An Adaptive Resampling Test for Detecting the Presence of Significant Predictors

Ian W. McKeague and Min Qian

This article 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 nonstandard 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 family-wise 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 nonstandard 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.

KEY WORDS: Bootstrap; Family-wise error rate; Marginal regression; Nonregular 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 datasets 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, Ma, and Zhang 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 cannot 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 article 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 nonnormal 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 "nonregularity" causes nonuniform 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 articles by the same authors) established that nonuniform 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 nonregularity occurs in various other settings as well, for example, 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 nonstandard 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 article, 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

Ian W. McKeague (E-mail: *im2131@columbia.edu*) is Professor, and Min Qian (E-mail: *mq2158@columbia.edu*) is Assistant Professor, Department of Biostatistics, Columbia University, New York, NY 10027. Research of the first author is supported by NIH Grant R01GM095722-01 and NSF Grant DMS-1307838. Research of the second author is supported by NSF Grant DMS-1307838.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1095099

is consistent, thus circumventing the nonregularity mentioned above. Under nonlocal 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, Meier, and Bühlmann (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, Tsybakov, and Verzelen (2010) and Arias-Castro, Candès, and Plan (2011) had 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, Chernozhukov, and Hansen (2013), Bühlmann (2013), Zhang and Zhang (2014), and Ning and Liu (2015) had 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 article is organized as follows. We formulate the problem and discuss the issue of nonregularity 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 NONREGULARITY

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 nonzero. 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:

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

and let $\alpha_0 + \theta_0 X_{k_0}$ be the best linear predictor based on X_{k_0}, that is,

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

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-square estimates of α_0, θ_0, and k_0, respectively:

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

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 $\sqrt{n}(\hat{\theta}_n - \theta_0)$ does not converge uniformly with respect to θ_0, as mentioned in the Introduction. The nonuniformity occurs in the neighborhood of $\theta_0 = 0$. Specifically, there exists a bounded continuous function $h : \mathbb{R} \to \mathbb{R}$ such that $f_n(\theta_0) \equiv \mathrm{E}h(\sqrt{n}(\hat{\theta}_n - \theta_0))$ does not converge uniformly in any neighborhood of $\theta_0 = 0$, despite converging pointwise. To see this, first note that under mild conditions

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

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). 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 $\theta_0 = 0$ (i.e., the nonregularity 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 nonlinear 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 (2015) to provide robust confidence intervals for adaptive lasso. As already noted, the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ 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 nonregular 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

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

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

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

where $\theta_n = \operatorname{cov}(X_{k_n}, Y)/\operatorname{var}(X_{k_n})$ 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

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

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 $\boldsymbol{\beta}_0 \neq \mathbf{0}$), then $k_n \to k_0$, and θ_n is asymptotically bounded away from zero (a nonlocal 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, nonuniqueness of the maximally correlated predictor does not adversely affect power.

Theorem 1. Suppose that $k_0 = \bar{k}(\boldsymbol{\beta}_0)$ is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, and $\bar{k}(\boldsymbol{b}_0)$ is unique when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 \neq \mathbf{0}$. Then, under the local model (2),

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

where $K = \arg\max_{k=1,\ldots,p}[Z_k(\mathbf{0}) + C_k^T\boldsymbol{b}_0]^2/V_k$, $C_k = \operatorname{cov}(X_k, \mathbf{X})$, 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

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

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

The nonregularity at $\boldsymbol{\beta}_0 = \mathbf{0}$ is explained by the dependence of the limiting distribution on the (nonidentifiable) 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 A.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 $\sqrt{n}(\hat{\theta}_n - \theta_n)$, as discussed in the Introduction for the nonlocal case.

To get around this problem, we decompose $\sqrt{n}(\hat{\theta}_n - \theta_n)$ 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,

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

where $\mathbb{Z}_{n,k} = \mathbb{G}_n[\epsilon(X_k - \mathbb{P}_n X_k)]$, $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_n)$ 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 $\mathrm{P}(|T_n| > \lambda_n) \to 1_{\boldsymbol{\beta}_0\neq\mathbf{0}}$. 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 $k_n = \bar{K}(\boldsymbol{b}_0)$, where

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

and

$$\bar{K}(\boldsymbol{b}) = \arg\max_{k=1,\ldots,p} \frac{[\operatorname{cov}(X_k, \mathbf{X}^T\boldsymbol{b})]^2}{\operatorname{var}(X_k)},$$

so

$$\begin{aligned} \mathbb{V}_n(\boldsymbol{b}) &= \frac{\mathbb{Z}_{n,\mathbb{K}_n(\boldsymbol{b})} + \widehat{\operatorname{cov}}(X_{\mathbb{K}_n(\boldsymbol{b})}, \mathbf{X}^T\boldsymbol{b})}{\widehat{\operatorname{var}}(X_{\mathbb{K}_n(\boldsymbol{b})})} \\ &\quad - \frac{\operatorname{cov}(X_{\bar{K}(\boldsymbol{b})}, \mathbf{X}^T\boldsymbol{b})}{\operatorname{var}(X_{\bar{K}(\boldsymbol{b})})}. \end{aligned} \tag{4}$$

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 $|T_n^*| \le \lambda_n$ and $|T_n| \le \lambda_n$.

Here and throughout the article, 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 $\sqrt{n}(\hat{\theta}_n - \theta_n)$ 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 $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$ as $n \to \infty$. Then, under the local model (2),

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

converges to the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ 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

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

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 $\boldsymbol{b}_0 = \mathbf{0}$ under the null, so the critical values can be readily computed from A_n^*. In contrast, Laber and Murphy (2015) proposed 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 pretest (for $\theta_n = 0$) with asymptotically negligible Type I error rate: $\lim_{n\to\infty} \mathrm{P}(|T_n| > \lambda_n | \theta_n = 0) = 0$. On the other hand, if λ_n increases too quickly, the pretest 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 $\lambda > 0$,

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

Hence, the pretest Type I error rate can be asymptotically controlled below level γ, without sacrificing consistency, by choosing

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

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., sec. 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 nonzero. 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 $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ is used to estimate the upper and lower quantiles of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, 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) $Y = \epsilon$, (ii) $Y = X_1/4 + \epsilon$, and (iii) $Y = \sum_{k=1}^{p} \beta_k X_k + \epsilon$, where $\beta_1 = \cdots = \beta_5 = 0.15$, $\beta_6 = \cdots = \beta_{10} = -0.1$, 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 $j \neq k$,

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

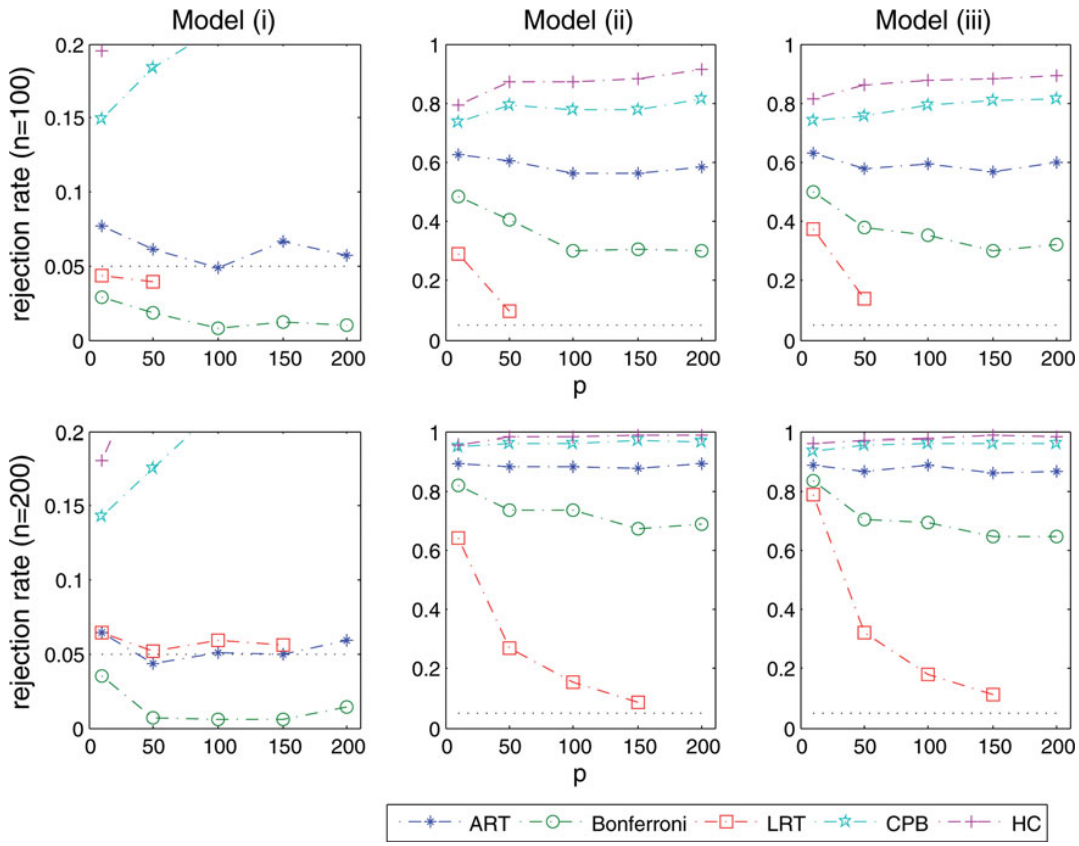

Figure 1. Empirical rejection rates based on 1000 samples generated from models (i), (ii), and (iii) as the dimension ranges from $p = 10$ to $p = 200$, for $n = 100$ (top row) and $n = 200$ (bottom row), and $\rho = 0.8$.

where ρ takes values 0, 0.5, and 0.8, and the noise $\epsilon \sim N(0, 1)$ 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 1000. To specify the threshold λ_n in ART, the double bootstrap is implemented by generating 1000 bootstrap estimates $\hat{\theta}_n^*$, then choosing λ_n so that 5% of the ARTs (based on 1000 nested bootstrap samples) with test statistic $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ reject.

Empirical rejection rates based on 1000 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), nonuniqueness 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 $Y = (n^{-1/2}b_0)X_1 + \epsilon$, 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

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

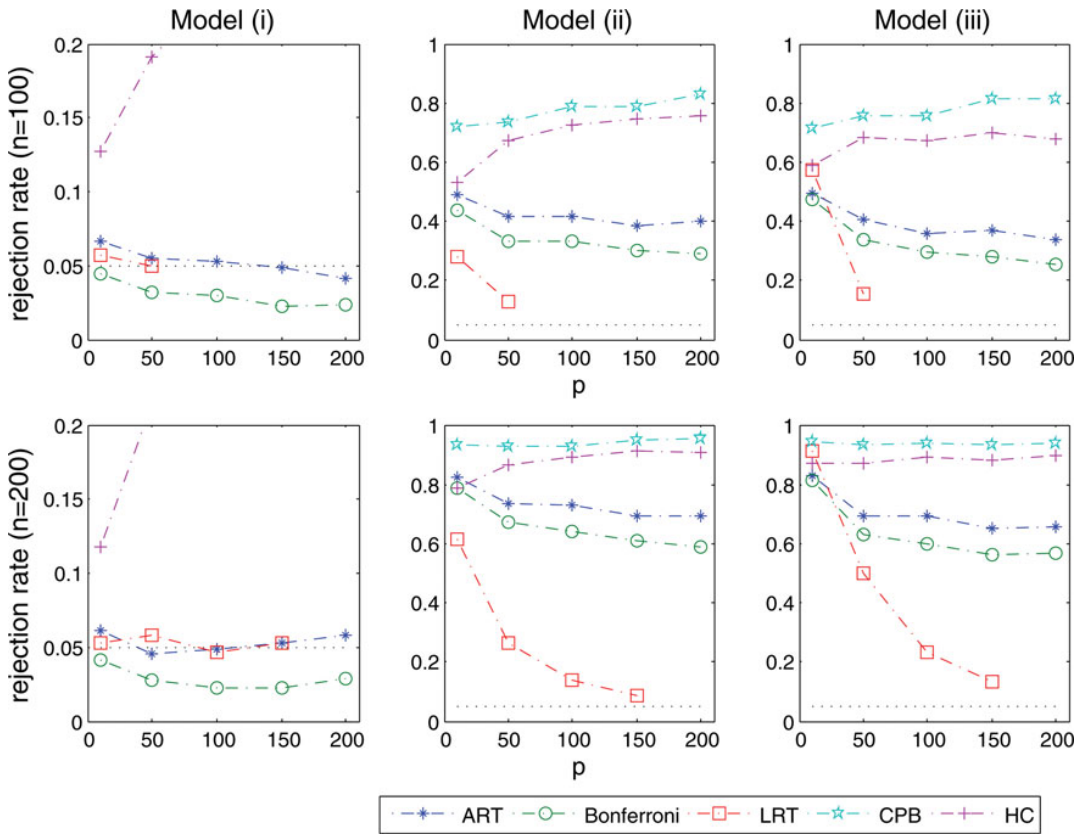

Figure 2. Empirical rejection rates as in Figure 1 except with lower correlation between predictors: $\rho = 0.5$.

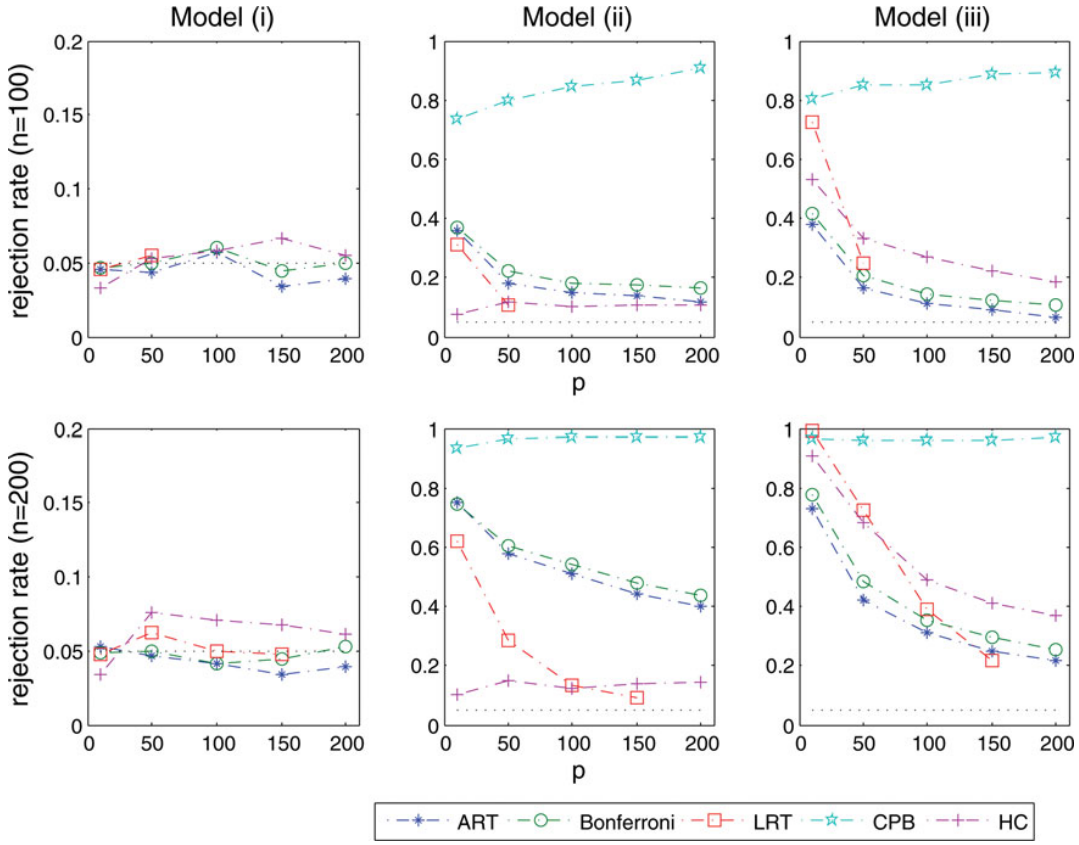

Figure 3. Empirical rejection rates as in Figure 1 except for independent predictors.

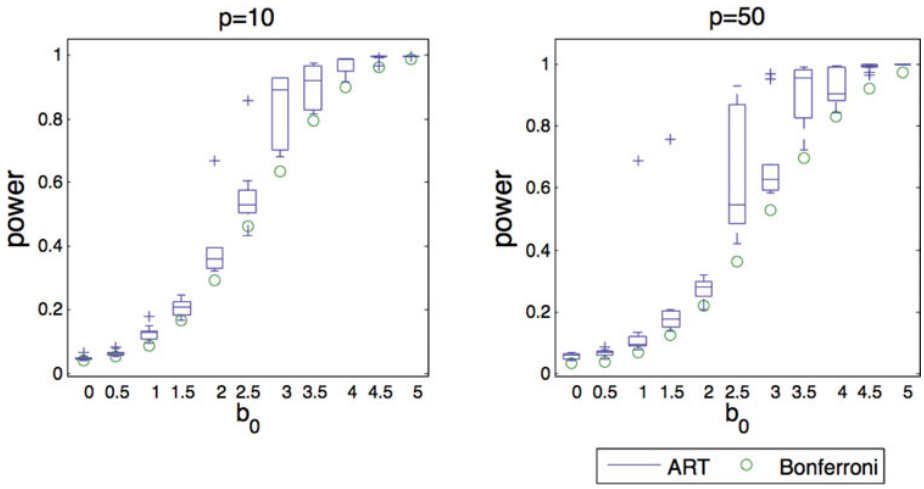

Figure 4. Asymptotic Type I error and power of ART (boxplots) compared with Bonferroni (circles) as a function of the local parameter b_0, for $p = 10$ and 50, $\rho = 0.5$, calculated using Steps 1–3 in Section 4.2.

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 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 5000 draws from the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ 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.

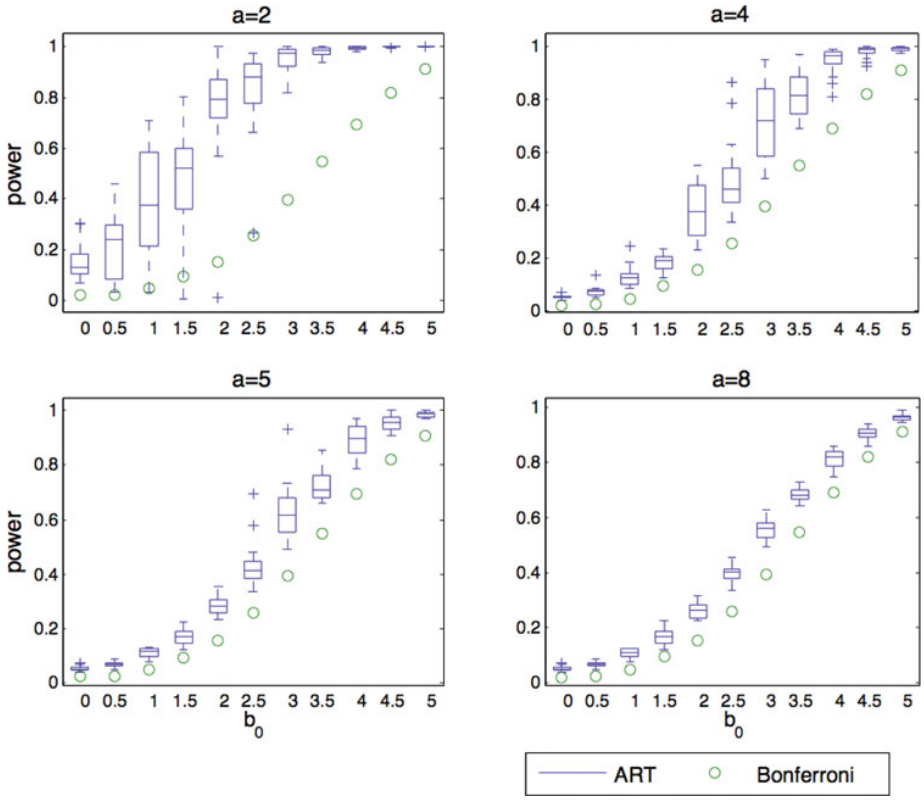

Figure 5. Asymptotic Type I error and power of ART compared with Bonferroni for $p = 1000$ and $\rho = 0.5$, where ART is implemented using a fixed threshold λ_n specified by $a = 2, 4, 5, 8$, and each boxplot is based on 20 independent replications with $n = 10{,}000$.

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

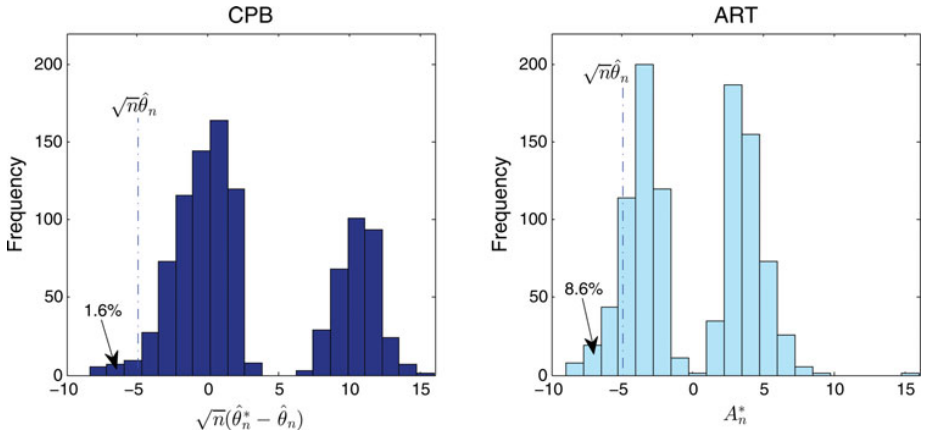

Figure 6. Gene expression example. Left panel: histogram of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ showing that the two-sided CPB p-value is 3.2%. Right panel: histogram of A_n^* showing that the two-sided ART p-value is 17.2%.

2. To assess the asymptotic power of ART at each given b_0, we generate 10 independent large samples (with $n = 5000$) 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 1000 samples each of size $n = 5000$.

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,

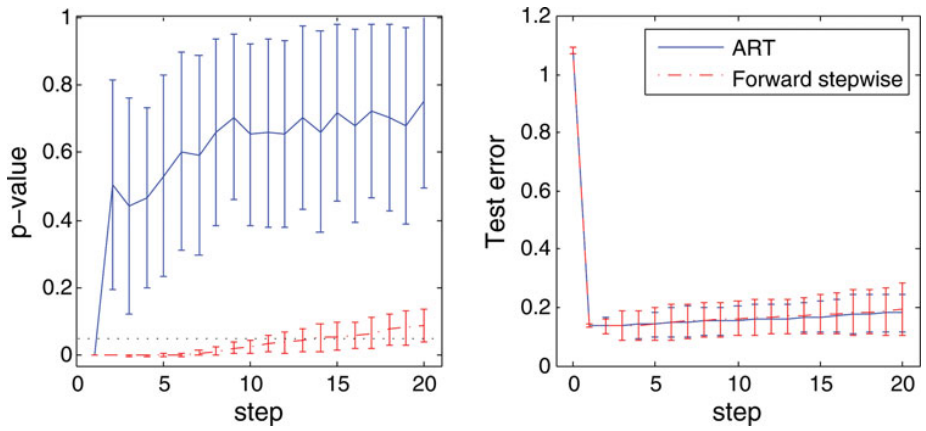

Figure 7. HIV drug resistance example. Left panel: training set p-values (mean $\pm$ SD) over 50 random splits of the data for forward stepwise ART (solid line), standard forward stepwise regression (dash-dot line), and the 0.05 alpha level (dotted). Right panel: test set error for the corresponding models (including all previously selected variables); the two lines are almost indistinguishable.

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

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 $\pm$ SD) for the newly entered predictor at each step, over the 50 random splits, and the right panel shows the corresponding prediction errors (mean $\pm$ 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 article, we have developed an adaptive resampling test (ART) for detecting the existence of a significant predictor, X_{k_0}, from among predictors $X_1, \ldots, X_p$. The procedure is designed to adjust to the nonregular 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 nonregularity: if there is at least one significant predictor, it reduces to a centered percentile bootstrap, otherwise it mimics the local (nonuniform) 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 articles.

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 nonregularity under diverging p are not yet available, except for higher criticism in the case that the predictors are nearly uncorrelated, as established by Ingster, Tsybakov, and Verzelen (2010) and Arias-Castro, Candès, and Plan (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 $\hat{\rho}_n = \max_{k=1,\ldots,p} |\widehat{\mathrm{Corr}}(X_k, Y)|$ and thus devise an asymptotically correct calibration: if $p = p_n \to \infty$ at subexponential 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.

APPENDIX: PROOF

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

$$\begin{aligned}
\sqrt{n}\hat{\theta}_k &= \frac{\sqrt{n}\widehat{\mathrm{cov}}(X_k, Y)}{\widehat{\mathrm{var}}(X_k)} \\
&= \frac{\sqrt{n}\widehat{\mathrm{cov}}(X_k, \mathbf{X}^T)\boldsymbol{\beta}_n + \mathbb{G}_n[\epsilon(X_k - \mathbb{P}_n X_k)]}{\widehat{\mathrm{var}}(X_k)} \\
&= \frac{(\mathbb{G}_n X_k \mathbf{X}^T - P_n X_k \mathbb{G}_n \mathbf{X}^T - \mathbb{G}_n X_k \mathbb{P}_n \mathbf{X}^T)\boldsymbol{\beta}_n}{\widehat{\mathrm{var}}(X_k)} \\
&\quad + \frac{\mathbb{G}_n[\epsilon(X_k - P_n X_k)] - \mathbb{P}_n \epsilon \mathbb{G}_n X_k + \sqrt{n}\mathrm{cov}(X_k, \mathbf{X}^T)\boldsymbol{\beta}_n}{\widehat{\mathrm{var}}(X_k)},
\end{aligned} \tag{A.1}$$

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

$$\widehat{R}_k \equiv \mathbb{P}_n[Y - \hat{\alpha}_k - \hat{\theta}_k X_k]^2 = \widehat{\mathrm{var}}(Y) - \widehat{\mathrm{var}}(X_k)\hat{\theta}_k^2. \tag{A.2}$$

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 (nonregular) asymptotic behavior of $\hat{\theta}_n$ when there are no active predictors.

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

Proof. Denote $\widehat{\mathbf{R}} \equiv (\widehat{R}_1, \ldots, \widehat{R}_p)^T$. When $\boldsymbol{\beta}_0 \neq \mathbf{0}$, $\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}_0) > 0$. By the strong law of large numbers (SLLN)

$$\frac{\widehat{\mathrm{var}}(Y) - \widehat{\mathbf{R}}}{\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}_0)} \xrightarrow{\text{a.s.}} \Big(\mathrm{Corr}^2(X_1, \mathbf{X}^T\boldsymbol{\beta}_0), \ldots, \mathrm{Corr}^2(X_p, \mathbf{X}^T\boldsymbol{\beta}_0)\Big)^T.$$

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

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

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

where the second equality uses $\hat{k}_n \xrightarrow{\text{a.s.}} k_0$ and $k_n \to k_0$ as $n \to \infty$, and the third equality follows from the law of large numbers (LLN) and $\mathrm{cov}(\epsilon, X_{k_0}) = 0$. Similarly, $\widehat{\mathrm{var}}(\hat{X}) \xrightarrow{P_n} V_{k_0} \equiv \mathrm{var}(X_{k_0})$. The proof is completed using Slutsky's lemma and the central limit theorem (CLT). □

Lemma A.2. If all conditions in Theorem 1 hold and $\boldsymbol{\beta}_0 = \mathbf{0}$, then $\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} Z_K(\mathbf{0})/V_K + (C_K/V_K - C_{\bar{k}(\boldsymbol{b}_0)}/V_{\bar{k}(\boldsymbol{b}_0)})^T\boldsymbol{b}_0$.

