HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2016 April 10.

Published in final edited form as:

J Am Stat Assoc. 2015 ; 110(512): 1422–1433. doi:10.1080/01621459.2015.1095099.

An adaptive resampling test for detecting the presence of significant predictors

Ian W. McKeague[*] and **Min Qian**[†]

[*] Department of Biostatistics, Columbia University. im2131@columbia.edu.

[†]Department of Biostatistics, Columbia University. mq2158@columbia.edu.

Abstract

This paper investigates marginal screening for detecting the presence of significant predictors in high-dimensional regression. Screening large numbers of predictors is a challenging problem due to the non-standard limiting behavior of post-model-selected estimators. There is a common misconception that the oracle property for such estimators is a panacea, but the oracle property only holds away from the null hypothesis of interest in marginal screening. To address this difficulty, we propose an adaptive resampling test (ART). Our approach provides an alternative to the popular (yet conservative) Bonferroni method of controlling familywise error rates. ART is adaptive in the sense that thresholding is used to decide whether the centered percentile bootstrap applies, and otherwise adapts to the non-standard asymptotics in the tightest way possible. The performance of the approach is evaluated using a simulation study and applied to gene expression data and HIV drug resistance data.

Keywords

Bootstrap; Family-wise error rate; Marginal regression; Non-regular asymptotics; Screening covariates

1 Introduction

The problem of selecting significant predictors is a central aspect of scientific discovery, and has become increasingly important in an era in which massive data sets are readily available (Fan and Li, 2006). Much of the modern statistical literature in this area focuses on consistency of variable selection in high-dimensional settings based on machine learning and data mining techniques (e.g., Fan and Li 2001; Zou and Hastie 2005; Huang et al. 2008; Fan and Lv 2008; Genovese et al. 2012). A major gap in this literature, however, has been the scarcity of formal hypothesis testing procedures that take variable selection into account; the oracle property enjoyed by many variable selection methods in the presence of high dimensionality can not be applied *directly* for testing whether a post-model-selected variable is significant. In bioinformatics, for example, variable selection techniques based on penalization (such as lasso, scad, etc) are routinely used to produce lists of differentially-expressed genes that are most related to disease risk, but few methods for obtaining valid p-values have been developed.

A more traditional approach to the selection of significant predictors is multiple testing to control either family-wise error rate (FWER), or false-discovery rate (Benjamini and Hochberg 1995; Dudoit et al. 2003; Efron 2006; Dudoit and van der Laan 2008; Efron 2010). Procedures that control FWER (e.g., Bonferroni, or Holm's procedure) are often criticized as being too conservative (in the sense of having low power). False-discovery rate methods, on the other hand, although having greater power, incur the cost of inflated FWER. Our aim in the present paper is to introduce a more powerful *single* test that can be used as an alternative screening procedure to detect the presence of *some* significant predictor while rigorously controlling FWER.

The proposed procedure uses marginal linear regression to select the predictor (from among covariates $X_1, \ldots, X_p$) that has maximal sample correlation with a scalar outcome Y (as in marginal screening or correlation learning, Genovese et al. 2012). The test is based on $\hat{\theta}_n$ the estimated marginal regression coefficient of the selected predictor. If there is a unique predictor, say X_{k_0}, maximally correlated with the outcome, then the selection procedure consistently estimates k_0, and $\hat{\theta}_n$ is asymptotically normal; if all predictors are uncorrelated with the outcome, then the selected predictor does not converge (in probability) and $\hat{\theta}_n$ has a non-normal limiting distribution. In particular, the limiting distribution is discontinuous (at zero) as a function of the regression coefficient of X_{k_0} (where k_0 is not identifiable), and this "non-regularity" causes non-uniform convergence.

Breiman (1992) drew early attention to the issue of invalid post-model-selection inference, calling it the "quiet scandal" of Statistics; even earlier references are mentioned in Berk et al. (2013). Samworth (2003) gave a detailed account of the inaccuracy of bootstrap methods applied to super-efficient estimators. Leeb and Pötscher (2006) (and other papers by the same authors) established that non-uniform limiting behavior of post-model-selected estimators is at the root of the problem, and that estimates of asymptotic null distributions in such settings can give a misleading picture of finite-sample performance. In particular, calibrating a test based on $\hat{\theta}_n$ in a way that does not adapt to the implicit post-model-selection will be extremely inaccurate. This type of non-regularity occurs in various other settings as well, e.g., when a nuisance parameter is only defined under an alternative hypothesis (Davies, 1977), and when the parameter of interest under the null hypothesis is on the boundary of the parameter space (Andrews, 2000). McCloskey (2012) surveyed non-standard testing problems in econometrics, and introduced some Bonferroni-based size-correction methods designed to improve power. As far as we know, however, there is not yet a resolution of these issues for marginal screening.

In this paper we introduce an *adaptive resampling test* (ART) for marginal screening that adapts to the small sample behavior of $\hat{\theta}_n$ in terms of a local model. Under local alternatives, we find an explicit representation of the asymptotic distribution of $\hat{\theta}_n$ and construct a suitable bootstrap estimator of this distribution that is consistent, thus circumventing the non-regularity mentioned above. Under non-local alternatives, we show that the critical values obtained in this way agree asymptotically with those used by the oracle (who is given knowledge of k_0), so ART can be expected to provide good power as well.

Several new approaches to post-model selection inference for linear regression have been proposed in recent years. Meinshausen et al. (2009) introduced a random sample splitting procedure in the high-dimensional setting to obtain (conservative) Bonferroni-adjusted p-values following variable selection. Chatterjee and Lahiri (2011) developed a modified bootstrap method that provides an asymptotically valid confidence region for the regression parameters based on the lasso estimator; this method depends on the presence of at least one active predictor, so it is not applicable to marginal screening (under the null hypothesis there is no active predictor).

More relevant to marginal screening, the covariance test recently introduced by Lockhart et al. (2014) uses a forward stepwise lasso procedure to test for active predictors entering a sparse linear model under the assumption of normal errors. Also in the sparse linear model setting with normal errors, but further assuming that the predictors are nearly uncorrelated, Ingster et al. (2010) and Arias-Castro et al. (2011) have studied the detection boundary and optimality properties of general classes of multiple testing procedures (including Bonferroni and Higher Criticism). Berk et al. (2013) developed a valid method of post-model selection inference that is feasible for up to about $p = 20$ predictors, also assuming normal errors. In various sparse high-dimensional settings, Belloni et al. (2013), Bühlmann (2013), Zhang and Zhang (2014) and Ning and Liu (2015) have established asymptotically valid confidence intervals for a preconceived regression parameter after variable selection on the remaining predictors, but this does not apply to marginal screening (where no regression parameter is singled-out *a priori*).

This paper is organized as follows. We formulate the problem and discuss the issue of non-regularity in Section 2. In Section 3, we develop the ART procedure and establish the consistency of the underlying bootstrap. Simulation studies and applications to gene expression data and HIV drug resistance data are presented in Section 4. Concluding discussion appears in Section 5, and proofs are collected in the Appendix.

2 Marginal regression and non-regularity

Consider a scalar outcome Y and a p-dimensional vector of covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$ such that the marginal variance of each covariate is finite and non-zero. Marginal regression consists in using separate linear models to predict Y from each X_k. Let k_0 be the label of a covariate that maximizes the absolute correlation with Y:

$$k_0 \in arg \max_{k=1,\ldots,p} |Corr(X_k, Y)|,$$

and let $\alpha_0 + \theta_0 X_{k_0}$ be the best linear predictor based on X_{k_0}, i.e.,

$$(\alpha_0, \theta_0) = arg \min_{\alpha,\theta\in\mathbb{R}} E(Y - \alpha - \theta X_{k_0})^2 = \left(EY - \theta_0 EX_{k_0}, \frac{Cov(X_{k_0}, Y)}{Var(X_{k_0})} \right). \quad (1)$$

We are interested in testing whether at least one of the covariates is correlated with Y, for which it suffices to check whether X_{k_0} and Y are correlated. This is equivalent to testing

$$H_0{:}\theta_0{=}0 \qquad \text{versus} \qquad H_a{:}\theta_0 \neq 0.$$

Given an iid sample of size n, let $\hat{\alpha}_n$, $\hat{\theta}_n$, and $\hat{k_n}$ be the least squares estimates of α_0, θ_0, and k_0, respectively:

$$\hat{\alpha}_n{=}\mathbb{P}_n Y - \hat{\theta}_n \mathbb{P}_n X_{\hat{k}_0}, \quad \hat{\theta}_n{=}\frac{\widehat{Cov}\left(X_{\hat{k}_n}, Y\right)}{\widehat{Var}\left(X_{\hat{k}_n}\right)}, \quad \hat{k}_n \in arg\max_{k=1,\ldots,p} |\widehat{Corr}\left(X_k, Y\right)|,$$

where $\mathbb{P}_n$ is the empirical distribution, and the hats indicate sample versions. It is natural to base the test on $\hat{\theta}_n$ but calibration is problematic because the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ does not converge uniformly with respect to θ_0, as mentioned in the Introduction. The non-uniformity occurs in the neighborhood of $\theta_0 = 0$. Specifically, there exists a bounded continuous function $h{:}\mathbb{R} \to \mathbb{R}$ such that $f_n(\theta_0) \equiv \mathrm{E}h\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right)$ does not converge uniformly in any neighborhood of $\theta_0 = 0$, despite converging pointwise. To see this, first note that under mild conditions

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} U \equiv \begin{cases} Z_{k_0}/V_{k_0} & \text{if} \quad \theta_0 \neq 0, \\ Z_K/V_K & \text{if} \quad \theta_0{=}0, \end{cases}$$

where $V_k = \mathrm{Var}(X_k)$, $K{=}arg\,max_{k=1,\ldots,p} Z_k^2/V_k$, and $(Z_1, \ldots, Z_p)^T$ is a mean-zero normal random vector with covariance matrix depending on parameters of the full linear model (this is a special case of Theorem 1 below). From the form of the distribution of U, we can choose h so that $f_\infty(\theta_0) \equiv \mathrm{E}h(U)$ is discontinuous at $\theta_0 = 0$ (this is the non-regularity mentioned in the Introduction). If f_n were to converge uniformly to f_∞ on some compact neighborhood of zero, we would have a contradiction because each f_n is continuous, and the uniform limit of a sequence of continuous functions on a compact interval is continuous.

To address this problem, in the next section we develop a formal test procedure (ART) inspired by work of Cheng (2008, 2015) concerning robust confidence intervals for non-linear regression parameters in the presence of weak-identifiability. Other variations of this approach have been used by Laber and Murphy (2011) to construct a confidence interval for the classification error, by Laber et al. (2014) in a sequential decision making problem, and by Laber and Murphy (2013) to provide robust confidence intervals for adaptive lasso. As already noted, the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ does not converge uniformly in the neighborhood of $\theta_0 = 0$, so its small sample behavior can be very far from normal when the true parameter is close to zero. Therefore an understanding of the asymptotic behavior of $\hat{\theta}_n$ under local alternatives plays a crucial role in devising a suitable test, or more generally in providing robust confidence intervals for θ_0.

3 Adaptive resampling test

In this section, we develop the proposed ART procedure for detecting the presence of a significant predictor. The idea is to adapt to the inherent non-regular behavior of the post-model-selected estimator $\hat{\theta}_n$ in a way that accurately captures its asymptotic behavior in $\sqrt{n}$-neighborhoods of the null hypothesis.

