2006; Zhang and Huang 2008) also satisfies (A1) and (A2) under suitable conditions. Other examples include, assuming appropriate conditions, L_2 boosting (Friedman 2001; Bühlmann 2006), orthogonal matching pursuit (Tropp and Gilbert 2007), and sure independence screening (Fan and Lv 2008).

We typically use the Lasso (and extensions thereof) as a screening method. Other algorithms are possible as well. Wasserman and Roeder (2009) studied various scenarios under which these two properties are satisfied for the Lasso, depending on the choice of the regularization parameter. We refrain from repeating these and similar arguments, and operate on the assumption that we have a selection procedure, $\tilde{S}$, that satisfies both the *screening property* and the *sparsity property*.

3.2 Family-Wise Error Rate Control

We propose two versions of multiplicity-adjusted p-values: $Q_j(\gamma)$, as defined in (2.2), which relies on a choice of $\gamma \in (0, 1)$, and the adaptive version P_j defined in (2.3), which makes an adaptive choice of γ. We show that both quantities are multiplicity-adjusted p-values providing asymptotic FWER error control.

Theorem 3.1. Assume that (A1) and (A2) apply. Let $\alpha, \gamma \in (0, 1)$. If the null hypothesis $H_{0,j} : \beta_j = 0$ gets rejected whenever $Q_j(\gamma) \le \alpha$, then the FWER is asymptotically controlled at level α, that is,

$$\limsup_{n\to\infty} \mathbb{P}\Big[\min_{j\in N} Q_j(\gamma) \le \alpha\Big] \le \alpha,$$

where $\mathbb{P}$ is with respect to the data sample and the statement holds for any of the B random sample splits.

The proof is given in the Appendix.

Theorem 3.1 is valid for any predefined value of the quantile γ. However, the adjusted p-values, $Q_j(\gamma)$, involve the somehow arbitrary choice of γ, which could possibly pose a problem for practical applications. Thus we propose the adjusted p-values, P_j, that search for the optimal value of γ adaptively.

Theorem 3.2. Assume that (A1) and (A2) apply. Let $\alpha \in (0, 1)$. If the null hypothesis $H_{0,j} : \beta_j = 0$ is rejected whenever $P_j \le \alpha$, then the FWER is asymptotically controlled at level α, that is,

$$\limsup_{n\to\infty} \mathbb{P}\Big[\min_{j\in N} P_j \le \alpha\Big] \le \alpha,$$

where the probability $\mathbb{P}$ is as in Theorem 3.1.

The proof is given in the Appendix.

A brief remark regarding the asymptotic nature of the results seems to be in order. The proposed error control relies on all truly important variables being selected in the screening stage with very high probability. This is our *screening property* (A1). Let $\mathcal{A}$ be the event $S \subseteq \tilde{S}$. The results for the example in Theorem 3.2 can be formulated in a nonasymptotic way as $\mathbb{P}[\mathcal{A} \cap \{\min_{j\in N} P_j \le \alpha\}] \le \alpha$, and $P(\mathcal{A}) \to 1$, typically exponentially fast, for $n \to \infty$. Analogous remarks apply to Theorems 3.1 and 3.3.

3.3 False Discovery Rate Control

The adjusted p-values can be used for FDR control, as laid out in Section 2.3. The set of selected variables, $\hat{S}_{multi;FDR}$, was defined in (2.8). Here we show that FDR is indeed controlled at the desired rate with this procedure.

Theorem 3.3. Assume that (A1) and (A2) apply. Let $\tilde{q} > 0$ and $\hat{S}_{multi;FDR}$ be the set of selected variables, as defined in (2.8), with a cutoff value of $q = \tilde{q}/\sum_{i=1}^{p} i^{-1}$ in (2.7). Let $V = |\hat{S}_{multi;FDR} \cap N|$ and $R = |\hat{S}_{multi;FDR}|$. The FDR (2.5) with $Q = V/\max\{1, R\}$ is then asymptotically controlled at level $\tilde{q}$, that is,

$$\limsup_{n\to\infty} \mathbb{E}(Q) \le \tilde{q}.$$

The proof is given in the Appendix.

As with FWER control, we could use, for any fixed value of γ, the values $Q_j(\gamma), j = 1, \ldots, p$ instead of $P_j, j = 1, \ldots, n$. We refrain from giving the full details here, because in our experience, the foregoing adaptive version works reliably and does not require an a priori choice of the quantile γ that is necessary otherwise.

3.4 Model Selection Consistency

If we let level $\alpha = \alpha_n \to 0$ for $n \to \infty$, then the probability of falsely including a noise variable vanishes because of the preceding results. To get the property of consistent model selection, we must analyze the asymptotic behavior of the power. It turns out that this property is inherited from the single-split method.

Corollary 3.1. Let $\hat{S}_{single}$ be the selected model of the single-split method. Assume that $\alpha_n \to 0$ can be chosen for $n \to \infty$ at a rate such that $\lim_{n\to\infty} \mathbb{P}[\hat{S}_{single} = S] = 1$. Then, for any

$\gamma_{\min}$ [see (2.3)], the multisplit method is also model selection–consistent for a suitable sequence α_n; that is, for $\hat{S}_{multi} = \{j \in \tilde{S}; P_j \le \alpha_n\}$, it holds that

$$\lim_{n\to\infty} \mathbb{P}[\hat{S}_{multi} = S] = 1.$$

Wasserman and Roeder (2009) discussed conditions that ensure that $\lim_{n\to\infty} \mathbb{P}[\hat{S}_{single} = S] = 1$ for various variable selection methods, such as the Lasso or some forward variable selection scheme.

