Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Comment

Emre Barut & Huixia Judy Wang

To cite this article: Emre Barut & Huixia Judy Wang (2015) Comment, Journal of the American Statistical Association, 110:512, 1442-1445, DOI: 10.1080/01621459.2015.1100619

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1100619

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

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

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

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

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

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]