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

$$\frac{(Z_j(\mathbf{0}) + C_j^T\boldsymbol{b}_0)^2}{V_j} \neq \frac{(Z_k(\mathbf{0}) + C_k^T\boldsymbol{b}_0)^2}{V_k} \text{ for any } j \neq k \text{ a.s.} \tag{A.3}$$

So K is unique a.s.

Denote $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. 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 (A.1) 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 (A.2), we have

$$n\big[\widehat{\mathrm{var}}(Y) - \widehat{\mathbf{R}}\big] = (\sqrt{n}\hat{\boldsymbol{\theta}}) \odot (\sqrt{n}\hat{\boldsymbol{\theta}}) \odot \big(\widehat{\mathrm{var}}(X_1), \ldots, \widehat{\mathrm{var}}(X_p)\big)^T,$$

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

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

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 (A.3) and since $\sqrt{n}\hat{\theta}_n = \sqrt{n}\hat{\boldsymbol{\theta}}^T h\big(n[\widehat{\mathrm{var}}(Y) - \widehat{\mathbf{R}}]\big)$, the result follows by applying the continuous mapping theorem to the above display. □

Lemma A.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 $f(z_1, \ldots, z_p, \boldsymbol{b})_k = (z_k + C_k^T\boldsymbol{b})^2/V_k$ for $k = 1, \ldots, p$, and using (A.3).

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 reexpressing (4), along with $\bar{K}(\mathbf{b})$ 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

$$\mathbb{Z}_{n,k}^* = \mathbb{G}_n^*[\hat{\epsilon}_n(X_k - \mathbb{P}_n^* X_k)] = \mathbb{G}_n^*[\hat{\epsilon}_n X_k] - [\mathbb{G}_n^*\hat{\epsilon}_n][\mathbb{P}_n^* X_k], \tag{A.4}$$

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 $1_{|T_n^*| > \lambda_n}$ or $|T_n| > \lambda_n} \xrightarrow{P^M} 1_{\boldsymbol{\beta}_0 \neq \mathbf{0}}$ and $1_{|T_n^*| \leq \lambda_n} 1_{|T_n| \leq \lambda_n} \xrightarrow{P^M} 1_{\boldsymbol{\beta}_0 = \mathbf{0}}$ conditionally (on the data) in probability. This together with Lemmas A.4 and A.5 implies the result.

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

$$\begin{aligned}
\sqrt{n}\hat{\theta}_k^* &= \frac{\sqrt{n}[\mathbb{P}_n^* X_k Y - (\mathbb{P}_n^* X_k)(\mathbb{P}_n^* Y)]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2} \\
&= \frac{\mathbb{G}_n^* X_k Y - \mathbb{G}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{G}_n^* Y) + \sqrt{n}[\mathbb{P}_n X_k Y - (\mathbb{P}_n X_k)(\mathbb{P}_n Y)]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2} \\
&= \frac{\mathbb{G}_n^* X_k Y - \mathbb{G}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{G}_n^* Y) + \sqrt{n}\hat{\theta}_k[\mathbb{P}_n X_k^2 - (\mathbb{P}_n X_k)^2]}{\mathbb{P}_n^* X_k^2 - (\mathbb{P}_n^* X_k)^2}.
\end{aligned} \tag{A.5}$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, by Lemma A.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,

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

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

Hence,

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

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

Lemma A.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 $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{d} Z_{k_0}(\boldsymbol{\beta}_0)/V_{k_0}$ conditionally (on the data) in probability.

Proof. It follows from (A.5), the SLLN and Slutsky's lemma that, when $\boldsymbol{\beta}_0 \neq \mathbf{0}$,

$$\begin{aligned} \widehat{\mathrm{var}}^*(X_k)\hat{\theta}_k^* &= n^{-1/2}\big[\mathbb{G}_n^* X_k Y - \mathbb{G}_n^* X_k \mathbb{P}_n^* Y - (\mathbb{P}_n X_k)(\mathbb{G}_n^* Y)\big] \\ &\quad + \hat{\theta}_k[\mathbb{P}_n X_k^2 - (\mathbb{P}_n X_k)^2] \xrightarrow{P^M} C_k^T \boldsymbol{\beta}_0 \end{aligned}$$

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

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

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

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

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

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

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

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

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

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)[\mathbb{P}_n^* X_{\hat{k}_n^*}^2 - (\mathbb{P}_n^* X_{\hat{k}_n^*})^2] \\ &= \sqrt{n}[\mathbb{P}_n^* X_{\hat{k}_n^*} Y - (\mathbb{P}_n^* X_{\hat{k}_n^*})(\mathbb{P}_n^* Y) - \hat{\theta}_n(\mathbb{P}_n^* X_{\hat{k}_n^*}^2 - (\mathbb{P}_n^* X_{\hat{k}_n^*})^2)] \\ &= \sqrt{n}(\mathbb{P}_n^* X_{\hat{k}_n^*} \hat{\epsilon}_n - \mathbb{P}_n^* X_{\hat{k}_n^*} \mathbb{P}_n^* \hat{\epsilon}_n) \\ &\quad + \sqrt{n}\hat{\theta}_n\left[(\mathbb{P}_n^* X_{\hat{k}_n^*})^2 - \mathbb{P}_n^* X_{\hat{k}_n^*}^2 + \mathbb{P}_n^* X_{\hat{k}_n^*} \hat{X} - (\mathbb{P}_n^* X_{\hat{k}_n^*})(\mathbb{P}_n^* \hat{X})\right] \\ &= \mathbb{G}_n^* \hat{\epsilon}_n (X_{\hat{k}_n^*} - P_n X_{\hat{k}_n^*}) - \mathbb{G}_n^* X_{\hat{k}_n^*}(\mathbb{P}_n^* - \mathbb{P}_n)\hat{\epsilon}_n \\ &\quad - \mathbb{G}_n^* \hat{\epsilon}_n (\mathbb{P}_n - P_n) X_{\hat{k}_n^*} \\ &\quad + \sqrt{n}\hat{\theta}_n\left[(\mathbb{P}_n^* X_{\hat{k}_n^*})^2 - \mathbb{P}_n^* X_{\hat{k}_n^*}^2 + \mathbb{P}_n^* X_{\hat{k}_n^*} \hat{X} - (\mathbb{P}_n^* X_{\hat{k}_n^*})(\mathbb{P}_n^* \hat{X})\right]. \end{aligned} \tag{A.6}$$

The last term in (A.6) 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 (A.6) converges in distribution to $Z_{k_0}(\boldsymbol{\beta}_0)$ conditionally (on the data) in probability. By Lemma A.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 $\hat{\alpha}_n \xrightarrow{P_n} \alpha_0 + E\mathbf{X}^T\boldsymbol{\beta}_0 - \theta_0 E X_{k_0}$. 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

$$\begin{aligned} &\mathbb{G}_n^* \hat{\epsilon}_n [(X_{\hat{k}_n^*} - P_n X_{\hat{k}_n^*} - (X_{k_0} - P_n X_{k_0})] + \mathbb{G}_n^*[(\hat{\epsilon}_n - \bar{\epsilon})(X_{k_0} - P_n X_{k_0})] \\ &\quad + \mathbb{G}_n^*[\bar{\epsilon}(X_{k_0} - P_n X_{k_0})]. \end{aligned} \tag{A.7}$$

The first term in (A.7) is $o_{P^M}(1)$ a.s. since $\hat{k}_n^* \xrightarrow{P^M} k_0$. The second term in (A.7) can be written as

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

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 (A.7) converges in distribution to $Z_{k_0}(\boldsymbol{\beta}_0)$ conditionally (on the data) in probability.

Similarly, the second and third terms in (A.6) and $\mathbb{P}_n^* X_{\hat{k}_n^*}^2 - (\mathbb{P}_n^* X_{\hat{k}_n^*})^2 - \mathrm{var}(X_{k_0})$ can be shown to be $o_{P^M}(1)$ in probability. The result then follows from Slutsky's lemma. □

Lemma A.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 $\sqrt{n}(\hat{\theta}_n - \theta_n)$ 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 $\mathbb{Z}_{n,k} = \mathbb{G}_n[\epsilon(X_k - \mathbb{P}_n X_k)]$,

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

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

$$\frac{\widehat{\mathrm{cov}}(X_k, \mathbf{X}^T\boldsymbol{b}) + \mathbb{Z}_{n,k}}{\widehat{\mathrm{var}}(X_k)} - \frac{\mathrm{cov}(X_j, \mathbf{X}^T\boldsymbol{b})}{\mathrm{var}(X_j)}.$$

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 $\boldsymbol{\beta}_0 = \mathbf{0}$,

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

Moreover, the second equality in the above display also holds for the bootstrap version. Writing the bootstrapped version of $\mathbb{Z}_{n,k}$ in (A.4) as

$$\begin{aligned} \mathbb{Z}_{n,k}^* &= \mathbb{G}_n^*[\epsilon(X_k - P_n X_k)] + \mathbb{G}_n^*[(\hat{\epsilon}_n - \epsilon)(X_k - P_n X_k)] \\ &\quad + [(P_n - \mathbb{P}_n^*)X_k][\mathbb{G}_n^* \hat{\epsilon}_n], \end{aligned}$$

and using arguments similar to those in the proof Lemma A.4 for handling (A.7), 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, $(\hat{D}_n'(\boldsymbol{b}_0), \mathbb{W}_n^*(\boldsymbol{b}_0), \mathbb{M}_n^*(\boldsymbol{b}_0)) \xrightarrow{d} (D'(\boldsymbol{b}_0), \mathbb{W}(\boldsymbol{b}_0), \mathbb{M}(\boldsymbol{b}_0))$ 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,

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

using similar arguments to those at the end of the proof of Lemma A.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. □

[Received May 2014. Revised September 2015.]

REFERENCES

Andrews, D. (2000), "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405. [1422]

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), "Global Testing Under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism," *The Annals of Statistics*, 39, 2533–2556. [1423,1430]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. [1423]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society,* Series B, 57, 289–300. [1422]

Berk, R., Brown, L. D., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics*, 41, 802–837. [1422,1423]

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754. [1422]

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [1423]

Cai, T. T., and Jiang, T. (2012), "Phase Transition in Limiting Distributions of Coherence of High-dimensional Random Matrices," *Journal of Multivariate Analysis*, 107, 24–39. [1430]

Cancer Genome Atlas Research Network, (2008), "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways," *Nature*, 455, 1061–1068. [1430]

Chatterjee, A., and Lahiri, S. N. (2011), "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 106, 608–625. [1423]

Cheng, X. (2008), "Robust Confidence Intervals in Nonlinear Regression Under Weak Identification," unpublished manuscript, Department of Economics, University of Pennsylvania. Available at *https://economics.sas.upenn.edu/events/robust-confidence-intervals-nonlinear-regression-under-weak-identification* [1423]

——— (2015), "Robust Inference in Nonlinear Models With Mixed Identification Strength," *Journal of Econometrics*, 189, 207–228. [1423]

Davies, R. B. (1977), "Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64, 247–254. [1422]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [1425]

——— (2015), "Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects," *Statistical Science*, 30, 1–25. [1425]

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103. [1422]

Dudoit, S., and van der Laan, M. J. (2008), *Multiple Testing Procedures With Applications to Genomics*, New York: Springer. [1422]

Efron, B. (2006), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of American Statistical Association*, 99, 96–104. [1422]

——— (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge, UK: Cambridge University Press. [1422]

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap (Monographs on Statistics & Applied Probability)*, Boca Raton, FL: Chapman & Hall/CRC. [1425]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1422]

——— (2006), "Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians* (Vol. III), eds. M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, Zurich: European Mathematical Society, pp. 595–622. [1422]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [1422]

Genovese, C., Jin, J., Wasserman, L., and Yao, Z. (2012), "A Comparison of the Lasso and Marginal Regression," *Journal of Machine Learning Research*, 13, 2107–2143. [1422]

HIV Drug Resistance Database (2014), "Genotype-Phenotype Datasets," Stanford University, available at *http://hivdb.stanford.edu/pages/genopheno.dataset.html* [1430]

Huang, J., Ma, S., and Zhang, C.-H. (2008), "Adaptive Lasso for High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618. [1422]

Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010), "Detection Boundary in Sparse Regression," *Electronic Journal of Statistics*, 4, 1476–1526. [1423,1430]

Johnson, R. A., and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis* (6th ed.), Upper Saddle River, NJ: Prentice Hall. [1425]

Laber, E., Lizotte, D., Qian, M., and Murphy, S. A. (2014), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225–1272. [1423]

Laber, E., and Murphy, S. A. (2011), "Adaptive Confidence Intervals for the Test Error in Classification" (with discussion), *Journal of the American Statistical Association*, 106, 904–913. [1423]

——— (2015), "Adaptive Inference After Model Selection," under review. [1423,1425]

Leeb, H., and Pötscher, B. M. (2006), "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554–2591. [1422]

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014), "A Significance Test for the Lasso," *The Annals of Statistics*, 42, 413–468. [1423,1425,1430]

McCloskey, A. (2012), "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," Working Paper. Available at *http://www.econ.brown.edu/fac/adam_mccloskey/Research_files/McCloskey_BBCV.pdf* [1422]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [1423]

Ning, Y., and Liu, H. (2015), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," available at *http://arxiv.org/abs/1412.8765* [1423]

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003), "Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database," *Nucleic Acids Research*, 31, 298–303. [1430]

Samworth, R. (2003), "A Note on Methods of Restoring Consistency to the Bootstrap," *Biometrika*, 90, 985–990. [1422]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press. [1432]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society,* Series B, 76, 217–242. [1423]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society,* Series B, 67, 301–320. [1422]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]

Comment

A. CHATTERJEE and S. N. LAHIRI

1. INTRODUCTION

This is an interesting article dealing with the important issue of consistency of bootstrap approximations for distributions of nonregular estimators under local asymptotics. Our discussion of the article (referred to as [MQ] in the following to save space) will focus on two aspects: (1) the use of bootstrap on a nonregular test statistic and (2) an alternative solution to the present testing problem where the issue of nonregularity can be bypassed, allowing the naive bootstrap to be used without any modification. Since the nonregular behavior of the test statistic is present only in a neighborhood of $\boldsymbol{\beta}_0 = \mathbf{0}$ in model (2) of [MQ], we shall set $\boldsymbol{\beta}_0 = \mathbf{0}$ and restrict attention to the local asymptotic structure:

$$Y_i = \alpha_0 + \mathbf{X}_i'\big[n^{-1/2}\mathbf{b}_0\big] + \epsilon_i,\ i = 1, \ldots, n \tag{1}$$

for $\alpha \in \mathbb{R}$ and $\mathbf{b}_0 \in \mathbb{R}^p$, where Y_i, ϵ_i and $\mathbf{X}_i$ are as in [MQ] and where $\mathbf{A}'$ denotes the transpose of a matrix $\mathbf{A}$. On (2), it is observed that the alternative test has attractive power properties particularly under diffuse alternatives, that is, when the components of $\mathbf{b}_0$ in model (1) are small nonzero numbers. Further, the alternative test is computationally simple and it can be easily extended to the high-dimensional case. On the other hand, the ART is better suited to identify the most significant predictor which the alternative test cannot do.

The rest of the discussion is organized as follows. In Section 2, we discuss some important aspects of the ART methodology and its implications in the general context of bootstrapping nonregular estimators. In Section 3, we describe a simple alternative solution to the testing problem dealt with in the article. In Section 4, we compare the adaptive resampling test (ART) and the alternative test in a simulation study. Finally, some concluding remarks are given in Section 5.

2. BOOTSTRAPPING A NONEGULAR ESTIMATOR

It has been known for a long time (see Andrews 2000; Samworth 2003) that for nonregular estimators, such as the Hodges' estimator, a naive application of the bootstrap method fails to capture their limit distributions. The present article considers one such example where the test-statistic of interest is nonregular and also, its limit distributions over the parameter space has a discontinuity at the origin. The naive Bootstrap method that resamples randomly and with replacement from the observables $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ has a random variation of the order $n^{-1/2}$ around the true joint distribution of $(\mathbf{X}_1, Y_1)$. As a result, defining the Bootstrap version of a nonregular statistic by simply replacing the $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ with the resampled values $\{(\mathbf{X}_i^*, Y_i^*) : i = 1, \ldots, n\}$ (say) propagates this $n^{-1/2}$-order random fluctuations into the distribution of a nonregular estimator. This, in turn, makes it impossible for the naive bootstrap approach to recover the discontinuous limit laws of the nonregular estimator with probability tending to 1. One of the major contributions of the article is to propose a novel approach toward dealing with this issue which is applicable more generally in similar inference problems involving bootstrapping nonregular estimators. Existing approaches to dealing with inconsistency of the bootstrap include: (1) resampling from especially constructed distribution estimators (e.g., weighted empirical distribution function in place of the empirical distribution function in linear regression models (see Lahiri 1992), (2) resampling fewer observations than the original sample size or the 'm out of n Bootstrap' (see Athreya 1987; Bickel, Götze, and van Zwet 1997), among others. In contrast, the approach proposed here is strikingly different from existing approaches—it seeks to overcome the inconsistency of the naive bootstrap by carefully modifying the definition of the bootstrapped estimator itself, keeping the resampling scheme (in this case, the paired bootstrap—see Freedman (1981)) unchanged. Specifically, denoting the resampled observations as $\{(X_i^*, Y_i^*) : i = 1, \ldots, n\}$, generated by simple random sampling with replacement from $\{(X_i, Y_i) : i = 1, \ldots, n\}$, the naive bootstrap version of $\sqrt{n}(\hat{\theta}_n - \theta)$ is

$$\sqrt{n}\big(\hat{\theta}_n^* - \hat{\theta}_n\big)$$

where $\hat{\theta}_n^*$ is obtained by replacing $\{(X_i, Y_i) : i = 1, \ldots, n\}$ with $\{(X_i^*, Y_i^*) : i = 1, \ldots, n\}$. As discussed earlier, since $\hat{\theta}_n$ is a nonregular estimator (see Theorem 1 of [MQ]), this naive version of the bootstrap fails. The approach developed by [MQ] is to carefully redefine the bootstrap version of $\sqrt{n}(\hat{\theta}_n - \theta)$ so that it is able to recognize and adapt to the discontinuity in the limit laws of $\sqrt{n}(\hat{\theta}_n - \theta)$. The modified bootstrap version of $R_n \equiv \sqrt{n}(\hat{\theta}_n - \theta)$, given in Theorem 2 of [MQ], is

$$R_n^*[\mathrm{MQ}] = \begin{cases} \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) & \text{if } \max\{|T_n^*|, |T_n|\} > \lambda_n \\ \mathbb{V}_n(\mathbf{b}_0) & \text{if } \max\{|T_n^*|, |T_n|\} \le \lambda_n. \end{cases}$$

A. Chatterjee, Stat-Math Unit, Indian Statistical Institute, New Delhi, India (E-mail: *cha@isid.ac.in*). S. N. Lahiri, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: *snlahiri@ncsu.edu*). Research partially supported by NSF grant DMS 1310068.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1102143

Table 1. Empirical probability of rejecting H_0 by the ART test at different choices of b_1 and corresponding standard deviation (in parentheses). Here, $n = 200$ and $p = 10$ and $\epsilon_1 \sim \chi^2_1 - 1$

	b_1					
λ_n	0	1/4	1	2	5	10
0	0.013	0.007	0.008	0.08	0.715	1
	(0.02)	(0.01)	(0.01)	(0.1)	(0.24)	(0)
1	0.063	0.058	0.138	0.389	0.867	1
	(0.07)	(0.05)	(0.14)	(0.3)	(0.22)	(0)
2.5	0.027	0.049	0.209	0.577	0.989	1
	(0.03)	(0.04)	(0.13)	(0.19)	(0.02)	(0)
5	0.035	0.066	0.265	0.663	0.996	1
	(0.03)	(0.05)	(0.12)	(0.14)	(0)	(0)

Here, the truncation of T_n and T_n^* at (a suitable level) λ_n is used for capturing the different limit behaviors of R_n in two scenarios—(1) away from the origin (i.e., for $\boldsymbol{\beta}_0 \neq 0$) and (2) in $n^{-1/2}$-compact neighborhoods of the origin (i.e., for $\boldsymbol{\beta}_0 = 0$ and $\boldsymbol{\beta}_n = n^{-1/2}\mathbf{b}_0$ in (2) of [MQ]). It is important to note that the definition of R_n^* depends on the (local) parameter $\mathbf{b}_0$ which would be known in testing problems with simple null hypotheses, for example, as in the ART for $H_0 : \theta_n = \mathbf{0}$ or equivalently, for $H_0 : \mathbf{b}_0 = \mathbf{0}$.

The choice of the truncation level λ_n plays a crucial role in determining the performance of the ART in finite samples. While the asymptotic results are valid for any sequence $\{\lambda_n\}$ with $\lambda_n^{-1} + \lambda_n/\sqrt{n} = o(1)$, data-based choices of λ_n are preferable. In their article, [MQ] suggests a double-bootstrap method for choosing λ_n. Although this would provide a reasonable choice, computationally simpler alternatives such as the jackknife-after-bootstrap method of Efron (1992) may also be used to find the desired quantiles (see Lahiri 2005). In particular, the JAB can be used to generate computationally simpler estimates of the MSE of the bootstrap quantiles corresponding to a set of λ_n values which would then be minimized to produce a choice of λ_n.

3. TESTING $H_0 : \theta_n = 0$

We now formulate an alternative test procedure for testing $H_0 : \theta_n = 0$, that is, if any of the p predictors is relevant. Here,

Table 2. Empirical probability of rejecting H_0 by the ART test at different choices of Δ and corresponding standard deviation (in parentheses). Here, $n = 200$ and $p = 10$ and $\epsilon_1 \sim \chi^2_1 - 1$

	Δ					
λ_n	0	1/4	1	2	5	10
0	0.012	0.008	0.005	0.024	0.533	0.996
	(0.02)	(0.01)	(0.01)	(0.04)	(0.28)	(0.03)
1	0.062	0.057	0.09	0.238	0.777	0.996
	(0.06)	(0.05)	(0.09)	(0.22)	(0.31)	(0.03)
2.5	0.027	0.041	0.128	0.364	0.969	1
	(0.03)	(0.04)	(0.09)	(0.18)	(0.07)	(0)
5	0.036	0.056	0.168	0.444	0.989	1
	(0.03)	(0.04)	(0.09)	(0.15)	(0.01)	(0)

Table 3. Empirical probability of rejecting H_0 by the ART test at different choices of Δ and corresponding standard deviation (in parentheses). Here, $n = 200$ and $p = 150$ and $\epsilon_1 \sim \chi^2_1 - 1$

	Δ					
λ_n	0	1/4	1	2	5	10
0	0.076	0.054	0.016	0.01	0.318	0.972
	(0.09)	(0.07)	(0.03)	(0.02)	(0.3)	(0.08)
1	0.177	0.146	0.098	0.137	0.52	0.988
	(0.18)	(0.14)	(0.09)	(0.21)	(0.39)	(0.04)
2.5	0.074	0.077	0.119	0.276	0.875	0.995
	(0.17)	(0.16)	(0.18)	(0.23)	(0.24)	(0.03)
5	0.117	0.137	0.231	0.45	0.973	1
	(0.25)	(0.26)	(0.27)	(0.26)	(0.06)	(0)

θ_n is as in [MQ], that is, $\theta_n = \eta_{k_n}$ where $\eta_k = \frac{\mathrm{Cov}(X_k, Y)}{\mathrm{Var}(X_k)}$ and $k_n = \mathrm{argmax}_k |\eta_k|$. Note that

$$H_0 : \theta_n = 0 \text{ is equivalent to } H_0 : \eta_1 = \cdots = \eta_p = 0. \quad (2)$$

The test in [MQ] is based on the maximum of the sample versions of the η_k's, which targets the ℓ_∞ norm of $(\eta_1, \ldots, \eta_p)$. We can, alternatively, consider the ℓ_γ norm of $(\eta_1, \ldots, \eta_p)$ for any $\gamma \in (0, \infty)$, as testing $H_0 : \theta_n = 0$ is equivalent to testing $H_0 : \|(\eta_1, \ldots, \eta_p)'\|_\gamma = 0$ where $\|\cdot\|_\gamma$ denotes the ℓ_γ norm of a vector. For simplicity, we restrict attention to $\gamma = 2$, corresponding to the standard Euclidean norm. Let $\hat{\eta}_k = \frac{\widehat{\mathrm{cov}}(X_k, Y)}{\widehat{\mathrm{var}}(X_k)} = \frac{n^{-1}\sum_{i=1}^n (X_{i,k} - \bar{X}_k)(Y_i - \bar{Y})}{n^{-1}\sum_{i=1}^n (X_{i,k} - \bar{X}_k)^2}$ where a bar $\bar{}$ over a symbol denotes averaging over its n-values. We define the test statistic

$$\Lambda_n = \sum_{j=1}^{p} t_j^2$$

where $t_j = \hat{\eta}_j / s_j$ is the t-statistic for the jth marginal regression with $s_j^2 = \frac{n^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}{\sum_{i=1}^n (X_{i,j} - \bar{X}_{n,j})^2}$. Here, we would reject H_0 in (2) for large values of Λ_n, Note that calibration of the test statistic Λ_n is straightforward. Indeed, it is easy to check that the limit distribution of Λ_n is given by a weighted sum of p chi-squared random variables, where the weights are determined by the covariance structure of $(X_1, \ldots, X_p)$ that can be well approximated by the naive bootstrap.

We now point out some of the advantages of the above testing procedure. Note that the test statistic Λ_n is regular and unlike $\hat{\theta}$, its limit distributions do not suffer from noncontinuity issues. Second, the naive bootstrap method can be used without any modifications to find the critical points of the test. Third, it is computationally simple and does not require selection of the truncation parameter λ_n. Indeed, the test based on Λ_n has a natural extension to the high-dimensional setting, as given by Gregory et al. (2015). Compared with the LRT that has a poor performance in high dimensions as demonstrated by the simulation results of [MQ], the Λ_n-based test does not suffer from this problem, as it completely avoids inversion of a high-dimensional covariance matrix. See Gregory et al. (2015) for more details on the properties of the Λ_n-based test in high dimensions.