The reverse of Corollary 3.1 is not necessarily true. The multisplit method can be consistent if the single-split method is not. A necessary condition for consistency of the single-split method is $\limsup_{n\to\infty} \mathbb{P}[P_j^{(b)} \le \alpha] = 1$ for all $j \in S$, where the probability is with respect to both the data and the random split-point, because otherwise there is a positive probability that variable j will not be selected with the single-split approach. For the multisplit method, on the other hand, we need only a bound on quantiles of $P_j^{(b)}$ over $b = 1, \ldots, B$. We refrain from going into more detail here and instead show, with numerical results, that the multisplit method is indeed more powerful than the single-split analog. We also remark that the Bonferroni correction in (2.1), multiplying the raw p-values by the number, $|\tilde{S}^{(b)}|$, of selected variables, possibly could be improved using ideas of Hothorn, Bretz, and Westfall (2008), further increasing the power of the procedure.

4. NUMERICAL RESULTS

In this section we compare the empirical performance of the different estimators on simulated and real data sets. Simulated data allow a thorough evaluation of the model selection properties. The real data set demonstrates that we can find signals in data with our proposed method that would not be picked up by the single-split method. We use a default value of $\alpha = 0.05$ everywhere.

4.1 Simulations

We use the following simulation settings:

(A) Simulated data set with $n = 100$, $p = 100$, and a Toeplitz design matrix coming from a centered multivariate normal distribution with covariance $\rho^{|j-k|}$ between variables j and k, with $\rho = 0.5$.
(B) As in (A), but with $n = 100$ and $p = 1000$.
(C) Real data set with $n = 71$ and $p = 4088$ for the design matrix $\mathbf{X}$ and artificial response $\mathbf{Y}$.

The data set in (C) is from gene expression measurements in *Bacillus subtilis*. The $p = 4088$ predictor variables are log-transformed gene expressions, and there is a response measuring the logarithm of the production rate of riboflavin in *B. subtilis*. The data were kindly provided by DSM Nutritional Products, Switzerland. Because the true variables are not known, we consider a linear model with design matrix from real data and simulate a sparse parameter vector $\boldsymbol{\beta}$ as follows. In each simulation run, a new parameter vector $\boldsymbol{\beta}$ is created by either "uniform" or "varying-strength" sampling. Under uniform sampling, $|S|$ randomly chosen components of $\boldsymbol{\beta}$ are set to 1, and the remaining $p - |S|$ components are set to 0. Under varying-strength sampling, $|S|$ randomly chosen components of $\boldsymbol{\beta}$ are set to values $1, \ldots, |S|$. The error variance σ^2 is adjusted such that the signal-to-noise ratio (SNR) is maintained at a desired level at each simulation run. We perform 50 simulations for each setting.

The sample-splitting is done such that the model is trained on a data set of size $\lfloor (n-1)/2 \rfloor$, and the p-values are calculated on the remaining data set. This slightly unbalanced scheme precludes situations where the full model might be selected on the first data set. Calculation of p-values would not be possible on the remaining data in such a situation. We use a total of $B = 50$ sample splits for each simulation run. Following Wasserman and Roeder (2009), we compute p-values for all procedures using a normal approximation. The results are qualitatively similar when using a t distribution instead.

We compare the average number of true positives and the FWER for the single-split and multisplit methods for the three simulation settings (A)–(C), using SNRs of 0.25, 1, 4, and 16 (corresponding to population R^2 values of 0.2, 0.5, 0.8, and 0.94, respectively). The number of relevant variables, $|S|$, is either 5 or 10. As the initial variable selection or screening method, $\tilde{S}$, we use three approaches, all based on the Lasso (Tibshirani 1996). The first approach, denoted by $\tilde{S}_{fixed}$, uses the Lasso and selects those $\lfloor n/6 \rfloor$ variables that appear most often in the regularization path when varying the penalty parameter. The constant number of $\lfloor n/6 \rfloor$ variables is chosen, somewhat arbitrarily, to ensure a reasonably large set of selected coefficients on the one hand and on the other hand, to ensure that least squares estimation will work reasonably well on the second half of the data with sample size $\lfloor n/2 \rfloor$. While the choice seems to work well in practice and can be implemented very easily and efficiently, it is still slightly arbitrary. Avoiding any such choices of non–data-adaptive tuning parameters, the second method, $\tilde{S}_{cv}$, uses the Lasso with penalty parameter chosen by 10-fold cross-validation, selecting the variables whose corresponding estimated regression coefficients are different than 0. The third method, $\tilde{S}_{adap}$, is the adaptive Lasso of Zou (2006), in which regularization parameters are chosen based on 10-fold cross-validation, with the Lasso solution used as the initial estimator for the adaptive Lasso. The selected variables are again those whose corresponding estimated regression parameters are different than 0.

Figures 2 and 3 show results for both the single-split and multisplit methods with the default setting $\gamma_{\min} = 0.05$. Using the multisplit method, the average number of true positives (i.e., the variables in S which are selected) typically is slightly increased, while the FWER (i.e., the probability of including variables in N) is reduced sharply. The single-split method often has a FWER above the level $\alpha = 0.05$ at which it is asymptotically controlled, while for the multisplit method, the FWER is above the nominal level in only a few scenarios. The asymptotic control seems to give a good control in finite-sample settings with the multisplit method, possibly apart from the method $\tilde{S}_{fixed}$ on the very high-dimensional data set (C). The single-split method, in contrast, selects too many noise variables, exceeding the desired FWER sometimes substantially, in nearly all settings. This suggests that the asymptotic error control seems to work better for finite sample sizes for the multisplit method. Even though the multisplit method is more conservative than the single-split method (having a substantially lower FWER), the number of