We frame the problem in terms of the general local linear model

$$Y=\alpha_0+\mathbf{X}^T\boldsymbol{\beta}_n+\epsilon, \quad (2)$$

where $\alpha_0 \in \mathbb{R}$, $\boldsymbol{\beta}_n \in \mathbb{R}^p$, the noise ε has mean 0, finite variance, and is uncorrelated with $\mathbf{X}$, and $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{b}_0$, where $\boldsymbol{b}_0 \in \mathbb{R}^p$ is the local parameter. The distributions of ε and $\mathbf{X}$ are assumed to be fixed, so only the distribution of Y depends on n (although we suppress n in the notation for Y). The relevant hypotheses are now

$$H_0:\theta_n=0 \qquad \text{versus} \qquad H_a:\theta_n \neq 0,$$

where $\theta_n = \mathrm{Cov}(X_{k_n}, Y)/\mathrm{Var}(X_{k_n})$ and k_n is the label of a component of $\mathbf{X}$ that maximizes absolute correlation with Y.

Our first result gives the asymptotic distribution of $\hat{\theta}_n$. To state the result, we need the notation

$$\bar{k}(\boldsymbol{b}) \equiv arg\max_{k=1,\ldots,p} |Corr\left(X_k, \mathbf{X}^T\boldsymbol{b}\right)$$

for any $\boldsymbol{b} \in \mathbb{R}^p$. Note that $k_n = \bar{k}(\boldsymbol{\beta}_n)$ under the local model. If $k_0 \equiv \bar{k}(\boldsymbol{\beta}_0)$ is unique (so $\boldsymbol{\beta}_0 \neq 0$), then $k_n \to k_0$, and θ_n is asymptotically bounded away from zero (a non-local alternative). On the other hand, if $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\bar{k}(\boldsymbol{b}_0)$ is unique, then $k_n = \bar{k}(\boldsymbol{b}_0)$; also θ_n is in the neighborhood of zero and represents a local alternative. Finally, if $\boldsymbol{\beta}_0 = \boldsymbol{b}_0 = \mathbf{0}$, then k_n is not well-defined and the null hypothesis $\theta_n = 0$ holds. We need the uniqueness of the most active predictor k_0 (away from the null hypothesis), but this seems to be a very mild condition because the likelihood that there would be two or more predictors having *exactly* the same maximal correlation with Y seems remote in practice. Even in practice, as we will see in the simulation study, non-uniqueness of the maximally correlated predictor does not adversely affect power.

Theorem 1

Suppose that $k_0 = \bar{k}(\boldsymbol{b}_0)$ is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, and $\bar{k}(\boldsymbol{b}_0)$ is unique when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{b}_0 \neq \mathbf{0}$. Then, under the local model (2),

$$\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) \xrightarrow{d} \begin{cases} Z_{k_0}(\boldsymbol{\beta}_0)/V_{k_0} & if \quad \boldsymbol{\beta}_0 \neq \mathbf{0}, \\ Z_K(\mathbf{0})/V_K + \left(C_K/V_K - C_{\bar{k}(\boldsymbol{b}_0)}/V_{\bar{k}(\boldsymbol{b}_0)}\right)^T \boldsymbol{b}_0 & if \quad \boldsymbol{\beta}_0 = \mathbf{0}, \end{cases}$$

where $K = arg\,max_{k=1,\ldots,p}\left[Z_k(\mathbf{0}) + C_k^T \boldsymbol{b}_0 \right]^2 / V_k$, $C_k = Cov(X_k, \mathbf{X})$, and $(Z_k(\boldsymbol{\beta}))_{k=1}^p$ is a mean-zero normal random vector with covariance matrix $\Sigma(\boldsymbol{\beta})$ given by that of the random vector with components

$$\left((\mathbf{X} - E\mathbf{X})^T \boldsymbol{\beta} - (\mathbf{X_k} - \mathbf{EX_k})\, \mathbf{C_k^T} \boldsymbol{\beta} / \mathbf{V_k} + \epsilon \right) (X_k - EX_k),$$

for k = 1, . . . , p, and $\Sigma(\boldsymbol{\beta}_0)$ is assumed to exist.

The non-regularity at $\boldsymbol{\beta}_0 = \mathbf{0}$ is explained by the dependence of the limiting distribution on the (non-identifiable) local parameter $\boldsymbol{b}_0$. The limiting distribution is nevertheless continuous as a function of $\boldsymbol{b}_0 \in \mathbb{R}^p$ into the space of distribution functions (this is a simple consequence of Lemma 3 in the Appendix), and the convergence is uniform over compact subsets of $\mathbb{R}^p$, unlike the limiting behavior discussed in the previous section, so finite-sample accuracy should be less of an issue when designing a screening test using this result. On the other hand, naive resampling methods that do not take into account the local asymptotic behavior will fail to provide consistent estimates of the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$, as discussed in the Introduction for the non-local case.

To get around this problem, we decompose $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ in a way that isolates the possibility that $\boldsymbol{\beta}_0 \neq 0$ by comparing $|T_n|$ to some threshold λ_n (to be specified later), where $T_n = \hat{\theta}_n / s_n$ is the post-model-selected t-statistic and s_n is the standard error of the slope estimator when regressing Y on $X_{\hat{k}_n}$. Specifically,

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_n - \theta_n\right) &= \sqrt{n}\left(\hat{\theta}_n - \theta_n\right) 1_{|T_n| > \lambda_n \quad \text{or} \quad \boldsymbol{\beta}_0 \neq \mathbf{0}} + \sqrt{n}\left(\hat{\theta}_n - \theta_n\right) 1_{|T_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}} \\ &= \sqrt{n}\left(\hat{\theta}_n - \theta_n\right) 1_{|T_n| > \lambda_n \quad \text{or} \quad \boldsymbol{\beta}_0 \neq \mathbf{0}} + \left[\frac{\mathbb{Z}_{n,\hat{k}_n} + \widehat{Cov}\left(X_{\hat{k}_n}, \mathbf{X}^T \boldsymbol{b}_0\right)}{\widehat{Var}\left(X_{\hat{k}_n}\right)} - \frac{Cov\left(X_{k_n}, \mathbf{X}^T \boldsymbol{b}_0\right)}{Var\left(X_{k_n}\right)} \right] 1_{|T_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}}, \end{aligned} \tag{3}$$

where $\mathbb{Z}_{n,k} = \mathbb{G}_n\left[\epsilon (X_k - \mathbb{P}_n X_k)\right]$, $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_n)$ is the empirical process, and P_n is the distribution of $(\mathbf{X}, Y)$. It is clear that the nonparametric bootstrap is consistent for the first term in (3) if $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, since it is easily shown that $P(|T_n| > \lambda_n) \to 1_{\boldsymbol{\beta}_0 \neq \mathbf{0}}$. The second term is more problematic though because $\hat{k}_n$ does not converge in probability to k_0 when $\boldsymbol{\beta}_0 = \mathbf{0}$. Denote the term in the square brackets by $\mathbb{V}_n(\boldsymbol{b})$, indexed by $\boldsymbol{b} = \boldsymbol{b}_0 \in \mathbb{R}^p$. Note that when this term is active (under $\boldsymbol{\beta}_0 = \mathbf{0}$), $\hat{k}_n = \mathbb{K}_n(\boldsymbol{b}_0)$ and $k_n = K(\boldsymbol{b}_0)$, where

$$\mathbb{K}_n(\boldsymbol{b}) = arg\,\max_{k=1,\ldots,p} \frac{\left[\mathbb{Z}_{n,k} + \widehat{\mathbf{Cov}}\left(X_k, \mathbf{X}^T \boldsymbol{b}\right)\right]^2}{\widehat{Var}(X_k)}$$

and

$$\bar{K}(\boldsymbol{b}) = arg\max_{k=1,\ldots,p} \frac{\left[Cov\left(X_k, \mathbf{X}^T\boldsymbol{b}\right)\right]^2}{Var\left(X_k\right)},$$

so

$$\mathbb{V}_n(\boldsymbol{b}) = \frac{\mathbb{Z}_{n,\mathbb{I}_n(\boldsymbol{b})} + \widehat{Cov}\left(X_{\mathbb{I}_n(\boldsymbol{b})}, \mathbf{X}^T\boldsymbol{b}\right)}{\widehat{Var}\left(X_{\mathbb{I}_n(\boldsymbol{b})}\right)} - \frac{Cov\left(X_{\bar{K}(\boldsymbol{b})}, \mathbf{X}^T\boldsymbol{b}\right)}{Var\left(X_{\bar{K}(\boldsymbol{b})}\right)}. \quad (4)$$

All parts of $\mathbb{V}_n(\boldsymbol{b})$ are now seen to be smooth functions of $\mathbb{P}_n$, so it is reasonable to expect that a consistent bootstrap can be constructed by replacing $\mathbb{P}_n$ by its nonparametric bootstrap $\mathbb{P}_n^*$, and replacing P_n by $\mathbb{P}_n$. In such a construction, the event indicated in the second term of (3) is naturally replaced by the event that $|T_n^*| \leq \lambda_n$ and $|T_n| \leq \lambda_n$.

Here and throughout the paper, a superscript * is used to indicate the nonparametric bootstrap (sometimes called "bootrapping in pairs" in regression settings, to distinguish it from the residual bootstrap). The above arguments lead to our main result showing that $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ can indeed be consistently bootstrapped under the general local model. The precise definition of $\mathbb{V}_n^*$ is given at the start of the proof.

Theorem 2

Suppose all assumptions in Theorem 1 hold, and the tuning parameter λ_n satisfies $\lambda_n = o\left(\sqrt{n}\right)$ and $\lambda_n \to \infty$ as $n \to \infty$. Then, under the local model (2),

$$\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) 1_{|T_n^*|>\lambda_n \quad \text{or} \quad |T_n|>\lambda_n} + \mathbb{V}_n^*(\boldsymbol{b}_0)\, 1_{|T_n^*|\leq\lambda_n, |T_n|\leq\lambda_n}$$

converges to the limiting distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ conditionally (on the data) in probability.

ART procedure

ART provides a bootstrap calibration for the test statistic $\sqrt{n}\hat{\theta}_n$ based on a special case of the above theorem. Under H_0 we have the simplification $\mathbb{V}_n^*(\boldsymbol{b}_0) = \mathbb{V}_n^*(\mathbf{0})$. For some nominal level γ, let c_l and c_u be the lower and upper $\gamma/2$ quantiles, respectively, of

$$A_n^* = \sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) 1_{|T_n^*|>\lambda_n \quad \text{or} \quad |T_n|>\lambda_n} + \mathbb{V}_n^*(\mathbf{0})\, 1_{|T_n^*|\leq\lambda_n, |T_n|\leq\lambda_n}.$$

If $\sqrt{n}\hat{\theta}_n$ falls outside the interval $[c_l, c_u]$, then we reject H_0 and conclude that there is at least one significant predictor.

Before applying ART, it is advisable to standardize all the variables X_k and Y (by sample mean and standard deviation), which has the advantage of making the procedure scale invariant ($\hat{\theta}_n$ is then the maximal sample correlation); our results naturally extend, but we develop the theory only for the unstandardized variables to keep the presentation simple.

Robust confidence intervals

The above theorem also allows the construction of a robust confidence interval for θ_n by treating $\boldsymbol{b}_0$ as unknown, then finding the widest bootstrap quantiles over all $\boldsymbol{b}_0$. Here by "robust" we mean asymptotically valid uniformly over $\boldsymbol{b}_0$. For testing purposes, however, this approach would be too conservative and also computationally intensive (grid search over $\mathbb{R}^p$ is needed); for this reason, in ART we set $\boldsymbol{b}_0 = 0$ under the null, so the critical values can be readily computed from A_n^*. In contrast, Laber and Murphy (2013) propose using *almost sure* bounds over their local parameter $\boldsymbol{b}_0$ to find robust confidence intervals for adaptive lasso; this involves less computation than distributional bounds, but is still computationally intensive, and it produces more conservative confidence intervals than the distributional approach.