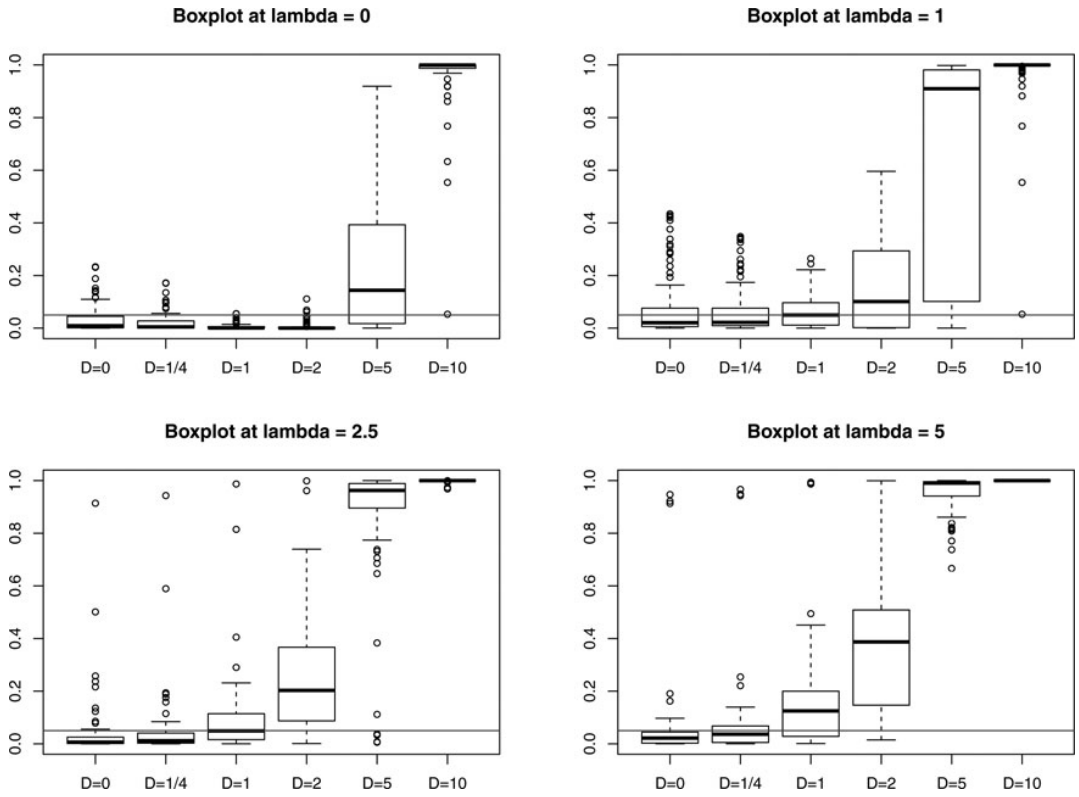

Figure 1. $n = 200$ and $p = 50$. Error df $= \chi^2_1 - 1$. Power of the ART with different choices of Δ (denoted as D in the x-axis) at $\lambda_n \in \{0, 1, 2.5, 5\}$. Boxplots are based on 100 samples. Horizontal blue line denotes the 5% level.

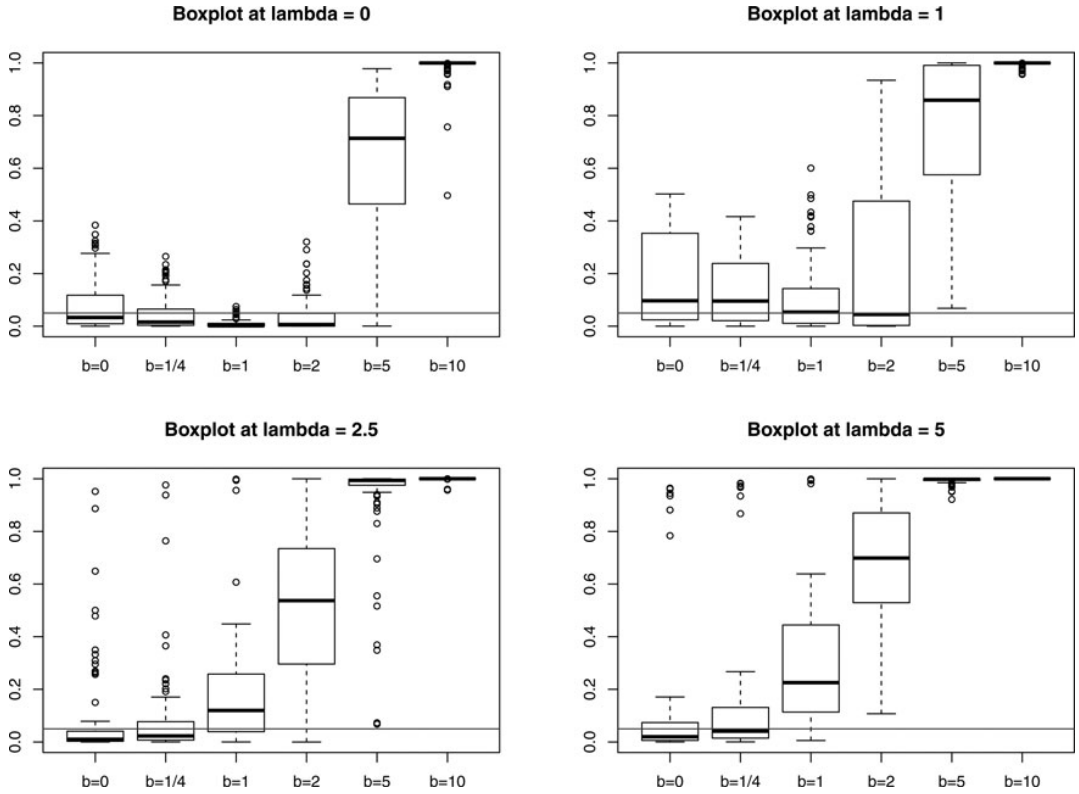

Figure 2. $n = 200$ and $p = 150$. Error df $= \chi^2_1 - 1$. Power of the ART with different choices of b_1 at $\lambda_n \in \{0, 1, 2.5, 5\}$. Boxplots are based on 100 samples. Horizontal blue line denotes the 5% level.

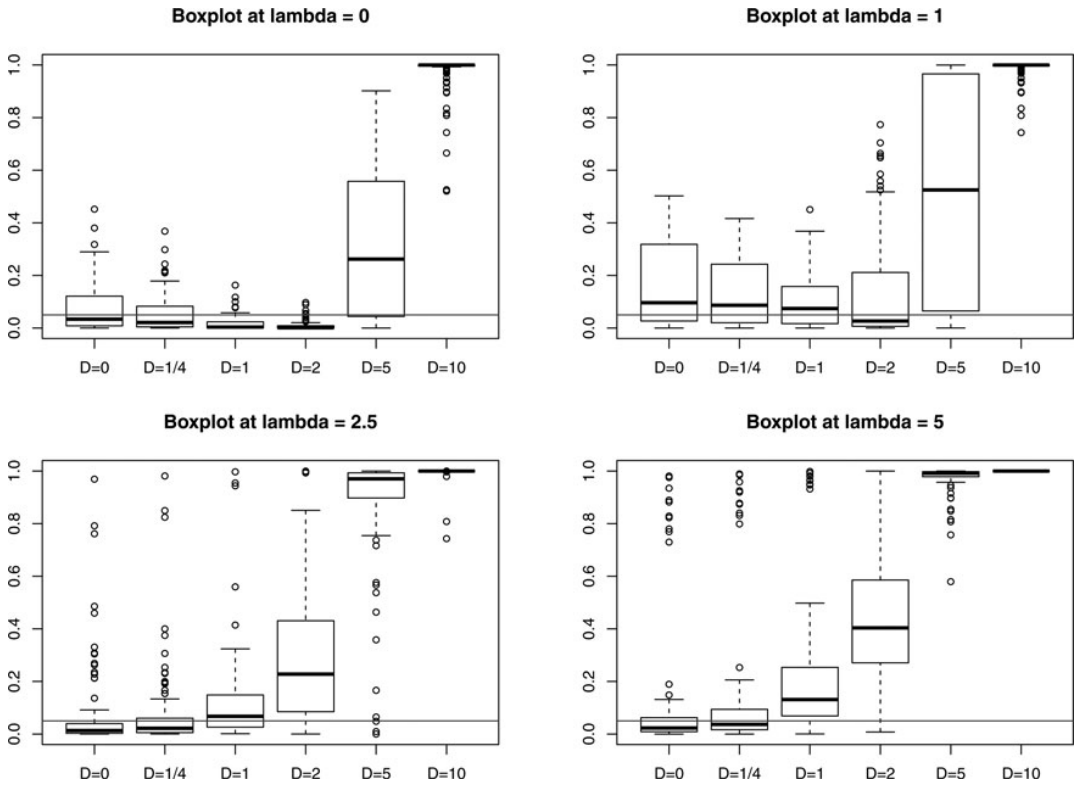

Figure 3. $n = 200$ and $p = 150$. Error df $= \chi^2_1 - 1$. Power of the ART with different choices of Δ (denoted as D in the x-axis) at $\lambda_n \in \{0, 1, 2.5, 5\}$. Horizontal blue line denotes the 5% level.

In the next section, we report results from a simulation study that compares the ART and the Λ_n-based test.

4. NUMERICAL RESULTS

We considered model (1) with iid (0, 1) errors ϵ_i such that $\mathbf{X} = (X_1, \ldots, X_p)$ and ϵ are independent. The covariate vectors $\mathbf{X}_i$ are iid realizations from $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (1 - \rho)\mathbf{I_p} + \rho\mathbf{1}\mathbf{1}'$, for $\rho = 0.8$. We considered two possible choices of the error distribution.

(E1) ϵ_i are iid $N(0, 1)$.
(E2) ϵ_i are iid $\chi^2_1 - 1$.

Further, we considered two specific choices of $\mathbf{b}$.

1. $\mathbf{b} = (b_1, 0, \ldots, 0)'$, with $b_1 \in \{0, 1/4, 1, 2, 5, 10\}$, and
2. $\mathbf{b} = p^{-1}\Delta \cdot \mathbf{1}$, with $\Delta \in \{0, 1/4, 1, 2, 5, 10\}$ where $\mathbf{1}$ is a p-vector of 1s.

Note that in the first case, the alternative is a spiked signal while in the other case, it is a diffuse signal. Choices of (n, p) were given by $(n, p) = (200, 10), (200, 50), (200, 150)$. For the ART, we also considered several values of λ_n, given by $\lambda_n \in \{0, 1/4, 1/2, 1, 2, 5, 10\}$. The last one is not shown in the plots due to similarity with $\lambda_n = 5$.

Tables 1 and 2 show the mean and standard deviations (sd-s) for the power computed over 100 samples (i.e., simulation runs) at $p = 10$ for the two alternatives driven by b_1 (spiked signal) and Δ (diffuse signal), respectively. The same is shown in Table 3 for $p = 150$ for the diffuse case. In all three cases, the error distribution is $\chi^2(1) - 1$. Table 1 suggest a good performance of the ART procedure in case of the spiked signal with a correct choice of λ_n. As seen from Table 2, the performance is slightly worse for the diffuse signal compared to the spiked signal. Table 3 for $p = 150$ shows a clear increase in the sd's of the power values.

Figure 1 shows that the variability of the empirical power of the ART increases rapidly when $p = 50$, even though a large choice of λ_n seemingly improves the performance for the same choice of Δ. Figures 2 and 3 show that the ART procedure is quite erratic when $p = 150$, especially in the diffuse case. Also, the Type I error rate can go much higher than the nominal value of 0.05. The high variability is an important concern while applying this test for moderately large values of p. The plot in Figure 4 shows the power curve for the alternative test procedure suggested in Section 3, at $p = 150$. The plots for $p = 10$ and $p = 50$ are not shown, since they are very similar. Besides the computational ease, the test maintains good power even for alternatives very close to the null, with a slightly worse performance under the $\chi^2(1) - 1$ error distribution compared to the $N(0,1)$ case for all three choices of p.

5. CONCLUDING REMARKS

[MQ] presents a novel approach to dealing with inconsistency of the bootstrap for nonregular estimators. Although the development here is given in the specific context of a testing problem in a regression model, the same approach can be applied more generally to other problems involving nonregular estimators. For the specific testing problem considered here, the Λ_n-based

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

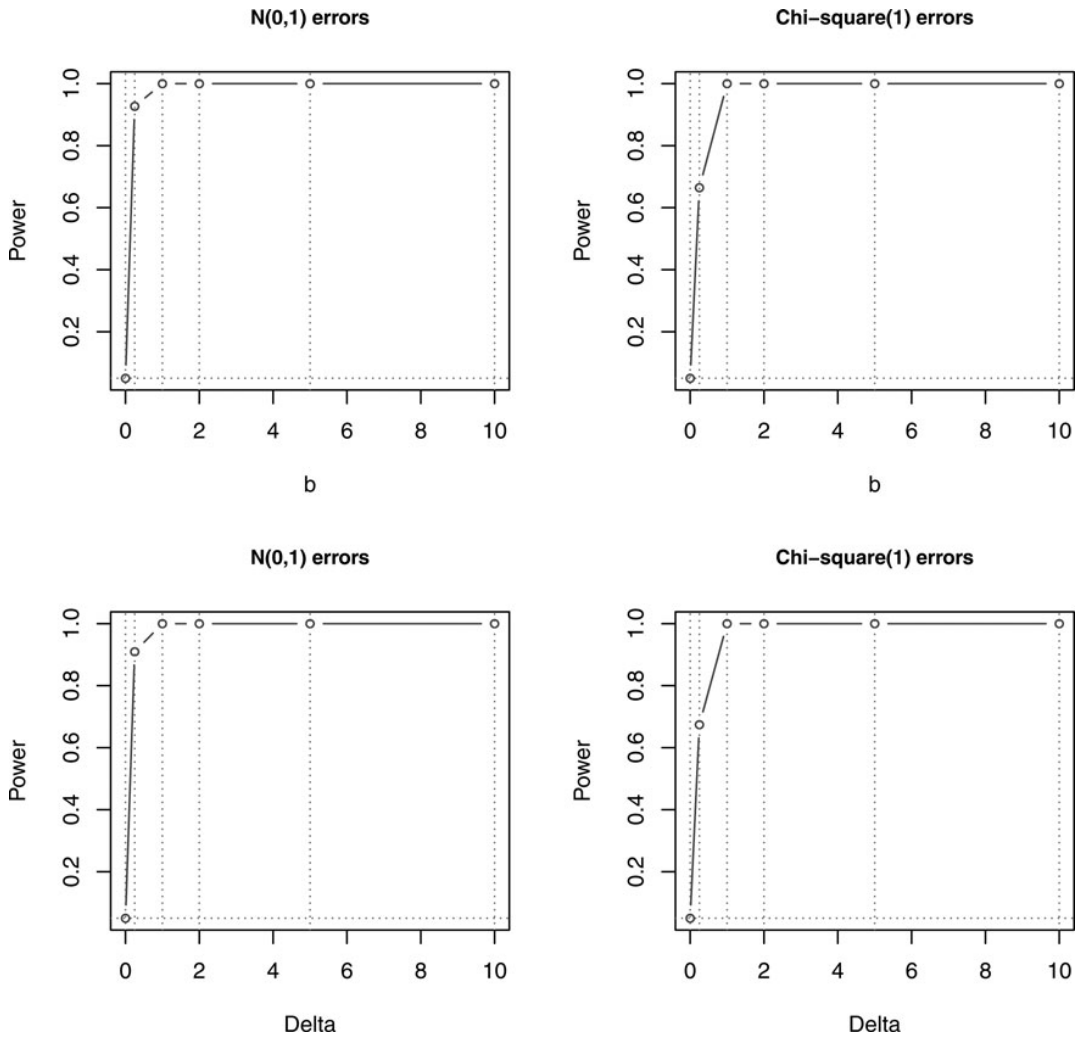

Figure 4. $n = 200$ and $p = 150$. Power of the alternative test procedure at different values of b and Δ for both $N(0, 1)$ and $(\chi^2_1 - 1)$ errors. The values of b and Δ are $\{0, 1/4, 1, 2, 5, 10\}$. Plots are based on 10,000 samples. Horizontal dotted line at bottom denotes the 5% level.

test given in Section 3 provides a simple alternative and seems to have some desirable properties, at least in the case of diffuse alternatives. However, it must be noted that the Λ_n-based test is not suitable for some of the other uses of the ART outlined in the article, such as identification of the most significant predictor, variable selection based on forward stepwise ART, etc.

[Received September 2013. Revised July 2014.]

REFERENCES

Andrews, D. W. K. (2000), "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405. [1434]

Athreya, K. B. (1987), "Bootstrap of the Mean in the Infinite Variance Case," *Annals of Statistics*, 15, 724–731. [1434]

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997), "Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses," *Statistica Sinica*, 7, 1–31. [1434]

Efron, B. (1992), "Jackknife-After-Bootstrap Standard Errors and Influence Functions," *Journal of The Royal Statistical Society,* Series A, 54, 83–127. [1435]

Freedman, D. A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218–1228. [1434]

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015), "A Two-Sample Test For Equality of Means in High Dimension," *Journal of the American Statistical Association*, 110, 837–849. [1435]

Lahiri, S. N. (1992), "Bootstrapping M-Estimators of a Multiple Linear Regression Parameter," *The Annals of Statistics*, 20, 1548–1570. [1434]

——— (2005), "Consistency of the Jackknife-After-Bootstrap Variance Estimator for the Bootstrap Quantiles of a Studentized Statistic," *The Annals of Statistics*, 33, 2475–2506. [1435]

Samworth, R. (2003), "A Note on Methods of Restoring Consistency to the Bootstrap," *Biometrika*, 90, 985–990. [1434]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]

Comment

Rajen D. SHAH and Richard J. SAMWORTH

1. INTRODUCTION

We are grateful for the opportunity to discuss this new test, based on marginal screening, of a global null hypothesis in linear models. Marginal screening has become a very popular tool for reducing dimensionality in recent years, and a great deal of work has focused on its variable selection properties (e.g., Fan and Lv 2008; Fan, Samworth, and Wu 2009). Corresponding inference procedures are much less well developed, and one of the interesting contributions of this article is the observation that the limiting distribution (here and throughout, we use the same notation as in the article) of $n^{1/2}(\hat{\theta}_n - \theta_0)$ is discontinuous at $\theta_0 = 0$. Such nonregular limiting distributions are well known to cause difficulties for the bootstrap (e.g., Beran 1997; Samworth 2003). Although in some settings, these issues are an artefact of the pointwise asymptotics of consistency usually invoked to justify the bootstrap (Samworth 2005), there are other settings where some modification of standard bootstrap procedures is required. Two such examples include bootstrapping Lasso estimators (Chatterjee and Lahiri 2011) and certain classification problems (Laber and Murphy 2011), where thresholded versions of the obvious estimators are bootstrapped, in an analogous fashion to the approach in this article.

2. STANDARDIZED OR UNSTANDARDIZED PREDICTORS?

Theorem 1 of the article reveals that the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$ may be quite complicated, even under the global null. To see this, consider a setting where $p = 2$, where X_1 and X_2 are highly correlated, but $\mathrm{var}(X_1) \ll \mathrm{var}(X_2)$. In this case, it is essentially a coin toss as to which predictor has the greater sample correlation with Y, but if $\hat{k}_n = 1$ then $|\hat{\theta}_n|$ will be tend to be large, while if $\hat{k}_n = 2$ then $|\hat{\theta}_n|$ will be tend to be small. The unfortunate consequence for the power of the procedure is that even for large sample sizes, we will only have a reasonable chance of rejecting the global null if we select X_1 (in particular, the power will be not much greater than 50% even when the signal is relatively large). For instance, consider the situation where $n = 100$, $p = 2$, $X_1 \sim N(0, 1)$, $X_2 = 20X_1 + \eta$, where $\eta \sim N(0, 1)$ is independent of X_1, and

$$Y = X_1 + \epsilon, \tag{2.1}$$

where $\epsilon \sim N(0, 1)$ is independent of X_1 (and η). Instead of using adaptive resampling test (ART) to obtain the critical value for the test of size $\alpha = 0.05$, we simply simulated from the null model where (X_1, X_2) are as above, but $Y = \epsilon \sim N(0, 1)$. A density plot of the values of $\hat{\theta}_n$ computed over 10,000 repetitions is shown in the top-left panel of Figure 1; note that the spike around 0 is due mainly to the 5017 occasions where X_2 happened to have higher absolute correlation with Y (i.e., $\hat{k}_n = 2$). The critical value for the test was taken to be the $100(1 - \alpha)$th quantile of the realizations of $|\hat{\theta}_n|$, namely, 0.171. Under the alternative specified by (2.1), $\hat{\theta}_n$ has a highly bimodal distribution as illustrated in the bottom-left panel of Figure 1. The only occasions when we were able to reject the null were when X_1 had higher absolute correlation with Y, yielding a power of 59.8%.

Fortunately, it is straightforward to construct a slightly modified test statistic that can yield great improvements. Indeed, it is standard practice in variable selection contexts to standardize each predictor X_k so that $\hat{\mathrm{E}}(X_k) = 0$ and $\widehat{\mathrm{var}}(X_k) = n$, and likewise to standardize the response so that $\hat{\mathrm{E}}(Y) = 0$ and $\widehat{\mathrm{var}}(Y) = n$. This amounts to using the test statistic $|\tilde{\theta}_n|$, where

$$\tilde{\theta}_n = \widehat{\mathrm{Corr}}(X_{\hat{k}_n}, Y).$$

Note that the definition of $\tilde{\theta}_n$ does not depend on whether the predictors and the response have been standardized or not, and that we have the simple expression

$$|\tilde{\theta}_n| = \max_{j=1,\ldots,p} |\widehat{\mathrm{Corr}}(X_j, Y)|.$$

For the example above, the top-right panel of Figure 1 gives a density plot of $\tilde{\theta}_n$ under the null; the critical value for our modified test was 0.198. Under the alternative, $\tilde{\theta}_n$ tends to be inflated, regardless of whether $\hat{k}_n = 1$ or $\hat{k}_n = 2$; in fact, we obtain an empirical power of 100%.

We emphasize that the problems described in this section are not observed in the simulation study of the article because there all of the predictors have equal variance. In the next section, we consider predictors and response standardized as above, and consider alternative approaches to calibrate the test statistic $n^{1/2}|\tilde{\theta}_n|$, as well as another test statistic proposed in Goeman, van de Geer, and van Houwelingen (2006).

Rajen D. Shah (E-mail: *r.shah@statslab.cam.ac.uk*) and Richard J. Samworth (E-mail: *r.samworth@statslab.cam.ac.uk*), Statistical Laboratory, University of Cambridge, Cambridge CB2 1TN, United Kingdom. The second author is supported by an Engineering and Physical Sciences Research Council Fellowship and a grant from the Leverhulme Trust.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1102142

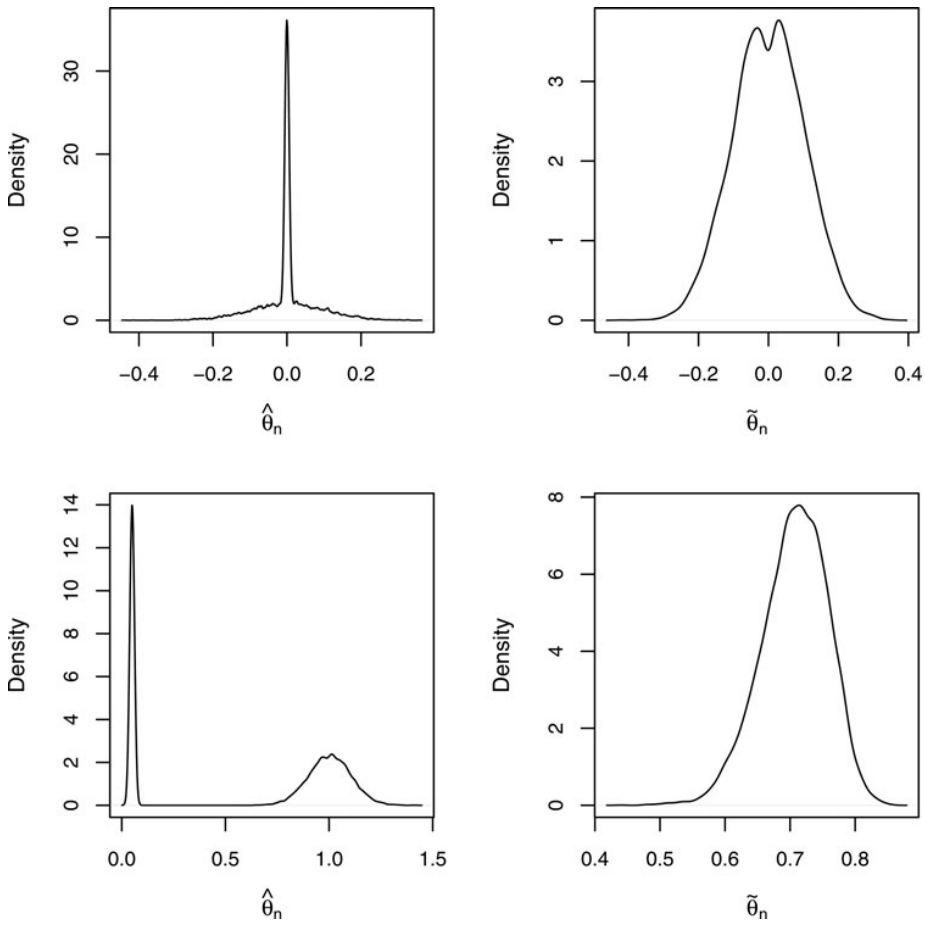

Figure 1. Top row: density plots of $\hat{\theta}_n$ (left) and $\tilde{\theta}_n$ (right) under the global null hypothesis for the example in Section 2. Bottom row: corresponding density plots of $\hat{\theta}_n$ (left) and $\tilde{\theta}_n$ (right) under the alternative specified in (2.1).

3. ALTERNATIVE APPROACHES

Although the nonregularities in the problem considered here make the construction of a confidence interval for θ_0 a challenging task, the particularly simple form of the global null hypothesis makes the testing problem amenable to several other approaches. Under the global null, $\mathbf{X}$ and Y are independent, so by the central limit theorem,

$$n^{1/2}\begin{pmatrix}\widehat{\text{Corr}}(X_1, Y)\\ \vdots \\ \widehat{\text{Corr}}(X_p, Y)\end{pmatrix} \xrightarrow{d} N_p(0, \Theta),$$

as $n \to \infty$, where $\Theta_{jk} = \text{Corr}(X_j, X_k)$. Then by the continuous mapping theorem,

$$n^{1/2}|\tilde{\theta}_n| \xrightarrow{d} \max_{j=1,\ldots,p} |Z_j|,$$

where $(Z_1, \ldots, Z_p)^T \sim N_p(0, \Theta)$. Since the distribution on the right does not depend on the distribution of Y, we can simulate $n^{1/2}|\tilde{\theta}_n|$ under the distribution of Y being (a) the empirical measure of the data $Y_1, \ldots, Y_n$, or (b) $N(0, 1)$, for example, to calibrate the test statistic. Figures 2 and 3 display the results of using these approaches in the numerical experiments of Section 4.1 in the article. Method (a) appears to yield a test with size not exceeding its nominal level and with similar power to the ART procedure. When the error distribution is normal, the size of the test from method (b) is exactly equal to the nominal level, up to Monte Carlo error; again the power in similar to that of ART.

An alternative approach to calibration is via permutations. Making the dependence of $\tilde{\theta}_n$ on $Y_1, \ldots, Y_n$ explicit, we note that the law of $\tilde{\theta}_n(Y_1, \ldots, Y_n)$ is the same as that of $\tilde{\theta}_n(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ for any permutation π of $\{1, \ldots, n\}$. The permutation test has the advantage over (a) and (b), of having its size not exceeding the nominal level regardless of the distribution of Y. Its power performance also seems close to that of ART.

Although it may seem natural to base test statistics on $\tilde{\theta}_n$, there are other possibilities. For example, Goeman, van de Geer, and van Houwelingen (2006) constructed a locally most powerful test for high-dimensional alternatives under the global null. We compare the power of their *globaltest* procedure with the approaches discussed above, in Figures 2 and 3. Overall, its performance is similar to that of ART, though in certain settings it seems to have a slight advantage and in others a slight disadvantage.