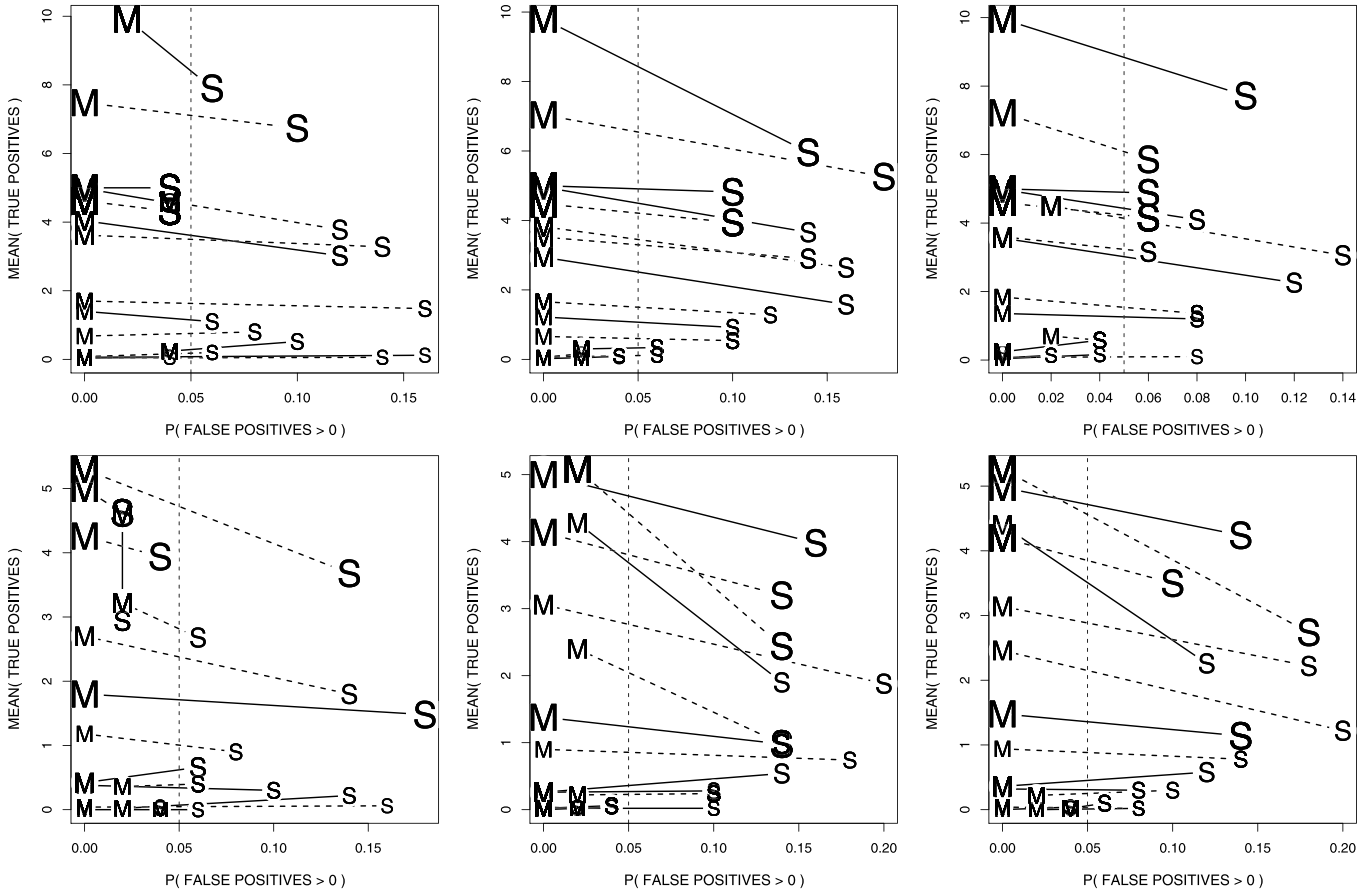

Figure 2. Simulation results for setting (A) in the top and (B) in the bottom row. Average number of true positives vs. the family-wise error rate (FWER) for the single split method ('S') against the multi-split version ('M'). FWER is controlled (asymptotically) at $\alpha = 0.05$ for both methods and this value is indicated by a broken vertical line. From left to right are results for $\tilde{S}_{fixed}$, $\tilde{S}_{cv}$ and $\tilde{S}_{adap}$. Results of a unique setting of SNR, sparsity and design are joined by a line, which is solid if the coefficients follow the 'uniform' sampling and broken otherwise. Increasing SNR is indicated by increasing symbol size.

true discoveries often is increased. We note that for data (C), with $p = 4088$, and in general for low SNRs, the number of true positives is low, because we control the very stringent family-wise error criterion at a significance level of $\alpha = 0.05$. As an alternative, controlling less conservative error measures is possible, as discussed in Section 5.

4.2 Comparisons With the Adaptive Lasso

Here we compare the multisplit selector with the adaptive Lasso (Zou 2006). We have used the adaptive Lasso as a variable selection method in our proposed multisplit method. Usually, the adaptive Lasso is used by itself. A few choices must

