Testing for the presence of significant covariates through conditional marginal regression

By YANLIN TANG

School of Mathematical Sciences, Tongji University, 1239 Siping Road, Shanghai 200092, China

tangyl@tongji.edu.cn

HUIXIA JUDY WANG AND EMRE BARUT

Department of Statistics, The George Washington University, 801 22nd St NW, Washington, District of Columbia 20052, U.S.A.

judywang@gwu.edu   barut@gwu.edu

SUMMARY

Researchers sometimes have a priori information on the relative importance of predictors that can be used to screen out covariates. An important question is whether any of the discarded covariates have predictive power when the most relevant predictors are included in the model. We consider testing whether any discarded covariate is significant conditional on some pre-chosen covariates. We propose a maximum-type test statistic and show that it has a nonstandard asymptotic distribution, giving rise to the conditional adaptive resampling test. To accommodate signals of unknown sparsity, we develop a hybrid test statistic, which is a weighted average of maximum- and sum-type statistics. We prove the consistency of the test procedure under general assumptions and illustrate how it can be used as a stopping rule in forward regression. We show, through simulation, that the proposed method provides adequate control of the familywise error rate with competitive power for both sparse and dense signals, even in high-dimensional cases, and we demonstrate its advantages in cases where the covariates are heavily correlated. We illustrate the application of our method by analysing an expression quantitative trait locus dataset.

Some key words: Adaptive resampling; Conditional marginal regression; Forward selection; Hybrid test.

1. INTRODUCTION

Researchers are often interested in the effect of covariates X on a response Y, conditional on other covariates Z. This problem occurs when the effects of Z have been verified or are assumed to be significant through prior knowledge or previous studies. Classical examples of such variables are age, gender and patient stratum in psychological or biological studies.

To test the significance of X in the presence of Z, Lan et al. (2014) proposed a partial F-test, which has a simple asymptotic null distribution and can be applied with $p > n$. However, the test focuses on the joint effect of X and hence cannot identify which covariates in X are significant conditional on Z; it is also not powerful for sparse signals. We propose a maximum-type test to detect the effects of X in the presence of Z by considering conditional marginal regression, where the response variable is separately regressed on each component of X conditional on Z. Conditional marginal regression has numerous benefits in the high-dimensional regime: it can

help to recover hidden significant covariates and to reduce false positive and false negative rates (Barut et al., 2016). The proposed test based on conditional regression differs from the post-selection inference procedures in Leeb & Pötscher (2005) and Belloni et al. (2015), where the inference focuses on the effect of Z after the selection of variables in X. In addition, we assume that the effects of variables in the conditioned set, Z, have either been validated through numerous previous studies or that their inclusion is necessary for statistical analysis.

The proposed test differs from the unconditional test in McKeague & Qian (2015) in three major aspects. First of all, our test is conditional, designed to check for the presence of significant variables given a set of conditioned variables. Second, it is based on a scale-invariant t-statistic, whereas that in McKeague & Qian (2015) is based only on the maximum slope estimator. In particular, our method selects the most predictive covariate through maximizing $(\hat{\theta}_j^Z)^2 \hat{V}_{j|Z}$, where $\hat{\theta}_j^Z$ is the estimated coefficient of X_j when regressing Y on Z and X_j together, and $\hat{V}_{j|Z}$ is the estimated conditional variance of X_j given Z. Because of the covariate-specific $\hat{V}_{j|Z}$ in our set-up, pre-standardization of covariates is not sufficient to make the type of statistic in McKeague & Qian (2015) scale-free, so it is important to incorporate the scale of $\hat{\theta}_j^Z$ in the test statistic, and this requires new computation and theory. Finally, to accommodate signals of unknown sparsity, we propose a new hybrid test statistic, a weighted average of maximum- and sum-type statistics. The hybrid test not only is powerful for detecting sparse and large signals but also can have good power against dense alternatives. For test calibration, we adapt the idea in McKeague & Qian (2015) and develop a conditional adaptive resampling test procedure by using the asymptotic properties of the maximum-type test statistic under the local model.

In applications, one can use the proposed test as a first step to assess the overall significance of the effect of X on Y in the presence of Z. When the null hypothesis is rejected, one natural question is which variables in X should be included in the model. To answer this question, we can apply our proposed test in the context of forward regression, where at each step we add the selected covariate to the conditioning set Z until no more significant covariates are identified.