4. EXTENSIONS

In our view, the main attraction of ART is that it can be used to construct confidence intervals for θ_n. It would be interesting

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

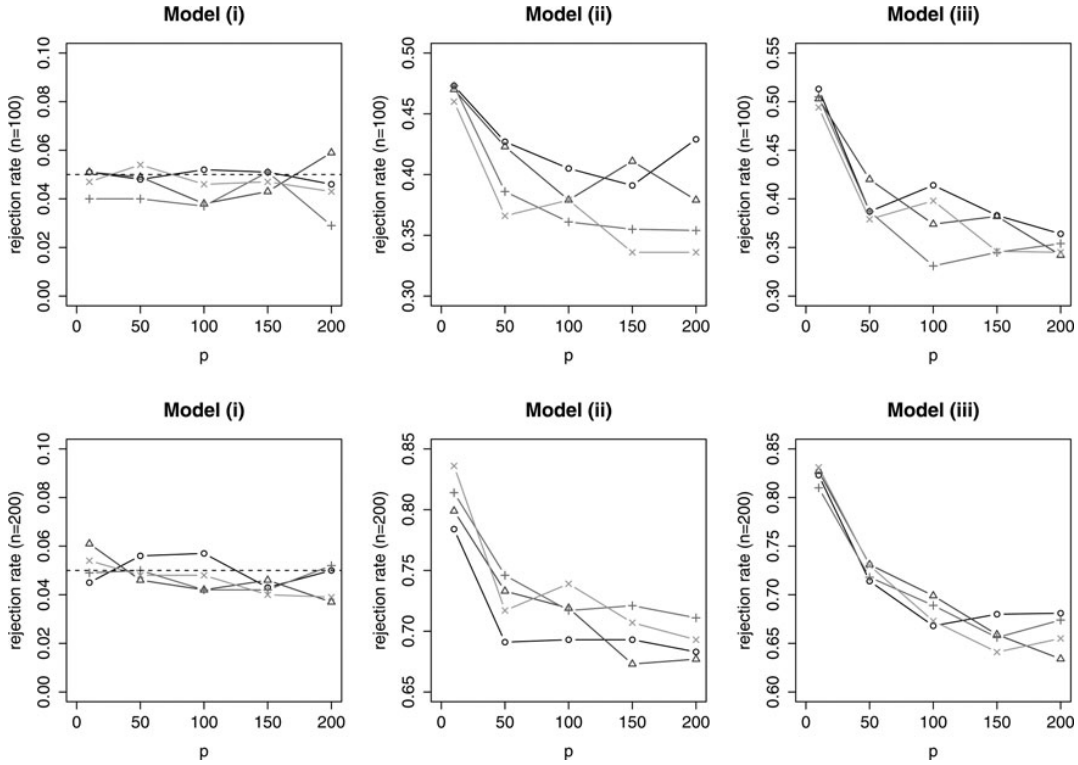

Figure 2. The same graphs as in Figure 1 ($\rho = 0.5$) of the original article but for *globaltest* (black circles), method (a) (green crosses), method (b) (red plus signs), and the permutation test (blue triangles). Note model (i) is the null model. (For interpretation of the references to color in this caption and that of Figure 3, the reader is referred to the web version of the article.)

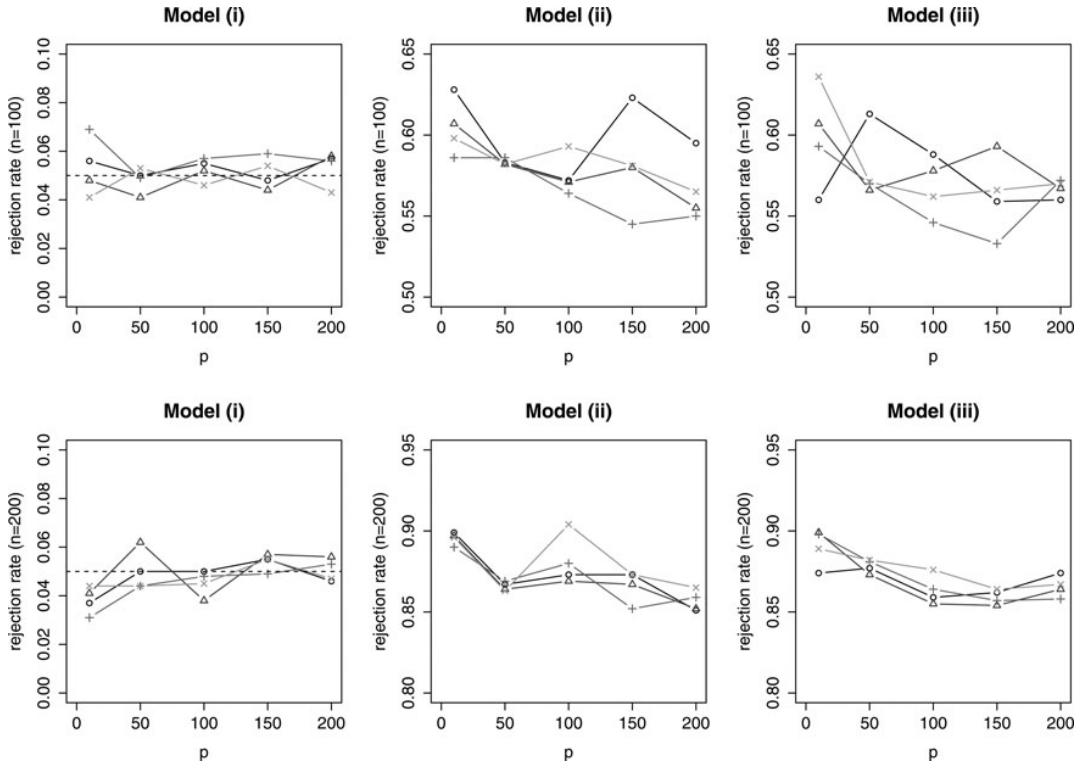

Figure 3. The same graphs as in Figure 2 ($\rho = 0.8$) of the original article but for *globaltest* (black circles), method (a) (green crosses), method (b) (red plus signs), and the permutation test (blue triangles).

to study empirically the coverage properties and lengths of these intervals. Another interesting related question would be to try to provide some form of uncertainty quantification for the variable having greatest absolute correlation with the response. The ideas of stability selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2013) provide natural quantifications of variable importance through empirical selection probabilities over subsets of the data. However, it is not immediately clear how to use these to provide, say, a (nontrivial) confidence set of variable indices that with at least $1 - \alpha$ probability contains all indices of variables having largest absolute correlation with the response (in particular this would be set full set $\{1, \ldots, p\}$ of indices under the global null).

Although understanding marginal relationships between variables and the response is useful in certain contexts, in other situations, the coefficients from multivariate regression are of more interest. It would be interesting to see whether the ART methodology can be extended to provide confidence intervals for the largest regression coefficients in absolute value.

[Received September 2013. Revised July 2014.]

REFERENCES

Beran, R. J. (1997), "Diagnosing Bootstrap Success," *Annals of the Institute of Statistical Mathematics*, 4, 1–24. [1439]

Chatterjee, A., and Lahiri, S. N. (2011), "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 106, 608–625. [1439]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society*, Series B, 70, 849–912. [1439]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038. [1439]

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006), "Testing Against a High Dimensional Alternative," *Journal of the Royal Statistical Society*, Series B, 68, 477–493. [1439,1440]

Laber, E., and Murphy, S. A. (2011), "Adaptive Confidence Intervals for the Test Error in Classification" (with discussion), *Journal of the American Statistical Association*, 106, 904–913. [1439]

Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection" (with discussion), *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [1442]

Samworth, R. (2003), "A Note on Methods of Restoring Consistency to the Bootstrap," *Biometrika*, 90, 985–990. [1439]

——— (2005), "Small Confidence Sets for the Mean of a Spherically Symmetric Distribution," *Journal of the Royal Statistical Society*, Series B, 67, 343–361. [1439]

Shah, R. D., and Samworth, R. J. (2013), "Variable Selection With Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society*, Series B, 75, 55–80. [1442]

Comment

Emre BARUT and Huixia Judy WANG

We congratulate Ian McKeague and Min Qian for a stimulating, timely, and interesting article on the important topic of hypothesis testing and post-selection inference in high-dimensional regression.

The authors developed an adaptive resampling test (ART) procedure for detecting the presence of significant predictors through marginal regression. In statistical applications, identifying the important predictors is at least as important as detecting their significance. For this purpose, the authors suggested a forward stepwise ART method, where in after identifying the first significant predictor, the ART procedure is successively applied by treating residuals from the previous stage as the new response until no more significant predictors are detected. The authors showed that this stepwise method performs very well in the cross-validation study of the HIV drug data. In the first section of our discussion, we carry out a small-scale simulation experiment to compare the performance of the forward stepwise ART method with other procedures built for high-dimensional inference. In these simulation experiments, it is seen that, unsurprisingly, the performance of ART (as well as other inference procedures) declines as the correlation between covariates increases.

It is well known in the literature that increased correlation between the variables can deteriorate the performance of variable selection procedures. However, we speculate that the performance of ART can be improved by extending ART to forward regression, in which the coefficients of already included variables are refit at each step. This would yield different results than the current forward stepwise ART procedure, which uses the residuals as the response at each stage; and hence is more susceptible to problems due to high correlation. This new forward-regression-based ART procedure will certainly require new theoretical developments as well as changes to the bootstrapping procedure.

As correlation between the important and the nonimportant variables increases, marginal-regression-based methods are known to be susceptible to the problem of "unfaithfulness" (Genovese et al. 2012): high correlation between the inactive variables and the active variables can cause (1) marginal coefficients of active variables to be close to zero and hence much harder to detect, (2) the marginal coefficients of inactive variables might be large because of their correlation to other important active variables. In the second section of our discussion, we argue that conditional marginal regression (e.g.,

Emre Barut is Assistant Professor (E-mail: *barut@gwu.edu*) and Huixia Judy Wang (E-mail: *judywang@gwu.edu*) is Associate Professor, Department of Statistics, George Washington University, Washington, DC 20052. The research is partially supported by the NSF CAREER Award DMS-1149355.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1100619

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

forward regression) may help alleviate some of the issues due to faithfulness.

1. FORWARD STEPWISE ART

In this section, we carry out a small-scale simulation study to compare the performance of the forward stepwise ART method, with the single sample splitting method (denoted by "sSplit") of Wasserman and Roeder (2009), and the multiple sample splitting method (denoted by "mSplit") of Meinshausen, Meier, and Bühlmann (2009). Both sSplit and mSplit use the three-stage stepwise regression, where data are randomly split into three parts to be used for screening, cross-validation and cleaning, respectively. The p-values from the mSplit method are calculated using 50 random sample splitting. All three methods are based on marginal regression of responses or residuals on each covariate separately in each step.

We generate data from the model $Y_i = \sum_{k=1}^{100} X_{ik}\beta_k + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, 1)$. Four cases are considered. The coefficients are set as $\beta_1 = \beta_2 = \beta_3 = 1$ in Cases 1 and 3, $(\beta_1, \beta_2, \beta_3) = (1, 2/3, 1/3)$ in Cases 2 and 4, and $\beta_k = 0$ for $k = 4, \ldots, 100$. The covariates $X_{ik}, k = 1, \ldots, 100$ are independent standard normal random variables in Cases 1 and 2, and they are from a multivariate normal with mean zero, variance 1, and an exchangeable correlation of 0.5 in Cases 3 and 4.

Table 1 summarizes the simulation results for $n = 99$ and $n = 300$. In Cases 1–2 with independent covariates, the ART method is the most effective one; it shows higher chance to identify the correct model while controlling the false positive rate close to the nominal level of 0.05. When covariates are correlated, all three marginal-regression-based methods have difficulty identifying the correct model especially for situations with small sample sizes or weak signals. For instance in Case 4, all three methods have difficulty selecting the third covariate with weaker coefficient $\beta_3 = 1/3$ even with larger sample size $n = 300$. Relatively speaking, in Cases 3 and 4, the ART method is competitive to mSplit and both work better than sSplit for smaller samples.

2. FAITHFULNESS AND CONDITIONAL MARGINAL REGRESSION

In this section, in an effort to understand the effects of correlation on forward stepwise ART's performance, we study the variable selection properties of forward regression. More specifically, we provide sufficient conditions for consistent variable selection of forward regression assuming some set $\mathcal{C}$ has already been recruited. We compare these conditions to those of Lasso and show that there are strong similarities.

We consider the setting in which the responses are generated from the following model,

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

and $\mathbf{Y}$ is a n-dimensional vector, $\mathbf{X}$ is an $n \times p$ matrix and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_n)$. We do not place any distributional assumptions on $\mathbf{X}$. Instead, we assume that it is deterministic, and the columns of $\mathbf{X}$, X_j, are normalized and each column has mean zero and variance 1. We define the Gram matrix $\mathbf{G}$ as $\mathbf{G} := n^{-1}\mathbf{X}^T\mathbf{X}$. For clarity, we do not use any notation based on n, although the variables, for example, $\mathbf{Y}$, $\mathbf{G}$, all depend on n.

We consider conditional marginal regression (Barut, Fan, and Verhasselt 2015), in which a predetermined set of conditioning variables $\mathcal{C} \subset \{1, 2, \ldots, p\}$ are included with each marginal regression. We let $\mathcal{P} = \{1, \ldots, p\}$ and define

$$\hat{\beta}_j^{\mathcal{C}} = \arg\min_{\beta_j^{\mathcal{C}}} \left(\min_{\hat{\beta}^{\mathcal{C}}} \left\| Y - X_{\mathcal{C}}\hat{\beta}^{\mathcal{C}} - X_j\hat{\beta}_j^{\mathcal{C}} \right\|_2^2 \right), \quad \text{for } j \in \mathcal{P}\backslash\mathcal{C}.$$

After the conditional marginal coefficients are estimated, one can perform screening by recruiting variables for which the conditional marginal coefficient is above a threshold value, t, that is, by screening out the set $\{j : |\hat{\beta}_j^{\mathcal{C}}| < t\}$. In the forward regression framework, one adds the variable with the highest coefficient to the set $\mathcal{C}$ (after adjusting for correlation) and repeats this over several iterations. Therefore, consistency results on conditional marginal screening can be extended to forward regression.

We assume that X_{ij} are bounded, although generalizations can be made to nonbounded but concentrated X_{ij} as in Fan and Song (2010). By the sub-Gaussianity of noise, using simple concentration arguments (Boucheron, Lugosi, and Massart 2013), it holds with high probability that

$$\left\| \beta_j^{\mathcal{C}} - \hat{\beta}_j^{\mathcal{C}} \right\|_\infty \le c_1\sigma^2 \sqrt{\frac{|\mathcal{C}| \log(p - |\mathcal{C}|)}{n}}, \tag{1}$$

where $\beta_j^{\mathcal{C}}$ are the noiseless (population) conditional marginal regression (CMR) coefficients, $|\mathcal{C}|$ is the cardinality of the set $\mathcal{C}$ and c_1 is a constant. The constant c_1 is inversely proportional to the minimum eigenvalue of the $|\mathcal{C}| + 1$ sized sub-blocks of $\mathbf{G}$.

To make the following presentation better, we introduce variable-specific partitions of the set $\mathcal{P}$. For a given $\mathcal{C}$ and j, we denote the set of other covariates by $\mathcal{O}$:

$$\mathcal{O} = \mathcal{P}\backslash(\mathcal{C} \cup j).$$

Furthermore, the Gram matrix $\mathbf{G}$ is partitioned as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\mathcal{C}\mathcal{C}} & \mathbf{G}_{\mathcal{C}j} & \mathbf{G}_{\mathcal{C}\mathcal{O}} \\ \mathbf{G}_{\mathcal{C}j}^T & G_{jj} & \mathbf{G}_{j\mathcal{O}} \\ \mathbf{G}_{\mathcal{C}\mathcal{O}}^T & \mathbf{G}_{j\mathcal{O}}^T & \mathbf{G}_{\mathcal{O}\mathcal{O}} \end{bmatrix},$$

where $\mathbf{G}_{AB} = \frac{1}{n}X_A^T X_B$. In addition, due to standardization, it holds that $G_{kk} = 1$ for all $k \in \{1, \ldots, p\}$.

It is trivial to show that the noiseless CMR coefficients $\beta_j^{\mathcal{C}}$ are given by

$$\beta_j^{\mathcal{C}} = \beta_j^\star + \frac{1}{\kappa_j^2}\left(\mathbf{G}_{j\mathcal{O}} - \mathbf{G}_{\mathcal{C}j}^T\mathbf{G}_{\mathcal{C}\mathcal{C}}^{-1}\mathbf{G}_{\mathcal{C}\mathcal{O}}\right)\beta_{\mathcal{O}}^\star, \tag{2}$$

where $\kappa_j^2 = 1 - \mathbf{G}_{\mathcal{C}j}^T\mathbf{G}_{\mathcal{C}\mathcal{C}}^{-1}\mathbf{G}_{\mathcal{C}j} < 1$, that is, the conditional variance of X_j given $X_{\mathcal{C}}$. The second term in $\beta_j^{\mathcal{C}}$ can be expressed as the "correlation of j and $\mathcal{O}$, conditional on $\mathcal{C}$." That is, conditional on $\mathcal{C}$, if the jth variable is not significantly correlated to other variables, the second term will be small. This is not a surprising result, since any active variables that are not included in $\mathcal{C}$ will not "disrupt" the estimation of $\beta_j^{\mathcal{C}}$ if they do not have any correlation with X_j conditional on $\mathcal{C}$.

We next present the conditions for which, given some set $\mathcal{C}$, forward regression recruits an active variable with high probability. We use $\mathcal{S}$ to represent the set of active variables, that is, $\mathcal{S} = \{j \in \mathcal{P} : \beta_j^\star \neq 0\}$, and we use $\mathcal{N} = \mathcal{P}\backslash\mathcal{S}$ to denote the complement of $\mathcal{S}$.

Table 1. Simulation results for three stepwise regression methods

		$n = 99$					$n = 300$				
Case	Method	OracleP	FP	TP1	TP2	TP3	OracleP	FP	TP1	TP2	TP3
1	ART	1.00	0.05	1.00	1.00	1.00	1.00	0.05	1.00	1.00	1.00
	sSplit	0.71	0.00	0.87	0.86	0.88	1.00	0.01	1.00	1.00	1.00
	mSplit	0.99	0.00	0.99	1.00	1.00	1.00	0.00	1.00	1.00	1.00
2	ART	0.31	0.04	1.00	1.00	0.31	0.98	0.04	1.00	1.00	0.98
	sSplit	0.02	0.00	0.97	0.62	0.03	0.46	0.00	1.00	1.00	0.46
	mSplit	0.00	0.00	1.00	0.71	0.00	0.57	0.00	1.00	1.00	0.57
3	ART	0.49	0.01	0.83	0.83	0.83	1.00	0.00	1.00	1.00	1.00
	sSplit	0.25	0.10	0.58	0.60	0.58	1.00	0.00	1.00	1.00	1.00
	mSplit	0.75	0.01	0.92	0.90	0.92	1.00	0.00	1.00	1.00	1.00
4	ART	0.01	0.06	1.00	0.88	0.07	0.16	0.00	1.00	1.00	0.16
	sSplit	0.00	0.06	0.79	0.36	0.03	0.25	0.01	1.00	0.98	0.26
	mSplit	0.00	0.01	0.98	0.50	0.02	0.23	0.00	1.00	1.00	0.23

Notes: OracleP is the proportion of selecting the correct active covariates, FP is the false positive rate (i.e., the proportion of selecting at least one inactive covariates), and TP1, TP2, and TP3 are the proportions of selecting the first three active covariates, respectively.

Condition 1 (Beta-min). For the active variables it holds that,

$$\min_{j\in\mathcal{S}} \left|\beta_j^\star\right| > c_{\beta\min} > 0.$$

The constant $c_{\beta\min}$ can depend on n and/or p. In the literature, $c_{\beta\min}$ is often assumed to be on the order of $\sqrt{\log p/n}$.

Condition 2 (Beta-max). Active variables that are not included in $\mathcal{C}$ are bounded above in magnitude, that is

$$\max_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\beta_j^\star\right| = \left\|\beta_{\mathcal{S}\setminus\mathcal{C}}^\star\right\|_\infty \le c_{\beta\max}.$$

Remark 1. Although the Beta-min condition is plausible, and almost always necessary in a high-dimensional framework, the Beta-max condition is much more restrictive as it requires that all of the variables with large coefficients are contained in the set $\mathcal{C}$. However, in practice, one would expect that bigger variables are easier to "spot," and the Beta-max condition is not very restrictive for such situations. Note that, there are no assumptions about the other elements of $\mathcal{C}$. The conditioning set can include nonactive variables and the results continue to hold as long as the largest active coefficients are included in $\mathcal{C}$.

The variables recruited with the conditional set $\mathcal{C}$ will be in the active set, if it holds that

$$\min_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\hat{\beta}_j^{\mathcal{C}}\right| > \max_{j\notin\mathcal{S}\cup\mathcal{C}} \left|\hat{\beta}_j^{\mathcal{C}}\right|.$$

Conditioning on the high probability set in which Equation (1) holds, we can write sufficient conditions as

$$\begin{aligned}&\min_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\beta_j^{\mathcal{C}} \pm C\sigma^2\sqrt{\frac{|\mathcal{C}|\log(p-|\mathcal{C}|)}{n}}\right| \\ &> \max_{j\notin\mathcal{S}\cup\mathcal{C}} \left|\beta_j^{\mathcal{C}} \pm C\sigma^2\sqrt{\frac{|\mathcal{C}|\log(p-|\mathcal{C}|)}{n}}\right| \Leftarrow \\ &\min_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\beta_j^{\mathcal{C}}\right| > \max_{j\notin\mathcal{S}\cup\mathcal{C}} \left|\beta_j^{\mathcal{C}}\right| + 2\underbrace{C\sigma^2\sqrt{\frac{|\mathcal{C}|\log(p-|\mathcal{C}|)}{n}}}_{\Delta_{\mathcal{C},n,p}}. \quad (3)\end{aligned}$$

If condition (3) holds, then with high probability, forward regression recruits an active variable, one that is in $\mathcal{S}$.

Plugging in the expression for $\beta_j^{\mathcal{C}}$ in (2), we rewrite the condition (3). First, we create matrix $A \in \mathbb{R}^{q\times q}$, where $q = p - |\mathcal{C}|$, and set $A_{jj} = 0$ for all j. The remaining terms in the jth row of A are given by

$$A_{j,-j} = \left[\frac{1}{\kappa_j^2}\left(G_{j\mathcal{O}} - G_{\mathcal{C}j}^T G_{\mathcal{C}\mathcal{C}}^{-1} G_{\mathcal{C}\mathcal{O}}\right)\right],$$

where $\mathcal{O}$ is implicitly dependent on j. Entries of matrix A can be thought as the conditional covariances (conditional on $\mathcal{C}$) between covariates not in $\mathcal{C}$. It then follows that, condition (3) is equivalent to

$$\min_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\beta_j^\star + A_{j,-j}^T\beta_{\mathcal{O}}^\star\right| > \max_{j\notin\mathcal{S}\cup\mathcal{C}} \left|A_{j,-j}^T\beta_{\mathcal{O}}^\star\right| + \Delta_{\mathcal{C},n,p}. \quad (4)$$

Next, we obtain a lower bound for the LHS, and an upper bound for the RHS of the equation. The LHS of (4) can be bounded below as,

$$\begin{aligned}\min_{j\in\mathcal{S}\setminus\mathcal{C}} \left|\beta_j^\star + A_{j,-j}^T\beta_{\mathcal{O}}^\star\right| &\ge \min_{j\in\mathcal{S}\setminus\mathcal{C}} \left(\left|\beta_j^\star\right| - \left|A_{j,-j}^T\beta_{\mathcal{O}}^\star\right|\right) \\ &\ge c_{\beta\min} - \max_{j\in\mathcal{S}\setminus\mathcal{C}} \left|A_{j,-j}^T\beta_{\mathcal{O}}^\star\right|. \quad (5)\end{aligned}$$

With Condition 2, the last term in Equation (5) can be bounded above by

$$\begin{aligned}\max_{j\in\mathcal{S}\setminus\mathcal{C}} \left|A_{j,-j}^T\beta_{\mathcal{O}}^\star\right| &= \max_{j\in\mathcal{S}\setminus\mathcal{C}} \left|A_{j,(\mathcal{S}\setminus(C\cup j))}^T\beta_{(\mathcal{S}\setminus(C\cup j))}^\star\right| \\ &\le \max_{j\in\mathcal{S}\setminus\mathcal{C},\|\nu\|_\infty\le c_{\beta\max}} \left|A_{j,(\mathcal{S}\setminus(C\cup j))}^T\nu\right| \\ &\le c_{\beta\max}\left\|A_{(\mathcal{S}\setminus C),(\mathcal{S}\setminus C)}\right\|_\infty,\end{aligned}$$

where the norm $\|\cdot\|_\infty$ is defined as the maximum of the absolute sum of the rows of the matrix. Similarly, the other term in (4) can be bounded with the same norm

$$\begin{aligned}\max_{j\notin\mathcal{S}\cup\mathcal{C}} \left|A_{j,-j}^T\beta_{\mathcal{O}}^\star\right| &\le \max_{j\in\mathcal{N}\setminus\mathcal{C},\|\nu\|_\infty\le c_{\beta\max}} \left|A_{j,(\mathcal{S}\setminus(C\cup j))}^T\nu\right| \\ &\le c_{\beta\max}\left\|A_{(\mathcal{N}\setminus\mathcal{C}),(\mathcal{S}\setminus C)}\right\|_\infty.\end{aligned}$$

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]

We now state our main result.

Lemma 1. Given some conditioning set $\mathcal{C}$, the forward regression recruits an active variable with high probability if Conditions 1 and 2 hold and

$$c_{\beta\min} > c_{\beta\max}\left(\left\|A_{(\mathcal{N}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty + \left\|A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty\right) + \Delta_{\mathcal{C},n,p}. \quad (7)$$

Remark 2. A stronger statement can be made: if the conditions hold, the first $|\mathcal{S}\setminus C|$ coefficients selected by conditional marginal regression will be active. If the conditional correlation coefficient to other active variables is small, one can work with much more general conditions. In fact, if conditional on $\mathcal{C}$, none of the active variables are correlated (for instance in an equal correlation design in which $\mathcal{C}$ includes one element), the condition (6) simply becomes

$$c_{\beta\min} > \left\|A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\beta^\star_{\mathcal{S}\setminus\mathcal{C}}\right\|_\infty + \Delta_{\mathcal{C},n,p} = 0 + \Delta_{\mathcal{C},n,p} = \Delta_{\mathcal{C},n,p}.$$

As given in Genovese et al. (2012), three sufficient conditions for the variable selection consistency of Lasso are:

- Minimum eigenvalue condition: $\lambda_{\min}(\mathbf{G}_{\mathcal{S},\mathcal{S}}) \geq c_2 > 0$,
- Irrepresentability condition: $\|\mathbf{G}_{\mathcal{N}\mathcal{S}}\mathbf{G}_{\mathcal{S}\mathcal{S}}^{-1}\|_\infty < 1$,
- Tuning parameter condition: $c_{\beta\min} > \lambda\|\mathbf{G}_{\mathcal{S},\mathcal{S}}^{-1}\|_\infty$,

where λ is the tuning parameter for the penalty term in Lasso and needs to be taken on the order of $\sqrt{\log p/n}$. For ease of comparison, we rewrite the sufficient condition (6) of our Lemma as follows:

$$\begin{aligned} c_{\beta\min} &> \eta_1 c_{\beta\max}\left\|A_{(\mathcal{N}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty, \\ c_{\beta\min} &> \eta_2 c_{\beta\max}\left\|A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty, \\ c_{\beta\min} &> \eta_3\Delta_{\mathcal{C},n,p}, \end{aligned}$$

where $\eta_1+\eta_2+\eta_3 \leq 1$. We compare the condition (6) to reconstruction conditions for Lasso in Table 2. As can be seen from Table 2, the conditions are comparable. The minimum eigenvalue condition is replaced by a minimum eigenvalue condition on the submatrices of $\mathbf{G}$. This condition is necessary to ensure that the conditional coefficients converge to their true (population) values.