Choice of the tuning parameter λ_n

The above theorem requires that $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$ as $n \to \infty$. Under this condition, the thresholding provides a consistent pre-test (for $\theta_n = 0$) with asymptotically negligible type I error rate: $\lim_{n\to\infty} P(|T_n| > \lambda_n | \theta_n = 0) = 0$. On the other hand, if λ_n increases too quickly, the pre-test will be conservative. One simple choice would be to set $\lambda_n = \sqrt{a \log n}$, for some constant $a > 0$, but it is also desirable that λ_n increase with p, see Section 5 for discussion about the null limiting behavior of T_n as both p and $n \to \infty$. To that end, note that by Theorem 1 in the special case that ε and $\mathbf{X}$ are independent, under $\theta_n = 0$ (or $\boldsymbol{b}_0 = 0$ and $\boldsymbol{\beta}_0 = 0$) we have $T_n \xrightarrow{d} \tilde{Z}_K$, where $K = arg\,max_{k=1,\ldots,p} \tilde{Z}_k^2$, and $(\tilde{Z_1}, \ldots, \tilde{Z_p})^T$ is a vector of standard normal random variables. Thus, for any fixed $\lambda > 0$,

$$P(|T_n| > \lambda | \theta_n = 0) \to P\left(\max_{k=1,\ldots,p} |\tilde{Z}_k| > \lambda \right) \le \sum_{k=1}^{p} P\left(|\tilde{Z}_k| > \lambda \right).$$

Hence the pre-test type I error rate can be asymptotically controlled below level γ, without sacrificing consistency, by choosing

$$\lambda_n = max\left\{ \sqrt{a \log n}, \text{upper} \quad \gamma/(2p) \text{ - quantile of } \quad N(0,1) \right\}. \quad (5)$$

In the simulation study below we describe a way of specifying the constant a via the double bootstrap, and this is used whenever we refer to ART in the sequel.

Forward stepwise ART

If we find a significant predictor using ART, it would be reasonable to continue applying the procedure in a forward stepwise fashion until no more significant predictors are detected. That is, in successive stages the residual $Y - \hat{\alpha}_n - \hat{\theta}_n X_{\hat{k}_n}$ is treated as a new outcome

variable and marginal regression carried out on the remaining predictors. Although it would be challenging to extend our theoretical results to this procedure, we find that in real data applications it performs well, and in a similar way to the covariance test of Lockhart et al. (2014), as we discuss in the HIV drug resistance example considered in the next section.

4 Numerical studies

In this section, we study the performance of the proposed ART procedure using simulated data, and give illustrations of the approach in two real data examples.

4.1 Finite sample simulations

We compare the performance of ART with four procedures that are commonly used for detecting the presence of a significant predictor:

Likelihood ratio test (LRT)—This test is based on assuming a full linear model involving all of the covariates, and is applicable when $n > p$. Under the null hypothesis, all the regression coefficients are zero. The reduction in the residual sum of squares is compared to the residual sum of squares for the full model using an F-ratio [see, e.g. Section 7.4 of Johnson and Wichern (2007)]. When the full linear model holds, it can be seen that both null and alternative hypotheses are identical to those used in ART.

Multiple testing with Bonferroni correction—As in ART, marginal linear models are used to predict Y from each X_k. A t-test with Bonferroni correction is then carried out to detect whether each regression coefficient is non-zero. The intersection of the p null hypotheses coincides with the null used in ART.

Centered percentile bootstrap (CPB)—This procedure is similar to ART, except $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ is used to estimate the upper and lower quantiles of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$, providing critical values for the test statistic $\sqrt{n}\hat{\theta}_n$, see Efron and Tibshirani (1993).

Higher Criticism (HC)—This is a test originally proposed by John Tukey for determining the overall significance of a collection of independent p-values. We apply the statistic HC_N^+ developed by Donoho and Jin (2004, 2015), which is expected to perform well if the predictors are nearly uncorrelated.

We consider three examples for the data generating model: **i)** $Y = \varepsilon$, **ii)** $Y = X_1/4 + \varepsilon$, and **iii)** $Y = \sum_{k=1}^{p} \beta_k X_k + \epsilon$, where $\beta_1 = \ldots = \beta_5 = 0.15$, $\beta_6 = \ldots = \beta_{10} = -0.1$, and $\beta_k = 0$ for $k = 11, \ldots, p$. In the first example, there is no active predictor, in the second there is a single active predictor, and in the third there are 10 active predictors and the maximally correlated predictor is not unique. The covariate vector $\mathbf{X}$ is distributed as p-dimensional normal with each component $X_k \sim N(0, 1)$, an exchangeable correlation structure $\mathrm{Corr}(X_j, X_k) = \rho$ for $j \neq k$, where ρ takes values 0, 0.5 and 0.8, and the noise $\varepsilon \sim N(0, 1)$ is independent of $\mathbf{X}$.

We consider two sample sizes ($n = 100$ and 200), and five values of the dimension ($p = 10$, 50, 100, 150 and 200). A nominal 5% significance level is used throughout. The bootstrap

sample size is taken as 1,000. To specify the threshold λ_n in ART the double bootstrap is implemented by generating 1,000 bootstrap estimates $\hat{\theta}_n^*$, then choosing λ_n so that 5% of the ARTs (based on 1,000 nested bootstrap samples) with test statistic $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ reject.

Empirical rejection rates based on 1,000 Monte Carlo replications are reported in Figures 1–3. For model **i)**, the figures provide type I error rates, which should be compared with the 5% nominal rate; for models **ii)** and **iii)**, the figures provide the power of each test. The ART procedure has good control of the type I error rate throughout (compared to all the other methods), while consistently maintaining relatively high power. Comparing the results of models **ii)** and **iii)**, non-uniqueness of the maximally correlated predictor has no adverse effect on the power of ART.

Bonferroni is highly conservative when $\rho = 0.5$ and 0.8, see the left panels of Figures 1 and 2. The CPB method is highly anti-conservative, with empirical type I error rates exceeding 15% for both sample sizes (and thus out of range for most of the panels on the left). The LRT effectively controls the type I error rate at around the nominal 5% level when it is applicable, but it has very low power compared with all the other methods, except under model **iii)** in the "classical case" of small numbers of predictors that are not highly correlated, see the right panels of Figures 2 and 3. Higher Criticism fails to control type I error except when the predictors are independent (Figure 3), in which case it is slightly anti-conservative and has excellent power under model **iii)**, but very poor under model **ii)**. That is, HC performs well (under zero correlation) when there are multiple active predictors, but not in the sparse case of only one active predictor. Except in the case of independent predictors, when Bonferroni is slightly better, ART outperforms all the competing procedures when both type I error and power are taken into account, and the improvement increases with the correlation between predictors.

4.2 Asymptotic power

In this section, we carry out a simulation study to assess the asymptotic power of ART compared with that of the Bonferroni procedure. The computational expense of implementing ART is high because of the double bootstrap, so our full simulation study of the previous section is only feasible for small sample sizes. Nevertheless, we are able to assess asymptotic power by making use of our results on the local model in Section 3.

Consider the local model $Y = (n^{-1/2}b_0)X_1 + \varepsilon$, where $b_0 \in \mathbb{R}$. Here $\mathbf{X}$ and ε are generated in the same way as Section 4.1, but now we only consider $\rho = 0.5$. The local parameter $\boldsymbol{b}_0$ takes the special form $(b_0, 0, \ldots, 0)^T$, and we allow b_0 to vary over a grid in [0, 5], in increments of 0.5. We set $\boldsymbol{\beta}_0 = 0$, $\boldsymbol{b}_0 = (b_0, 0, \ldots, 0)^T$ and make use of the given covariance structure of $\mathbf{X}$ and the explicit form of the limiting distribution in Theorem 1 to generate a draws from the asymptotic distribution of $\sqrt{n}\hat{\theta}_n$. Specifically, we carry out the following steps:

1. For each value of b_0 on the grid, take 5,000 draws from the limiting distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ given in Theorem 1 (this distribution only depends on b_0 and the given distribution of $(\mathbf{X}, Y)$), then add b_0 to obtain draws from the limiting

distribution of $\sqrt{n}\hat{\theta}_n$. Based on these draws, we can obtain the (approximate) rejection rate of the test statistic $\sqrt{n}\hat{\theta}_n$ for any given rejection region. In particular, the asymptotic rejection rate of ART (for any given b_0 on the grid) can be calculated by referring to the rejection rate corresponding to the particular critical values c_l and c_u generated by ART.

2. To assess the asymptotic power of ART at each given b_0, we generate 10 independent large samples (with $n = 5{,}000$) from the local model, find c_l and c_u for each sample, and display in a boxplot the corresponding asymptotic rejection rates (using the results of step 1).

3. For comparison, we also plot the asymptotic power of the Bonferroni procedure, which is approximated using 1,000 samples each of size $n = 5{,}000$.

The results are presented in Figure 4 for $p = 10$ and 50. The main source of variation within each boxplot is due to randomness over the 10 independent samples drawn from the local model, rather than bootstrap randomness (in view of bootstrap consistency and the large sample size $n = 5{,}000$). The median of each boxplot provides a suitable reference point to compare with the asymptotic power of Bonferroni (indicated by the circle). Note that ART provides accurate control of asymptotic type I error, and, as expected, Bonferroni is slightly conservative. In terms of median power, ART always outperforms Bonferroni, and can provide an additional 25% power (e.g., at $b_0 = 3$ for $p = 10$, and at $b_0 = 3.5$ for $p = 50$).

The cost of implementing the double bootstrap part of ART makes it prohibitive to extend the results in Figure 4 to larger p, but if we fix λ_n, then it becomes practical to run the simulations for $p = 1000$. Figure 5 shows how the asymptotic power of ART compares with Bonferroni as the constant a used to specify λ_n takes values 2, 4, 5 and 8 (the corresponding λ_n are 4.3, 6.1, 6.8 and 8.6). Note that as a increases (going from one panel to the next), ART becomes more stable and provides more accurate type I error control, but the overall power decreases. At small values of a, ART behaves like the CPB, which is anti-conservative (as we have already seen in the previous section), whereas at larger values the influence of CPB is diluted. For the CPB (which corresponds to setting $\lambda_n = 0$), the plot (not shown) appears very similar to that for $a = 2$; also, for $a > 8$ the plots appear very similar to $a = 8$. The best choice of a, therefore, is a trade-off between type I error control and power; comparing with Figure 4, ART with double bootstrapping appears to achieve a satisfactory balance in this regard. Also note that, even at the largest value $a = 8$, ART can provide an additional 20% power over Bonferroni, and thus outperform Bonferroni by a considerable margin in high-dimensional settings as well, at least when there is a high degree of correlation among the components of $\mathbf{X}$.

4.3 Gene expression example

We consider gene expression profiles from the tumors of $n = 156$ patients diagnosed with a common type of adult brain cancer (glioblastoma), collected as part of the Cancer Genome Atlas pilot project (TCGA, 2008). Our analysis is based on log gene expression levels $\mathbf{X}$ at $p = 181$ loci along chromosome 1. We are interested in detecting the presence of a gene that is significantly related to log-survival time Y.

