P-Values for High-Dimensional Regression

Nicolai Meinshausen * † Lukas Meier * ‡ Peter Bühlmann ‡

November 13, 2008

Abstract

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 (2008) which 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. It involves, however, a one-time random split of the data. Results are sensitive to this arbitrary choice: it 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 keeping asymptotic control over the inclusion of noise variables. In addition, the proposed aggregation is shown to improve power, while reducing the number of falsely selected variables substantially.

Keywords: High-dimensional variable selection, data splitting, 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 work-flow 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 were not available in high-dimensional situations, except for the proposal of Wasserman and Roeder (2008). An ad-hoc solution for assigning relevance

*These authors contributed equally to this work
†Department of Statistics, University of Oxford, UK
‡Seminar für Statistik, ETH Zurich, Switzerland

is to use the bootstrap to analyze the stability of the selected predictors and to focus on those which are selected most often (or even always). Bach (2008) shows for the Lasso that this leads to a consistent model selection procedure under fewer restrictions than for the non-bootstrap case.

More recently, some progress has been achieved to obtain error control (Wasserman and Roeder, 2008; Meinshausen and Bühlmann, 2008). Here, we build upon the approach of Wasserman and Roeder (2008) and show that an extension of their 'screen and clean' algorithm leads to a more powerful variable selection procedure. Moreover, desired error rates can be controlled, while Wasserman and Roeder (2008) focus on variable selection rather than assigning significance via p-values.

This article is organized as follows. We discuss the single-split method of Wasserman and Roeder (2008) briefly in Section 2, showing that results can strongly depend on the arbitrary choice of a random sample splitting. We propose a multi-split method, removing this dependence. In Section 3 we prove error control of the multi-split method under identical conditions as in Wasserman and Roeder (2008), and we show in Section 4 numerically for simulated and real datasets that the method is more powerful than the single-split version while reducing substantially the number of false discoveries. Some possible extensions of the proposed methodology are outlined in Section 5.

2 Sample Splitting and High-Dimensional Variable Selection

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

$$Y = X\beta + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \ldots \varepsilon_n)$ is a random error vector with ε_i iid. $\mathcal{N}(0, \sigma^2)$ and $\beta \in \mathbb{R}^p$ is the parameter vector. Extensions to other models are outlined 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 (X_i, Y_i), $i = 1, \ldots, n$. We allow for potentially high-dimensional designs, i.e. $p \gg n$. This makes statistical inference very challenging. An approach proposed by Wasserman and Roeder (2008) is to split the data into two parts, reducing the dimensionality of predictors on one part to a manageable size of predictors (keeping the important variables with high probability), and then to assign p-values and making a final selection on the second part of the data, using classical least squares estimation.

2.1 The Single-Split Method

The procedure of Wasserman and Roeder (2008) relies on sample-splitting, performing variable selection and dimensionality reduction on one part of the data and classical significance testing on the remaining part. The data are splitted randomly into two disjoint groups $D_{in} = (X_{in}, Y_{in})$ and $D_{out} = (X_{out}, Y_{out})$ of equal size. Let $\tilde{S}$ be a variable selection or screening procedure which 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}, i.e. 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} by 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. $\tilde{S} \supseteq S$, 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 $\alpha \in (0, 1)$,

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

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

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

i.e. control of the family-wise error rate. The method is easy to implement and yields the asymptotic control under weak assumptions. The single-split method relies, however, on an arbitrary split into D_{in} and D_{out}. Results can change drastically if this split is chosen differently. This in itself is unsatisfactory since results are not reproducible.

2.2 The New Multi-Split 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. It is not obvious, though, how to combine and aggregate the results. Here, we give a possible answer. We will later show empirically that, maybe unsurprisingly, the resulting procedure is more powerful than the single-split method. The multi-split method also makes results reproducible, at least approximately if the number of random splits is chosen to be very large.