The irrepresentability condition of Lasso is also replaced with a very similar condition. The Lasso condition limits the covariance of the active and nonactive variables, while the same condition for conditional screening limits the conditional covariance of the active and nonactive variables, conditioned on $\mathcal{C}$. If conditioning helps reduce some of the correlation between the variables, conditional covariance will be significantly smaller. Hence, in practical applications with highly correlated variables, one would expect this condition to be easier to satisfy than the irrepresentability condition of Lasso.

Table 2. Comparison of variable selection consistency conditions for Lasso and conditional marginal screening.

Lasso condition	Related condition for conditional screening
$\lambda_{\min}(\mathbf{G}_{\mathcal{S},\mathcal{S}}) \geq c_2 > 0$	$\min_{j\in\mathcal{P}\setminus\mathcal{C}}\lambda_{\min}(\mathbf{G}_{\mathcal{S}\cup j,\mathcal{S}\cup j}) > \frac{1}{c_1}$
$\left\|\mathbf{G}_{\mathcal{N}\mathcal{S}}\mathbf{G}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty < 1$	$\eta_1\left\|A_{(\mathcal{N}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty < \frac{c_{\beta\min}}{c_{\beta\max}}$
$c_{\beta\min} > \lambda\left\|\mathbf{G}_{\mathcal{S},\mathcal{S}}^{-1}\right\|_\infty$	$c_{\beta\min} > \eta_2 c_{\beta\max}\left\|A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\right\|_\infty$

Finally, the tuning parameter condition is analogous to the condition on $A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}$. The tuning parameter λ is generally taken on the order of $O(\sqrt{\log p/n})$. If the conditioned set can be chosen to ensure $c_{\beta\max} = O(\sqrt{\log p/n})$, which will happen if large variables are easily recognizable, these two conditions are very similar. In addition, by conditioning on more variables, one would expect $\|A_{(\mathcal{S}\setminus\mathcal{C}),(\mathcal{S}\setminus\mathcal{C})}\|_\infty$ to decrease. Therefore, as is the case with the other conditions, the recovery conditions for conditional regression are often less stricter than those of Lasso.

These results suggest that forward regression can be a very powerful method for variable selection. In fact, forward regression can overperform Lasso, if in the early stages forward regression recruits variables that are large in magnitude (so that $c_{\beta\max}$ is small) and/or if recruited variables have high correlation with others.

3. CONCLUSION

We would like to once again congratulate the authors for their timely and beautiful results on this important topic. We expect that some readers may be cautious in implementing ART, thinking that unfaithfulness causes issues with marginal regression. To ease such concerns, we have shown forward regression will select important variables under conditions that are comparable to those of Lasso. It would be very interesting to see an adaptation of ART for forward regression, and we hope that the results presented in this discussion are encouraging for such a method. We conclude by thanking the authors for this inspirational and stimulating article.

[Received September 2013. Revised July 2014.]

REFERENCES

Barut, E., Fan, J., and Verhasselt, A. (2015), "Conditional Sure Independence Screening," *Journal of the American Statistical Association*, to appear. [1443]

Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford, UK: Oxford University Press. [1443]

Fan, J., and Song, R. (2010), "Sure Independence Screening In Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1443]

Genovese, C., Jin, J., Wasserman, L., and Yao, Z. (2012), "A Comparison of the Lasso and Marginal Regression," *Journal of Machine Learning Research*, 13, 2107–2143. [1442,1445]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "*P*-value for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [1443]

Wasserman, L., and Roeder, K. (2009), "High-Dimensional Variable Selection," *The Annals of Statistics*, 37, 2178–2201. [1443]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]

Comment

Lawrence D. BROWN and Daniel MCCARTHY

1. INTRODUCTION

The following comments relate to material in McKeague and Qian (2015) (henceforth referred to as M&Q). We have found this material to be very worthwhile reading and we thank the organizers for the opportunity to provide these comments.

M&Q study a maximum statistic, $\hat{\theta}_n$, that can be used to test for the null hypothesis of no effect of covariates in a linear-model analysis. Their major focus is on the development of a bootstrap style procedure that serves the dual purpose of estimating the distribution of $\hat{\theta}_n$ and then using this estimate as the basis for a test of the null hypothesis. Our perspective is that their procedure is composed of two almost separate components—a simulation of the distribution under the null and a bootstrap estimate that is valid away from the null. Our comments focus on the testing component of their procedure, and on alternate tests for this hypothesis.

The first part of our discussion deals with this perspective of their formulation. It points to concerns that are treated in the remainder of our discussion and, at the end, raises a few questions for the authors. Section 2 of our discussion sketches a generalization of their basic statistical model that we have treated elsewhere in more detail, and suggests that a modification of their test may be suitable also for this generalization. Section 3 discusses a related, though different, test of their null hypothesis that is embedded in Berk et al. (2013). Section 4 describes bootstrap ideas that do yield a valid test of the null hypothesis without attempting to estimate the distribution of $\hat{\theta}_n$. That section concludes with some prospective remarks about our ongoing research into bootstrap methods for this and related problems.

1.1 Formulation

The essential structure of the observed data is implicit in the first sentence of Section 2 and in Equation (2) of McKeague and Qian (2015) (henceforth referred to as M&Q). The observations are a sample $\{\mathbf{X}_i, Y_i : i = 1, \ldots, n\}$ from a population whose distribution has the property that

$$Y = \alpha_0 + \mathbf{X}^T \beta + \varepsilon. \tag{1.1}$$

Here $\mathbf{X}$ is a p-vector of covariates and β is a p-vector of parameters. Since $\mathbf{X}$ is random we refer to this formulation as having a random-covariate structure in contrast to the traditional structure in which the elements of each $\mathbf{X}_i$ are viewed as fixed constants. The marginal distribution of $\mathbf{X}$ is not known or constrained (except that it is assumed to have a finite covariance matrix). The residual variables $\{\varepsilon_i : i = 1, \ldots, n\}$ are an iid sample and independent of $\{\mathbf{X}_i\}$. Their distribution is not specified in advance, except that they are assumed to have a finite variance, and hence be homoscedastic.

A primary goal of M&Q is to develop a test of the null hypothesis $H_0 : \beta = \mathbf{0}$ in (1.1) that is based on a maximal statistic drawn from simple regressions rather than from the multiple regression analysis. In order to describe this statistic, and to prepare for further discussion, some additional notation may be helpful. The following quantities are discussed in M&Q but not given explicit notations. The population and the sample slope coefficients from the simple (one-dimensional, marginal) regressions are

$$\beta_k^1 = \frac{\operatorname{cov}(\mathbf{X}_k, Y)}{\operatorname{var}(\mathbf{X}_k)}, \quad \hat{\beta}_k^1 = \frac{\widehat{\operatorname{cov}}(\mathbf{X}_k, Y)}{\widehat{\operatorname{var}}(\mathbf{X}_k)}, \quad k = 1, \ldots, p.$$

The corresponding t-statistics are

$$T_k^1 = \frac{\hat{\beta}_k^1}{\widehat{\operatorname{SE}}\left(\hat{\beta}_k^1\right)},$$

where $\widehat{\operatorname{SE}}^2(\hat{\beta}_k^1) = \sum (Y_i - \bar{Y} - \hat{\beta}_k^1 (X_{ki} - \bar{X}_k))^2 / \sum \left(X_{ki} - \bar{X}_k\right)^2$.

M&Q then define

$$\hat{k}_n = \arg\max |\widehat{\operatorname{corr}}(\mathbf{X}_k, Y)| = \arg\max \left|T_k^1\right|, \tag{1.2}$$

where the last equality results from standard textbook manipulations. They then consider the statistic

$$\hat{\theta}_n = \hat{\beta}_{\hat{k}_n}^1 = \frac{\widehat{\operatorname{cov}}\left(\mathbf{X}_{\hat{k}_n}, Y\right)}{\widehat{\operatorname{var}}\left(\mathbf{X}_{\hat{k}_n}\right)}. \tag{1.3}$$

As M&Q aptly point out, direct bootstrap methods are not appropriate if one wishes to use this statistic to test H_0 (the flaw in attempting to use a bootstrap here is evident in the simulations in Section 4: The conventional bootstrap procedure leads to a strongly anti-conservative test of H_0.) For a test of H_0 one needs the null distribution of $\hat{\theta}_n$. This null distribution cannot accurately be obtained via a standard bootstrap since the distribution of $\hat{\theta}_n$ does not converge uniformly in a $1/\sqrt{n}$ neighborhood of the null. But it is possible to simulate the distribution of $\hat{\theta}_n$ under the null. The simulated random variable is given in the paper as $\mathbf{V}_n^*(\mathbf{0})$. This random quantity depends on an independent simulation of a mean-zero multivariate normal random vector, $\mathbf{Z}(\mathbf{0})$, whose covariance matrix is described in a display in Theorem 1. (The covariance matrix for this vector could be estimated directly from the data. In their computer

Lawrence Brown (E-mail: *lbrown@wharton.upenn.edu*) and Daniel McCarthy (E-mail: *danielmc@wharton.upenn.edu*), Department of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut St., Philadelphia, PA 19104-6340.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1099536

code (which they have kindly shared with us) M&Q appear to use a bootstrap to estimate this covariance matrix. This may be more accurate in practice than a direct estimate of the covariance, but the paper provides no evidence on this secondary issue.)

The more general bootstrap result in Theorem 2 is the basis for the author's ART procedure. This should be viewed as a marriage of the simulation described above and a more standard bootstrap estimate of the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$. The bootstrap is used when the data convincingly reject the null hypothesis (whenmax $\left(|T_n|\,,\ \left|T_n^*\right|\right) > \lambda_n$) and otherwise the simulation at $b_0 = \beta_0 = \mathbf{0}$ is used. Theorem 2 should not be interpreted as providing a reliable bootstrap estimate for the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ at every true parameter (which M&Q correctly remark does not exist). Such a global bootstrap based on this theorem (and hence without knowledge of b_0) would require a consistent estimate of b_0, but such an estimate does not exist. (One would also need to know, or assume, a priori that β_0 satisfies the special assumption of Theorem 1 as expressed in the paragraph above the Theorem. This would not, for example, be the case if the true β in (1.1) were sparse.)

We have two additional questions about the formal results.

(i) The ART simulation and test is based on the distribution of the selected slope statistic in (1.3). But selection is based on the maximal $|t|$-statistic as in (1.2). The maximal $|t|$-statistic is $\left|T^1_{\hat{k}_n}\right|$, and, in general, $\left|T^1_{\hat{k}_n}\right| \neq \hat{\theta}_n$. Such a test statistic would have the advantage of being invariant under coordinate-wise affine transformations of the X-variables, whereas the statistic in (1.3) is not. Did you investigate the performance of such a procedure? In the simulations of Section 4 there would be very little difference in numerical value or performance since the X-coordinates there are independent with equal variance, but differences might be more noticeable in other settings.

(ii) Theorems 1 and 2 are proved within a formulation in which p remains fixed as $n \to \infty$. Yet, the simulations in Section 4 address some cases in which p is comparable to n, or even larger. (It is an advantage of the ART procedure over more familiar procedures based on inference about the full vector β in (1.1) that it can numerically deal with such situations.) The ART procedure appears to perform satisfactorily even in these large p cases. Is this perhaps only because the choice of standard normality for the distributions of $\mathbf{X}$ and ε are so favorable to ART, or is this a more general phenomenon? Is there any asymptotic theory to justify the simulation at the heart of ART when $p \to \infty$ along with n?

2. ASSUMPTION-LEAN MODELS

Buja et al. (2015) contains a detailed exposition of an "assumption-lean" regression formulation. In such a formulation one need only assume that the observed variables are a random sample $\{\mathbf{X}_i, Y_i : i = 1, \ldots, n\}$ from a joint distribution possessing low-order moments. The target of inference is the population slope; this can be defined in any one of several equivalent ways. A straightforward version is to define the population slope vector via $\beta^\bullet = \arg\min_\gamma E\,(Y - \mathbf{X}\gamma)^2$. An alternate form that is more analogous to (1.1) involves writing

$$Y = \alpha_0 + \mathbf{X}^T\beta^\bullet + \varepsilon \quad \text{where} \quad \text{cov}\,(\mathbf{X},\ \varepsilon) = 0. \tag{2.1}$$

(We use the notation $\beta^\bullet$ here, rather than just β in order to distinguish this formulation from that of (1.1). (2.1) is more general that (1.1), but if the assumptions of (1.1) hold and the population model is full-rank then $\beta^\bullet = \beta$ and $\beta^{\bullet 1} \neq \beta^1$.)

This formulation shares with (1.1) the feature that $\mathbf{X}$ is a random vector whose marginal distribution is unknown (except for the existence of low order moments). But it is otherwise much broader and assumption-lean. The randomness of $\mathbf{X}$ (in both (1.1) and (2.1)) has an important side benefit in that in general it justifies the asymptotic use of an $\mathbf{X}$-Y bootstrap such as that in Theorem 2 when $\beta^\bullet$ is not in a $\sqrt{n}$ neighborhood of 0, and otherwise satisfies the assumptions in that theorem. But, as noted above, the core of the ART procedure as a test is really a simulation of the null distribution of the statistic $\hat{\theta}_n$. We believe that this simulation should also be valid in the assumption-lean setting of (2.1), and should hence lead to a useful test of $H_0 : \beta^\bullet = 0$. Here's why.

Buja et al. (op. cit.) describes interpretations and inference for $\beta^\bullet$ from data as in (2.1), and several other aspects of such a formulation. The sandwich estimator of Huber (1967) and White (1980a,b; 1982) plays a key role in such inference. For M&Q a key ingredient of the simulation in ART is the covariance matrix in Theorem 1 at $\beta = \mathbf{0}$. This relates to a form of the sandwich estimator for the covariance matrix of the vector of marginal sample slopes, $\hat{\beta}^1$. The appropriate sandwich estimator would be

$$\left[\text{diag}\left(\{\widehat{\text{var}}\,(\mathbf{X}_k)\}\right)\right]^{-1}\mathbf{M}\left[\text{diag}\left(\{\widehat{\text{var}}\,(\mathbf{X}_k)\}\right)\right]^{-1} \quad \text{where}$$
$$\mathbf{M}_{k\ell} = n^{-1}\sum\nolimits_i \left(Y_i - \bar{Y} - \hat{\beta}^1_k\left((\mathbf{X}_i)_k - \bar{\mathbf{X}}_k\right)\right)^2\left(\left((\mathbf{X}_i)_k - \bar{\mathbf{X}}_k\right)\right)^2. \tag{2.2}$$

(If β^1 is assumed to lie in a $1/\sqrt{n}$ neighborhood of $\mathbf{0}$ then the term $\hat{\beta}^1_k\left((\mathbf{X}_i)_k - \bar{\mathbf{X}}_k\right)$ in (2.2) is asymptotically negligible and can be ignored.) The matrix $\mathbf{M}$ is very similar to the covariance matrix described in Theorem 1; we believe they are asymptotically equivalent when β^1 is assumed to lie in a $1/\sqrt{n}$ neighborhood of $\mathbf{0}$. (The inverse diagonal matrix terms do not appear in the covariance expression described in Theorem 1, but are instead accommodated in the first denominator of (4).)

In summary, we believe that the simulation idea embodied within ART can be directly applied to testing the null hypothesis $H_0 : \beta^\bullet = 0$. Thus the simulation component in ART may turn out to be more flexible and robust than appears from the specific formulation via (1.1). (The bootstrap idea in M&Q for the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ away from the null and other special points excluded by the assumptions of Theorem 1 is also almost automatically valid under (2.1).)

3. POSI

Berk et al. (2013) provided a simultaneous confidence interval procedure for estimates of slope coefficients in a setting like that of (1.1), but with the elements of $\mathbf{X}$ treated as fixed constants, rather than as random variables. To be more precise, the setting for that paper involves observation of an $n \times 1$ vector $\mathbf{Y}$ satisfying

$$\mathbf{Y} = \alpha_0\mathbf{1} + \mathbf{X}_{n\times p}\beta^\circ + \varepsilon \quad \text{where} \quad \varepsilon \sim N_n\left(0,\ \sigma^2\mathbf{I}_{p\times p}\right). \tag{3.1}$$

Here, the design matrix, $\mathbf{X}_{n\times p}$, is composed of observed constants.

Primary interest in Berk et al. (2013) focuses on finding simultaneous confidence levels for the slope coefficients in the family of all possible submodels composed of subsets of columns of $\mathbf{X}$. However, that paper also discusses cases in which the family or the targeted slope coefficients are restricted in some fashion. One possibility is to restrict the sub-models to consist of only one predictor at a time (see possibility (2) in Section 4.4 of that paper). The family of confidence intervals is thus composed of confidence intervals for simple, marginal regression coefficients. These coefficients can be combined into a vector $\beta^{\circ 1}$, where the notation is analogous to that used following (1.1). Hence, the POSI procedure provides a valid test of $H_0^\circ : \beta^{\circ 1} = \mathbf{0}$ relative to the model (3.1).

This POSI test is also valid for testing the null hypothesis of M&Q—$H_0 : \beta^1 = \mathbf{0}$ within the model (1.1) with Gaussian errors (see comment (iii) below on this issue). In most cases results within (1.1) and (3.1) do not transfer so directly. But in the case of this null hypothesis the direct carryover is justified because under the null hypothesis (but not otherwise) the distribution of $\mathbf{X}$ is an ancillary statistic. Thus a test that is conditionally valid (i.e., is valid under (3.1)) is also unconditionally valid (i.e., under (1.1)). Buja et al. (2015) contains an extensive discussion of ancillarity issues in random design models like (1.1) and (2.1).

The POSI test described here is not a clone of the test provided by the simulation in ART because the critical value in Berk et al. (2013) is drawn from the distribution of $\left|T^1_{\hat{k}_n}\right|$ rather than from the distribution of $\hat{\theta}_n$ as defined in (1.3) (see our comment 1(i)). In other respects the simulations in the two procedures are very similar. Comment (ii) below points to one additional structural difference but otherwise there seem to be only minor technical differences that are asymptotically insignificant.

Some additional comments may be helpful.

(i) R-code is available for computing the critical constant for the POSI test. This code has an explicit option for the restriction to marginal submodels as described above. See Buja (2015).

(ii) The POSI test involves an estimate for the residual variance, σ^2, in (3.1). The POSI software and the theory supporting it draw this estimate from the full model Sum of Squares for Error. The simulation portion of ART draws an estimate for the analogous purpose via the sandwich style expression in Theorem 1 at $\beta = \mathbf{0}$ combined with (4). This is asymptotically equivalent (under suitable assumptions) to what would appear at the corresponding step of the POSI algorithm if one were to draw the estimate of σ^2 under the assumption that H_0 is true (i.e., from the restricted model rather than from the full model). The POSI software can be modified to proceed in this fashion. Unless $n - p$ is small we would not recommend proceeding in this fashion because the resulting test will not be similar (even under assumptions of normality). But for small or negative values of $n - p$ such a path would be desirable.

(iii) The residual distributions in (1.1) are more general than in (3.1); in (3.1) they are required to be Gaussian whereas in (1.1) they need only be iid (with finite variance). Berk et al. (2013) does not explicitly discuss such an extension of the model in (3.1). However, in retrospect, after reading M&Q we realize that the considerations in the POSI paper appear to be asymptotically valid under such an iid assumption for the coordinates of ε in (3.1). We conjecture that this is so, and hence that the POSI test of H_0° and H_0 is asymptotically valid.

4. A BOOTSTRAP TEST

A too casual reading of M&Q might incline one to feel that a bootstrap test of their H_0 is not possible unless the test also includes a simulation component, as does their ART. This is not so. What is true is that a pure bootstrap estimate of the distribution of a statistic like their $\hat{\theta}_n$ would be flawed, and this would also be the case for a statistic like $\left|T^1_{\hat{k}_n}\right|$ discussed above. A valid bootstrap test of H_0 requires a different structure.

Here is an outline of a simple bootstrap test of H_0 that is (asymptotically) valid under the model (1.1). Consider a family of confidence sets for β^1 of rectangular form:

$$\text{Rect}_C\left(\hat{\beta}^1\right) = \left\{\beta^1 : \left|\beta^1_k - \hat{\beta}^1_k\right| \le C, \quad k = 1, \ldots, p\right\}. \quad (4.1)$$

Use a bootstrap to determine the constant, C_{boot}, for which these rectangles have the desired estimated coverage, $1 - \alpha$. Then reject H_0 if $\mathbf{0} \notin \text{Rect}_{C_{boot}}$. This procedure has the desired asymptotic coverage as $n \to \infty$ for fixed p, and provides asymptotically satisfactory performance. (There is no claim here of any optimality for this test. The particular form for the rectangles in (4.1) is suggested here only for expositional convenience, and as a parallel to the focus of M&Q on $\hat{\theta}_n$.) In this simple setting, the asymptotic properties follow from standard bootstrap theory, but see Buja and Rolke (2015—and earlier) for a full, general treatment of such procedures.

Although this procedure does not attempt to discover the true distribution of a maximum statistic like $\hat{\theta}_n$ it does involve the distribution of $\hat{\theta}_n - \theta_n$. The form of the rectangles in (4.1) was chosen because of its relation to the simulation in ART. Other forms of confidence region may yield more satisfactory performance. For example, one could choose the sides of the rectangle to be proportional to the values of sandwich estimates of $\text{SD}\left(\hat{\beta}^1_k\right)$. This bootstrap procedure can be converted to create yield an asymptotically valid statement about the distribution of $\hat{\theta}_n - \theta_n$ under the null hypothesis. As such, it could be used in place of the simulation in ART involving $\mathbf{V}^*_n(\mathbf{0})$.

Along with collaborators including K. Zhang, we are preparing a methodological study of bootstrap confidence intervals for the slope coefficients in the assumption-lean model (see McCarthy et al. 2015). The bootstrap estimator we are proposing is a double bootstrap, with the second level of bootstrap improving the calibration of intervals provided by the first level. Asymptotic theory in our study suggests that such a double bootstrap can have better performance than a single bootstrap. (Based on helpful dialog with M&Q we note that that our proposal involves a more sophisticated, and more computer intensive, style of bootstrap than the CPB bootstrap used in their simulations—of course, this does not negate the objection to using a bootstrap of any sort to estimate the distribution of a statistic such as $\hat{\theta}_n$.) Our research to date has been focused on intervals for prechosen coordinates $\beta^\bullet_k$ (or $\beta^{\bullet 1}_k$). But, after reading M&Q, we realize that the methodology in our study can be adapted to the simultaneous confidence problem described here, and can also yield more evolved forms of confidence rectangles than those in (4.1). We intend to pursue such issues in future.

Both space and time constrain us from going into further detail here. But we are indebted to M&Q for indirectly providing the motivation to study such an issue as well as for the very interesting treatment and results involving their ART procedure.

[Received 20Aug2015. Revised 14Sep2015.]

REFERENCES

Berk, R., Brown, L. D., Buja, A., Zhang, K., and Zhao, L. (2013), ''Valid Post-Selection Inference,'' *Annals of Statistics,* 41, 802–837. [1446,1447,1448]

Buja, A. (2015), ''Software for Computing the POSI Constant,'' available at *http://www-stat.wharton.upenn.edu/~buja./* [1448]

Buja, A., and Rolke, W. (2015), ''Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference With Applications to Function Estimation and Functional Data,'' available at *https://statistics.wharton.upenn.edu/profile/555/research/?pubFilter= publishedPaper*. [1448]

Buja, A., Berk, R., Brown, L. D., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2015), ''Models as Approximations—A Conspiracy of Random Regressors and Model Deviations Against Classical Inference in Regression,'' submitted. Available from *http://www-stat.wharton.upenn.edu/~buja/*. [1447,1448]

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, CA: University of California Press, 221–233. [1447]

McCarthy, D., Zhang, K., Berk, R., Brown, L. D., Buja, A., George, E., and Zhao, L. (2015), "Calibrated Percentile Double Bootstrap for Robust Linear Regression Inference," available at *arXiv.org>stat>arXiv:1511.00273*. [1448]

McKeague, I. W., and Qian, M. (2015), ''An Adaptive Resampling Test for Detecting the Presence of Significant Predictors,'' *Journal of the American Statistical Association*, 110, this issue [1446]

White, H. (1980a), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170. [1447]

——— (1980b), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838. [1447]

——— (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1447]

Comment

Alexandre BELLONI and Victor CHERNOZHUKOV

Professors McKeagan and Qian provided us with an interesting, thought-provoking article on detecting the presence of significant predictors in high-dimensional settings which is an important issue in modern applications. They contributed to an emerging literature on formal hypothesis testing that takes into account variable selection. In addition to the results presented in the article they are also to be congratulated for bringing forth two important messages: (1) the importance of procedures that are valid uniformly over a substantial region of the parameter space, and (2) the key role that subsampling (in particular bootstrap) procedures can play in high-dimensional estimation.

Much of the research in detecting significant predictors in linear models focuses on consistency of variable selection. This approach relies on having enough data to separate signal from noise and implicitly requires a high signal-to-noise ratio. Although this could describe data-generating processes from signal processing and other engineering applications, this seems less suitable to represent data-generating processes from biology and economics. In those settings, where the signal and noise can be of similar order of magnitude, model selection mistakes are to some extent unavoidable.

In the article, the authors face the challenges of dealing with nonstandard limit behavior of estimators that arise from misspecification due to model selection mistakes. In model selection settings, we typically observe a discontinuity of the limit distribution for the values of the parameter vector with zero components (precisely the components we would like to exclude from the model). However, such discontinuity is not revealed in pointwise asymptotic analysis, where $\beta_n = \beta_0$ is fixed as $n \to \infty$. (It is important to stress that this occurs in low-dimensional cases, where the dimension p is fixed, and in high-dimensional cases where the dimension $p > n$.) This motivates the authors to consider a local asymptotic analysis in a $\sqrt{n}$-neighborhood of the null hypothesis, of the form $\beta_n = \beta_0 + n^{-1/2}b_0$, to provide a better approximation for the finite sample behavior of the estimator. An alternative approach would be to pursue results that are valid for any sequence of β_n in a region Θ_n so that results are valid uniformly over all the data-generating processes induced by Θ_n.

It has been documented that asymptotic results that are uniformly valid are more reflective of finite sample behavior than results that are valid only pointwise. Broadly speaking uniformly valid procedures rely less on the specific data-generating process which in turn translates into higher reliability in finite sample. Somewhat counterintuitive at first, a sequence of negative results derived by Leeb and Pötscher (2008) established that procedures that achieve pointwise model selection consistency lack of uniform validity. A direct consequence is that to construct a uniformly valid procedures one needs (necessarily) to violate the model selection consistency which has been actively advocated in the early literature of model selection. (As a side note, these negative results also imply that there are no procedures that can achieve model selection consistency uniformly over the set of models induced by $\Theta_n = \{\beta_n \in \mathbb{R}^p : \|\beta_n\| \leq C\}$.)

