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

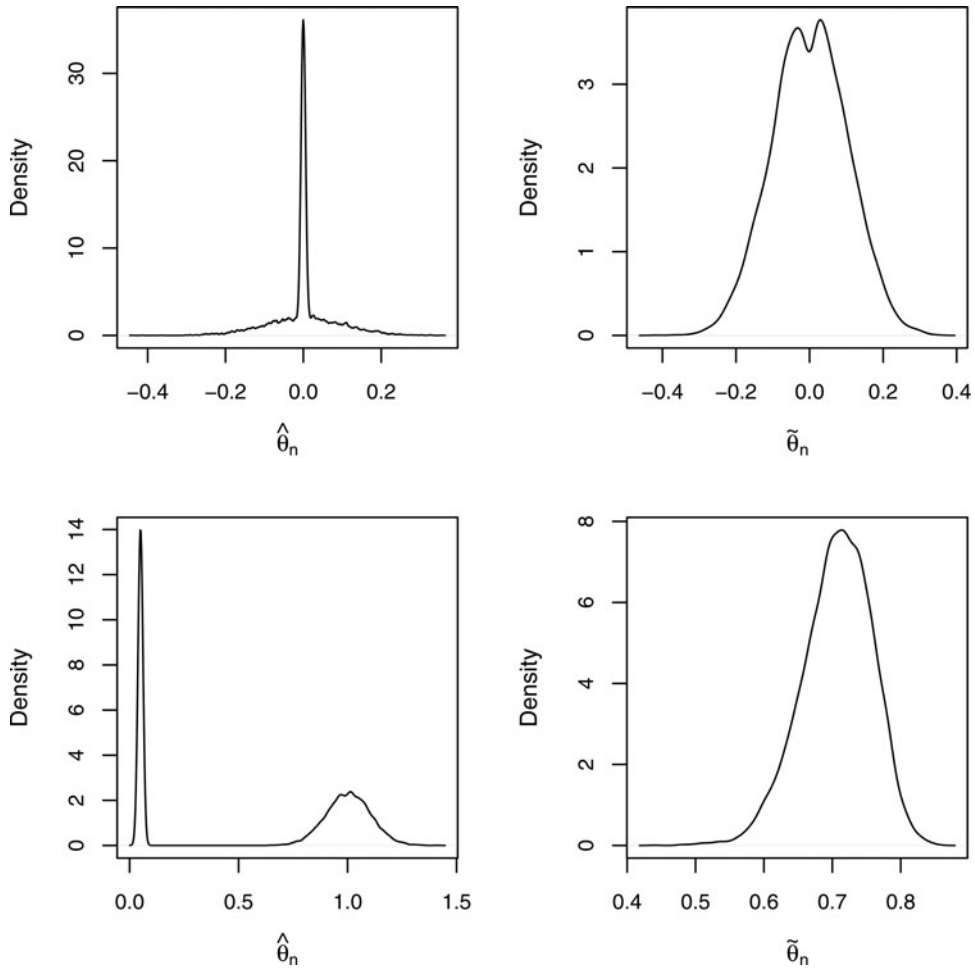

Figure 1. Top row: density plots of $\hat{\theta}_n$ (left) and $\tilde{\theta}_n$ (right) under the global null hypothesis for the example in Section 2. Bottom row: corresponding density plots of $\hat{\theta}_n$ (left) and $\tilde{\theta}_n$ (right) under the alternative specified in (2.1).

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,

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

as $n \to \infty$, where $\Theta_{jk} = \mathrm{Corr}(X_j, X_k)$. 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