We compare the results from applying the Bonferroni, CPB and ART procedures; LRT is not applicable since $p > n$. The three methods yield very different p-values. The smallest Bonferroni adjusted p-value is 40.8%, suggesting that no gene is significantly related to Y. The CPB and ART p-values are 3.2% and 17.2%, respectively, from 1000 bootstrap samples. Figure 6 shows how these p-values are calculated. Thus the CPB method suggests the presence of a significant genetic effect, whereas ART does not.

4.4 HIV drug resistance example

Our second example uses data from the HIV Drug Resistance Database (2014), an important public resource for understanding how HIV-1 mutation patterns cause resistance to antiretroviral drugs (Rhee et al., 2002). We will compare our results with those of Lockhart et al. (2014), who applied their covariance test to data on the susceptibility (a measure of drug resistance) of the nucleotide reverse transcriptase inhibitor lamivudine (3TC). We code susceptibility on a log-scale (Y), and each predictor X_j is taken as indicating the presence/absence of a mutation at a given sequence position. The viral sequence positions are indexed by j. Excluding missing data and rare mutations resulted in data on $p = 103$ positions and a total of 1266 isolates.

We randomly split the data 50 times into a training set of size $n = 126$ and a test set of size 1140. For each split, we carry out 20 steps of forward stepwise ART and standard forward stepwise regression using the training data, and calculate the corresponding prediction error (including all previously selected variables) using the test data. The left panel of Figure 7 shows the training data p-values (mean ± SD) for the newly entered predictor at each step, over the 50 random splits, and the right panel shows the corresponding prediction errors (mean ± SD). Forward stepwise ART detects one very highly significant mutation, but no more, as confirmed by the test set error plot, and this result is roughly consistent with the findings of Lockhart et al. (2014). Standard forward stepwise regression picks out at least 10 mutations, but there is no improvement in test set error after the first predictor enters the model; moreover, the test error almost exactly coincides with ART.

5 Discussion

In this paper we have developed an adaptive resampling test (ART) for detecting the existence of a significant predictor, X_{k0}, from among predictors $X_1, \ldots, X_p$. The procedure is designed to adjust to the non-regular limiting behavior of the estimated marginal regression coefficient $\hat{\theta}_n$ of the selected predictor. This is done by using a thresholded version of the bootstrap that adapts to the non-regularity: if there is at least one significant predictor, it reduces to a centered percentile bootstrap, otherwise it mimics the local (non-uniform) asymptotic behavior of $\hat{\theta}_n$. We have shown that in simulation studies, ART performs favorably compared with standard methods such as Bonferroni, but also compared with more sophisticated methods such as Higher Criticism. The advantage of ART may stem from it being designed to take into account correlations between predictors, while also avoiding distributional assumptions (the nonparametric bootstrap steps in ART are essentially distribution free). We have restricted attention to linear models, but our approach

has much wider applicability (e.g., generalized linear models, quantile regression, and censored time-to-event outcomes), and these will be studied in future papers.

Although our simulation results suggest that ART is useful and remarkably stable in "large p, small n" settings, the asymptotic theory that we have used to calibrate ART relies on assuming a fixed p, with n tending to infinity. In view of the conservative nature of the Bonferroni procedure in high-dimensional settings, there is a pressing need for more powerful tests in this area. In future work it would be of interest to develop the asymptotic theory of ART for the case of p growing with n, although this would be very challenging. As far as we know, formal testing procedures that provably control FWER and adjust to non-regularity under diverging p are not yet available, except for Higher Criticism in the case that the predictors are nearly uncorrelated, as established by Ingster et al. (2010) and Arias-Castro et al. (2011). In the only other instance we know of, under the strong assumption that $X_1, \ldots, X_p$, Y are iid $N(0, 1)$, results of Cai and Jiang (2012) can be used to find the weak limit of $\hat{\rho}_n = max_{k=1,\ldots,p} |\widehat{Corr}(X_k, Y)$ and thus devise an asymptotically correct calibration: if $p = p_n \to \infty$ at sub-exponential rate, $\log(p)/n \to 0$, then $\hat{\rho}_n \to_p 0$ and $n\hat{\rho}_n^2 - 2\log p + \log\log p \xrightarrow{d} F$ where $F(y) = e^{-e^{-y/2}/(2\sqrt{\pi})}$. In the super-exponential case, $\log(p)/n \to \infty$, then $\hat{\rho}_n \to_p 1$ and there is a similar weak limit.

Another interesting direction for future work would be to study the forward stepwise version of ART discussed in Section 3. Modifications to ART when applied stepwise in this way would be needed to adjust for the implicit dependence among the new outcomes. By repeating such a procedure until no more significant predictors are detected, the aim would be to correctly identify all active predictors.

Acknowledgments

Research supported by NIH Grant R01GM095722-01 and NSF Grant DMS-1307838.

Appendix: Proofs

Proof of Theorem 1

For $k = 1, \ldots, p$, let $\left(\hat{\alpha}_k, \hat{\theta}_k\right) = arg\, min_{(\alpha,\theta)} \mathbb{P}_n (Y - \alpha - \theta X_k)^2$. Then $\hat{k}_n = arg\, min_{k=1,\ldots,p} \mathbb{P}_n \left(Y - \hat{\alpha}_k - \hat{\theta}_k X_k\right)^2$ and $\left(\hat{\alpha}_n, \hat{\theta}_n\right) = \left(\hat{\alpha}_{\hat{k}_n}, \hat{\theta}_{\hat{k}_n}\right)$. It is easy to verify that $\hat{\alpha}_k = \mathbb{P}_n \left(Y - \hat{\theta}_k X_k\right)$,

$$\begin{aligned} \sqrt{n}\hat{\theta}_k &= \frac{\sqrt{n}\widehat{Cov}(X_k, Y)}{\widehat{Var}(X_k)} \\ &= \frac{\sqrt{n}\widehat{Cov}\left(X_k, \mathbf{X}^T\right)\boldsymbol{\beta}_n + \mathbb{G}_n[\epsilon(X_k - \mathbb{P}_n X_k)]}{\widehat{Var}(X_k)} \\ &= \frac{\left(\mathbb{G}_n X_k \mathbf{X}^T - P_n X_k \mathbb{G}_n \mathbf{X}^T - \mathbb{G}_n X_k \mathbb{P}_n \mathbf{X}^T\right)\beta}{n} \widehat{Var}(X_k) + \frac{\mathbb{G}_n[\epsilon(X_k - P_n X_k)] - \mathbb{P}_n \epsilon \mathbb{G}_n X_k + \sqrt{n} Cov\left(X_k, \mathbf{X}^T\right)\boldsymbol{\beta}_n}{\widehat{Var}(X_k)}; \end{aligned} \tag{6}$$

where P_n is the distribution of $(Y, \mathbf{X})$, and the mean residual squared error

$$\hat{R}_k \equiv \mathbb{P}_n\left[Y - \hat{\alpha}_k - \hat{\theta}_k X_k\right]^2 = \widehat{Var}(Y) - \widehat{Var}(X_k)\,\hat{\theta}_k^2. \quad (7)$$

The result then follows immediately from the following two lemmas. The first lemma verifies the oracle property for marginal regression under the assumption that there is at least one active predictor; the proof is included for completeness. The second lemma gives the (non-regular) asymptotic behavior of $\hat{\theta}_n$ when there are no active predictors.

Lemma 1

If all conditions in Theorem 1 hold and $\boldsymbol{\beta}_0 \neq \mathbf{0}$, then $\hat{k}_n \xrightarrow{a.s.} k_0$ and $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) \xrightarrow{d} Z_{k_0}(\boldsymbol{\beta}_0)/V_{k_0}$, where Z_{k_0} is defined in Theorem 1.

Proof—Denote $\hat{\mathbf{R}} \equiv (\hat{R_1}, \ldots, \hat{R_p})^T$. When $\boldsymbol{\beta}_0 \neq 0$, $\mathrm{Var}(\mathbf{X}^T\boldsymbol{\beta}_0) > 0$. By the SLLN

$$\frac{\widehat{Var}(Y) - \hat{\mathbf{R}}}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)} \xrightarrow{a.s.} \left(Corr^2\left(X_1, \mathbf{X}^T\beta_0\right), \ldots, Corr^2\left(X_p, \mathbf{X}^T\boldsymbol{\beta}_0\right)\right)^T.$$

Since $\hat{k}_n = arg\,max_{k=1,\ldots,p}\left[\widehat{Var}(Y) - \hat{R}_k\right]/Var\left(\mathbf{X}^T\boldsymbol{\beta}_0\right)$ and $\mathrm{Corr}^2(X_k, \mathbf{X}^T\boldsymbol{\beta}_0)$ is maximized at $k = k_0$, it follows immediately that $\hat{k}_n \xrightarrow{a.s.} k_0$.

Next, denote $\hat{X} = X_{\hat{k}_n}$ and $X_n = X_{k_n}$. Since $\mathbb{P}_n\left[Y - \mathbb{P}_n Y - \hat{\theta}_n\left(\hat{X} - \mathbb{P}_n\hat{X}\right)\right]\hat{X} = 0$ and $Y = \alpha_0 + \mathbf{X}^T\boldsymbol{\beta}_n + \varepsilon$, we have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)\widehat{Var}\left(\hat{X}\right) &= \sqrt{n}\widehat{Cov}\left(\hat{X}, \mathbf{X}^T\right)\boldsymbol{\beta}_n + \sqrt{b}\mathbb{P}_n\left(\epsilon\left(\hat{X} - \mathbb{P}_n\hat{X}\right)\right) - \sqrt{n}\widehat{Var}\left(\hat{X}\right)\frac{Cov(X_n, \mathbf{X})^T\boldsymbol{\beta}_n + Cov(X_n, \epsilon)}{Var(X_n)} \\
&= \sqrt{n}\widehat{Cov}\left(X_{k_0}, \mathbf{X}^T\right)\boldsymbol{\beta}_n + \sqrt{n}\mathbb{P}_n\left(\epsilon\left(X_{k_0} - \mathbb{P}_n X_{k_0}\right)\right) - \sqrt{n}\widehat{Var}\left(X_{k_0}\right)\frac{Cov\left(X_{k_0}, \mathbf{X}\right)^T\boldsymbol{\beta}_n + Cov\left(X_{k_0}, \epsilon\right)}{Var\left(X_{k_0}\right)} + o_{P_n}(\\
&= \mathbb{G}_n\left[\left(\epsilon + (\mathbf{X} - P_n\mathbf{X})^T\boldsymbol{\beta}_0 - \frac{Cov\left(X_{k_0}, \mathbf{X}\right)^T\boldsymbol{\beta}_0}{Var\left(X_{k_0}\right)}\left(X_{k_0} - P_n X_{k_0}\right)\right)\left(X_{k_0} - P_n X_{k_0}\right)\right] + o_{P_n}(1),
\end{aligned}$$

where the second equality uses $\hat{k}_n \xrightarrow{a.s.} k_0$ and $k_n \to k_0$ as $n \to \infty$, and the third equality follows from the LLN and $\mathrm{Cov}(\varepsilon, X_{k_0}) = 0$. Similarly, $\widehat{Var}\left(\hat{X}\right) \xrightarrow{P_n} V_{k_0} \equiv Var(X_{k_0})$. The proof is completed using Slutsky's lemma and the CLT.