The multi-split 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, i.e.

$$\tilde{P}_j^{(b)} = 1,\ j \notin \tilde{S}^{(b)}.$$

4. Define the adjusted (non-aggregated) p-values as

$$P_j^{(b)} = \min\Big(\tilde{P}_j^{(b)}|\tilde{S}^{(b)}|,\ 1\Big),\ j = 1, \ldots, p \tag{2.1}$$

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

The 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

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

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 $0 < \gamma < 1$. We will show in Section 3 that this is an asymptotically correct p-value, adjusted for multiplicity. A typical choice for γ would be to use the median, i.e. $\gamma = 0.5$. The value $Q_j(0.5)$ corresponds to twice the median value among $P_j^{(b)}$, $b = 1, \ldots, B$. As we will see later, the choice $\gamma = 0.5$ can be too restrictive. A proper selection of γ may be difficult. Error control is not guaranteed anymore if we search for the best value of γ.

We propose to use instead an adaptive version which 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

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

The extra correction factor $1 - \log \gamma_{min}$ ensures that the family-wise error rate remains controlled at level α despite of the adaptive search for the best quantile, see Section 3. For the recommended choice of $\gamma_{\min} = 0.05$, this factor is upper bounded by 4; in fact, $1 - \log(0.05) \approx 3.996$. The selected subset consists now of all variables whose p-value is below a specified significance level $\alpha \in (0, 1)$,

$$\hat{S}_{multi} = \{j \in \tilde{S} : P_j \leq \alpha\}.$$

Figure 1 shows an example. The left panel contains the 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 4.2. The single split

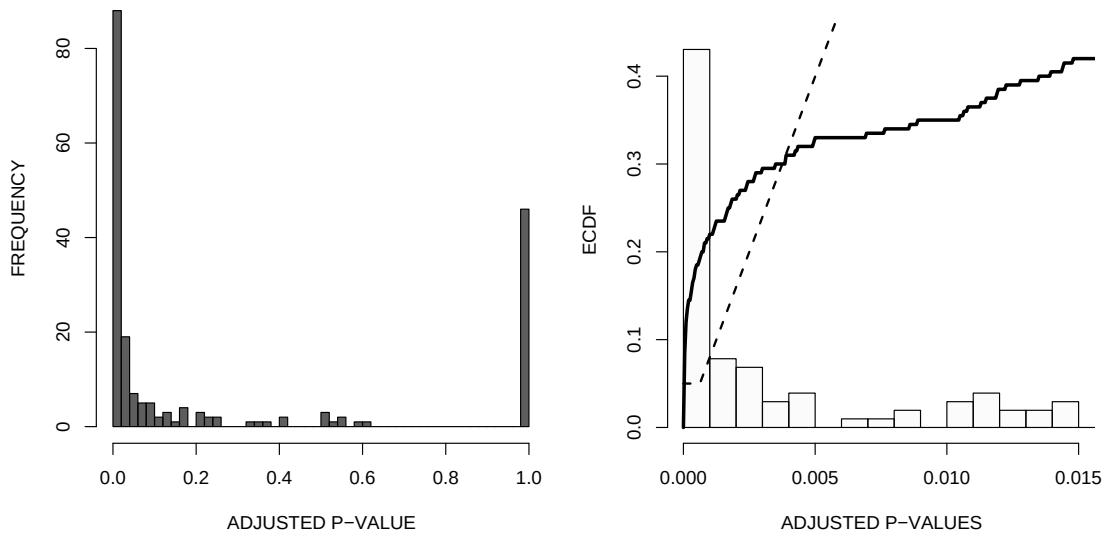

Figure 1: *Left: a histogram of adjusted p-values $P_j^{(b)}$ for the selected variable in the motif regression data example of Section 4.2. The single split method picks randomly one of these p-values (a "p-value lottery") and rejects if it is below α. For the multi-split method, we reject if and only 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 $\alpha = 0.05$. For the given example, the bound is indeed exceeded and the variable is thus selected.*

method is equivalent to picking one of these p-values randomly and selecting the variable if this randomly picked p-value is sufficiently small. To avoid this "p-value lottery", the multi-split method computes the empirical distribution of all p-values $P_j^{(b)}$ for $b = 1, \ldots, B$ and rejects if the empirical distribution crosses the broken line in the right panel of Figure 1. A short derivation of the latter is as follows. Variable j is selected if and only if $P_j \leq \alpha$, which happens if and only if there exists some $\gamma \in (0.05, 1)$ such that $Q_j(\gamma) \leq \alpha/(1 - \log 0.05) \approx \alpha/3.996$. Equivalently, using definition (2.2), the γ-quantile of the adjusted p-values, $q_\gamma(P_j^{(b)})$, has to be smaller than or equal to $\alpha\gamma/3.996$. This in turn is equivalent to the event that 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 the right panel of Figure 1.