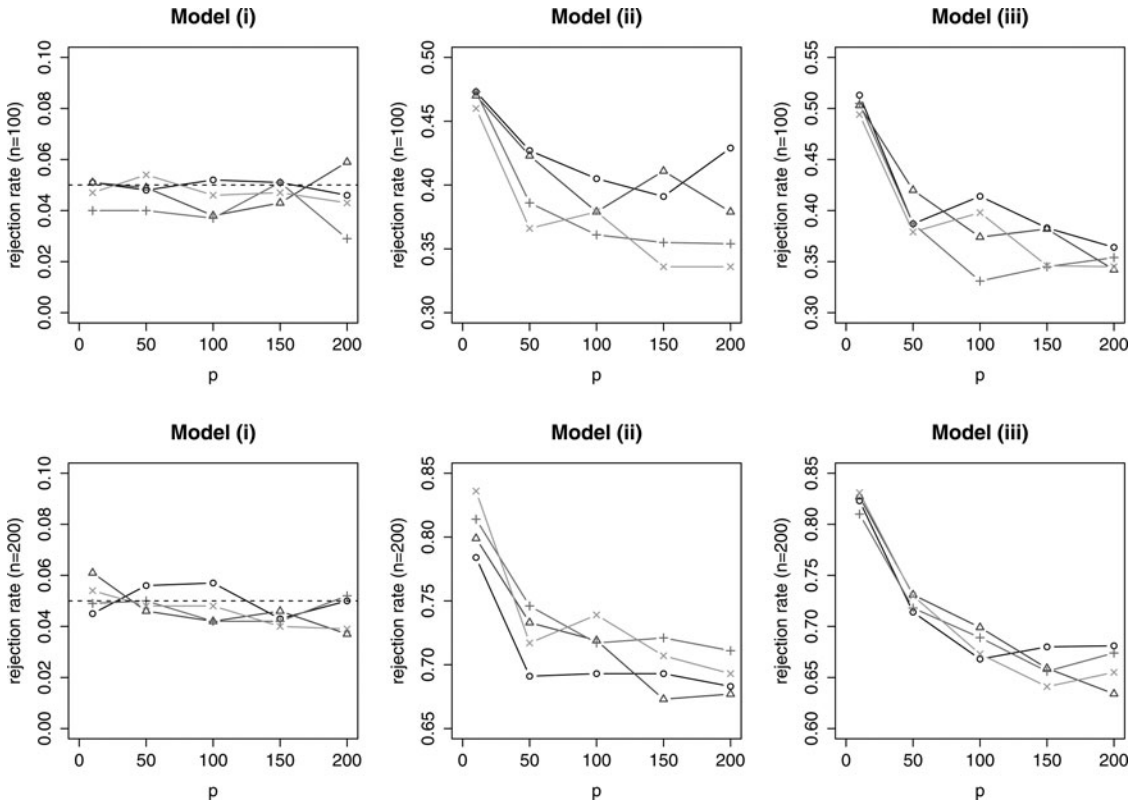

Figure 2. The same graphs as in Figure 1 ($\rho = 0.5$) of the original article but for *globaltest* (black circles), method (a) (green crosses), method (b) (red plus signs), and the permutation test (blue triangles). Note model (i) is the null model. (For interpretation of the references to color in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

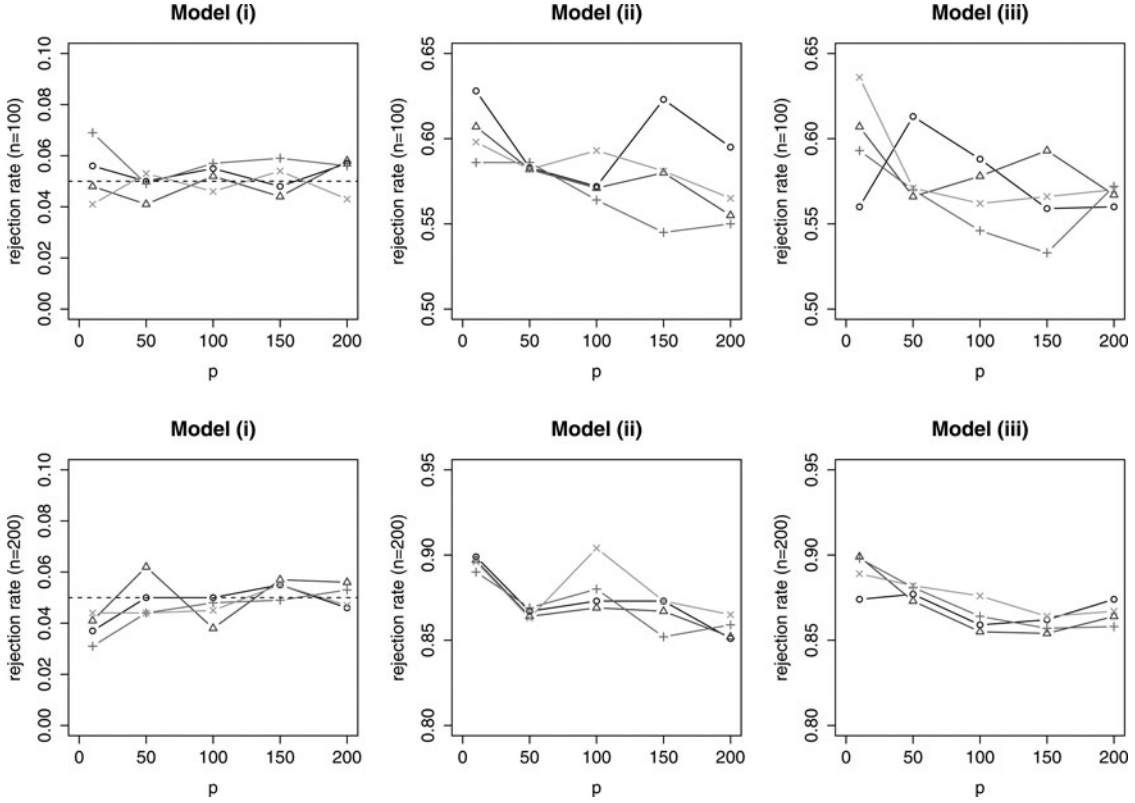

Figure 3. The same graphs as in Figure 2 ($\rho = 0.8$) of the original article but for *globaltest* (black circles), method (a) (green crosses), method (b) (red plus signs), and the permutation test (blue triangles).

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.

DOI: 10.1080/01621459.2015.1100619