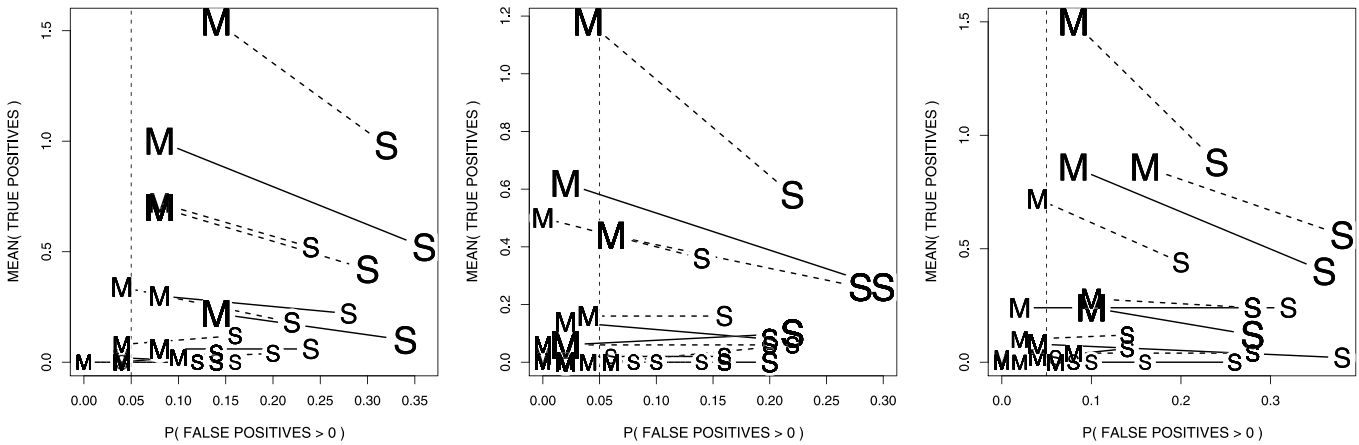

Figure 3. Results of simulation setup (C).

Table 1. Comparison of the multisplit method with CV-Lasso selection, $\tilde{S}_{adap}$, and the selection made using the adaptive Lasso and a CV choice of the involved penalty parameters for a setting with $n = 100$ and $p = 200$

			E(true positives)		E(false positives)		P(false positives > 0)	
Uniform sampling	$\|S\|$	SNR	Multisplit	Adaptive Lasso	Multisplit	Adaptive Lasso	Multisplit	Adaptive Lasso
NO	10	0.25	0.00	2.30	0	9.78	0	0.76
NO	10	1	0.58	6.32	0	20.00	0	1
NO	10	4	4.14	8.30	0	25.58	0	1
NO	10	16	7.20	9.42	0.02	30.10	0.02	1
YES	10	0.25	0.02	2.52	0	10.30	0	0.72
YES	10	1	0.10	7.46	0.02	21.70	0.02	1
YES	10	4	2.14	9.96	0	28.46	0	1
YES	10	16	9.92	10.00	0.04	30.66	0.04	1
NO	5	0.25	0.06	1.94	0	11.58	0	0.84
NO	5	1	1.50	3.86	0.02	19.86	0.02	1
NO	5	4	3.52	4.58	0.02	23.56	0.02	1
NO	5	16	4.40	4.98	0	27.26	0	1
YES	5	0.25	0.02	2.22	0	12.16	0	0.8
YES	5	1	0.82	4.64	0.02	22.18	0.02	1
YES	5	4	4.90	5.00	0	24.48	0	1
YES	5	16	5.00	5.00	0	28.06	0	1

be made when using the adaptive Lasso; we make the same choices as previously. The initial estimator is obtained as the Lasso solution with a 10-fold cross-validation (CV) choice of the penalty parameter. The adaptive Lasso penalty is also obtained by 10-fold CV.

Despite desirable asymptotic consistency properties (Huang, Ma, and Zhang 2008), the adaptive Lasso does not offer error control in the same way as Theorem 3.1 does for the multisplit method. In fact, the FWER (i.e., the probability of selecting at least one noise variable) is very close to 1 with the adaptive Lasso in all of the simulations that we have seen. In contrast, our multisplit method offers asymptotic control, which was very well matched by the empirical FWER in the vicinity of $\alpha = 0.05$. Table 1 compares the simulation results for the multisplit method using $\tilde{S}_{adap}$ and the adaptive Lasso by itself for a simulation setting with $n = 100$, $p = 200$, and the same settings as in (A) and (B) otherwise. The adaptive Lasso selects roughly 20 noise variables (out of $p = 200$ variables), even though the number of truly relevant variables is just 5 or 10. The average number of false positives is at most 0.04 and often simply 0 with the proposed multisplit method.

There is clearly a price to pay for controlling the FWER. Our proposed multisplit method detects fewer truly relevant variables than the adaptive Lasso on average. The difference is most pronounced for very low SNRs. The multisplit method generally selects neither correct nor incorrect variables for SNR $=$ 0.25, while the adaptive Lasso averages between 2 and 3 correct selections, among 9–12 wrong selections. Depending on the objectives of the study, either outcome is preferred. For larger SNRs, the multisplit method detects almost as many truly important variables as the adaptive Lasso, while still reducing the number of falsely selected variables from 20 or more to roughly 0.

The multisplit method seems to be beneficial in settings where the cost of making an erroneous selection is rather high. For example, expensive follow-up experiments are usually required to validate results in biomedical applications, and stricter error control will channel more of the available resources into experiments more likely to be successful.

4.3 Motif Regression

We apply the multisplit method to a real data set related to motif regression (Conlon et al. 2003). For a total of $n = 287$ DNA segments, we have the binding intensity of a protein to each of the segments. These are our response values, $Y_1, \ldots, Y_n$. Moreover, for $p = 195$ candidate words ("motifs"), we have scores, x_{ij}, that measure how well the jth motif is represented in the ith DNA sequence. The motifs are typically 5- to 15-bp-long candidates for the true binding site of the protein. The hope is that the true binding site is included in the list of significant variables with the strongest relationship between motif score and binding intensity. Using a linear model with $\tilde{S}_{adap}$, the multisplit method identifies one predictor variable at the 5% significance level. In contrast, the single-split method cannot identify a single significant predictor. In view of the asymptotic error control and the empirical results in Section 4, there is substantial evidence indicating that the selected variable corresponds to a true binding site. For this specific application, it seems desirable to pursue a conservative approach with low FWER. As mentioned earlier, we could control other, less conservative error measures, as discussed in Section 5.