Besides better reproducibility and asymptotic family-wise error control, the multi-split version is, maybe unsurprisingly, more powerful than the single-split selection method. Before showing numerical evidence, we show that the proposed method provides indeed the desired error control.

3 Error Control and Consistency

3.1 Assumptions

To achieve asymptotic error control, a few assumptions are made in Wasserman and Roeder (2008) regarding the crucial requirements for the variable selection procedure $\tilde{S}$.

(A1) *Screening property*: $\lim_{n\to\infty} \mathbb{P}\Big[\tilde{S} \supseteq S\Big] = 1$.

(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, too, 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 which satisfies (A1) and (A2) under appropriate conditions discussed in Meinshausen and Bühlmann (2006), Zhao and Yu (2006), van de Geer (2008) and Bickel et al. (2008). The adaptive Lasso (Zou, 2006; Zhang and Huang, 2008) satisfies (A1) and (A2) as well under suitable conditions. Other examples include, assuming appropriate conditions, L_2 Boosting (Friedman, 2001; Bühlmann, 2006), orthogonal matching pursuit (Tropp and Gilbert, 2007) or sure independence screening (Fan and Lv, 2008).

We will typically use the Lasso (and extensions thereof) as screening method. Other algorithms would be possible. Wasserman and Roeder (2008) 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, just working on the assumption that we have a selection procedure $\tilde{S}$ at hand which satisfies both the *Screening property* and the *Sparsity property*.

3.2 Error Control

We proposed two versions for multiplicity-adjusted p-values. One is $Q_j(\gamma)$ as defined in (2.2) which relies on a choice of $\gamma \in (0, 1)$. The second is 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 error control.

Theorem 3.1. *Assume (A1) and (A2). Let $\alpha, \gamma \in (0, 1)$. If the null-hypothesis $H_{0,j} : \beta_j = 0$ gets rejected whenever $Q_j(\gamma) \le \alpha$, the family-wise error rate is asymptotically controlled at level α, i.e.*

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

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

Theorem 3.1 is valid for any pre-defined value of the quantile γ. However, the adjusted p-values $Q_j(\gamma)$ involve the somehow arbitrary choice of γ which might pose a problem for practical applications. We therefore proposed the adjusted p-values P_j which search for the optimal value of γ adaptively.

Theorem 3.2. *Assume (A1) and (A2). Let $\alpha \in (0,1)$. If the null-hypothesis $H_{0,j} : \beta_j = 0$ gets rejected whenever $P_j \leq \alpha$, the family-wise error rate is asymptotically controlled at level α, i.e.*

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

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

We comment briefly on the relation between the proposed adjustment to false discovery rate (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) or family-wise error (Holm, 1979) controlling procedures. While we provide a family-wise error control and as such use union bound corrections as in Holm (1979), the definition of the adjusted p-values (2.3) and its graphical representation in Figure 1 are vaguely reminiscent of the false discovery rate 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 only taken for one predictor variable, though, which is either in S or N. This would correspond to a multiple testing situation where we are testing a single hypothesis with multiple statistics.

3.3 Model Selection Consistency

If we let level $\alpha = \alpha_n \to 0$ for $n \to \infty$, the probability of falsely including a noise variable vanishes because of the preceding results. In order to get the property of consistent model selection, we have to 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 γ_{min} (see (2.3)), the multi-split method is also model selection consistent for a suitable sequence α_n, i.e. for $\hat{S}_{multi} = \{j \in \tilde{S}; P_j \leq \alpha_n\}$ it holds that*

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

Wasserman and Roeder (2008) discuss conditions which 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 the Corollary above is not necessarily true. The multi-split method can be consistent if the single-split method is not. One could imagine for example a scenario, where the p-values of

a variable $j \in S$ satisfy for some $c \in (0,1)$ that $\limsup_{n\to\infty} \mathbb{P}[P_j^{(b)} \leq c] < 1$, where the probability is with respect to both the data and the random split-point. In this case, the single-split method cannot be consistent, as there is a positive probability of variable $j \in S$ not being selected when $\alpha_n \to 0$. On the other hand, some quantiles of the distribution of $P_j^{(b)}$ under repeated random split-point selection can converge to 0, which would make the multi-split method consistent even though the single-split method is not. We refrain from going into more details here and rather show with numerical results that the multi-split method is indeed more powerful than the single-split analogue.