The article considers the use of marginal screening for detecting the presence of significant predictors in a linear models of the form $Y = X\beta_n + \varepsilon$, in a local asymptotic regime where $\beta_n = \beta_0 + n^{-1/2}b_0$ for some fixed p-dimensional vectors β_0 and b_0, and ε independent of X. The authors distinguish

Alexandre Belloni, Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, NC 27708 (E-mail: *abn5@duke.edu*). Victor Chernozhukov, Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142 (E-mail: *vchern@mit.edu*).

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1098545

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

two cases of main interest: (i) $\bar{k}(\beta_0)$ being unique (in particular $\beta_0 \neq 0$); and (ii) $\beta_0 = 0$ with $\bar{k}(b_0)$ being unique; where $\bar{k}(b) = \arg\max_{j=1,\ldots,p} |\mathrm{Corr}(X_j, X^T b)|$. Their test is adaptive as it attempts to discover which regime we are facing. There are several important features associated with this approach. The approach does not rely on model selection consistency. The local asymptotic regime considered by the authors yields a limit distribution that is continuous with respect to b_0. Finally, they rely on a (thresholded version of the) bootstrap procedure to estimate the limiting distribution under the null hypothesis of $\beta_n = 0$. Such approach should provide a better approximation to the finite sample behavior relative to (classical) pointwise asymptotic analysis (where $\beta_n = \beta_0$) and (asymptotic) analytic approximations.

This work contributes to a growing literature on the construction of valid hypothesis testing after model selection that does not rely on model selection consistency. In a variety of different settings Belloni et al. (2012), Belloni, Chernozhukov, and Hansen (2013, 2014), and Belloni, Chernozhukov, and Kato (2014) have proposed the use of orthogonal score functions to estimate some target parameters of interest in high-dimensional settings, building upon the classical orthogonalization ideas of Neyman (1979). Alternative approaches have been proposed by Zhang and Zhang (2014), Geer et al. (2014), Javanmard and Montanari (2014) and others.

To illustrate the use of orthogonal score functions consider the high-dimensional linear regression model

$$y_i = x_i'\beta_n + \epsilon_i, \quad \mathrm{E}[\epsilon_i x_i] = 0,$$

where x_i contains a constant and the observations (y_i, x_i') are iid for simplicity. Consider the partition of regressors $x_i' = (d_i, z_i')$ and the parameter values $\beta_n' = (\mu_n, \vartheta_n')$. Here, d_i is the regressor of interest whose regression coefficient μ_n we would like to estimate and construct confidence intervals. By Frisch–Waugh–Lovell theorem,

$$y_i - z_i'\eta_n = \mu_n(d_i - z_i'\gamma_n) + \epsilon_i, \quad \mathrm{E}[\epsilon_i(d_i - z_i'\gamma_n)] = 0,$$

and therefore μ_n can be recovered in population from the bivariate regression of $\tilde{y}_i = y_i - z_i'\eta_n$ on $\tilde{d}_i = d_i - z_i'\gamma_n$, where

$$\eta_n \in \arg\min_{\eta} \mathrm{E}[(y_i - z_i'\eta)^2], \quad \gamma_n \in \arg\min_{\gamma} \mathrm{E}[(d_i - x_i'\gamma)^2].$$

Note that $\tilde{y}_i$ and $\tilde{d}_i$ are the "residuals" that are left after partialling out the effects of z_i from y_i and d_i. The projection parameters η_n and γ_n are the nuisance high-dimensional parameters.

A score function ψ for the target parameter value μ_n, which is orthogonal with respect to the nuisance parameters, is

$$\psi(w_i, \mu, \eta, \gamma) = (y_i - z_i'\eta - \mu(d_i - z_i'\gamma))(d_i - z_i'\gamma),$$

where $w_i = (y_i, d_i, z_i')'$. Note that the following conditions hold

$$\mathrm{E}[\psi(w_i, \mu_n, \eta_n, \gamma_n)] = 0$$

and

$$\partial_{\eta,\gamma} \mathrm{E}[\psi(w_i, \mu_n, \eta, \gamma)]|_{\eta=\eta_n, \gamma=\gamma_n} = 0,$$

where the former yields identification of μ_n and the latter is the orthogonality property with respect to the nuisance parameters η_n and γ_n. The orthogonality condition reduces sensitivity of the estimation of μ_n with respect to the nuisance parameters, which must be estimated by nonregular estimators in very high-dimensional problems.

Provided that the vectors η_n and γ_n are approximately sparse, we can obtain their estimates $\hat{\eta}$ and $\hat{\gamma}$ using ℓ_1-penalized methods (e.g., lasso or square-root lasso) and post-ℓ_1-penalized estimators (e.g., post-lasso). We apply any of these estimation methods to data $(y_i, z_i')_{i=1}^n$ and $(d_i, x_i')_{i=1}^n$ to estimate η_n and γ_n, respectively. The resulting estimator $\hat{\mu}$ of μ_n, minimizes over μ the criterion function

$$\left(\frac{1}{n}\sum_{i=1}^n \psi(w_i, \mu, \hat{\eta}, \hat{\gamma})\right)^2.$$

Under mild conditions $\hat{\mu}$ obeys

$$\Omega^{-1/2} J \sqrt{n}(\hat{\mu} - \mu_n) \Rightarrow N(0, 1),$$

for $J = \mathrm{E}[\tilde{d}_i^2]$ and $\Omega = \mathrm{E}[\epsilon_i^2 \tilde{d}_i^2]$. Under homoscedasticity, $\hat{\mu}$ is actually semiparametrically efficient, achieving the known semiparametric efficiency bounds for estimating μ_n; as pointed out in Belloni, Chernozhukov, and Hansen (2014, 2013).

The score given above for the linear model is also known as "doubly robust score," which arise in treatment effect analysis. In the context of heterogeneous treatment effects, linear model is no longer appropriate, but doubly robust scores, which involve propensity score and regression functions as nuisance parameters, are available and could be used for treatment effect evaluation in high-dimensional setting, where adjustment for covariates is needed either to reduce variance or gain identification. The papers by Belloni, Chernozhukov, and Hansen (2014) (Section 5) and Belloni et al. (2013) developed the estimation and inference theory based on the use of "doubly robust scores" (see Robins and Rotnitzky 1995; Hahn 1998).

Belloni, Chernozhukov, and Kato (2014) investigated the use of orthogonal score functions for generic (possibly nonsmooth) Z-estimators. Moreover, they construct joint (rectangle) confidence regions for all p parameters of a vector β_n, essentially by applying the orthogonal scores to each of the coefficients of interest. In similar spirit to the work of McKeague and Qian, the confidence regions are uniformly valid and are constructed based on a multiplier bootstrap procedure for the calculation of correct critical values (provided technical conditions and $\log^7 p = o(n)$ hold). The procedure is computationally fast because it only involves resampling the estimated scores.

The articles above intend to construct approximate (but nonconservative) confidence intervals for the coefficients of β_n. In contrast, the work of McKeague and Qian aims to test the hypothesis of $H_0 : \theta_n = 0$ versus $H_a : \theta_n \neq 0$ where $\theta_n = \max_j |\mathrm{cov}(X_j, Y)|/\mathrm{var}(X_j)$. That is, they aim to design a single test to detect the presence of some significant predictor while controlling for familywise error rate. Although the questions are related, the former needs to avoid collinearity while in the former avoiding collinearity is not necessary. This is crucial for the validity of marginal regressions as described above. In fact by exploiting correlations among the covariates the marginal regression approach can have more power when coefficients β_n are small. Therefore providing an interesting complimentary approach to the literature above.

It is clear that several extensions of the adaptive resampling test (ART) are interesting. As mentioned by the authors, a natural extension is to allow p to grow with the sample size. This

regime can shed further light on the finite sample behavior of the estimator. Although not covered by the theory, their simulation results already suggest that ART can be useful in the $p > n$ setting (as shown in the simulations for $p = 200$ and $n = 100$). Another important direction would be to consider the generalization of the results to achieve uniform validity over $\Theta_n = \{\beta_n : \|\beta_n\| \leq C\}$. This includes the local asymptotic results discussed here but encompasses cases in which two component β_{nj} are β_{nk} are close, for example, $|\beta_{nj} - \beta_{nk}| = O(n^{-1/2})$. Finally, it would be interesting to understand how the results would carry over to more general models. For example, the case of Z-estimators which have been considered in the post-model selection discussed above. A natural starting point would be a logistic model, where the outcome Y is binary.

The recent literature on formal hypothesis testing accounting for misspecification that arise from model selection mistakes is still in its initial stages. For instance, several procedures that have been recently proposed are asymptotically equivalent but enjoy very different finite sample performances. Clearly, much research is still needed to better understand the finite sample behavior of estimators. Although it is unlikely one procedure will dominate others in all regimes it is important to better characterize their performance. Indeed, the wealth of different asymptotic regimes and different uniformity guarantees can be used as potential guidance for practitioners on which estimators they should focus on. The work of McKeague and Qian definitely contribute to this debate and will certainly stimulate future research.

[Received 20Aug2015. Revised 14Sep2015.]

REFERENCES

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2430. [1450]

Belloni, A., Chernozhukov, V., and Hansen, C. (2013), "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society*, 3, 245–295. [1450]

——— (2014), "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *The Revirw of Economics Studies*, 81, 608–650. [1450]

Belloni, A., Chernozhukov, V., and Kato, K. (2014), "Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other Z-Estimation Problems," *Biometrika*, pp. asu056. [1450]

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013), "Program Evaluation With High-Dimensional Data," *arXiv:1311.2645*. [1450]

Geer, S. V., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1450]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [1450]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [1450]

Leeb, H., and Pötscher, B. M. (2008), "Recent Developments in Model Selection and Related Areas," *Econometric Theory*, 24, 319–322. [1449]

Neyman, J. (1979), "$C(\alpha)$ Tests and Their Use," *Sankhya*, 41, 1–21. [1450]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [1450]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low-Dimensional Parameters With High-Dimensional Data," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1450]

Comment

Yichi ZHANG and Eric B. LABER

1. INTRODUCTION

We thank the editors for organizing this discussion of a timely and thought-provoking article. We congratulate McKeague and Qian (hereafter, M&Q) for illustrating and offering an effective solution to an important technical problem associated with marginal screening in high-dimensional linear regression. We believe that their work will serve as a catalyst for new statistical methodologies for screening and inference with nonregular functionals. In our discussion, we present two potential screening procedures constructed by modifying the adaptive resampling test (ART): (1) a "parametric bootstrap" analog of ART; and (2) an ART-inspired adaptive testing procedure designed to be more powerful against dense, weak alternatives. The parametric bootstrap procedure avoids the tuning parameter used in ART and thus eliminates potentially computationally burdensome tuning. Furthermore, we show that the proposed parametric bootstrap procedure has a desirable invariance property under local alternatives. However, we show that both ART and our proposed parametric bootstrap analog can have poor power against dense, weak alternatives. We propose a class of adaptive procedures that reduce to our parametric bootstrap version of ART under strong, sparse signals and reduce to a sum of squares criteria under weak, dense signals.

The proposed modifications to ART are not intended as improvements or as criticisms but rather as a demonstration of the potential of the framework proposed by M&Q. We conclude the article with a short philosophical point about the nonregularity associated with the ART procedure.

2. PARAMETRIC BOOTSTRAP ANALOG OF ART

As in M&Q we consider the local linear model

$$Y = a_0 + n^{-1/2}\mathbf{X}^{\intercal}\mathbf{b}_0 + \epsilon, \tag{1}$$

where $\alpha_0 \in \mathbb{R}$, $\mathbf{b}_0 \in \mathbb{R}^p$, $\mathbf{X} \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and $\epsilon \in \mathbb{R}$ is independent of $\mathbf{X}$, mean zero, and has finite variance σ_ϵ^2. Let $\Sigma_{\mathbf{X}}$

Yichi Zhang (E-mail: *yzhang52@ncsu.edu*), and Eric Laber (E-mail: *laber@stat.ncsu.edu*), Department of Statistics, North Carolina State University, 2311 Stinson Drive, SAS Hall, Raleigh, NC 27695.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1106403

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

denote the variance–covariance matrix of $\mathbf{X}$, and V_k denote the variance of X_k, $k = 1, \ldots, p$. For $k = 1, \ldots, p$ let $\widehat{\rho}_k$ denote the sample correlation coefficient between X_k and Y, $\widehat{V}_k$ the sample variance of X_k, and $\widehat{\sigma}_Y^2$ the sample variance of Y. Then, $\widehat{\beta}_k = \widehat{\text{cov}}(X_k, Y)/\widehat{V}_k$ is the estimated slope in the marginal regression of Y on X_k. The test statistic proposed by M&Q is

$$\widehat{\theta}_n = \sum_{k=1}^{p} \widehat{\beta}_k I\left\{|\widehat{\rho}_k| \geq \max_j |\widehat{\rho}_j|\right\},$$

where $I\{\cdot\}$ denotes an indicator function, and we assume that $\arg\max_k \widehat{\rho}_k$ is a singleton. Theorem 1 of M&Q showed that under the null hypothesis $\mathbf{b}_0 = 0$, $\sqrt{n}\widehat{\theta}_n$ converges in distribution to Z_K/V_K where $(Z_1, \ldots, Z_p)^\intercal$ is normally distributed with mean zero and covariance $\Sigma_{\mathbf{Z}} = \sigma_\epsilon^2 \Sigma_{\mathbf{X}}$, and $K = \arg\max_k Z_k^2/V_k$. Thus, given an estimator $\widehat{\Sigma}_{\mathbf{Z}}$ of $\Sigma_{\mathbf{Z}}$, one approach to approximate the null distribution of $\sqrt{n}\widehat{\theta}_n$ is to construct values Z_K/V_K using draws from a normal distribution with mean zero and variance–covariance $\widehat{\Sigma}_{\mathbf{Z}}$. We prove in the Appendix that distribution of Z_K/V_K is a smooth function of $\Sigma_{\mathbf{Z}}$.

The preceding parametric bootstrap procedure is computationally simple and avoids having to use the double bootstrap to select a tuning parameter. We now suggest a slight modification that avoids estimation of the residual variance σ_ϵ^2 and adds a scale-invariance property that is not present in ART nor the preceding parametric bootstrap procedure. Let $\widehat{t}_k$ denote the t-statistic $\sqrt{n}\widehat{\beta}_k \left\{\widehat{\sigma}_Y^2/\widehat{V}_k - \widehat{\beta}_k^2\right\}^{-1/2}$ for testing $\beta_k = 0$. We base our modified procedure on the statistic

$$\begin{aligned}\widehat{\xi}_n &= \sum_{k=1}^{p} \widehat{t}_k I\left\{|\widehat{\rho}_k| \geq \max_j |\widehat{\rho}_j|\right\} \\ &= \sum_{k=1}^{p} \widehat{t}_k I\left\{|\widehat{t}_k| \geq \max_j |\widehat{t}_j|\right\}, \end{aligned} \tag{2}$$

where the second inequality follows from $\widehat{t}_k^2 = \widehat{\rho}_k^2/(1 - \widehat{\rho}_k^2)$. A slight modification of the proof of Theorem 1 in M&Q shows that under the local linear model (1) $\sqrt{n}\widehat{\xi}_n$ converges in distribution to $(Z_K + \mathbf{C}_K^\intercal \mathbf{b}_0)/(\sigma_\epsilon V_K^{1/2})$, where $(\mathbf{C}_k)_j = \text{Cov}(X_k, X_j)$ $j = 1, \ldots, p$, and $K = \arg\max_k (Z_k + \mathbf{C}_k^\intercal \mathbf{b}_0)^2/V_k$. Thus, the limiting distribution of $\sqrt{n}\widehat{\xi}_n$ is scale-invariant, that is, replacing X_k with ςX_k and $(\mathbf{b}_0)_k$ with $(\mathbf{b}_0)_k/\varsigma$ for some $\varsigma > 0$ does not change the limiting distribution of $\sqrt{n}\widehat{\xi}_n$. On the other hand, $\sqrt{n}\widehat{\theta}_n$ converges in distribution to $(Z_K + \mathbf{C}_k^\intercal \mathbf{b}_0)/V_K$ which is not scale-invariant as replacing X_k with ςX_k and $(\mathbf{b}_0)_k$ with $(\mathbf{b}_0)_k/\varsigma$ scales the limiting distribution by $1/\varsigma$. Note that the estimators in the preceding expression have not been centered and thus take a slightly different form than the (centered) limits presented by M&Q. Because $\Sigma_{\mathbf{Z}} = \sigma_\epsilon^2 \Sigma_{\mathbf{X}}$, it can be seen that under the null $H_0 : \mathbf{b}_0 = 0$, the limiting distribution of $\sqrt{n}\widehat{\xi}_n$ does not depend on the residual variance σ_ϵ^2. Thus, the null distribution of $\sqrt{n}\widehat{\xi}_n$ can be approximated by constructing values Z_K/V_K from random draws taken from a normal distribution with mean zero and variance–covariance matrix $\widehat{\Sigma}_{\mathbf{X}}$, where $\widehat{\Sigma}_{\mathbf{X}}$ is the estimator of $\Sigma_{\mathbf{X}}$. In simulation experiments, this parametric bootstrap procedure provides similar power to ART while being orders of magnitude faster to compute (see Section 3).

3. ADAPTING TO DENSE AND WEAK SIGNALS

From (2), it follows that $\widehat{\xi}_n^2 = \max_j \widehat{t}_j^2$. Test statistics based on the maximum of marginal test statistics are well-studied in high-dimensional testing problems (e.g., Donoho and Jin 2004; Cai, Liu, and Xia 2014); however, while such methods are powerful for strong, sparse, and weakly correlated signals, they may fail to detect dense but weak signals. On the other hand, a sum of squares statistic, $\widehat{\zeta}_n = \sum_k \widehat{t}_k^2$, has the capability of aggregating many small signals and thereby might perform well for dense but weak signals while losing power to detect sparse signals. In this section, we use a numerical study to investigate the effect of sparsity on the power of tests based on $\widehat{\xi}_n$ and $\widehat{\zeta}_n$. This numerical study motivates a new test that attempts to adapt the unknown sparsity level of the signal.

3.1 Numerical Study

We use the asymptotic distributions of $\widehat{\xi}_n$ and $\widehat{\zeta}_n$ under the local linear model (1) to compare their power. Under the local linear model, $n\zeta_n$ converges in distribution to $\sigma_\epsilon^{-2}(\mathbf{Z} + \Sigma_{\mathbf{X}}\mathbf{b}_0)^\intercal D^{-1}(\mathbf{Z} + \Sigma_{\mathbf{X}}\mathbf{b}_0)$, where $D = \text{diag}(V_1, \ldots, V_p)$. We set $\sigma_\epsilon = 1$ and $V_k = 1$ for $k = 1, \ldots, p$ so that the signal is $\Sigma_{\mathbf{X}}\mathbf{b}_0$. Thus, both the covariance structure of the predictors and the regression coefficients contribute to the sparsity level. We consider the following covariance structures: (I) independent, $(\Sigma_{\mathbf{X}})_{ij} = I\{i = j\}$; (A) autoregressive, $(\Sigma_{\mathbf{X}})_{ij} = (0.8)^{|i-j|}$; and (E) equicorrelated, $(\Sigma)_{ij} = (1 + I\{i = j\})$ (see Figure 2 of M&Q). For each covariance structure, we consider three coefficient settings: (S) sparse, $(\mathbf{b}_0)_i = \gamma I\{i = 1\}$; (M) moderate, $(\mathbf{b}_0)_i = \gamma I\left\{i \bmod \lfloor\sqrt{p}\rfloor = 0\right\}$; and (D) dense, $(\mathbf{b}_0)_i = \gamma I\{i \leq \lfloor p/3\rfloor\} - (\gamma/3) I\{i \leq \lfloor 2p/3\rfloor\}$, where $\lfloor u\rfloor$ is the greatest integer below u and $\gamma \in \mathbb{R}$ controls the size of nonzero components of $\mathbf{b}_0$. For each combination of covariance matrix and regression coefficients we generate data of dimension $p = 10, 100, 500$, and 1000; furthermore, for each covariance matrix and regression coefficient combination we select γ so that the maximum power across all dimensions and both tests is 0.8.

Table 1 shows the asymptotic power for $\widehat{\xi}_n$ and $\widehat{\zeta}_n$. In settings with a large, sparse signal, for example, models IS and AS, the test based on $\widehat{\xi}_n$ has markedly better power. On the other hand, in settings with a weak, dense signal, for example, models ID, AD, and ED, the test based on $\widehat{\zeta}_n$ has more power. A heuristic justification for this trend is as follows. Both $\widehat{\xi}_n$ and $\widehat{\zeta}_n$ are functions of $\widehat{\mathbf{t}} = (\widehat{t}_1, \ldots, \widehat{t}_p)^\intercal$. For the purpose of exposition, assume that $\widehat{\mathbf{t}}$ is drawn from its asymptotic distribution. We also assume that p diverges with n, the smallest eigenvalue of $\Sigma_{\mathbf{X}}$ is bounded away from zero, and the largest eigenvalue of $\Sigma_{\mathbf{X}}$ is bounded above. Under the null, $\widehat{\xi}_n^2$ is the maximum of p dependent χ_1^2 random variables, thus, $n\widehat{\xi}_n^2 - 2\log(p) + \log\{\log(p)\}$ converges in distribution to an extreme-value distribution as p diverges (Cai et al. 2014, Theorem 1). Thus, for the test based on $\widehat{\xi}_n$ to be powerful, $||\Sigma_{\mathbf{X}}\mathbf{b}_0||_\infty$ should be of order $\{\log(p)\}^{1/2}$. In contrast, $n\widehat{\zeta}_n$ has mean $||\Sigma_{\mathbf{X}}\mathbf{b}_0||_2^2 + \text{tr}(\Sigma_{\mathbf{X}})$, and variance $4(\Sigma_{\mathbf{X}}^\intercal \mathbf{b}_0)^\intercal \Sigma_{\mathbf{X}}(\Sigma_{\mathbf{X}}\mathbf{b}_0) + 2\,\text{tr}\left(\Sigma_{\mathbf{X}}^2\right)$. Thus, provided $\liminf_p p^{-1}||\Sigma_{\mathbf{X}}\mathbf{b}_0||_2^2$ is nonzero, $\widehat{\zeta}_n$ will be able to separate the null from the alternative as p diverges. In settings with weak, dense signals $||\Sigma_X\mathbf{b}_0||_2^2$ may be considerably larger than $||\Sigma_{\mathbf{X}}\mathbf{b}_0||_\infty$; in such cases, $\widehat{\zeta}_n$ may yield higher power than $\widehat{\xi}_n$.

Table 1. Asymptotic power of $\hat{\xi}_n$ and $\hat{\zeta}_n$ under the local linear model (1)

		$p = 10$		$p = 50$		$p = 200$		$p = 1000$	
$\Sigma_{\mathbf{X}}$	$\mathbf{b}_0$	$\hat{\xi}_n$	$\hat{\zeta}_n$	$\hat{\xi}_n$	$\hat{\zeta}_n$	$\hat{\xi}_n$	$\hat{\zeta}_n$	$\hat{\xi}_n$	$\hat{\zeta}_n$
	S	0.80	0.69	0.65	0.33	0.51	0.16	0.36	0.09
I	M	0.48	0.59	0.49	0.70	0.46	0.75	0.44	0.80
	D	0.07	0.08	0.09	0.14	0.10	0.30	0.11	0.80
	S	0.80	0.69	0.63	0.35	0.47	0.17	0.32	0.10
A	M	0.64	0.73	0.55	0.80	0.29	0.66	0.18	0.58
	D	0.05	0.05	0.08	0.10	0.15	0.29	0.20	0.80
	S	0.80	0.75	0.72	0.69	0.67	0.67	0.65	0.67
E	M	0.06	0.06	0.09	0.10	0.20	0.23	0.75	0.80
	D	0.05	0.05	0.05	0.05	0.08	0.08	0.75	0.80

3.2 Adaptive Parametric Bootstrap

The tests based on $\widehat{\xi}_n$ and $\widehat{\zeta}_n$ are powerful under complementary sparsity signatures; it is of interest to see if they can be combined to construct a test that is adaptive to the unknown level of sparsity. Let $\widehat{t}^2_{(1)}, \ldots, \widehat{t}^2_{(p)}$ be the order statistics of $\widehat{t}^2_1, \ldots, \widehat{t}^2_p$ and for each $k = 1, \ldots, p$ define $\widehat{\psi}_{n,k} = \sum_{j=1}^{k} \widehat{t}^2_{(p-j+1)}$. Thus, $\widehat{\xi}^2_n = \widehat{\psi}^2_{n,1}, \widehat{\zeta}_n = \widehat{\psi}^2_{n,p}$, and $\widehat{\psi}^2_{n,k}, k = 2, \ldots, p-1$ are test statistics that interpolate between these two extremes. The goal is to use the data to choose k to maximize power. In our developments, we assume that p is fixed.

For any fixed k, $\widehat{\psi}_{k,n}$ is a continuous function of $\widehat{\mathbf{t}}$. Let $\mathbf{U} = (U_1, \ldots, U_p)^{\intercal}$ be normally distributed with mean zero and variance–covariance matrix $D^{-1/2}\Sigma_{\mathbf{X}}D^{-1/2}$ and define $U^2_{(1)}, \ldots, U^2_{(p)}$ to be the order statistics of $U^2_1, \ldots, U^2_p$. Using the continuous mapping theorem, it can be seen that under the null the limiting distribution of $n\widehat{\psi}_{k,n}$ is equal in distribution to $\sum_{j=1}^{k} U^2_{(p-j+1)}$. For $k = 1, \ldots, p$, define

$$\widehat{\omega}_{n,k} = P\left(\sum_{j=1}^{k} U^2_{(p-j+1)} \geq n\widehat{\psi}_{n,k} \middle| \widehat{\psi}_{n,k} \right),$$

and subsequently $\widehat{\psi}_n = \min_k \widehat{\omega}_{n,k}$. The proposed adaptive test rejects if $\widehat{\psi}_n$ is below a critical value.

We derive a critical value for $\widehat{\psi}_n$ as follows. Let $h_k(\cdot)$ be the cumulative distribution function of $\sum_{j=1}^{k} U^2_{(p-j+1)}$ so that $\widehat{\omega}_{n,k} = 1 - h_k\{n\widehat{\psi}_{n,k}\}$. By the continuous mapping theorem, under the null, $n\widehat{\psi}_n$ is asymptotically equal in distribution to $\min_k[1 - h_k\{\sum_{j=1}^{k} U^2_{(p-j+1)}\}]$. We can simulate the distribution of $n\widehat{\psi}_n$ by generating draws $\mathbf{U}$ from a normal distribution with mean zero and variance–covariance matrix $\widehat{D}^{-1/2}\widehat{\Sigma}_{\mathbf{X}}\widehat{D}^{-1/2}$, where $\widehat{D}$, and $\widehat{\Sigma}_{\mathbf{X}}$ are estimates of D and $\widehat{\Sigma}_{\mathbf{X}}$.