4.4 Comparison With Standard Low-Dimensional False Discovery Rate Control

We mentioned that control of FDR can be an attractive alternative to FWER if a sizeable number of rejections is expected. Using the corrected p-values $P_1, \ldots, P_p$, a simple FDR-controlling procedure was derived in Section 2.3, and its asymptotic control of FDR was shown in Theorem 3.3. We now empirically evaluate the behavior of the resulting method and its power to detect truly interesting variables, using the standard Lasso with CV in the initial screening step. Turning again to the simulation setting (A), we vary the sample size n, the number

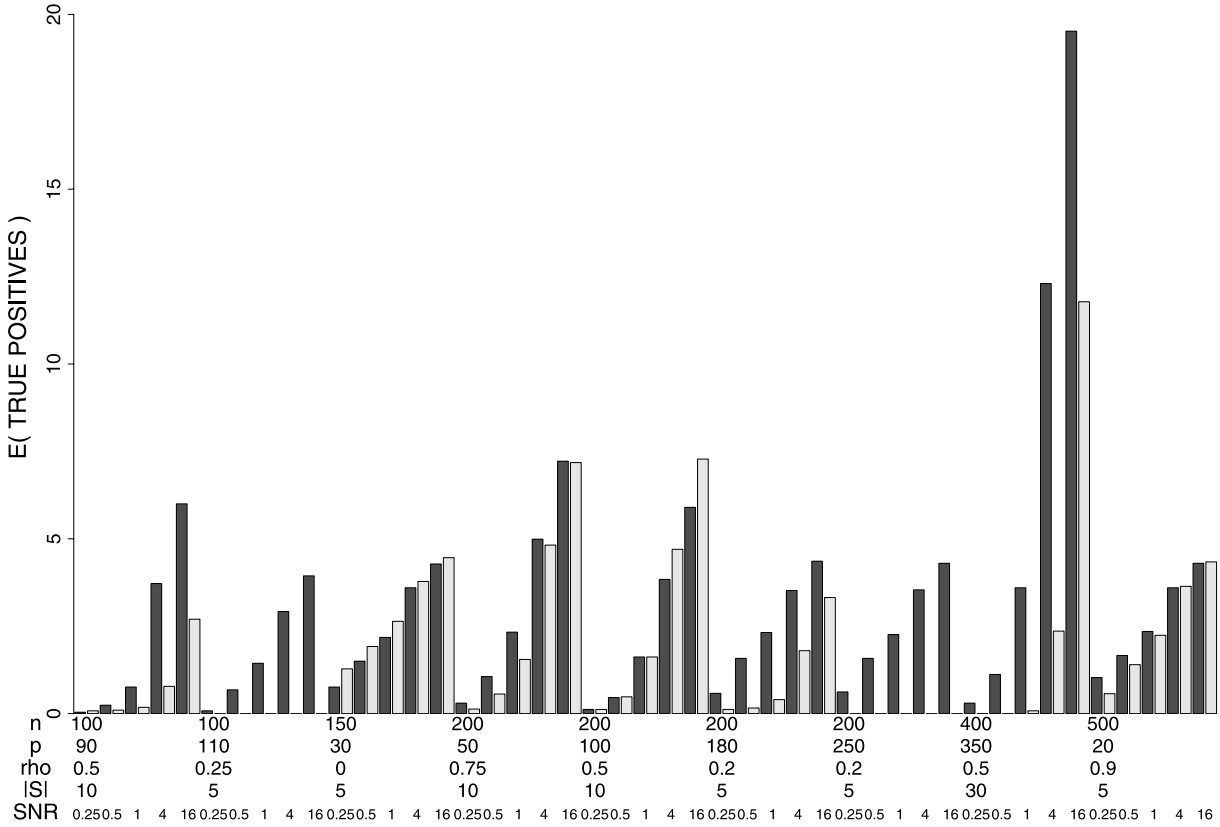

Figure 4. Results of FDR controlling simulations for the multisplit method (dark bar) and standard FDR control (light bar). The settings of $n, p, \rho, |S|$, and SNR are given below each simulation. The height of the bars corresponds to the average number of selected important variables. For $p > n$, the standard method breaks down, and the corresponding bars are set to height 0.

of variables p, the SNR, the correlation between neighboring variables ρ, and the number of truly interesting variables s.

We previously demonstrated that the multisplit method is preferable to the single-split method. Here we are more interested in a comparison with well-understood traditional FDR-controlling procedures. For $p < n$, the standard approach is to compute the least squares estimator once for the full data set. For each variable, a p-value is obtained, and the FDR-controlling procedure as in (2.6) can be applied. This approach obviously breaks down for $p > n$. Our proposed approach can be applied to both low-dimensional ($p < n$) and high-dimensional ($p \geq n$) settings.

In all settings, the empirical FDR of our method (not shown) is often close to 0 and always below the controlled value of $q = 0.05$ (where the correction factor, $\sum_{i=1}^{p} i^{-1}$, has already been taken into account). Results for power are shown in Figure 4 for control at $q = 0.05$.

Possibly unexpectedly, the multisplit method tracks the power of the standard FDR controlling procedure quite closely for low-dimensional data with $p < n$. In fact, the multi-split method is doing considerably better if n/p is below, say, 1.5 or the correlation among the tests is large. An intuitive explanation for this behavior is that, as p approaches n, the variance in each estimated coefficient vector under the ordinary least squares (OLS) estimate is increasing substantially. This in turn increases the variance of all OLS components $\hat{\beta}_j$, $j = 1, \ldots, p$, and diminishes the ability to select the truly important variables. The multisplit method, in contrast, trims the total number of variables to a substantially smaller number in one half of the samples and then suffers less from increased variance in the estimated coefficients in the second half of the samples. Repeating this over multiple splits thus leads to a surprisingly powerful variable selection procedure even for low-dimensional data. Nevertheless, we believe that the main application will be in high-dimensional data, for which the standard approach breaks down completely.