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 dataset shows 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 design matrix coming from a centered multivariate normal distribution with covariance structure $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$

(B) As (A) but with $n = 100$ and $p = 1000$.

(C) Real data set with $n = 71$ and $p = 4088$ for the design matrix X and artificial response 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 Bacillus Subtilis. The data is kindly provided by DSM (Switzerland). As the true variables are not known, we consider a linear model with design matrix from real data and simulating a sparse parameter vector β as follows. In each simulation run, a new parameter vector β is created by randomly setting $|S|$ components of β to 1 and the remaining $p - |S|$ components to 0. Placing the active components at random creates a more difficult problem than placing them in blocks (e.g. the first 10). Selecting a correlated predictor usually leads to a false decision in the random component case but not necessarily if the active components arise in blocks. 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 compare the average number of true positives and the family-wise error rate (FWER) for the single and multi-split methods for all three simulation settings (A)–(C) and vary in each the SNR to

1, 4 and 16 and the number $|S|$ of relevant variables between 5 and 10. As initial variable selection or screening method $\tilde{S}$ we use three approaches, which are all based on the Lasso (Tibshirani, 1996). The first one, denoted by $\tilde{S}_{fixed}$, uses the Lasso and selects those $\lfloor n/6 \rfloor$ variables which appear most often in the regularization path when varying the penalty parameter. The second method, denoted by $\tilde{S}_{cv}$, uses the Lasso with penalty parameter chosen by 10-fold cross-validation and selecting the variables whose corresponding estimated regression coefficients are different from zero. The third method, denoted by $\tilde{S}_{adap}$ is the adaptive Lasso of Zou (2006) where regularization parameters are chosen based on 10-fold cross-validation with the Lasso solution used as initial estimator for the adaptive Lasso. The selected variables are again the ones whose corresponding estimated regression parameters are different from zero. We use a total of $B = 50$ sample-splits for each simulation run.

Results are shown in Figures 2 and 3 for both the single-split method and the multi-split method with the default setting $\gamma_{\min} = 0.05$. Using the multi-split method, the average number of true positives (the variables in S which are selected) is typically slightly increased while the FWER (the probability of including variables in N) is reduced sharply. The single-split method has often a FWER above the level $\alpha = 0.05$ at which it is asymptotically controlled. For the multi-split method, the FWER is above the nominal level only in one scenario, for data (C) with a high SNR and the adaptive Lasso, and below α otherwise. The asymptotic control seems to give a good control in finite sample settings with the multi-split method, yet not the single-split method. Even though the multi-split method is more conservative than the single-split method (having substantially lower FWER), the number of true discoveries is often increased. We note that for data (C), with $p = 4088$, the number of true positives is low since we control the very stringent family-wise error criterion at $\alpha = 0.05$ significance level. As an alternative, controlling less conservative error measures would be possible and is discussed in Section 5.

We also experimented with using the value of $Q_j(\gamma)$ directly as an adjusted p-value, without the adaptive choice of γ but using a fixed value $\gamma = 0.5$ instead, i.e. looking at twice the median value of all p-values across multiple data splits. The results were not as convincing as for the adaptive choice and we recommend the adaptive version with $\gamma_{\min} = 0.05$ as a good default choice.