Lemma 2

If all conditions in Theorem 1 hold and $\boldsymbol{\beta}_0 = \mathbf{0}$, *then*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) \xrightarrow{d} Z_K(\mathbf{0})/V_K + \left(C_K/V_K - C_{-k(\boldsymbol{b}_0)}/V_{-k(\boldsymbol{b}_0)}\right)^T \boldsymbol{b}_0.$$

Proof—Since $(Z_1(\mathbf{0}), \ldots, Z_p(\mathbf{0}))^T$ is a normal random vector and $|\mathrm{Corr}(X_j, X_k)| < 1$ for $j \neq k$, it is easy to see that

$$\frac{\left(Z_j(\mathbf{0})+C_j^T\boldsymbol{b}_0\right)^2}{V_j} \neq \frac{\left(Z_k(\mathbf{0})+C_k^T\boldsymbol{b}_0\right)^2}{V_k} \quad \text{for any} \quad j \neq k \quad \text{a.s.} \quad (8)$$

So K is unique a.s.

Denote $\hat{\boldsymbol{\theta}}=\left(\hat{\theta}_1,\ldots,\hat{\theta}_p\right)^T$. Note that when $\boldsymbol{\beta}_0 = 0$, $\sqrt{n}\boldsymbol{\beta}_n=\boldsymbol{b}_0$. By the CLT and Slutsky's lemma, we see from (6) that

$$\sqrt{n}\hat{\boldsymbol{\theta}} \xrightarrow{d} \left(\frac{Z_1(\mathbf{0})+C_1^T\boldsymbol{b}_0}{V_1},\ldots,\frac{Z_p(\mathbf{0})+C_p^T\boldsymbol{b}_0}{V_p}\right)^T.$$

From (7), we have

$$n\left[\widehat{Var}(Y)-\hat{\mathbf{R}}\right]=\left(\sqrt{n}\hat{\theta}\right)\odot\left(\sqrt{n}\hat{\theta}\right)\odot\left(\widehat{Var}(X_1),\ldots,\widehat{Var}(X_p)\right)^T,$$

where $\odot$ denotes the elementwise (Hadamard) product, so, by the continuous mapping theorem and Slutsky's lemma,

$$\begin{pmatrix} \sqrt{n}\hat{\theta} \\ n\left[\widehat{Var}(Y)-\hat{\mathbf{R}}\right] \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \left(\left(Z_1(\mathbf{0})+C_1^T\boldsymbol{b}_0\right)/V_1,\ldots,\left(Z_p(\mathbf{0})+C_p^T\boldsymbol{b}_0\right)/V_p\right)^T \\ \left(\left(Z_1(\mathbf{0})+C_1^T\boldsymbol{b}_0\right)^2/V_1,\ldots,\left(Z_p(\mathbf{0})+C_p^T\boldsymbol{b}_0\right)^2/V_p\right)^T \end{pmatrix}.$$

Define $h(\mathbf{t}) = (1_{\arg\max_k t_k=1},\ldots,1_{\arg\max_k t_k=p})^T$, where $\mathbf{t}=(t_1,\ldots,t_p)^T \in \mathbb{R}^p$. Note that h is continuous at $\mathbf{t}$ if $\arg\max_k t_k$ is unique. Thus, using (8) and since $\sqrt{n}\hat{\theta}_n=\sqrt{n}\hat{\theta}^T h\left(n\left[\widehat{Var}(Y)-\hat{\mathbf{R}}\right]\right)$, the result follows by applying the continuous mapping theorem to the above display.

Lemma 3

Let $\mathbf{Z}$ *be a p-dimensional random vector and* $f:\mathbb{R}^{2p}\to\mathbb{R}^p$ *a function such that* $f(\mathbf{z},\cdot)$ *is continuous for every* $\mathbf{z}\in\mathbb{R}^p$, *and* $f(\mathbf{Z},\boldsymbol{b})_j \neq f(\mathbf{Z},\boldsymbol{b})_k$ a.s. for all $j\neq k$ and $\boldsymbol{b}\in\mathbb{R}^p$. *Then* $K(\boldsymbol{b}) \equiv \arg\max_{k=1,\ldots,p} f(\mathbf{Z},\boldsymbol{b})_k$ *is unique a.s. Also, if* $\boldsymbol{b}_l\to\boldsymbol{b}_0$, *then* $K(\boldsymbol{b}_l)=K(\boldsymbol{b}_0)$ *for l sufficiently large a.s.*

The proof is omitted. An immediate consequence of this lemma is the continuity of the limiting distribution in Theorem 1 as a function of $\boldsymbol{b}_0$; this is seen by setting $f(z_1,\ldots,z_p,\boldsymbol{b})_k=\left(z_k+C_k^T\boldsymbol{b}\right)^2/V_k$ for $k=1,\ldots,p$, and using (8).

Proof of Theorem 2

The notation $\hat{\theta}_n^*$ and $\hat{k}_n^*$ means that $\hat{\theta}_n$ and $\hat{k_n}$ are based on n iid observations taken from $\mathbb{P}_n$. The bootstrapped process $\mathbb{V}_n^*(\mathbf{b})$ in the statement of the theorem is defined by re-expressing (4), along with $K(\mathbf{b})$ and $\mathbb{K}_n(\mathbf{b})$, in terms of P_n and $\mathbb{P}_n$ operating on functions of ($\mathbf{X}$, Y), then replacing P_n by $\mathbb{P}_n$ and $\mathbb{P}_n$ by $\mathbb{P}_n^*$ throughout. In the case of $\mathbb{Z}_{n,k}$ in which ε is not observed, we also replace ε by $\hat{\epsilon}_n = \hat{\epsilon}_n(\mathbf{X}, Y) \equiv Y - \hat{\alpha}_n - \hat{\theta}_n \hat{X}$, resulting in

$$\mathbb{Z}_{n,k}^* = \mathbb{G}_n^*[\hat{\epsilon}_n(X_k - \mathbb{P}_n^* X_k)] = \mathbb{G}_n^*[\hat{\epsilon}_n X_k] - [\mathbb{G}_n^* \hat{\epsilon}_n][\mathbb{P}_n^* X_k] \quad (9)$$

where $\mathbb{G}_n^* = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ is the bootstrapped empirical process. As is conventional in empirical process theory, $\mathbb{P}_n^*$, $\mathbb{P}_n$ and P_n are assumed to operate only on functions that are defined on ($\mathbf{X}$, Y), explaining why $\mathbb{P}_n^* X_k$ can be separated in the above display.

Let E^M denote expectation conditional on the data, and let P^M be the corresponding probability measure. We will show that $1_{|T_n^*|>\lambda_n \text{ or } |T_n|>\lambda_n} \xrightarrow{P^M} 1_{\beta_0 \neq \mathbf{0}}$ and $1_{|T_n^*|\leq\lambda_n} 1_{|T_n|\leq\lambda_n} \xrightarrow{P^M} 1_{\beta_0=\mathbf{0}}$ conditionally (on the data) in probability. This together with Lemmas 4 and 5 below implies the result.

For $k = 1, \ldots, p$, the bootstrapped marginal regression coefficient $\hat{\theta}_k^*$ satisfies

$$\begin{aligned}
\sqrt{n}\hat{\theta}_k^* &= \frac{\sqrt{n}[\mathbb{P}_n^* X_k Y - (\mathbb{P}_n^* X_k)(\mathbb{P}_n^* Y)]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2} \\
&= \frac{\mathbb{P}_n^* X_k Y - \mathbb{P}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{P}_n^* Y) + \sqrt{n}[\mathbb{P}_n X_k Y - (\mathbb{P}_n X_k)(\mathbb{P}_n Y)]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2} \\
&= \frac{\mathbb{P}_n^* X_k Y - \mathbb{P}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{P}_n^* Y) + \sqrt{n}\hat{\theta}_k[\mathbb{P}_n X_k^2 - (\mathbb{P}_n X_k)^2]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2}.
\end{aligned} \quad (10)$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, by Lemma 2 and the condition that $\lambda_n \to \infty$ as $n \to \infty$, we have $T_n^*/\lambda_n \xrightarrow{P^M} 0$ in probability. When $\boldsymbol{\beta}_0 \neq \mathbf{0}$, it is easy to verify that $|\theta_n| \to |C_{k_0}^T \boldsymbol{\beta}_0|/V_{k_0}$, which is positive under the condition that k_0 is unique. Thus

$$\begin{aligned}
P^M(|T_n^*| \leq \lambda_n) &= P^M\left(\left|\left(\hat{\theta}_n^* - \hat{\theta}_n\right) + \left(\hat{\theta}_n - \theta_n\right) + \theta_n\right| \leq \lambda_n s_n^*\right) \\
&\leq P^M\left(|\theta_n| \leq \lambda_n s_n^* + |\hat{\theta}_n^* - \hat{\theta}_n| + |\hat{\theta}_n - \theta_n|\right)
\end{aligned}$$

tends to zero in probability when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, where the convergence follows from Lemma 1, Lemma 4 (below) and the condition that $\lambda_n = o(\sqrt{n})$. Hence

$$\begin{aligned}
E^M|1_{|T_n^*|\leq\lambda_m} - 1_{\boldsymbol{\beta}_0=\mathbf{0}}| &= E^M|1_{|T_n^*|>\lambda_n} - 1_{\boldsymbol{\beta}_0\neq\mathbf{0}}| \\
&= P^M(|T_n^*|>\lambda_n, \boldsymbol{\beta}_0=\mathbf{0}) + P^M(|T_n^*| \leq \lambda_n, \boldsymbol{\beta}_0 \neq \mathbf{0}) \\
&= P^M(|T_n^*|>\lambda_n | \boldsymbol{\beta}_0=\mathbf{0}) 1_{\boldsymbol{\beta}_0=\mathbf{0}} + P^M(|T_n^*| \leq \lambda_n | \boldsymbol{\beta} \neq \mathbf{0}) 1_{\boldsymbol{\beta}_0\neq\mathbf{0}}
\end{aligned}$$

tends to zero in probability. This implies that $1_{|T_n^*|>\lambda_n} \xrightarrow{P^M} 1_{\boldsymbol{\beta}_0 \neq \mathbf{0}}$ and $1_{|T_n^*|\leq\lambda_n} \xrightarrow{P^M} 1_{\boldsymbol{\beta}_0=\mathbf{0}}$ conditionally in probability. Since $1_{|T_n|\leq\lambda_n}$ converges to $1_{\boldsymbol{\beta}_0=\mathbf{0}}$ in probability, the result follows from Slutsky's lemma.