We examine the finite sample performance of $\widehat{\psi}_n, \widehat{\xi}_n, \widehat{\zeta}_n$, multiple testing with a Bonferroni correction, and ART with double bootstrap tuning as proposed in M&Q. In our implementation of the tests using $\widehat{\psi}_n, \widehat{\xi}_n$, and $\widehat{\zeta}_n$, we used the plug-in estimator of $\Sigma_{\mathbf{X}}$. When p is large it is possible that the operating characteristics of the proposed tests might be improved by using a regularized covariance estimator. We use the same generative models as proposed in Section 4.1 of M&Q with $\rho = 0.5$, so we will not repeat them here. We fix the sample size at $n = 100$ and set the nominal Type I error to 0.05. Estimated power is based on 1000 Monte Carlo replications. As noted by M&Q, results for ART could not be computed for $p = 1000$ as the double bootstrap is too computationally burdensome.

Table 2. Power comparisons for generative models in Section 4.1 of M&Q with $n = 100$

Model	p	ART	Bonferroni	$\hat{\xi}_n$	$\hat{\zeta}_n$	$\hat{\psi}_n$	Time (ART)	Time ($\hat{\psi}_n$)
	10	0.067	0.051	0.065	0.055	0.059	485	0.1
	50	0.067	0.040	0.054	0.057	0.064	747	0.2
(i)	200	0.054	0.028	0.052	0.046	0.046	3289	0.5
	1000	—	0.024	0.056	0.049	0.055	—	2.3
	10	0.510	0.439	0.496	0.492	0.509	489	0.1
	50	0.376	0.305	0.382	0.383	0.403	723	0.2
(ii)	200	0.376	0.277	0.380	0.398	0.403	4078	0.5
	1000	—	0.239	0.363	0.408	0.392	—	2.3
	10	0.496	0.461	0.501	0.505	0.526	502	0.1
	50	0.366	0.310	0.375	0.386	0.399	717	0.2
(iii)	200	0.337	0.279	0.350	0.375	0.383	4824	0.5
	1000	—	0.224	0.352	0.394	0.381	—	2.3

NOTES: Estimates are based on $B = 1000$ Monte Carlo replications. In model (i) the null hypothesis is true. In models (ii) and (iii), the alternative is true and we highlight methods with estimated power within one standard error (estimated conservatively as $1/\sqrt{2B} \approx 0.022$) of the highest observed power. The two rightmost columns denote the average runtime (seconds) to conduct a single test.

The results of the simulation are presented in Table 2. The proposed adaptive procedure attains good power while controlling Type I error. The adaptive procedure has better power than both $\widehat{\xi}_n$ and $\widehat{\zeta}_n$ in all but two examples and better power than ART in all but one example (where the difference is 0.001). The table also shows the runtime in seconds of ART with double bootstrap tuning and the test based on $\widehat{\psi}_n$. Computing $\widehat{\psi}_n$ is several orders of magnitude faster than ART and scales efficiently to large p. Thus, we believe that adaptive algorithms built on the framework of M&Q might be a fruitful research direction.

The adaptive testing procedure proposed here is one of many possibilities. For example, one extension would be to consider linear combinations of the order statistics of $\widehat{\mathbf{t}}$ of which the proposed procedure is a special case.

4. DISCUSSION

M&Q have made an important contribution to a challenging and timely problem. We have used their framework to propose a number of related tests that may reduce computation time or adapt to the underlying sparsity of the signal. We hope that these examples will serve to illustrate the large number of possible research directions opened by the framework proposed by M&Q. We would like to conclude our comment with a philosophical note about nonregularity and testing in regression problems. To streamline the discussion, we focus on a fixed (nonlocal) alternatives model.

In the standard linear model, $Y = \mathbf{X}^{\intercal}\beta^* + \epsilon$, the true regression coefficient, β^*, is a smooth functional of the generative model. However, as is well-known, $||\beta^*||_\infty$, is not a smooth functional. Thus, as M&Q note, inference based on plug-in estimator $||\widehat{\beta}_n||_\infty$ is difficult because of nonuniform convergence and special care must be taken to conduct hypothesis testing using this statistic. In this setting, the nonregularity seems to be

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

introduced by the choice of the sup-norm for testing $\beta^* \equiv 0$, rather than, say, some inherent nonsmoothness in the parameter of interest; for examples, where the nonsmoothness appears to be more entrenched in the estimand, see Hirano and Porter (2012); Laber et al. (2014). One smooth alternative to the sup-norm would be a soft-max $||\beta^*||_\tau = \tau^{-1}\log\{1+\sum_{j=1}^p \exp(\tau\beta_j^{*2})\}$, where $\tau > 0$ is a tuning parameter. Under mild regularity conditions, it is straightforward to derive the asymptotic distribution of $||\widehat{\beta}_n||_\tau$ and use it to conduct inference. Thus, one might argue for using the soft-max to avoid dealing with nonregular, complex asymptotic arguments. This raises the question of whether or not we should let our choice of estimand be dictated by the subsequent difficulty of inference. Our own opinion is that an estimand (or test statistic) should be selected for its operating characteristics irrespective of the difficulty of inference, though we would be interested in hearing the opinion of M&Q on this matter.

APPENDIX

Suppose $\mathbf{Z}(\Omega) = \{Z_1(\Omega), \ldots, Z_p(\Omega)\}^T$ is normal with mean zero and covariance Ω. Let $V_j(\Omega)$ be the jth diagonal element of Ω, $\mathbb{K}(\Omega) = \arg\max_j Z_j^2(\Omega)/V_j(\Omega)$, $Z_K(\Omega) = Z_{\mathbb{K}(\Omega)}(\Omega)$, and $V_K(\Omega) = V_{\mathbb{K}(\Omega)}(\Omega)$. If $\lambda_{\min}(\Omega)$, the smallest eigenvalue of Ω, is bounded away from zero and $\lambda_{\max}(\Omega)$, the largest eigenvalue of Ω, is bounded above, then the distribution of $Z_K(\Omega)/V_K(\Omega)$ depends on Ω continuously. That is,

$$\sup_{\alpha\in\mathbb{R}}\left|P\left(\frac{Z_K(\Omega)}{V_K(\Omega)} \le \alpha\right) - P\left(\frac{Z_K(\sim\Omega)}{V_K(\sim\Omega)} \le \alpha\right)\right| = O(\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2).$$

Proof. Denote equality in distribution by $=_d$. Let $\mathbf{U}$ be $N(0, I_p)$. Then, $\mathbf{Z}(\Omega) =_d \Omega^{1/2}\mathbf{U}$ and $\mathbf{Z}(\sim\Omega) =_d \sim\Omega^{1/2}\mathbf{U}$. For any j, we have

$$(\sim\Omega^{1/2}\mathbf{U})_j = (\Omega^{1/2}\mathbf{U})_j + \{(\sim\Omega^{1/2} - \Omega^{1/2})\mathbf{U}\}_j,$$
$$|\{(\sim\Omega^{1/2} - \Omega^{1/2})\mathbf{U}\}_j| \le \|(\sim\Omega^{1/2} - \Omega^{1/2})\mathbf{U}\|_2 \le \|\sim\Omega^{1/2} - \Omega^{1/2}\|_2\|\mathbf{U}\|_2.$$

Hence, we obtain

$$\left|\frac{(\Omega^{1/2}\mathbf{U})_j^2}{V_j(\Omega)} - \frac{(\sim\Omega^{1/2}\mathbf{U})_j^2}{V_j(\sim\Omega)}\right| \le C_1\|\mathbf{U}\|_2^2|V_j(\Omega) - V_j(\sim\Omega)| + C_2\|\mathbf{U}\|_2|(\Omega^{1/2}\mathbf{U})_j - (\sim\Omega^{1/2}\mathbf{U})_j| \le C_3\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2\|\mathbf{U}\|_2^2,$$

where C_i's are constants. Let $\mathbb{J}(\Omega) = \arg\max_j(\Omega^{1/2}\mathbf{U})_j^2/V_j(\Omega)$. Let set A be the event $\{\mathbb{J}(\Omega) = \mathbb{J}(\sim\Omega)\}$ and set B_{jk} be the event

$$\left\{\left|\frac{(\Omega^{1/2}\mathbf{U})_j^2}{V_j(\Omega)} - \frac{(\Omega^{1/2}\mathbf{U})_k^2}{V_k(\Omega)}\right| \le 2C_3\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2\|\mathbf{U}\|_2^2\right\}.$$

Then we have

$$P(A^c) \le P(B_{jk} \text{ for some } j < k) \le \sum_{j,k} P(B_{jk}).$$

Because Ω is invertible, it can be shown that $\{(\Omega^{1/2}\mathbf{U})_j^2/V_j(\Omega) - (\Omega^{1/2}\mathbf{U})_k^2/V_k(\Omega)\}/\|\mathbf{U}\|_2^2$ has bounded density. Thus, $P(B_{ij}) \le C_4\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2$ and $P(A^c) \le C_5\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2$.

Within the event A, let $J = \arg\max_j(\Omega^{1/2}\mathbf{U})_j^2/V_j(\Omega)$ be the common maximizer. Using the techniques above, we have

$$\left|\frac{(\Omega^{1/2}\mathbf{U})_J}{V_J(\Omega)} - \frac{(\sim\Omega^{1/2}\mathbf{U})_J}{V_J(\sim\Omega)}\right| \le \max_j\left|\frac{(\Omega^{1/2}\mathbf{U})_j}{V_j(\Omega)} - \frac{(\sim\Omega^{1/2}\mathbf{U})_j}{V_j(\sim\Omega)}\right| \le C_6\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2\|\mathbf{U}\|_2.$$

Hence, by Chebyshev's inequality, for any $\varepsilon > 0$, we have

$$P\left(\left|\frac{(\Omega^{1/2}\mathbf{U})_J}{V_J(\Omega)} - \frac{(\sim\Omega^{1/2}\mathbf{U})_J}{V_J(\sim\Omega)}\right| > \varepsilon\right) \le C_7\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2/\varepsilon.$$

Therefore, we obtain

$$P\left(\left|\frac{(\Omega^{1/2}\mathbf{U})_{\mathbb{J}(\Omega)}}{V_{\mathbb{J}(\Omega)}(\Omega)} - \frac{(\sim\Omega^{1/2}\mathbf{U})_{\mathbb{J}(\sim\Omega)}}{V_{\mathbb{J}(\sim\Omega)}(\sim\Omega)}\right| > \varepsilon\right) \le C_5\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2 + C_7\|\Omega^{1/2} - \sim\Omega^{1/2}\|_2/\varepsilon.$$

We conclude the proof by noting that $(\Omega^{1/2}\mathbf{U})_{\mathbb{J}(\Omega)}/V_{\mathbb{J}(\Omega)}(\Omega) =_d Z_K(\Omega)/V_K(\Omega)$ as well as $(\sim\Omega^{1/2}\mathbf{U})_{\mathbb{J}(\sim\Omega)}/V_{\mathbb{J}(\sim\Omega)}(\sim\Omega) =_d Z_K(\sim\Omega)/V_K(\sim\Omega)$. □

Remark. It is sufficient that $\sim\Omega$ be positive semidefinite instead of positive definite. Thus, the sample covariance matrix can be used even if $p > n$.

[Received 20Aug2015. Revised 14Sep2015.]

REFERENCES

Cai, T., Liu, W., and Xia, Y. (2014), "Two-Sample Test of High Dimensional Means Under Dependence," *Journal of the Royal Statistical Society*, Series B, 76, 349–372. [1452]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 962–994. [1452]

Hirano, K., and Porter, J. R. (2012), "Impossibility Results for Nondifferentiable Functionals," *Econometrica*, 80, 1769–1790. [1454]

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225. [1454]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]

Comment

Sai LI, Ritwik MITRA, and Cun-Hui ZHANG

We congratulate McKeague and Qian on their interesting and thought-provoking article. Modern statistics has seen rapid growth of regularized estimation methods for complex and high-dimensional problems. However, this development has not yet been matched by a similar profusion of tools for statistical inference, especially bootstrap-based methods. The problem under consideration is the inference of the marginal regression coefficient for the design variable having the highest population correlation with the response variable in linear regression. As is widely acknowledged, this is a very difficult problem, and we congratulate McKeague and Qian on taking up the topic. We have no doubt that the proposed test will be applied in real life problems and McKeague and Qian's ideas will be useful in the development of adaptive bootstrap methods in more general contexts.

Let μ_k and ρ_k be the population marginal regression and correlation coefficients of the kth design variable in a linear regression model. The parameter of interest in the article is

$$\theta = \mu_{k_{\max}}, \quad \text{where } k_{\max} = \arg\max_{k \le p} |\rho_k|.$$

Let $\widehat{\mu}_k$ and $\widehat{\rho}_k$ be the sample marginal regression and correlation coefficients associated with the kth design vector. The sample version of $k_{\max}$ and θ and the sample-size normalized estimation error are

$$\widehat{k} = \arg\max_{k \le p} |\widehat{\rho}_k|, \ \widehat{\theta} = \widehat{\mu}_{\widehat{k}} \ \text{ and } \ \xi = \sqrt{n}(\widehat{\theta} - \theta).$$

Let $\widehat{\mu}^*_k$ and $\widehat{\rho}^*_k$ be bootstrapped $\widehat{\mu}_k$ and $\widehat{\rho}_k$. A naive bootstrap scheme for ξ is

$$\widehat{k}^*_{\text{naive}} = \arg\max_{k \le p} |\widehat{\rho}^*_k|, \ \widehat{\theta}^*_{\text{naive}} = \widehat{\mu}^*_{\widehat{k}^*_{\text{naive}}} \ \text{ and}$$

$$\xi^*_{\text{naive}} = \sqrt{n}(\widehat{\theta}^*_{\text{naive}} - \widehat{\theta}).$$

This is fine when $\widehat{k}^*_{\text{naive}} = \widehat{k} = k_{\max}$ with large probability. However, the naive bootstrap scheme is not expected to work well under the null hypothesis $H_0 : \rho_k = 0 \ \forall k$. Let V^*_k and W^*_k be bootstrapped $\sqrt{n}(\widehat{\mu}_k - \mu_k)$ and $\sqrt{n}(\widehat{\rho}_k - \rho_k)$, respectively. Under the null, an appropriate bootstrap scheme is

$$\widehat{k}^*_{\text{null}} = \arg\max_{k \le p} |W^*_k| \ \text{ and } \ \xi^*_{\text{null}} = V^*_{\widehat{k}^*_{\text{null}}}.$$

Here the bootstrap method is generic and the centering of estimates is symbolic as our focus is on methodologies for using appropriately generated $\{\widehat{\mu}^*_k, \widehat{\rho}^*_k, V^*_k, W^*_k, 1 \le k \le p\}$. Actually, μ_k and ρ_k can be viewed as generic parameters in this discussion. As thresholding sample correlation is equivalent to t-test, McKeague and Qian's $\mathbb{V}^*_n(0)$ can be viewed as a version of ξ^*_{null}. We note that McKeague and Qian studied more general $\mathbb{V}^*_n(\boldsymbol{b})$, $\boldsymbol{b} \in \mathbb{R}^p$, for local models contiguous to H_0.

Suppose that the null hypothesis and an alternative H_a guaranteeing $\mathbb{P}\{\widehat{k}^*_{\text{naive}} = \widehat{k} = k_{\max}\} \to 1$ can be consistently tested by thresholding $\max_{k \le p} |\widehat{\rho}_k|$ at a level λ. A simple adaptive bootstrap scheme is

$$\xi^*_{\text{simple}} = \begin{cases} \xi^*_{\text{naive}}, & \max_{k \le p} |\widehat{\rho}_k| \ge \lambda, \\ \xi^*_{\text{null}}, & \max_{k \le p} |\widehat{\rho}_k| < \lambda. \end{cases}$$

However, the performance of such a scheme is unclear in the middle ground between the two hypotheses when the signal is not as strong as H_a depicts. The proposed adaptive bootstrap scheme is

$$\xi^*_{\text{proposed}} = \begin{cases} \xi^*_{\text{naive}}, & \max_{k \le p} \left(|\widehat{\rho}_k| \vee |\widehat{\rho}^*_k|\right) \ge \lambda, \\ \xi^*_{\text{null}}, & \max_{k \le p} \left(|\widehat{\rho}_k| \vee |\widehat{\rho}^*_k|\right) < \lambda. \end{cases}$$

McKeague and Qian developed elegant theoretical results under the null and alternative hypotheses and local models contiguous to H_0, and carried out simulation experiments for the size and power of the proposed bootstrap test. They pointed out that a robust confidence interval can be constructed based on $\mathbb{V}^*_n(\boldsymbol{b})$, and that such confidence intervals are conservative and requires a grid search over $\mathbb{R}^p$.

Our comments focus on the closely related problem of estimating the selected parameter and the coverage probability of confidence intervals for the simple and proposed adaptive bootstrap methods.

Statistical inference of the parameter of a selected population is an interesting problem in and of itself. In the context of the regression problem under consideration, the selected parameter and its sample-size normalized estimation error can be written as

$$\widetilde{\theta} = \mu_{\widehat{k}} \ \text{ and } \ \widetilde{\xi} = \sqrt{n}\big(\widehat{\theta} - \widetilde{\theta}\big),$$

and the naive, null, simple adaptive and proposed adaptive bootstrap schemes can be written as

$$\widetilde{\xi}^*_{\text{naive}} = \sqrt{n}\Big(\widehat{\mu}^*_{\widehat{k}^*_{\text{naive}}} - \widehat{\mu}_{\widehat{k}^*_{\text{naive}}}\Big),$$

$$\widetilde{\xi}^*_{\text{null}} = \xi^*_{\text{null}} = V^*_{\widehat{k}^*_{\text{null}}},$$

$$\widetilde{\xi}^*_{\text{simple}} = \begin{cases} \widetilde{\xi}^*_{\text{naive}}, & \max_{k \le p} |\widehat{\rho}_k| \ge \lambda, \\ \xi^*_{\text{null}}, & \max_{k \le p} |\widehat{\rho}_k| < \lambda, \end{cases}$$

Sai Li (E-mail: *sl1022@scarletmail.rutgers.edu*) is Graduate Student, and Cun-Hui Zhang (E-mail: *czhang@stat.rutgers.edu*) is Distinguished Professor, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854. Ritwik Mitra is Post-Doctoral Research Associate, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 (E-mail: *rmitra@princeton.edu*). Research supported by NSF Grants DMS-1129626 and DMS-1209014 and NSA Grant H98230-15-1-0040.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1106404

Table 1. Coverage probabilities of bootstrap confidence intervals based on 1000 replications. The confidence interval is constructed with 0.025 and 0.975 sample quantiles of $B = 1000$ bootstrap samples for each replication. The threshold level λ is chosen by simulation to guarantee approximately 5% Type I error, instead of double bootstrap, and $\xi^*_{\text{null}} = \mathbb{V}^*_n(0)$ as in McKeague and Qian's simulation study. The vector $\boldsymbol{\mu}$ gives the population marginal regression coefficients

Models and settings					Coverage probability			
					ξ^*_{proposed}	$\widetilde{\xi}^*_{\text{proposed}}$	ξ^*_{simple}	$\widetilde{\xi}^*_{\text{simple}}$
Model	p	ρ	λ	$\boldsymbol{\mu}$	θ_n	$\tilde{\theta}_n$	θ_n	$\tilde{\theta}_n$
	10	0	3.5	$\mathbf{1}_{10}0$	0.900	0.934	0.953	0.953
	200	0	4	$\mathbf{1}_{200}0$	0.819	0.886	0.954	0.954
I	10	0.5	3.5	$\mathbf{1}_{10}0$	0.891	0.922	0.944	0.944
	200	0.5	4	$\mathbf{1}_{200}0$	0.846	0.907	0.973	0.973
	10	0.8	3.5	$\mathbf{1}_{10}0$	0.898	0.925	0.943	0.943
	200	0.8	4	$\mathbf{1}_{200}0$	0.827	0.895	0.957	0.957
	10	0	3.5	$(0.25, \mathbf{1}_9 0)$	0.839	0.931	0.839	0.939
	200	0	4	$(0.25, \mathbf{1}_{199} 0)$	0.623	0.925	0.624	0.976
II	10	0.5	3.5	$(0.25, \mathbf{1}_9 0.125)$	0.945	0.932	0.945	0.938
	200	0.5	4	$(0.25, \mathbf{1}_{199} 0.125)$	0.922	0.906	0.923	0.927
	10	0.8	3.5	$(0.25, \mathbf{1}_9 0.2)$	0.944	0.939	0.944	0.941
	200	0.8	4	$(0.25, \mathbf{1}_{199} 0.2)$	0.939	0.935	0.942	0.936
	10	± 0.5	3.5	$(0.25, \mathbf{1}_9(-0.125))$	0.201	0.945	0.203	0.947
II*	200	± 0.5	4	$(0.25, \mathbf{1}_{199}(-0.125))$	0.048	0.902	0.048	0.920
	10	0	3.5	$(\mathbf{1}_5 0.15, \mathbf{1}_5(-0.1))$	0.638	0.949	0.646	0.951
	200	0	4	$(\mathbf{1}_5 0.15, \mathbf{1}_5(-0.1), \mathbf{1}_{190} 0)$	0.543	0.907	0.597	0.954
III	10	0.5	3.5	$(\mathbf{1}_5 0.2, \mathbf{1}_5 0.075)$	0.881	0.939	0.884	0.944
	200	0.5	4	$(\mathbf{1}_5 0.2, \mathbf{1}_5 0.075, \mathbf{1}_{190} 0.125)$	0.866	0.898	0.881	0.927
	10	0.8	3.5	$(\mathbf{1}_5 0.23, \mathbf{1}_5 0.18)$	0.921	0.939	0.922	0.940
	200	0.8	4	$(\mathbf{1}_5 0.23, \mathbf{1}_5 0.18, \mathbf{1}_{190} 0.2)$	0.910	0.921	0.914	0.925
IV	10	0	3.5	$(0.41, -0.40, \mathbf{1}_8 0)$	0.484	0.949	0.484	0.949
	200	0	4	$(0.41, -0.40, \mathbf{1}_{198} 0)$	0.502	0.953	0.502	0.963

and

$$\widetilde{\xi}^*_{\text{proposed}} = \begin{cases} \widetilde{\xi}^*_{\text{naive}}, & \max_{k \le p} \left(|\widehat{\rho}_k| \vee |\widehat{\rho}^*_k| \right) \ge \lambda, \\ \xi^*_{\text{null}}, & \max_{k \le p} \left(|\widehat{\rho}_k| \vee |\widehat{\rho}^*_k| \right) < \lambda. \end{cases}$$

Under the null hypothesis, the two problems are identical as $\theta = \widetilde{\theta}$ when $\rho_k = 0$ for all $k \le p$. When $\mathbb{P}\{\widehat{k}^*_{\text{naive}} = \widehat{k} = k_{\max}\} \to 1$, the two problems are nearly identical as $\mathbb{P}\{\widetilde{\theta} = \theta\} \to 1$. The interesting place is the middle ground where the signal is weak. We would like to point out that when the signal is nonzero but not detectable, the sign of θ is not tractable from the data, so that no bootstrap scheme will provide consistent estimation of the distribution of ξ. However, the simple and proposed adaptive schemes may still approximate the distribution of $\widetilde{\xi}$ well under proper assumptions, even when the signal is weak. It seems from this point of view that statistical inference about the "strongest population signal" θ is ideal but not attainable and inference about the "selected signal" $\widetilde{\theta}$ is the best we can achieve. Of course, the interpretation of the statistical inference about $\widetilde{\theta}$ is more complicated as it is a random parameter.

We report some simulation results of the bootstrap coverage probability to demonstrate our point, although McKeague and Qian's study is focused on hypothesis testing. In Table 1, models I, II, and III are identical to those considered by McKeague and Qian, and models II* and IV depict situations where the estimation of the distribution of $\widetilde{\xi}$ is feasible but that of ξ is not. The design variables are $N(0, 1)$ with a common correlation ρ, except for model II* where $\text{Corr}(X_1, X_j) = -|\rho|$ for $1 < j \le p$ and $\text{Corr}(X_j, X_k) = |\rho|$ for $1 < j < k \le p$. The noise vector is $N(0, \boldsymbol{I}_{n\times n})$. The coefficient vectors are $\boldsymbol{\beta} = \mathbf{0}$, $(0.25, \mathbf{1}_{p-1}0)$, $(0.25, \mathbf{1}_{p-1}0)$, $(\mathbf{1}_5 0.15, \mathbf{1}_5(-0.10), \mathbf{1}_{p-10}0)$, and $(0.41, -0.4, \mathbf{1}_{p-2}0)$, respectively, in models I, II, II*, III, and IV. As $\text{var}(X_k) = 1$, $\mu_k = \text{cov}(Y, X_k)/\text{var}(X_k)$ is proportional to $\rho_k = \text{Corr}(Y, X_k)$, so that the difficulty of selection is largely dependent on the separation of the largest $|\mu_k|$ from the rest. While both bootstrap schemes clearly cover the selected parameter $\widetilde{\theta}$ much better than the ideal θ, the simple adaptive bootstrap seems to have somewhat more accurate coverage probability in both problems in the experiment. Moreover, as Model II* demonstrates, the reasonably good coverage for θ in Model II in the case of $\rho = 0.5$ can be easily destroyed with a change of $\text{Corr}(X_1, X_j)$ to -0.5 for all $1 < j \le p$, while the coverage of $\widetilde{\theta}$ is much more robust against such changes. The same phenomenon is expected when Models II and III are changed in a similar way for $\rho = 0.8$.

Comment

Hannes LEEB

I congratulate Ian McKeague and Min Qian for this inspiring and creative piece of work. It is a substantial contribution to the development of inference procedures in a particularly important and statistically challenging scenario. Their findings also provide illuminating insights into a class of inferential problems of great contemporary interest, and they raise a couple of exciting questions for future research.

1. WHY IS IT HARD?

For designing hypothesis tests or confidence sets, a natural starting point is the cumulative distribution function (c.d.f.) of the (scaled) estimation error. In Theorem 1, McKeague and Qian derive the large-sample limit distribution of this quantity along sequences of parameters β_n of the form $\beta_n = \beta_0 + b_0/\sqrt{n}$. The limit always depends on β_0. And for $\beta_0 = 0$, the limit also depends on the local parameter b_0.

Phenomena like this occur in several challenging inferential problems that share some common features: Consider a generic parametric estimation problem where the underlying model is locally asymptotically normal around zero (as in the present case). Write $\varphi_n(\beta)$ for the quantity to be estimated at sample size n and under the true parameter β (here, the c.d.f. of the scaled estimation error). (The estimand can depend on additional nuisance parameters, which is not shown explicitly in the notation.) Assume that $\varphi_n(\beta)$ converges in a local neighborhood of zero, in the sense that, along sequences of parameters of the form $\beta_n = b_0/\sqrt{n}$, we have $\varphi_n(\beta_n) = \varphi_n(b_0/\sqrt{n}) \to g(b_0)$ as $n \to \infty$. Now if, as in the present case, the limit $g(b_0)$ is not constant in b_0, then no uniformly consistent estimator for the estimand $\varphi_n(\beta)$ exists. In other words, for any (possibly randomized) estimator $\hat{\varphi}_n$ for $\varphi_n(\beta)$, we have

$$\lim_{n\to\infty} \sup_{\beta} P_{n,\beta}\Big(d\left(\varphi_n(\beta), \hat{\varphi}_n\right) > \delta\Big) > 0 \qquad (1.1)$$

for sufficiently small $\delta > 0$, and the lower bound approaches $1/2$ as δ goes to zero. Moreover, if $\hat{\varphi}_n$ is a sequence of consistent estimators, then the expression on the left-hand side in the preceding display actually equals 1. (Here, $d(\cdot,\cdot)$ is some metric on the range-space of the estimand, like the absolute difference if the estimand is a c.d.f. at some point, or sup-norm of the difference if the estimand is the whole c.d.f.) In short, the problem is caused by the fact that the estimand and its large-sample limit depend on a quantity, namely, $b_0 = \sqrt{n}\beta_n$, that cannot be estimated with good accuracy at any sample size. See Leeb and Pötscher (2006) for a more detailed analysis of problems of this kind. Phenomena like this occur in several scenarios of contemporary interest, including inference problems involving post-model-selection estimators, shrinkage estimators, or situations where the parameter is close to the boundary of the identified region; see Leeb and Pötscher (2005).