4.2 Real Data

We apply the multi-split method to a real data set about 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 will be our response values $Y_1, \ldots, Y_n$. Moreover, for $p = 195$ candidate words ("motifs") we have scores x_{ij} which measure how well the jth motif is represented in the ith DNA sequence. The motifs are typically 5–15bp long candidates for the true binding site of the protein. The hope is that the true binding site is in the list of significant variables showing the strongest relationship between the motif score and the binding intensity. Using a linear model with $\tilde{S}_{adap}$, the multi-split method identifies one predictor variable at the 5% significance level. The single-split

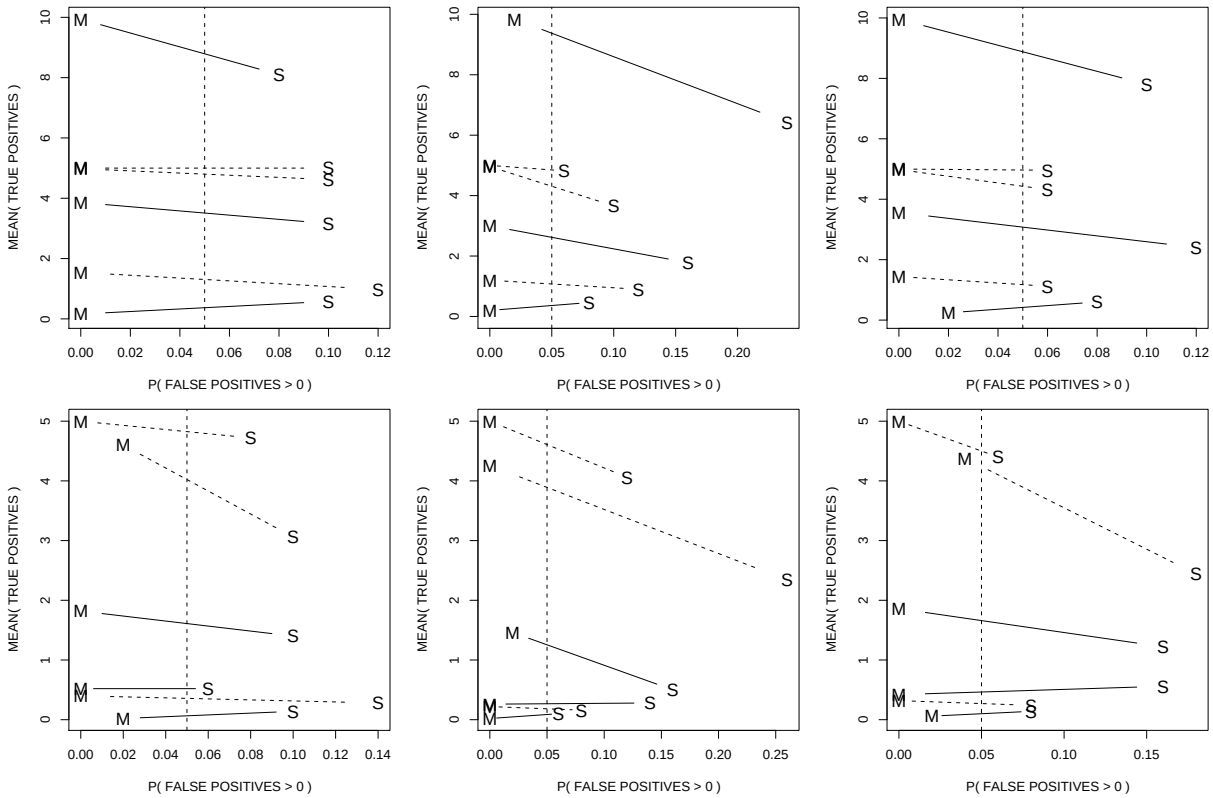

Figure 2: *Simulations for case (A) in the top and (B) in the bottom row. Average number of true positives vs. the family-wise error rate (FWER) for the single split method ("S") against the multi-split version ("M"). FWER is controlled (asymptotically) at* $\alpha = 0.05$ *for both methods and this value is indicated by a broken vertical line. From left to right are results for* $\tilde{S}_{fixed}$*,* $\tilde{S}_{cv}$ *and* $\tilde{S}_{adap}$*. Simulation results for the same SNR and same number* $|S|$ *of relevant variables are joined by a broken line if* $|S| = 5$ *and an unbroken line if* $|S| = 10$*.*