Lemma 4

If the conditions in Theorem 1 hold and $\boldsymbol{\beta}_0 \neq \mathbf{0}$, *then* $\hat{k}_n^* \xrightarrow{P^M} k_0$ *conditionally (on the data) a.s. and* $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) \xrightarrow{d} Z_{k_0}(\boldsymbol{\beta}_0)/V_{k_0}$ *conditionally (on the data) in probability.*

Proof—It follows from (10), the SLLN and Slutsky's lemma that, when $\boldsymbol{\beta}_0 \neq \mathbf{0}$,

$$\widehat{Var}^*(X_k)\,\hat{\theta}_k^* = n^{-1/2}\left[\mathbb{G}_n^* X_k Y - \mathbb{G}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{G}_n^* Y)\right] + \hat{\theta}_k\left[\mathbb{P}_n X_k^2 - (\mathbb{P}_n X_k)^2\right] \xrightarrow{P^M} C_k^T \boldsymbol{\beta}_0$$

and $\hat{\theta}_k^* \xrightarrow{P^M} C_k^T \boldsymbol{\beta}_0 / V_k$ a.s. for $k = 1, \ldots, p$. Denote the bootstrap mean squared error

$$\hat{R}_k^* \equiv \mathbb{P}_n^*\left[Y - \hat{\alpha}_k^* - \hat{\theta}_k^* X_k\right]^2 = \widehat{Var}^*(Y) - \left(\hat{\theta}_k^*\right)^2 \widehat{Var}^*(X_k),$$

where $\widehat{Var}^*(Y) = \mathbb{P}_n^* Y^2 - (\mathbb{P}_n^* Y)^2$ and $\widehat{Var}^*(X_k) = \mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2$. Then we can write

$$\hat{k}_n^* = arg\max_{k=1,\ldots,p} \frac{\widehat{Var}^*(Y) - \hat{R}_k^*}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)} = arg\max_{k=1,\ldots,p} \frac{\left(\hat{\theta}_k^*\right)^2 \widehat{Var}^*(X_k)}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)}$$

since the denominator plays no role. By Slutsky's lemma

$$\frac{\left(\hat{\theta}_k^*\right)^2 \widehat{Var}^*(X_k)}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)} \xrightarrow{P^M} Corr^2\left(X_k, \mathbf{X}^T\boldsymbol{\beta}_0\right)$$

a.s. for $k = 1, \ldots, p$, so we obtain

$$\begin{aligned} P^M\left(\hat{k}_n^* \neq k_0\right) &= P^M\left(\bigcup_{k:k\neq k_0}\left\{\frac{\left(\hat{\theta}_{k_0}^*\right)^2 \widehat{Var}^*(X_{k_0})}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)} \leq \frac{\left(\hat{\theta}_k^*\right)^2 \widehat{Var}^*(X_k)}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)}\right\}\right) \\ &\leq \sum_{k:k\neq k_0} P^M\left(\frac{\left(\hat{\theta}_{k_0}^*\right)^2 \widehat{Var}^*(X_{k_0})}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)} \leq \frac{\left(\hat{\theta}_k^*\right)^2 \widehat{Var}^*(X_{k_0})}{Var(\mathbf{X}^T\boldsymbol{\beta}_0)}\right) \\ &\to 0 \quad \text{a.s.}, \end{aligned}$$

where the convergence follows from the condition that k_0 is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$.

Recall that $\hat{\epsilon}_n \equiv Y - \hat{\alpha}_n - \hat{\theta}_n \hat{X}$, where $\hat{X} \equiv X_{\hat{k}_n}$. Note that $\mathbb{P}_n \hat{\epsilon}_n = 0$. By the definition of $\hat{\theta}_n^*$, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n^*-\hat{\theta}_n\right)\left[\mathbb{P}_n^*X_{\hat{k}_n^*}^2-\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right)^2\right] &= \sqrt{n}\left[\mathbb{P}_n^*X_{\hat{k}_n^*}Y-\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right)\left(\mathbb{P}_n^*Y\right)-\hat{\theta}_n\left(\mathbb{P}_n^*X_{\hat{k}_n^*}^2-\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right)^2\right)\right]\\ &= \sqrt{n}\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\hat{\epsilon}_n-\mathbb{P}_n^*X_{\hat{k}_n^*}\mathbb{P}_n^*\hat{\epsilon}_n\right)+\sqrt{n}\hat{\theta}_n\left[\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right)^2-\mathbb{P}_n^*X_{\hat{k}_n^*}^2+\mathbb{P}_n^*X_{\hat{k}_n^*}\hat{X}-\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right.\right.\\ &= \mathbb{G}_n^*\hat{\epsilon}_n\left(X_{\hat{k}_n^*}-P_nX_{\hat{k}_n^*}\right)-\mathbb{G}_n^*X_{\hat{k}_n^*}\left(\mathbb{P}_n^*-\mathbb{P}_n\right)\hat{\epsilon}_n-\mathbb{G}_n^*\hat{\epsilon}_n\left(\mathbb{P}_n-P_n\right)X_{\hat{k}_n^*}+\sqrt{n}\hat{\theta}_n\left[\left(\mathbb{P}_n^*X\right.\right.\end{aligned}$$

The last term in (11) is $o_{P^M}(1)$ a.s. because the first and last terms within the square bracket cancel asymptotically, similarly for the second and third terms, due to $\hat{k}_n^* \xrightarrow{P^M} k_0$ and $\hat{k_n} \to k_0$ a.s. We next show that the first term in (11) converges in distribution to $Z_{k_0}(\boldsymbol{\beta}_0)$ conditionally (on the data) in probability. By Lemma 1, it is easy to verify that $\hat{\theta}_n \xrightarrow{P_n} \theta_0 \triangleq C_{k_0}^T\boldsymbol{\beta}_0/V_{k_0}$ and $\hat{\alpha}_n \xrightarrow{P_n} \alpha_0+E\mathbf{X}^T\boldsymbol{\beta}_0-\theta_0EX_{k_0}$. Denote $\bar{\epsilon}=\epsilon+(\mathbf{X}-E\mathbf{X})^T\boldsymbol{\beta}_0-\theta_0(X_{k_0}-EX_{k_0})$. Then the first term can be decomposed as

$$\mathbb{G}_n^*\hat{\epsilon}_n\left[\left(X_{\hat{k}_n^*}-P_nX_{\hat{k}_n^*}-(X_{k_0}-P_nX_{k_0})\right]+\mathbb{G}_n^*\left[\left(\hat{\epsilon}_n-\bar{\epsilon}\right)(X_{k_0}-P_nX_{k_0})\right]+\mathbb{G}_n^*\left[\bar{\epsilon}\,(X_{k_0}-P_nX_{k_0})\right]. \quad (12)$$

The first term in (12) is $o_{P^M}(1)$ a.s. since $\hat{k}_n^* \xrightarrow{P^M} k_0$. The second term in (12) can be written as

$$\begin{aligned}&\left[\left(\alpha_0+E\mathbf{X}^T\boldsymbol{\beta}_0-\theta_0EX_{k_0}\right)-\hat{\alpha}_n\right]\mathbb{G}_n^*\left(X_{k_0}-P_nX_{k_0}\right)\\ &+\left(\mathbb{P}_n^*-\mathbb{P}_n\right)\left[\left(X_{k_0}-P_nX_{k_0}\right)\mathbf{X}^T\boldsymbol{b}_0\right]\\ &+\left(\theta_0-\hat{\theta}_n\right)\mathbb{G}_n^*\left[X_{k_0}\left(X_{k_0}-P_nX_{k_0}\right)\right]\\ &-\hat{\theta}_n\mathbb{G}_n^*\left[\left(\hat{X}-X_{k_0}\right)\left(X_{k_0}-P_nX_{k_0}\right)\right]\end{aligned}$$

which is $o_{P^M}(1)$ in probability by bootstrap consistency of the sample mean [see, e.g., Theorem 23.4 of van der Vaart (1998)], and the fact that $\hat{X}=X_{k_0}$ for n sufficiently large a.s. Bootstrap consistency of the sample mean also gives that the third term in (12) converges in distribution to $Z_{k_0}(\boldsymbol{\beta}_0)$ conditionally (on the data) in probability.

Similarly, the second and third terms in (11) and $\mathbb{P}_n^*X_{\hat{k}_n^*}^2-\left(\mathbb{P}_n^*X_{\hat{k}_n^*}\right)^2-Var\left(X_{k_0}\right)$ can be shown to be $o_{P^M}(1)$ in probability. The result then follows from Slutsky's lemma.

Lemma 5

If all conditions in Theorem 1 hold and $\boldsymbol{\beta}_0=\mathbf{0}$, *then* $\mathbb{V}_n^*(\boldsymbol{b}_0)$ *converges to the same limiting distribution as* $\sqrt{n}\left(\hat{\theta}_n-\theta_n\right)$ *conditionally (on the data) in probability.*

Proof—Define $\mathbb{Z}_n$, $\mathbb{M}_n(\boldsymbol{b})$ and $M(\boldsymbol{b})$ to be p-vectors with kth components given by $\mathbb{Z}_{n,k}=\mathbb{G}_n\left[\epsilon\left(X_k-\mathbb{P}_nX_k\right)\right]$,

$$\frac{\left[\widehat{Cov}\left(X_k, \mathbf{X}^T\boldsymbol{b}\right) + \mathbb{Z}_{n,k}\right]^2}{\widehat{Var}\left(X_k\right)} \quad \text{and} \quad \frac{\left[Cov\left(X_k, \mathbf{X}^T\boldsymbol{b}\right)\right]^2}{Var\left(X_k\right)},$$

respectively. Let $\mathbb{W}_n(\boldsymbol{b})$ be a $p \times p$ matrix with the (j, k)-th component given by

$$\frac{\widehat{Cov}\left(X_k, \mathbf{X}^T\boldsymbol{b}\right) + \mathbb{Z}_{n,k}}{\widehat{Var}\left(X_k\right)} - \frac{Cov\left(X_j, \mathbf{X}^T\boldsymbol{b}\right)}{Var\left(X_j\right)}.$$

Also, let $\mathbb{D}_n(\boldsymbol{b})$ and $D(\boldsymbol{b})$ be p-vectors of zeros, apart from a 1 in the entry that maximizes $\mathbb{M}_n(\boldsymbol{b})$ and $M(\boldsymbol{b})$, respectively. Then

$$\mathbb{V}_n(\boldsymbol{b}) = D'(\boldsymbol{b})^T \mathbb{W}_n(\boldsymbol{b}) \mathbb{D}_n(\boldsymbol{b}).$$

Similarly, define $\mathbb{M}(\boldsymbol{b})$, $\mathbb{W}(\boldsymbol{b})$ and $\mathbb{D}(\boldsymbol{b})$ (without indexing by n) to be processes of the same form as $\mathbb{M}_n(\boldsymbol{b})$, $\mathbb{W}_n(\boldsymbol{b})$ and $\mathbb{D}_n(\boldsymbol{b})$, except with $\mathbb{Z}_{n,k}$ replaced by $Z_k(\mathbf{0})$, and the sample variances/covariances replaced by their population versions.

Referring to the notation in (4), it is clear that when $\boldsymbol{\beta}_0 = \mathbf{0}$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) = \mathbb{V}_n(\boldsymbol{b}_0) = D'(\boldsymbol{b}_0)^T \mathbb{W}_n(\boldsymbol{b}_0) \mathbb{D}_n(\boldsymbol{b}_0) \xrightarrow{d} D'(\boldsymbol{b}_0)^T \mathbb{W}(\boldsymbol{b}_0) \mathbb{D}(\boldsymbol{b}_0).$$

Moreover, the second equality in the above display also holds for the bootstrap version. Writing the bootstrapped version of $\mathbb{Z}_{n,k}$ in (9) as

$$\mathbb{Z}^*_{n,k} = \mathbb{G}^*_n\left[\epsilon\left(X_k - P_n X_k\right)\right] + \mathbb{G}^*_n\left[\left(\hat{\epsilon}_n - \epsilon\right)\left(X_k - P_n X_k\right)\right] + \left[\left(P_n - \mathbb{P}^*_n\right) X_k\right] \mathbb{G}^*_n \hat{\epsilon}_n],$$

and using arguments similar to those in the proof Lemma 4 for handling (12), we have $\mathbb{Z}^*_n \xrightarrow{d} (Z_1(\mathbf{0}), \ldots, Z_p(\mathbf{0}))^T$ conditionally (on the data) in probability. As a result, $\left(\hat{D}'_n(\boldsymbol{b}_0), \mathbb{W}^*_n(\boldsymbol{b}_0), \mathbb{M}^*_n(\boldsymbol{b}_0)\right) \xrightarrow{d} \left(D'(\boldsymbol{b}_0), \mathbb{W}(\boldsymbol{b}_0), \mathbb{M}(\boldsymbol{b}_0)\right)$ conditionally (on the data) in probability, where $\hat{D}'_n(\boldsymbol{b})$ is the sample version of $D(\boldsymbol{b})$, and $\mathbb{W}^*_n(\boldsymbol{b})$ and $\mathbb{M}^*_n(\boldsymbol{b})$ are the bootstrap versions of $\mathbb{W}_n(\boldsymbol{b})$ and $\mathbb{M}_n(\boldsymbol{b})$, respectively. Finally, using similar arguments to those at the end of the proof of Lemma 2, along with the continuous mapping theorem, we conclude that

$$\mathbb{V}^*_n(\boldsymbol{b}_0) = \hat{D}'_n(\boldsymbol{b}_0)^T \mathbb{W}^*_n(\boldsymbol{b}_0) \mathbb{D}^*_n(\boldsymbol{b}_0) \xrightarrow{d} D'(\boldsymbol{b}_0)^T \mathbb{W}(\boldsymbol{b}_0) \mathbb{D}(\boldsymbol{b}_0)$$

conditionally (on the data) in probability.