In particular, for the finite-sample c.d.f. of the scaled estimation error considered in the article, that is, for the c.d.f. of $\sqrt{n}(\hat{\theta}_n - \theta_n)$, we see that no uniformly consistent estimator can exist.

2. WHY DOES IT WORK?

When testing a *simple* null hypothesis, like $H_0 : \beta = 0$ as in the present case with ART, the nonuniformity problem outlined in the preceding section is not an issue, however. Here, the quantity of interest is the distribution of the (scaled) estimation error under the null, and this distribution can often be estimated, either directly or (as in the present article) by estimating or approximating the large-sample limit distribution under the null. Indeed, in Theorem 1, the limit distribution in the case where $\beta_0 = b_0 = 0$ provides an approximation to the finite-sample distribution of interest. And while this limit distribution depends on nuisance parameters through the population variance/covariance structure, these can be consistently estimated by using sample versions as a plug-in.

Sampling from the limit distribution under the null (using sample covariances to replace population covariances) provides a simple and computationally efficient way to estimate critical values, and the size of the resulting test converges to the nominal size as sample size increases. This can be used instead of the bootstrap procedure given in Theorem 2 and also instead of the double bootstrap method used in the simulations, which both are computationally more expensive. (More generally, for any fixed β_0 and b_0, the finite-sample distribution along sequences of parameters the form $\beta_n = \beta_0 + b_0/\sqrt{n}$ as in Theorem 1 can be estimated by sampling from the corresponding limit distribution given in that theorem after replacing population covariances by estimates.) Also, if one insists on using the bootstrap, it is easy to see that, under the assumptions of Theorem 2 and under the null, also $\mathbb{V}_n^*(0)$ converges to the limiting distribution of interest.

The considerations presented so far in this section rely on the fact that the null hypothesis $H_0 : \beta = 0$ is simple. For *composite* null hypothesis, however, including those that occur in the stepwise ART procedure discussed in the article, nonuniformity can again become an issue. Consider a composite null hypothesis of the form $\beta \in B_0$. To control the size of a test of this null in large samples, one requires an estimator for the distribution of

Hannes Leeb is Professor, Department of Statistics, University of Vienna, 1010 Wien, Austria (E-mail: *hannes.leeb@univie.ac.at*). Research supported by FWF Projects P26354 and P28233-N32.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1109516

interest that is consistent uniformly over B_0. But if B_0 contains sequences of parameters of the form $\beta_n = \beta_0 + b_0/\sqrt{n}$ and if, along such a sequence, the estimand converges to a limit that is nonconstant in b_0, then no estimator can be consistent uniformly over B_0: Using the assumptions and notation of Section 1, write $\varphi_n(\beta)$ and $\hat{\varphi}_n$ for the quantity of interest and for some estimator, respectively. If B_0 contains sequences $\beta_n = \beta_0 + b_0/\sqrt{n}$ for which $\lim_{n\to\infty} \varphi_n(\beta_n) = g(b_0)$ is nonconstant in b_0, then (1.1) again holds with the supremum restricted to $\beta \in B_0$. It is yet to be seen whether the step-wise ART procedure suffers from a nonuniformity phenomenon, but I think it is rather likely that this is the case. For example, consider the case where B_0 contains a point β_0 for which k_0 is not unique, as well as a local neighborhood of β_0.

3. ON CONFIDENCE SETS

The confidence interval proposed in the article is obtained by taking the limit distribution from Theorem 1 for $\beta_0 = 0$ and some fixed b_0, estimating this limit distribution as in Theorem 2, and by then finding the widest confidence interval over all b_0. In other words, the interval is obtained by taking the widest from a collection of intervals, where each interval in the collection corresponds to the limit distribution along a sequence of parameters of the form $\beta_n = b_0/\sqrt{n}$ as in Theorem 1. For this procedure to be honest asymptotically, in the sense that the actual minimal coverage probability converges to the nominal one, one must consider *all* possible limit distributions along *arbitrary* sequences of parameters. (This is an asymptotic version of the simple idea of maximizing over unknown nuisance parameters. See, for example, Bickel and Doksum (1976, p. 170); variations of this idea are also discussed in Leeb and Pötscher (2015).)

In its current form, Theorem 1 does not provide accumulation points for arbitrary sequences of parameters. Therefore, the confidence sets proposed in the article are certainly more robust than the naive construction, in particular in local neighborhoods of zero, but it is not clear whether they actually are asymptotically honest. Asymptotically honest confidence intervals based on choosing the widest interval from possible limiting distributions are analyzed by Andrews and Guggenberger (2009). The methods proposed in that reference also provide conservative tests in case the c.d.f. of the test statistic under the null cannot be estimated in a uniformly consistent fashion (like in the situation discussed at the end of Section 2). A similar approach is taken by Laber and Murphy (2011). If one can extend the scope of Theorem 1 to cover arbitrary sequences of parameters β_n, then asymptotically honest confidence intervals can be constructed in the setting of the present article.

4. COMPETITORS

In the setting supported by Theorems 1 and 2, that is, for $p \ll n$, the likelihood ratio test can also be used. Comparing the two tests, it is tempting to expect that ART is more powerful if the signal is "spiked," in the sense that $\|\boldsymbol{\beta}_n\|$ is concentrated on a few components of $\boldsymbol{\beta}_n$, and that the likelihood ratio test is more powerful if that is not the case. This is consistent with the simulation results provided in the article, and it will be interesting to compare the two methods in a more extensive simulation study.

5. THE CASE WHERE p IS NOT SMALL RELATIVE TO n

One attractive feature of the proposed test statistic, and also of the proposed bootstrap estimator, is that they are computable also if $p > n$, while the likelihood ratio statistic is trivial in this case. The theory provided in the article, however, is based on large-sample asymptotics where p is fixed and $n \to \infty$ such that $p/n \to 0$, a situation that is not compatible with situations where p/n is not small. It is encouraging to see that ART performs well also if p/n is not small, and even if $p/n > 1$, at least in the simulation scenarios analyzed in the article. Unfortunately, more extensive simulations become prohibitively expensive if p is large, as there is a very high-dimensional parameter space to cover. This further underscores the need for theoretical results that allow for situations where p is not small relative to n. Such situations include scenarios where $p/n \to c$ with either $0 < c < 1$, with $1 < c < \infty$, and even with $c = \infty$. For the case where $0 < c < 1$, Leeb (2009) found, in a different inferential problem, that results from fixed-p-asymptotics can be misleading, and that alternative approximations are needed. Also, El Karoui and Purdom (2015) found that the bootstrap is not reliable in the case where $0 < c < 1$. It will be exciting to analyze such scenarios in future work.

6. SITUATIONS WHERE k_n IS NOT UNIQUE

Finally, I would like to comment on the case where k_n is not unique. With the exception of the case where $\beta_0 = b_0 = 0$, this nonunique case is ruled out in Theorems 1 and 2. But it is clear that new phenomenon will arise is the true parameter is such that k_n is not unique, or within a local neighborhood of such a point in parameter space. Given the machinery developed in the article so far, I do not think that this case will be difficult to deal with. In particular, one case where k_n is not unique is already covered, namely, the case where $\beta_0 = b_0 = 0$. I suspect that other cases, where k_n is not unique, can lead to asymptotic distributions that combine the features shown by the two types of limiting distributions shown in Theorem 1. For the construction of asymptotically honest confidence intervals outlined in Section 3, it will be instrumental to characterize all possible accumulation points of the finite-sample distributions along arbitrary sequences of parameters, which includes sequences along which k_n is not unique (or local to a point of nonuniqueness).

REFERENCES

Andrews, D. W. K., and Guggenberger, P. (2009), "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77, 721–762. [1458]

Bickel, P. J., and Doksum, K. A. (1976), *Mathematical Statistics: Basic Ideas and Selected Topics*, Oakland, CA: Holden-Day. [1458]

El Karoui, N., and Purdom, E. (2015), "Can We Trust the Bootstrap in High-Dimension?" Technical Report 824, Department of Statistics, University of California, Berkeley. [1458]

Laber, E. B., and Murphy, S. A. (2011), Adaptive Confidence Intervals for the Test Error in Classification," *Journal of the American Statistical Association*, 106, 904–913. [1458]

Leeb, H. (2009), "Conditional Predictive Inference Post Model Selection," *Annals of Statistics*, 37, 2838–2876. [1458]

Leeb, H., and Pötscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59. [1457]

——— (2006), "Performance Limits for Estimators of the Risk or Distribution of Shrinkage-Type Estimators, and Some General Lower Risk-Bound Results," *Econometric Theory*, 22, 69–97. [1457]

——— (2015), "Testing in the Presence of Nuisance Parameters: Some Comments on Tests Post-Model-Selection and Random Critical Values," arXiv:1209.4543. [1458]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

Rejoinder

Ian W. McKeague and Min Qian

We greatly appreciate all the hard work that the editors and the discussants put into providing enlightening comments. Their original perspectives on post-selection inference have led us to a deeper understanding of the problem. We have organized our rejoinder along the lines of their key questions. After recapping the main ideas in the adaptive resampling test (ART), we address the broad issues in order of increasing difficulty: the need for scale-invariance, calibration via simulation, robustness to model misspecification, the detection of weak dense signals, variable selection, and the problem of finding "honest" confidence sets.

ART is based on finding a suitable calibration for the test statistic $\sqrt{n}\hat{\theta}_n$, where

$$\hat{\theta}_n = \frac{\widehat{\mathrm{cov}}(X_{\hat{k}_n}, Y)}{\widehat{\mathrm{var}}(X_{\hat{k}_n})} \quad \text{and} \quad \hat{k}_n = \arg\max_{k=1,\ldots,p} |\widehat{\mathrm{Corr}}(X_k, Y)|$$

is the asymptotically unique index of the maximally correlated predictor. Our main result shows that it is possible to correct for the failure of the centered percentile bootstrap (CPB, or what many of the discussants call the "naive" bootstrap, Efron and Tibshirani 1993) in the neighborhood of the null hypothesis. This is achieved by adapting to evidence of nonregularity by resampling from an observed process $\mathbb{V}_n$ that is indexed by an (unidentifiable) local parameter $\boldsymbol{b}_0 \in \mathbb{R}^p$ representing uncertainty in the regression parameters at the $\sqrt{n}$-scale.

The central idea of ART is to calibrate the test statistic $\sqrt{n}\hat{\theta}_n$ by adaptive bootstrapping:

$$A_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)1_{\mathrm{reg}^*} + \mathbb{V}_n^*(\boldsymbol{b}_0)1_{\mathrm{nreg}^*},$$

where $\mathrm{reg}^* = \{\max(|T_n|, |T_n^*|) > \lambda_n\}$ indicates that the post-selected t-statistic along with its bootstrapped version exceed a threshold, so draws that agree with the CPB are "acceptable." On the complementary event, $\mathrm{nreg}^* = \{\max(|T_n|, |T_n^*|) \le \lambda_n\}$, there is evidence of a nonregular limit and the more sophisticated bootstrap $\mathbb{V}_n^*(\boldsymbol{b}_0)$ is needed to take into account the local asymptotic behavior of $\hat{\theta}_n$.

Theorem 2 shows that A_n^* consistently estimates the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ under arbitrary $\sqrt{n}$-scale perturbations of the regression parameters. At the null hypothesis we can set $\boldsymbol{b}_0 = 0$, so without having to cope with a function of $\boldsymbol{b}_0$, critical values are readily obtained. Our contention is that the problem of detecting the presence of significant predictors can be handled in a similar fashion for more sophisticated classes of marginal regression models; the tractability of the linear regression case, however, makes it an ideal testbed for the general approach.

1. SCALE-INVARIANCE

Several discussants raise the point that the test statistic $\sqrt{n}\hat{\theta}_n$ used in ART is not scale invariant. To compensate for this, Shah and Samworth (SS hereafter) recommend prestandardizing all variables before applying ART. They note that failure to do so could result in a substantial loss of power, as they show in a simple example. Although counterintuitive (since the fitting of linear regression is impervious to scale changes), scale-invariance is crucial in variable selection problems. Indeed, the standardization of predictors is routinely recommended when shrinkage methods are applied in high-dimensional regression, (see, e.g., Hastie, Tibshirani, and Friedman 2009, p. 63).

Zhang and Laber (ZL hereafter) suggested that ART should be based on the scale-invariant t-statistic $T_n = \hat{\theta}_n/s_n$, rather than $\sqrt{n}\hat{\theta}_n$, as did Brown and McCarthy (BM hereafter). ZL went on to discuss how our approach can be readily modified to apply to T_n (which they denoted $\hat{\xi}_n$), and noted that the resulting procedure is almost identical to ART (when Y and X_k have unit variance). Chatterjee and Lahiri (CL hereafter) suggested an alternative scale-invariant test statistic (denoted Λ_n) that we discuss later.

The expedient to the lack of scale invariance in ART that we prefer in practice is SS's suggestion of prestandardizing all variables. The reason we used the test statistic $\sqrt{n}\hat{\theta}_n$ (rather than maximal sample correlation) in ART is that the theory is simpler to explain (less cumbersome notation), the connection to robust CIs for the slope parameter more direct, and to make our results potentially relevant for more general marginal regression models. Our simulation studies used only standardized predictors, so the conclusions are not affected. To address the invariance issue, however, we have retrospectively added a comment in the article

Ian W. McKeague (E-mail: *im2131@columbia.edu*) is Professor and Min Qian (E-mail: *mq2158@columbia.edu*) is Assistant Professor, Department of Biostatistics, Columbia University, New York, NY 10027. Research of the first author is supported by NIH Grant R01GM095722-05 and NSF Grant DMS-1307838. Research of the second author is supported by NSF Grant DMS-1307838.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1107431

[Review Copy Only]

about the need to prestandardize (just after the description of the ART procedure).

2. CALIBRATION VIA SIMULATION FROM ESTIMATED NULL DISTRIBUTION

SS note that "under the global null that Y and $\mathbf{X}$ are independent" the limiting distribution of $\sqrt{n}\hat{\theta}_n$, after standardization of variables, does not depend on the distribution of Y when it can be assumed that ϵ and $\mathbf{X}$ are independent. In that case, simulation of $\sqrt{n}\hat{\theta}_n$ using $Y \sim N(0, 1)$, $Y \sim$ its empirical distribution, or Y-permutations, will indeed provide accurate calibration. Further, this would provide substantial computational savings over ART.

ZL had a similar suggestion: simulate from the estimated null limiting distribution of $T_n = \hat{\xi}_n = \hat{\theta}_n/s_n$, which they called a parametric bootstrap. This approach requires an estimate of $\text{cov}(\mathbf{X})$, and they propose that the sample covariance matrix $\widehat{\text{cov}}(\mathbf{X})$ (without regularization) is adequate for this purpose because the null limiting distribution of T_n is a smooth function of $\text{cov}(\mathbf{X})$.

We agree that these approaches provide substantial computational savings, but their validity depends on the highly restrictive assumption that ϵ and $\mathbf{X}$ are independent. On the other hand, our results justifying ART only require ϵ and $\mathbf{X}$ to be *uncorrelated*.

When ϵ and $\mathbf{X}$ are dependent, the null limiting distribution of $\sqrt{n}\hat{\theta}_n$ can depend on the distribution of Y, in which case the Y-permutation and other simulation methods suggested by SS break down. The method of ZL also breaks down since it no longer suffices to estimate $\text{cov}(\mathbf{X})$. As we show later using a simple simulation example, their approach can result in inflated Type I errors when ϵ and $\mathbf{X}$ are dependent. Moreover, by a simple extension of Theorem 1 of the article, to simulate draws from the null limiting distribution of T_n, moments of the form $E\epsilon^2 X_j X_k$ would need to be estimated. It is not clear how that could be done when ϵ and $\mathbf{X}$ are dependent. In fairness to the discussants, however, in the version of the manuscript that they initially saw, we inadvertently made the assumption of independence between ϵ and $\mathbf{X}$, even though in fact we only needed zero correlation.

A further difficulty with the direct simulation approach, which relies on having an accurate estimate of $\text{cov}(\mathbf{X})$ (not needed in ART), is that uncertainty about $\text{cov}(\mathbf{X})$ is not taken into account, and it is not clear how that could be done (although we admit that in the simulation examples studied by ZL there does not appear to be a problem in this regard). Another consideration is that in more complex types of marginal regression models (such as quantile regression), the limiting distribution can depend on nuisance parameters that are hard to estimate, so a bootstrap approach is desirable.

3. ROBUSTNESS TO MODEL MISSPECIFICATION

We are indebted to Brown and McCarthy (BM) for prompting us to reexamine the proofs of our main results to confirm that they still justify ART in the "assumption-lean" (Buja et al. in press) setting of ϵ and $\mathbf{X}$ just being uncorrelated, as discussed above. In reference to their query concerning sandwich estimators (in Section 2 of their discussion), we agree that there is a close parallel to our Theorem 1. Nevertheless, the Huber–White sandwich formula for the asymptotic variance of M-estimators only applies in regular settings, whereas our version also reflects nonregularity. More specifically, from our Theorem 1, the asymptotic variance of $\hat{\theta}_n$ when $\boldsymbol{\beta}_0 = \mathbf{0}$ does not reduce to a sandwich formula because K is random, so V_K^{-1} cannot be factored out of the expression.

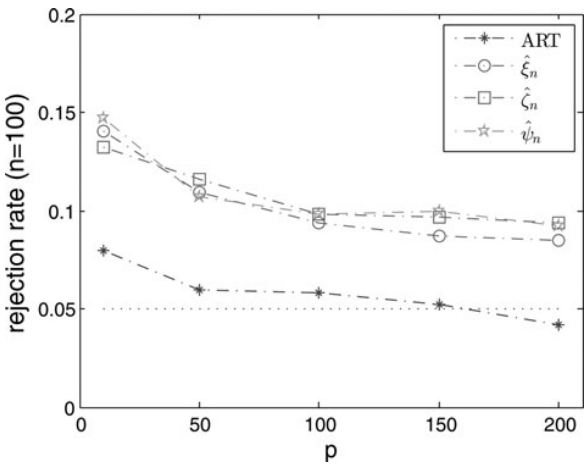

Figure 1. Empirical rejection rates based on 1000 samples generated from the heteroscedastic simulation model (1) as the dimension p ranges from 10 to 200, for $n = 100$.

We agree, however, that this parallel suggests that ART is much more flexible and robust to model misspecification than we originally thought. To examine this question, we devised the following simple simulation example in which we assess the Type I error control of ART when ϵ and $\mathbf{X}$ are not independent, just uncorrelated, and compare it with the "direct simulation" tests statistics $\hat{\xi}_n$, $\hat{\zeta}_n$, and $\hat{\psi}_n$ proposed by ZL. The following heteroscedastic model has no linear effects, so $H_0 : \theta_0 = 0$ holds:

$$Y = \epsilon \equiv X_1 X_2 + \delta, \tag{1}$$

where $X_k \sim N(0, 1)$, $\text{Corr}(X_j, X_k) = 0.2$, and $\delta \sim N(0, 1)$. Moreover, note that $E\epsilon X_k = E((X_1 X_2 + \delta)X_k) = 0$, so $\text{cov}(\epsilon, \mathbf{X}) = 0$ and ART should provide adequate Type I error control. Indeed, Figure 1 confirms this and shows that the direct simulation approach has inflated Type I error.

4. DETECTION OF WEAK DENSE SIGNALS

ZL proposed a test statistic $\hat{\zeta}_n$ for detecting weak dense signals (in contrast to a sparse signal), and provide simulation examples showing that it has better power than ART in such settings. Further, they proposed an adaptive parametric bootstrap test statistic that combines $\hat{\xi}_n$ and $\hat{\zeta}_n$ into a statistic $\hat{\psi}_n$ that adapts to an unknown level of sparsity.

Chatterjee and Lahiri (CL) make a similar proposal with their test statistic Λ_n, and suggest calibration by either naive bootstrap or direct simulation from the estimated null (which is a weighted sum of chi-squared random variables in this case). They report simulation results for ART under both spiked and weak-dense signals (in models with ϵ and $\mathbf{X}$ taken to be independent), and claim that ART performs "slightly worse" in the latter case, and that Λ_n has greater power. This is consistent with the simulation results presented by ZL. These are inventive proposals, but they appear to produce at most a borderline improvement in power over ART for weak dense signals (see Table 2 of ZL), and

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

the concern that they are not robust to model misspecification remains.

5. VARIABLE SELECTION

Barut and Wang (BW hereafter) investigate via simulation the variable selection performance of forward stepwise ART, and find that its performance declines as the correlation between covariates increases. This is not surprising as the proposed forward stepwise ART uses residuals from the previous stage as the new outcome, which essentially removes the effect of the remaining variables if they are highly correlated with already included variables. However, our argument is that although forward stepwise ART may not be variable selection consistent, it has high prediction accuracy, as we show in the real data example in Section 4.4. BW surmise that the variable selection performance of ART would be improved if it could be extended to forward regression by allowing the coefficients of already-included variables to be refit at each step (Barut, Fan, and Verhasselt 2015). We agree, and view this suggestion as a potentially fruitful direction for future research.

BW conclude their discussion with an illuminating analysis of conditions under which stepwise marginal screening has the property of "faithfulness," that is, being able to recruit active variables with high probability, and they compare with the analogous conditions for the Lasso. This relates to the broad and challenging problem of how to ensure variable selection consistency along with the provision of accurate post-selection inference.

6. CONFIDENCE INTERVALS

Several of the discussants, including Li, Mitra and Zhang (LMZ hereafter), SS, and Leeb, express interest in constructing CIs for marginal regression. In particular, LMZ provide a lucid explanation of how the bootstrap used in ART relates to various naive bootstrap procedures that are not expected to work. They also carry out a simulation study to assess various CIs that are related, though not identical, to what we discuss in the article. They compare coverage rates for the selected signal $\theta_{\hat{k}_n}$ and the "strongest population signal" θ_0, concluding that reliable inference for $\theta_{\hat{k}_n}$ is the best that can be achieved in the case of weak signals. In contrast to our proposed CI, none of the adaptive bootstrap procedures of LMZ involve maximization over a local parameter. We expect that maximization of quantiles over the local parameter, even though computationally expensive, along with the use of the double bootstrap for selecting the threshold, would result in better coverage of θ_0.

Leeb discusses the inherent difficulty of forming "honest" CIs when the limits of sampling distributions depend on local parameters $b_0 = \sqrt{n}\beta_n$ (in his notation), in which case the target parameter β_n cannot be estimated with good accuracy at any sample size (Leeb and Pötscher 2006). Our results extend to limit distributions along sequences of local parameters $b_n \to b_0$, and b_0 can even be infinite (corresponding to a nonlocal alternative), but it is not clear whether that is enough to produce honest CIs of the type that Leeb would like to see (Leeb and Pötscher 2014). Adapting to *arbitrary* sequences of parameters β_n having varying rates of convergence seems very challenging. Leeb also raises the interesting question of whether the uniqueness assumption for k_0 (the index of the strongest signal) could be relaxed in Theorem 1. Indeed, this can be done, although at the expense of a more complex limiting distribution.

Belloni and Chernozhukov discuss orthogonal score functions for constructing uniformly valid confidence sets for preconceived regression parameters (via a multiplier bootstrap procedure), where the uniformity is with respect to an underlying sparse model, see Belloni, Chernozhukov, and Kato (2014b). In related work, Javanmard and Montanari (2015) had developed accurate CIs for any given slope parameter in linear regression based on a de-biased Lasso estimator. In these approaches the dimension p is allowed to grow with n, but the resulting CIs are not suitable for the marginal screening of large numbers of predictors unless a Bonferroni-type correction is applied, which would be extremely conservative in high dimensions.

An interesting direction for further research would be to try to adapt these ideas to construct honest and computationally tractable CIs for θ_0 in marginal regression with growing dimension. The use of orthogonal score functions (as outlined in the discussion of Belloni and Chernozhukov) could potentially lead to an important extension of ART in which there is adjustment for high-dimensional controls that are automatically included in every marginal regression; this might be achieved by extending the approach in Belloni, Chernozhukov, and Hansen (2014a). At present, however, even formulating the type of asymptotic justification that would be needed under growing dimension seems challenging because post-selection is inevitably involved in the estimation of θ_0, and it appears difficult to find a normalization of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ that scales in a tractable fashion with dimension.

At the end of their discussion, ZL made the interesting suggestion that a target parameter such as the "soft-max" (that depends smoothly on the regression parameters) would offer a feasible alternative to θ_0 in terms of avoiding the need to handle complex asymptotic arguments need to justify the honesty of CIs. While we are sympathetic to this idea, we believe that the loss of interpretability in using a surrogate for θ_0 is too high a price to pay. Further, we would expect that the ad hoc nature of an estimand that depends on a tuning parameter would make the approach vulnerable to the same post-selection difficulties already inherent in θ_0.

We conclude with a philosophical point. In his famous essay *The Hedgehog and the Fox*, Isaiah Berlin drew attention to a dichotomy between the need to know many things, as with the fox, or to know one big thing, as with the hedgehog. That is, whether to prefer "a single, universal, organizing principle" on the one hand, or to "pursue many ends, often unrelated and even contradictory" on the other. By analogy, the fox has scattered knowledge about a vast collection of regression parameters, but (at least with some ART and the help of our gracious discussants) the hedgehog may know θ_0, the biggest of all.

REFERENCES

Barut, E., Fan, J., and Verhasselt, A. (2015), "Conditional Sure Independence Screening," *Journal of the American Statistical Association*, to appear, DOI: 10.1080/01621459.2015.1092974. [1461]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a), "Inference on Treatment Effects After Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. [1461]

Belloni, A., Chernozhukov, V., and Kato, K. (2014b), "Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other Z-Estimation Problems," *Biometrika*, 102, 77–94. [1461]

Buja, A., Berk, R., Brown, L. D., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (in press), "Models as Approximations—A Conspiracy of Random Regressors and Model Deviations Against Classical Inference in Regression," *Statistical Science*. [1460]

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap (Mono graphs on Statistics & Applied Probability)*, Boca Raton, FL: Chapman & Hall/CRC. [1459]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, New York: Springer. [1459]

Javanmard, A., and Montanari, A. (2015), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1461]

Leeb, H., and Pötscher, B. M. (2006), "Performance Limits for Estimators of the Risk or Distribution of Shrinkage-Type Estimators, and Some General Lower Risk-Bound Results," *Econometric Theory*, 22, 69–97. [1461]

——— (2014), "Testing in the Presence of Nuisance Parameters: Some Comments on Tests Post-Model-Selection and Random Critical Values," arXiv:1209.4543. [1461]

[Date: May 12, 2016 Received by Professor Todd Alan Kuffner for reviewing purposes only]

[Review Copy Only]