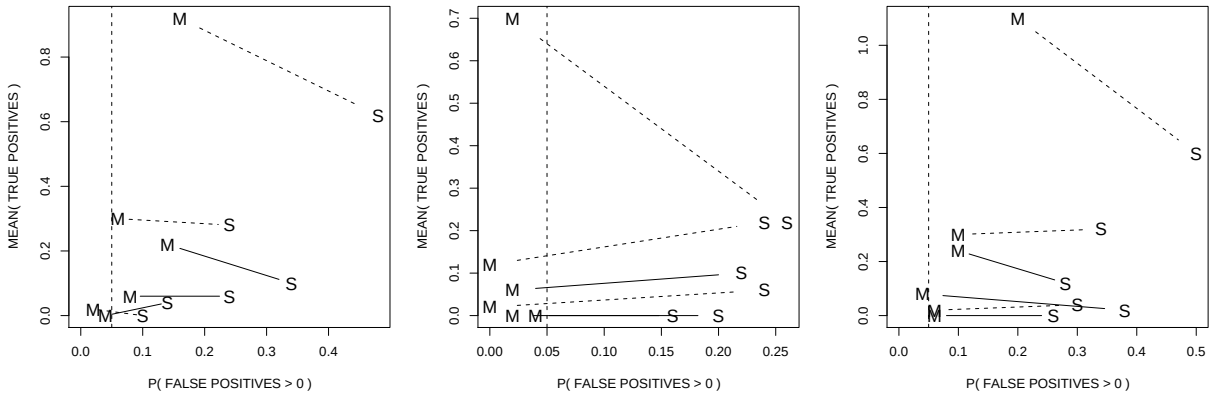

Figure 3: *Simulations for data (C). Average number of true positives vs. the family-wise error rate (FWER) for the single split method ("S") against the multi-split version ("M"). FWER is controlled (asymptotically) at* $\alpha = 0.05$ *for both methods and this value is indicated by a broken vertical line. From left to right are results for* $\tilde{S}_{fixed}$*,* $\tilde{S}_{cv}$ *and* $\tilde{S}_{adap}$*. Simulation results for the same SNR and same number* $|S|$ *of relevant variables are joined by a broken line if* $|S| = 5$ *and an unbroken line if* $|S| = 10$*.*

method is not able to identify a single significant predictor. As mentioned above, we could control other, less conservative error measures as discussed in Section 5.

5 Extensions

Due to 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 are generalized linear models (GLMs) or Gaussian Graphical Models. The dimensionality reduction step would typically involve some form of shrinkage estimation. An example for Gaussian Graphical Models would be the recently proposed 'Graphical Lasso' (Friedman et al., 2008). The second step would rely on classical (e.g. likelihood ratio) tests applied to the selected submodel, analogous to the methodology proposed for linear regression.

We have shown above how the family-wise error rate (FWER) can be controlled in a multi-split approach. We could control other error rates. Take for example the false-discovery rate (FDR). By directly aggregating the p-values $\tilde{P}_j^{(b)}$ (i.e. without applying a Bonferroni correction for the size of the selected model $\tilde{S}^{(b)}$), we can define "unadjusted" p-values $Q_j(\gamma)$ and P_j analogously to (2.2) and (2.3) with the property that

$$\limsup_{n\to\infty} \mathbb{P}[Q_j(\gamma) \leq \alpha] \leq \alpha,$$

and

$$\limsup_{n\to\infty} \mathbb{P}[P_j \leq \alpha] \leq \alpha.$$

Now, by using a multiplicity correction as in Benjamini and Hochberg (1995) or Benjamini and Yekutieli (2001) we can asymptotically control the FDR. However, the multiplicity correction involves here (for the smallest p-value) a factor p, which can be much larger than the average value of $|\tilde{S}^{(b)}|$ in high-dimensional settings, and it might be more powerful to work with the proposed FWER-controlling procedure.

In some settings, control of FWER at $\alpha = 0.05$ is too conservative. If the control of FDR, as alluded to above, is also too conservative, our method can easily be adjusted to control the expected number of false rejections. Take as an example the adjusted p-value P_j, defined in (2.3). Variable j is rejected if and only if $P_j \leq \alpha$. (For the following, assume that adjusted p-values, as defined in (2.1), are not capped at 1. This is a technical detail only as it does not modify the proposed FWER-controlling procedure.) Rejecting variable j if and only if $P_j \leq \alpha$ controls FWER at level α. Instead, one can reject variables if and only if $P_j/K \leq \alpha$, where $K > 1$ is a correction factor. Call the number of falsely rejected variables V,

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

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

6 Discussion

We proposed a multi-sample-split 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 method of Wasserman and Roeder (2008). Combining the results of multiple data-splits, based on quantiles as summary statistics, improves reproducibility compared to the single-split method. The multi-split method shares with the single-split method the property of asymptotic error control and model selection consistency. We argue empirically that the multi-split method usually selects much fewer false positives than the single-split method while the number of true positives is slightly increased. The method is very generic and can be used for a broad spectrum of error controlling procedures in multiple testing, including linear and generalized linear models.