5. EXTENSIONS

Because of the generic nature of our proposed methodology, extensions to any situation where (asymptotically valid) p-values, $\tilde{P}_j$, for hypotheses $H_{0,j}$ ($j = 1, \ldots, p$) are available are straightforward. An important class of examples comprises generalized linear models (GLMs), or Gaussian graphical models. The dimension-reduction step typically involves some form of shrinkage estimation. An example for Gaussian graphical models is the recently proposed "graphical Lasso" (Friedman, Hastie, and Tibshirani 2008). The second step relies on classical tests (e.g., likelihood ratio) applied to the selected submodel, analogous to the proposed methodology for linear regression.

In some settings, control of FWER at, say, $\alpha = 0.05$ is too conservative. One can either resort to controlling FDR, as alluded to earlier, or adjust FWER control to control the expected number of false rejections. As an example, consider the adjusted p-value P_j defined in (2.3). Variable j is rejected if and

only if $P_j \le \alpha$. [In what follows, assume that adjusted p-values, as defined in (2.1), are not capped at 1. This is a technical detail only; it does not modify the proposed FWER-controlling procedure.] Variable j is rejected if and only if $P_j \le \alpha$ controls FWER at level α. Alternatively, one can reject variables if and only if $P_j/K \le \alpha$, where $K > 1$ is a correction factor. Call the number of falsely rejected variables V, and calculate it as

$$V = \sum_{j \in N} 1\{P_j/K \le \alpha\}.$$

Then the expected number of false positives is controlled at level $\limsup_{n\to\infty} \mathbb{E}[V] \le \alpha K$. A proof of this result follows directly from the proof of Theorem 3.2. Of course, we can equivalently set $k = \alpha K$ and obtain a control, $\limsup_{n\to\infty} \mathbb{E}[V] \le k$. For example, setting $k = 1$ offers a much less conservative error control compared with controlling the FWER, if this is desired.

6. DISCUSSION

We have proposed a multisplit method for assigning statistical significance and constructing conservative p-values for hypothesis testing for high-dimensional problems where the number of predictor variables may be much larger than sample size. Our method is an extension of the single-split approach of Wasserman and Roeder (2009) and is extended to FDR control. Combining the results of multiple data splits, based on quantiles as summary statistics, improves reproducibility compared with the single-split method. The multisplit and single-split methods share the properties of asymptotic error control and model selection consistency. We argue empirically that the multisplit method usually selects much fewer false positives than the single-split method, with a slightly higher number of true positives. The main area of application will be high-dimensional data, where the number p of predictor variables exceeds sample size n, because standard approaches rely on least squares estimation and thus fail in this setting. We have shown that the multisplit method is also an interesting alternative to standard FDR and FWER control in lower-dimensional settings, because the proposed FDR control can be more powerful if p is reasonably large but smaller than sample size n. The method is very generic and can be used in a broad spectrum of error-controlling procedures in multiple testing, including linear models and GLMs.

APPENDIX: PROOFS

Proof of Theorem 3.1

For technical reasons, we define

$$K_j^{(b)} = P_j^{(b)} 1\{S \subseteq \tilde{S}^{(b)}\} + 1\{S \not\subseteq \tilde{S}^{(b)}\}, \tag{A.1}$$

where $K_j^{(b)}$ are the adjusted p-values if the estimated active set contains the true active set. Otherwise, all p-values are set to 1. Because of assumption (A1), for fixed B, $\mathbb{P}[K_j^{(b)} = P_j^{(b)}$ for all $b = 1, \ldots, B]$ on a set A_n with $\mathbb{P}[A_n] \to 1$. Thus we can define all of the quantities involving $P_j^{(b)}$ also with $K_j^{(b)}$, and under this slightly altered procedure, it is sufficient to show that

$$\mathbb{P}\Big[\min_{j \in N} Q_j(\gamma) \le \alpha\Big] \le \alpha.$$

In particular, here we can omit the limes superior.

For the proofs, we also omit the function $\min\{1, \cdot\}$ from the definitions of $Q_j(\gamma)$ and P_j in (2.2) and (2.3). The selected sets of variables are clearly unaffected, and the notation is simplified considerably.

Define for $u \in (0, 1)$ the quantity $\pi_j(u)$ as the fraction of bootstrap samples that yield $Kj^{(b)}$ less than or equal to u,

$$\pi_j(u) = \frac{1}{B} \sum_{b=1}^{B} 1\{K_j^{(b)} \le u\}.$$

Note that the events $\{Q_j(\gamma) \le \alpha\}$ and $\{\pi_j(\alpha\gamma) \ge \gamma\}$ are equivalent. Thus

$$\begin{aligned}\mathbb{P}\Big[\min_{j \in N} Q_j(\gamma) \le \alpha\Big] &\le \sum_{j \in N} \mathbb{E}\big[1\{Q_j(\gamma) \le \alpha\}\big] \\ &= \sum_{j \in N} \mathbb{E}\big[1\{\pi_j(\alpha\gamma) \ge \gamma\}\big]. \end{aligned} \tag{A.2}$$

Using a Markov inequality,

$$\sum_{j \in N} \mathbb{E}\big[1\{\pi_j(\alpha\gamma) \ge \gamma\}\big] \le \frac{1}{\gamma} \sum_{j \in N} \mathbb{E}[\pi_j(\alpha\gamma)].$$

By the definition of $\pi_j(\cdot)$,

$$\frac{1}{\gamma} \sum_{j \in N} \mathbb{E}[\pi_j(\alpha\gamma)] = \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \sum_{j \in N \cap \tilde{S}^{(b)}} \mathbb{E}\big[1\{K_j^{(b)} \le \alpha\gamma\}\big].$$