References

Andrews D. Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space. Econometrica. 2000; 68(2):399–405.

Arias-Castro E, Candès EJ, Plan Y. Global Testing Under Sparse Alter natives: ANOVA, Multiple Comparisons and the Higher Criticism. Ann. Statist. 2011; 39:2533–2556.

Belloni A, Chernozhukov V, Hansen C. Inference on Treatment Effects After Selection Amongst High-Dimensional Controls. Review of Economic Studies. 2014; 81(2):608–650.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A prac tical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Ser. B. 1995; 57:289–300.

Berk R, Brown LD, Buja A, Zhang K, Zhao L. Valid Post-Selection Inference. Annals of Statistics. 2013; 41:802–837.

Breiman L. The Little Bootstrap and Other Methods for Dimensionality Se lection in Regression: X-Fixed Prediction Error. Journal of the American Statistical Association. 1992; 87:738–754.

Bühlmann P. Statistical Significance in High-dimensional Linear Models. Bernoulli. 2013; 19:1212–1242.

Cai TT, Jiang T. Phase Transition in Limiting Distributions of Co herence of High-dimensional Random Matrices. Journal of Multivariate Analysis. 2012; 107:24–39.

Cancer Genome Atlas Research Network. Comprehensive Genomic Charac terization Defines Human Glioblastoma Genes and Core Pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]

Chatterjee A, Lahiri SN. Bootstrapping Lasso Estimators. Journal of the American Statistical Association. 2011; 106(494):608–625.

Cheng, X. Robust Confidence Intervals in Nonlinear Regression under Weak Identification. Department of Economics, University of Pennsylvania; 2008. Unpublished ManuscriptVersion posted in 2015: http://www.sas.upenn.edu/xucheng/papers/Cheng mixed id 19.pdf

Cheng X. Robust Inference in Nonlinear Models with Mixed Identification Strength. Journal of Econometrics. 2015 to appear.

Davies RB. Hypothesis Testing when a Nuisance Parameter Is Present Only under the Alternative. Biometrika. 1977; 64(2):247–254.

Donoho D, Jin J. Higher Criticism for Detecting Sparse Heterogeneous Mixtures. Annals of Statistics. 2004; 32(3):962–994.

Donoho D, Jin J. Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. Statistical Science. 2015; 30(1):1–25.

Dudoit S, Shaffer JP, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science. 2003; 18:71–103.

Dudoit, S.; van der Laan, MJ. Multiple Testing Procedures with Appli cations to Genomics. Springer; New York: 2008.

Efron B. Large-scale Simultaneous Hypothesis Testing: the Choice of a Null Hypothesis. Journal of American Statistical Association. 2006; 99:96–104.

Efron, B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press; 2010.

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability; 1993.

Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan, J.; Li, R. Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery.. In: Sanz-Sole, M.; Soria, J.; Varona, JL.; Verdera, J., editors. Proceedings of the International Congress of Mathematicians. Vol. III. European Mathematical Society; Zurich: 2006. p. 595-622.

Fan J, Lv J. Sure Independence Screening for Ultra-high Dimensional Feature Space. Journal of the Royal Statistical Society, Ser. B. 2008; 70:849–911. with discussion.

Genovese C, Jin J, Wasserman L, Yao Z. A Comparison of the Lasso and Marginal Regression. Journal of Machine Learning Research. 2012; 13:2107–2143.

HIV Drug Resistance Database. Genotype-Phenotype Datasets, Stanford Uni versity. 2014. http://hivdb.stanford.edu/pages/genopheno.dataset.html

Huang J, Ma S, Zhang C-H. Adaptive Lasso for High-dimensional Regression Models. Statistica Sinica. 2008; 18:1603–1618.

Ingster YI, Tsybakov AB, Verzelen N. Detection Boundary in Sparse Regression. Electron. J. Statist. 2010; 4:1476–1526.

Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. 6th Edition. Prentice Hall; New Jersey: 2007.

Laber E, Murphy SA. Adaptive Confidence Intervals for the Test Error in Classification (with discussion). Journal of the American Statistical Association. 2011; 106(495):904–913. [PubMed: 22053123]

Laber E, Murphy SA. Adaptive Inference after Model Selection. 2013 Under review.

Laber E, Lizotte D, Qian M, Murphy SA. Dynamic Treatment Regimes: Technical Challenges and Applications. Electronic Journal of Statistics. 2014; 8:1225–1272. [PubMed: 25356091]

Leeb H, Pötscher BM. Can One Estimate the Conditional Distribution of Post-model-selection Estimators? Annals of Statistics. 2006; 34(5):2554–2591.

Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A Significance Test for the Lasso. Annals of Statistics. 2014; 42(2):413–468. [PubMed: 25574062]

McCloskey, A. Bonferroni-based Size-correction for Nonstandard Test ing Problems. 2012. Working Paper, http://www.econ.brown.edu/fac/adam mccloskey/Research files/McCloskey BBCV.pdf

Meinshausen N, Meier L, Bühlmann P. P-values for High-dimensional Regression. Journal of the American Statistical Association. 2009; 104:1671–1681.

Ning, Y.; Liu, H. A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models. 2015. http://arxiv.org/abs/1412.8765

Rhee S-Y, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. 2003; 31(1):298–303. [PubMed: 12520007]

Samworth R. A Note on Methods of Restoring Consistency to the Bootstrap. Biometrika. 2003; 90:985–990.

van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; 1998.

Zhang C-H, Zhang S. Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. J. R. Stat. Soc. B. 2014; 76:217–242.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005; 67(2):301–320.

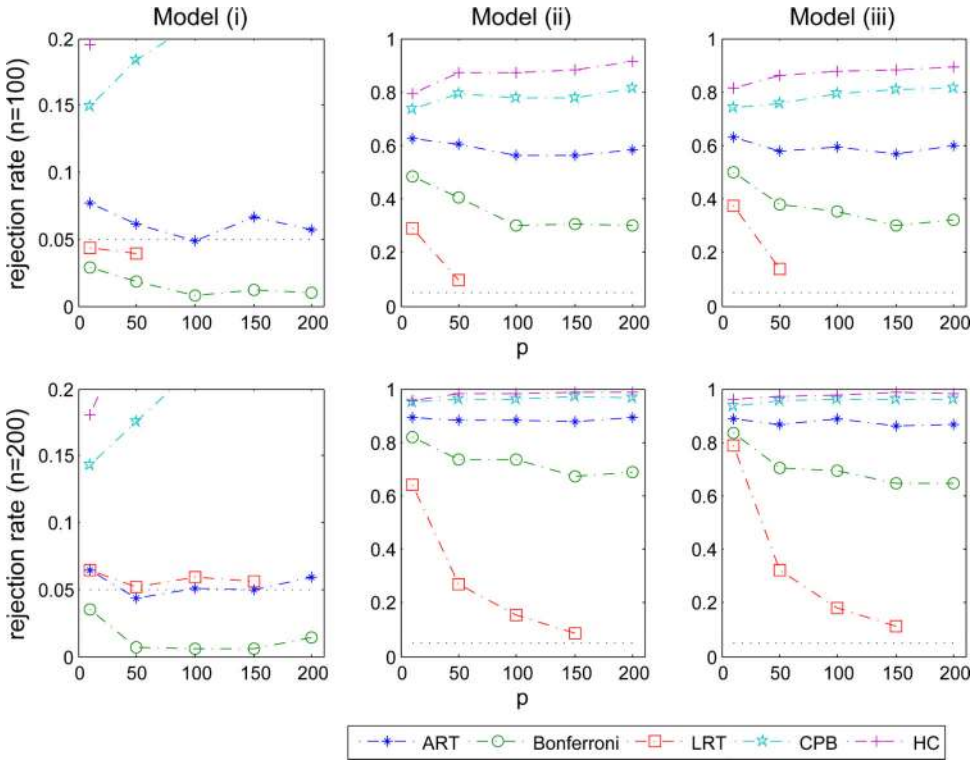

Figure 1.

Empirical rejection rates based on 1,000 samples generated from models **i), ii)** and **iii)** as the dimension ranges from $p = 10$ to $p = 200$, for $n = 100$ (top row) and $n = 200$ (bottom row), and $\rho = 0.8$.

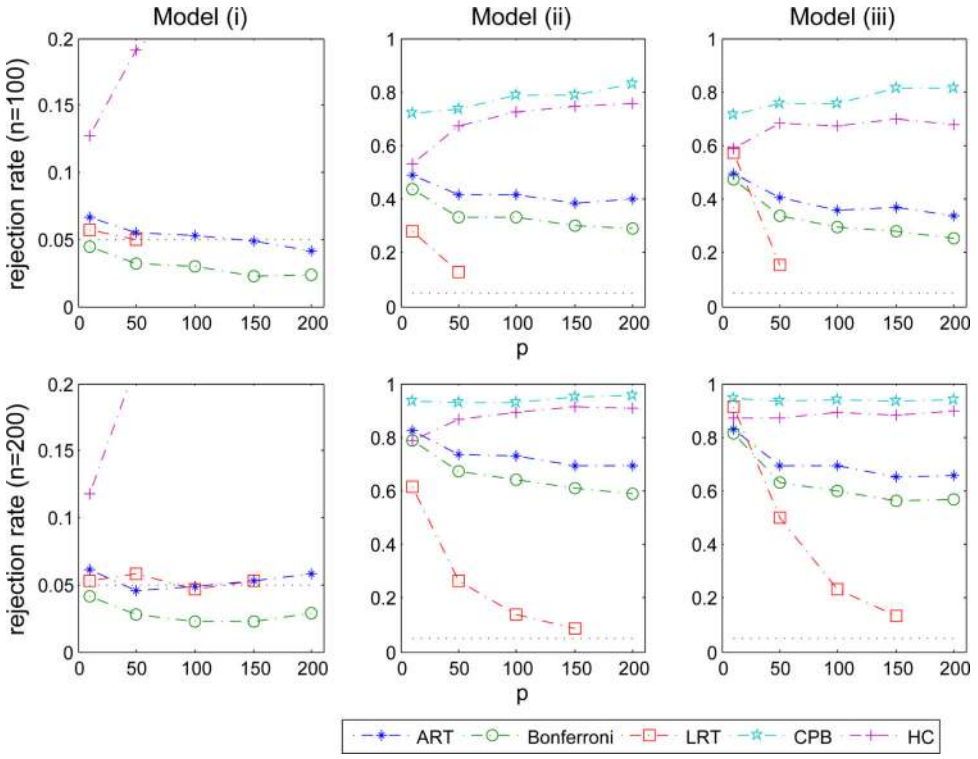

Figure 2.

Empirical rejection rates as in Figure 1 except with lower correlation between predictors: $\rho = 0.5$.