A Proofs

Proof of Theorem 3.1. For technical reasons we define

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

$K_j^{(b)}$ 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) and for fixed B, $\mathbb{P}[K_j^{(b)} = P_j^{(b)} \text{ for all } b = 1, \ldots, B]$ on a set A_n with $\mathbb{P}[A_n] \to 1$. Therefore, we can define all the quantities involving $P_j^{(b)}$ also with $K_j^{(b)}$, and it is sufficient to show under this slightly altered procedure that

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

In particular we can omit here the limes superior.

Define

$$\pi_j(\alpha) = \frac{1}{B} \sum_{b=1}^{B} 1\big\{K_j^{(b)}/\gamma \leq \alpha\big\}.$$

Note that the events $\{Q_j(\gamma) \leq \alpha\}$ and $\{\pi_j(\alpha) \geq \gamma\}$ are equivalent. Hence,

$$\mathbb{P}\Big[\min_{j \in N} Q_j(\gamma) \leq \alpha\Big] \leq \sum_{j \in N} \mathbb{E}\Big[1\big\{Q_j(\gamma) \leq \alpha\big\}\Big] = \sum_{j \in N} \mathbb{E}\Big[1\big\{\pi_j(\alpha) \geq \gamma\big\}\Big].$$

Using that $x/y \geq 1\{x \geq y\}$ for all $x, y > 0$,

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

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

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

Moreover,

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

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

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

□

Proof of Theorem 3.2. As in the proof of Theorem 3.1 we will 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 $\alpha \in (0,1)$,

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

Furthermore,

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

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

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

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

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

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

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

Analogously to (A.4),

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

Averaging over all bootstrap samples yields

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

Using that $x/y \ge 1\{x \ge y\}$ for all $x, y > 0$,

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

where we have used the same definition for $\pi_j(\cdot)$ as in the proof of Proposition 3.1.

Since the events $\{Q_j(\gamma) \leq \alpha\}$ and $\{\pi_j(\alpha) \geq \gamma\}$ are equivalent, it follows that

$$\mathbb{P}\Big[\min_{j \in N} \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma) \leq \alpha\Big] \leq \alpha(1 - \log \gamma_{\min}),$$

implying that

$$\mathbb{P}\Big[\min_{j \in N} \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)(1 - \log \gamma_{\min}) \leq \alpha\Big] \leq \alpha.$$

Using the definition of P_j in (2.3),

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

which completes the proof. □

Proof of Corollary 3.1. Because the single-split method is model selection consistent, it must hold that $\mathbb{P}[\max_{j \in S} \tilde{P}_j | \tilde{S}| \leq \alpha_n] \to 1$ for $n \to \infty$. Using multiple data-splits, this property holds for each of the B splits and hence $\mathbb{P}[\max_{j \in S} \max_b \tilde{P}_j^{(b)} |\tilde{S}^{(b)}| \leq \alpha_n] \to 1$, which implies 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 $n \to \infty$. □

References

Bach, F. R. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, New York, NY, USA, pp. 33–40. ACM.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188.

Bickel, P., Y. Ritov, and A. Tsybakov (2008). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*. To appear.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics*, 559–583.

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

Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B 70*(5), 849–911.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics 9*(3), 432.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis 52*(1), 374 – 393.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*(3), 1436–1462.

Meinshausen, N. and P. Bühlmann (2008). Stability selection. Preprint.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B 58*(1), 267–288.

Tropp, J. and A. Gilbert (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. *Annals of Statistics 36*(2), 614–645.

Wasserman, L. and K. Roeder (2008). High dimensional variable selection. *Annals of Statistics*. To appear.

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

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.