Moreover, using the definition of $K_j^{(b)}$ in (A.1),

$$\mathbb{E}\big[1\{K_j^{(b)} \le \alpha\gamma\}\big] \le \mathbb{P}\big[P_j^{(b)} \le \alpha\gamma \,|\, S \subseteq \tilde{S}^{(b)}\big] = \frac{\alpha\gamma}{|\tilde{S}^{(b)}|}.$$

This is a consequence of the uniform distribution of $\tilde{P}_j^{(b)}$ given $S \subseteq \tilde{S}^{(b)}$. Summarizing these results, we get

$$\mathbb{P}\Big[\min_{j \in N} Q_j(\gamma) \le \alpha\Big] \le \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}\Big[\sum_{j \in N \cap \tilde{S}^{(b)}} \frac{\alpha\gamma}{|\tilde{S}^{(b)}|}\Big] \le \alpha,$$

which completes the proof.

Proof of Theorem 3.2

As in the proof of Theorem 3.1, here we work with $K_j^{(b)}$ instead of $P_j^{(b)}$. Analogously, instead of $\tilde{P}_j^{(b)}$, we work with $\tilde{K}_j^{(b)}$. For any $\tilde{K}_j^{(b)}$ with $j \in N$ and $\alpha \in (0, 1)$,

$$\mathbb{E}\Big[\frac{1\{\tilde{K}_j^{(b)} \le \alpha\gamma\}}{\gamma}\Big] \le \alpha. \tag{A.3}$$

Furthermore,

$$\begin{aligned}\mathbb{E}\Big[\max_{j \in N} \frac{1\{K_j^{(b)} \le \alpha\gamma\}}{\gamma}\Big] &\le \mathbb{E}\Big[\sum_{j \in N} \frac{1\{K_j^{(b)} \le \alpha\gamma\}}{\gamma}\Big] \\ &\le \mathbb{E}\Big[\sum_{j \in N \cap \tilde{S}^{(b)}} \frac{1\{K_j^{(b)} \le \alpha\gamma\}}{\gamma}\Big]\end{aligned}$$

and thus, with (A.3) and using the definition (A.1) of $K_j^{(b)}$,

$$\mathbb{E}\Big[\max_{j \in N} \frac{1\{K_j^{(b)} \le \alpha\gamma\}}{\gamma}\Big] \le \mathbb{E}\Big[\sum_{j \in N \cap \tilde{S}^{(b)}} \frac{\alpha}{|\tilde{S}^{(b)}|}\Big] \le \alpha. \tag{A.4}$$

For a random variable U taking values in $[0, 1]$,

$$\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\{U \le \alpha\gamma\}}{\gamma} = \begin{cases} 0 & U \ge \alpha \\ \alpha/U & \alpha\gamma_{\min} \le U < \alpha \\ 1/\gamma_{\min} & U < \alpha\gamma_{\min}. \end{cases}$$

Moreover, if U has a uniform distribution on [0, 1], then

$$\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\{U\le\alpha\gamma\}}{\gamma}\right]=\int_0^{\alpha\gamma_{\min}}\gamma_{\min}^{-1}\,dx+\int_{\alpha\gamma_{\min}}^{\alpha}\alpha x^{-1}\,dx$$
$$=\alpha(1-\log\gamma_{\min}).$$

Thus, using the fact that $\tilde{K}_j^{(b)}$ has a uniform distribution on [0, 1] for all $j\in N$, conditional on $S\subseteq\tilde{S}^{(b)}$,

$$\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\{\tilde{K}_j^{(b)}\le\alpha\gamma\}}{\gamma}\right]\le\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\{\tilde{K}_j^{(b)}\le\alpha\gamma\}}{\gamma}\Big|S\subseteq\tilde{S}^{(b)}\right]$$
$$=\alpha(1-\log\gamma_{\min}).$$

Analogously to (A.4), we then can deduce that

$$\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\{K_j^{(b)}\le\alpha\gamma\}}{\gamma}\right]\le\alpha(1-\log\gamma_{\min}).$$

Averaging over all bootstrap samples yields

$$\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{(1/B)\sum_{b=1}^{B}1\{K_j^{(b)}/\gamma\le\alpha\}}{\gamma}\right]\le\alpha(1-\log\gamma_{\min}).$$

Again using a Markov inequality,

$$\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}1\{\pi_j(\alpha\gamma)\ge\gamma\}\right]\le\alpha(1-\log\gamma_{\min}),$$

where $\pi_j(\cdot)$ is defined as in the proof of Theorem 3.1.

Because the events $\{Q_j(\gamma)\le\alpha\}$ and $\{\pi_j(\alpha\gamma)\ge\gamma\}$ are equivalent, it follows that

$$\sum_{j\in N}\mathbb{P}\left[\inf_{\gamma\in(\gamma_{\min},1)}Q_j(\gamma)\le\alpha\right]\le\alpha(1-\log\gamma_{\min}),$$

which implies that

$$\sum_{j\in N}\mathbb{P}\left[\inf_{\gamma\in(\gamma_{\min},1)}Q_j(\gamma)(1-\log\gamma_{\min})\le\alpha\right]\le\alpha.$$

Using the definition of P_j in (2.3),

$$\sum_{j\in N}\mathbb{P}[P_j\le\alpha]\le\alpha,\tag{A.5}$$

and thus, by the union bound,

$$\mathbb{P}[\min_{j\in N}P_j\le\alpha]\le\alpha,$$

which completes the proof.

