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]

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

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

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

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.

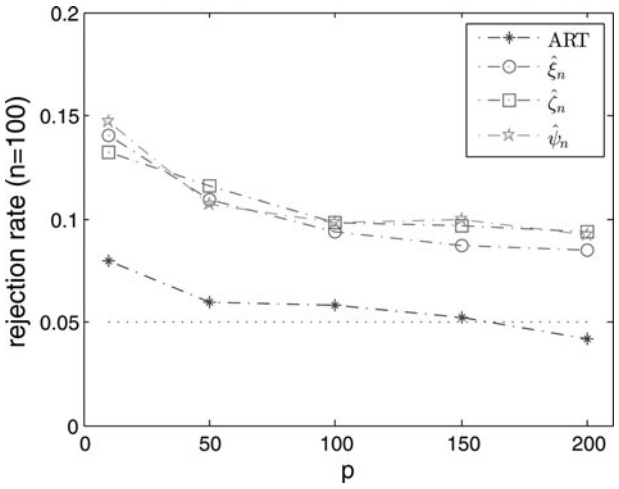

Figure 1. Empirical rejection rates based on 1000 samples generated from the heteroscedastic simulation model (1) as the dimension p ranges from 10 to 200, for $n = 100$.

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

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]