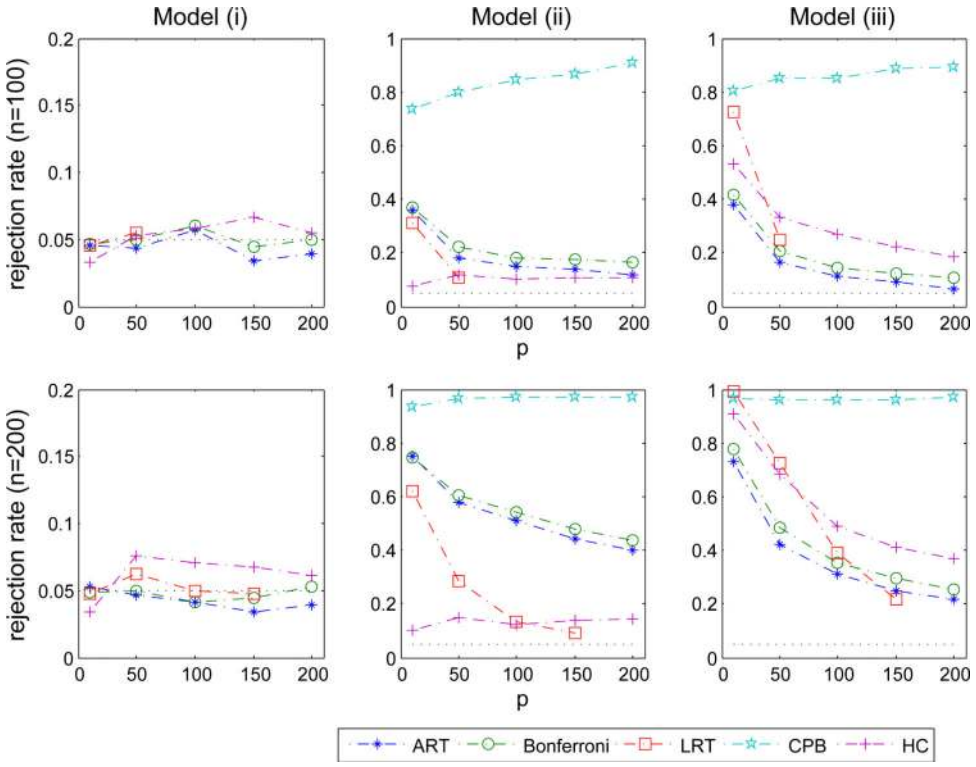

Figure 3.
Empirical rejection rates as in Figure 1 except for independent predictors.

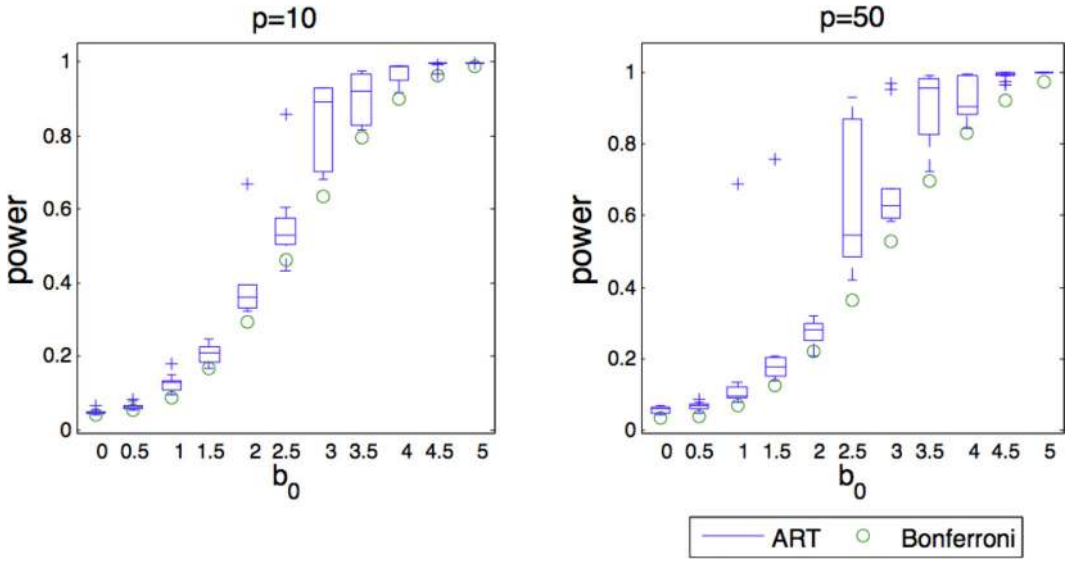

Figure 4.

Asymptotic type I error and power of ART (box plots) compared with Bonferroni (circles) as a function of the local parameter b_0, for $p = 10$ and 50, $\rho = 0.5$, calculated using steps 1–3 in Section 4.2.

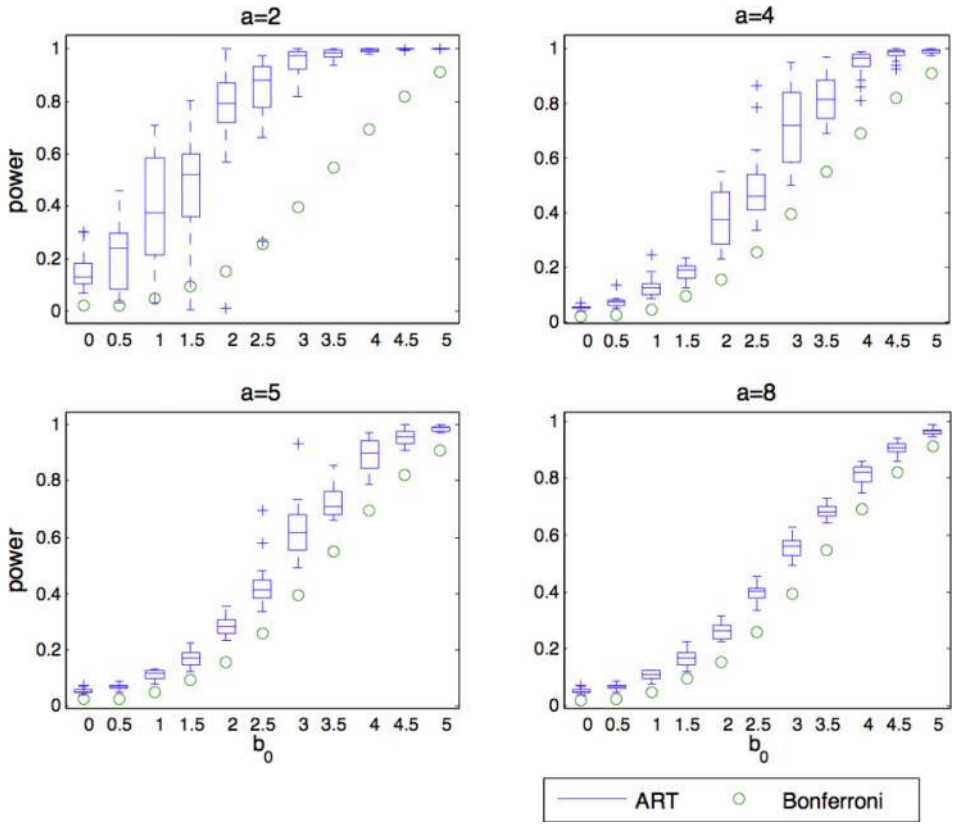

Figure 5.

Asymptotic type I error and power of ART compared with Bonferroni for $p = 1{,}000$ and $\rho = 0.5$, where ART is implemented using a fixed threshold λ_n specified by $a = 2, 4, 5, 8$, and each box plot is based on 20 independent replications with $n = 10{,}000$.

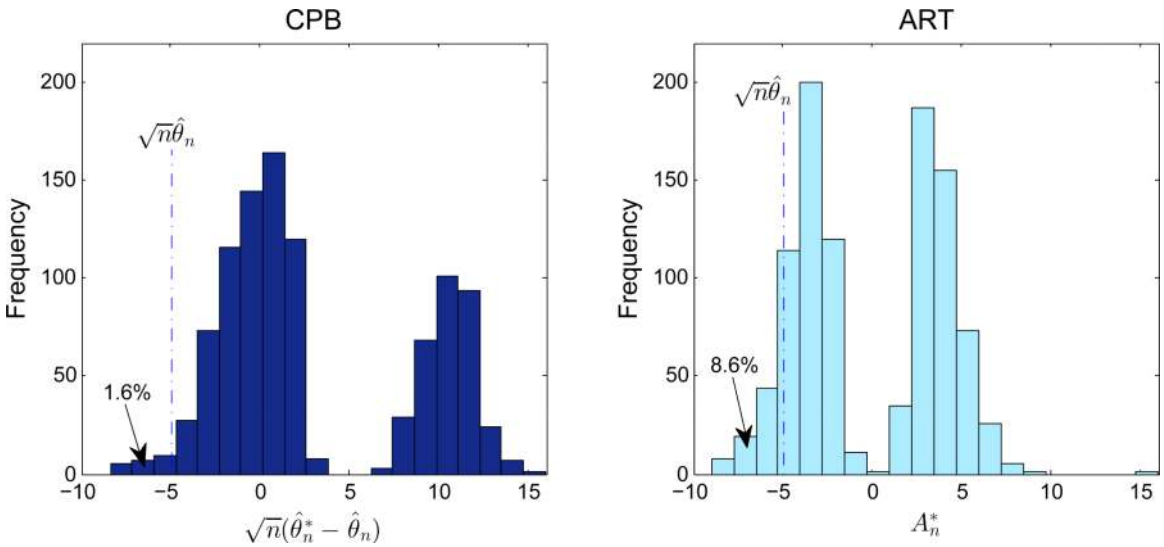

Figure 6.

Gene expression example. Left panel: histogram of $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ showing that the two-sided CPB p-value is 3.2%. Right panel: histogram of A_n^* showing that the two-sided ART p-value is 17.2%.

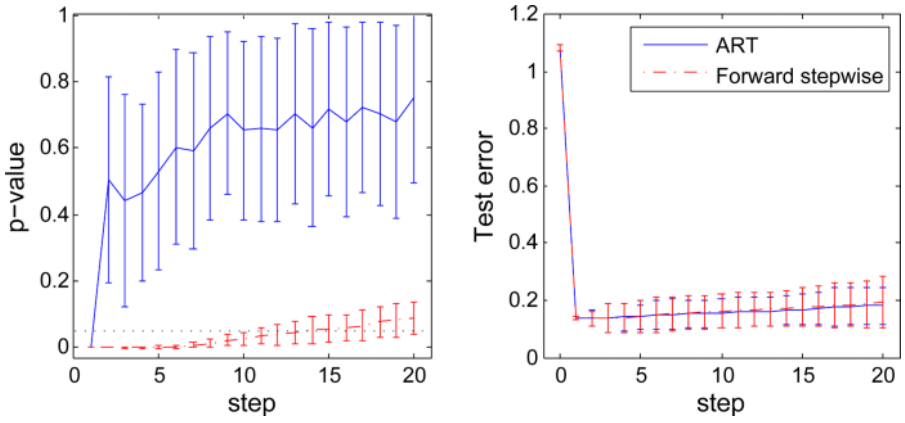

Figure 7.

HIV drug resistance example. Left panel: training set p-values (mean ± SD) over 50 random splits of the data for forward stepwise ART (solid line), standard forward stepwise regression (dash-dot line) and the 0.05 alpha level (dotted). Right panel: test set error for the corresponding models (including all previously selected variables); the two lines are almost indistinguishable.