Proof of Theorem 3.3

As in the proofs of Theorems 3.1 and 3.2, we implicitly use a correction as in (A.1) for all p-values. Otherwise, our notation is identical to that in the proof of theorem 1.3 of Benjamini and Yekutieli (2001). An exception is our use of the value q instead of q/m in the FDR-controlling procedure, because we are working with adjusted p-values. Let

$$p_{ijk}=\mathbb{P}\big(\{P_i\in[(j-1)q,jq]\}\text{ and }C_k^{(i)}\big),$$

where $C_k^{(i)}$ is the event that if variable i were rejected, then $k-1$ other variables were rejected as well. Now, as shown in eq. (10) as well as in eq. (28) of Benjamini and Yekutieli (2001),

$$\mathbb{E}(Q)=\sum_{i\in N}\sum_{k=1}^{p}\frac{1}{k}\sum_{j=1}^{k}p_{ijk}.$$

With this result, we use a argument similar to that of to Benjamini and Yekutieli (2001),

$$\mathbb{E}(Q)=\sum_{i\in N}\sum_{k=1}^{p}\frac{1}{k}\sum_{j=1}^{k}p_{ijk}=\sum_{i\in N}\sum_{j=1}^{p}\sum_{k=j}^{p}\frac{1}{k}p_{ijk}$$
$$\le\sum_{i\in N}\sum_{j=1}^{p}\sum_{k=j}^{p}\frac{1}{j}p_{ijk}\le\sum_{i\in N}\sum_{j=1}^{p}\frac{1}{j}\sum_{k=1}^{p}p_{ijk}$$
$$=\sum_{j=1}^{p}\frac{1}{j}\sum_{i\in N}\sum_{k=1}^{p}p_{ijk}.\tag{A.6}$$

We denote

$$f(j):=\sum_{i\in N}\sum_{k=1}^{p}p_{ijk},\qquad j=1,\dots,p.$$

Equation (A.6) can then be rewritten as

$$\mathbb{E}(Q)\le\sum_{j=1}^{p}\frac{1}{j}f(j)=f(1)+\sum_{j=2}^{p}\frac{1}{j}\left(\sum_{j'=1}^{j}f(j')-\sum_{j'=1}^{j-1}f(j')\right)\tag{A.7}$$
$$=\sum_{j=1}^{p-1}\left(\frac{1}{j}-\frac{1}{j+1}\right)\sum_{j'=1}^{j}f(j')+\frac{1}{p}\sum_{j'=1}^{p}f(j').\tag{A.8}$$

Note that, analogously to eq. (27) of Benjamini and Yekutieli (2001),

$$\sum_{k=1}^{p}p_{ijk}=P\left(\left\{P_i\in[(j-1)q,jq]\right\}\cap\left(\bigcup_{k}^{p}C_k^{(i)}\right)\right)$$
$$=P\big(P_i\in[(j-1)q,jq]\big)$$

and thus

$$f(j)=\sum_{i\in N}\sum_{k=1}^{p}p_{ijk}=\sum_{i\in N}P\big(P_i\in[(j-1)q,jq]\big),$$

from which it follows by (A.5) in the proof of Theorem 3.2 that

$$\sum_{j'=1}^{j}f(j')=\sum_{i\in N}P(P_i\le jq)\le jq.$$

Using this in (A.8), we obtain

$$\mathbb{E}(Q)\le\sum_{j=1}^{p-1}\left(\frac{1}{j}-\frac{1}{j+1}\right)jq+\frac{1}{p}pq$$
$$=\left(\sum_{j=1}^{p-1}\frac{1}{j(j+1)}j+1\right)q=q\sum_{j=1}^{p}\frac{1}{j}=\tilde{q},\tag{A.9}$$

which completes the proof.

Proof of Corollary 3.1

Because the single-split method is model selection–consistent, it must hold that $\mathbb{P}[\max_{j\in S}\tilde{P}_j|\tilde{S}|\le\alpha_n]\to1$ for $n\to\infty$. Using multiple data splits, this property holds for each of the B splits, and thus $\mathbb{P}[\max_{j\in S}\max_b\tilde{P}_j^{(b)}|\tilde{S}^{(b)}|\le\alpha_n]\to1$, implying that, with probability converging to 1 for $n\to\infty$, the quantile $\max_{j\in S}Q_j(1)$ is bounded from above by α_n. The maximum over all $j\in S$ of the adjusted p-values, $P_j=(1-\log\gamma_{\min})\inf_{\gamma\in(\gamma_{\min},1)}Q_j(\gamma)$, is thus bounded from above by $(1-\log\gamma_{\min})\alpha_n$, again with probability converging to 1 for $n\to\infty$.

[Received November 2008. Revised July 2009.]

REFERENCES

Bach, F. (2008), "Bolasso: Model Consistent Lasso Estimation Through the Bootstrap," in *ICML '08: Proceedings of the 25th International Conference on Machine Learning*, New York: ACM, pp. 33–40.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300.

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

Blanchard, G., and Roquain, E. (2008), "Two Simple Sufficient Conditions for FDR Control," *Electronic Journal of Statistics*, 2, 963–992.

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34, 559–583.

Conlon, E., Liu, X., Lieb, J., and Liu, J. (2003), "Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis," *Proceedings of the National Academy of Science*, 100, 3339–3344.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911.

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432.

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Hothorn, T., Bretz, F., and Westfall, P. (2008), "Simultaneous Inference in General Parametric Models," *Biometrical Journal*, 50, 346–363.

Huang, J., Ma, S., and Zhang, C.-H. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics and Data Analysis*, 52, 374–393.

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462.

——— (2008), "Stability Selection," preprint, University of Oxford.

Meinshausen, N., and Yu, B. (2009), "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," *The Annals of Statistics*, 37, 246–270.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Tropp, J., and Gilbert, A. (2007), "Signal Recovery From Random Measurements via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, 53 (12), 4655–4666.

van de Geer, S. (2008), "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36, 614–645.

Wasserman, L., and Roeder, K. (2009), "High Dimensional Variable Selection," *The Annals of Statistics*, 37, 2178–2201.

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.