This forward regression approach differs from the stepwise method in McKeague & Qian (2015), where at each stage the adaptive resampling test procedure is successively applied by treating residuals from the previous stage as the new response. Barut & Wang (2015) showed that the marginal regression-based adaptive resampling test method is susceptible to unfaithfulness when correlations between covariates increase, and that forward regression selects important variables under conditions comparable to those of the lasso. Our numerical studies suggest that using the adaptation of our test in forward regression leads to more accurate variable selection when covariates are highly correlated.

2. CONDITIONAL ADAPTIVE RESAMPLING TEST

2·1. *Conditional marginal regression*

Suppose that Y is a scalar response and $(Z^T, X^T)^T$ is a p-dimensional set of predictors, where Z is a q-dimensional set of variables that are known to be related to Y, and X is a d-dimensional set of remaining variables, with $d = p - q$. We consider the linear regression model

$$Y = \alpha_0 + Z^T \alpha_{Z,0} + X^T \beta_{X,0} + \varepsilon, \tag{1}$$

where α_0, $\alpha_{Z,0} = (\alpha_{1,0}, \ldots, \alpha_{q,0})^T$ and $\beta_{X,0} = (\beta_{1,0}, \ldots, \beta_{d,0})^T$ are unknown and ε has zero mean and finite variance and is uncorrelated with $(Z^T, X^T)^T$. We are interested in testing whether at least one of the predictors in X affects Y after accounting for the effects of Z.

We develop a testing method based on conditional marginal regression, which consists of fitting d separate linear models by regressing Y on each predictor in X conditional on Z. For $j = 1, \ldots, d$ we define $(\alpha_0, \theta_Z^j, \theta_j^Z) = \arg\min_{\alpha, \theta_Z, \theta_j} E\{(Y - \alpha - Z^\mathrm{T}\theta_Z - X_j\theta_j)^2\}$, where θ_j^Z measures the conditional effect of X_j on Y given Z. Write $V_{ZZ} = \mathrm{cov}(Z)$, $V_{Zj} = \mathrm{cov}(Z, X_j)$ and $V_{jj} = \mathrm{var}(X_j)$. We define the index of the most predictive covariate given Z as

$$k_0 = \operatorname*{arg\,max}_{j=1,\ldots,d} (\theta_j^Z)^2 V_{j|Z}, \qquad (2)$$

where $V_{j|Z} = V_{jj} - V_{Zj}^\mathrm{T} V_{ZZ}^{-1} V_{Zj}$. When $(Z^\mathrm{T}, X^\mathrm{T})^\mathrm{T}$ follows a multivariate normal distribution, $V_{j|Z}$ is the variance of X_j conditional on Z. With $\theta_0^Z = \theta_{k_0}^Z$, testing whether any predictor in X is correlated with Y given Z is equivalent to testing $H_0 : \theta_0^Z = 0$ versus $H_a : \theta_0^Z \neq 0$.

When $q = 1$, $k_0 = \arg\max_j |\mathrm{corr}(X_j, Y \mid Z)|$, i.e., the index of the predictor that maximizes the absolute partial correlation with Y given Z. For general cases, the following proposition gives two other equivalent definitions of k_0.

PROPOSITION 1. *For $j = 1, \ldots, d$, let $\eta_{j|Z} = V_{ZZ}^{-1} V_{Zj}$ denote the slope vector from regressing X_j on Z, and let $\tilde{X}_{j|Z} = X_j - Z^\mathrm{T}\eta_{j|Z}$. Under model* (1)*, k_0 in* (2) *is equivalent to*

$$k_0 = \operatorname*{arg\,max}_{j=1,\ldots,d} \left|\mathrm{corr}(\tilde{X}_{j|Z}, X^\mathrm{T}\beta_{X,0})\right| = \operatorname*{arg\,min}_{j=1,\ldots,d} \min_{\alpha, \theta_Z, \theta_j} E\{(Y - \alpha - Z^\mathrm{T}\theta_Z - \tilde{X}_{j|Z}\theta_j)^2\}.$$

Given a random sample $\{(Y_i, Z_i, X_i), i = 1, \ldots, n\}$ of (Y, Z, X), let $(\hat{\alpha}_j, \hat{\theta}_Z^j, \hat{\theta}_j^Z, \hat{k}_n)$ be the least squares estimators of $(\alpha_0, \theta_Z^j, \theta_j^Z, k_0)$:

$$(\hat{\alpha}_j, \hat{\theta}_Z^j, \hat{\theta}_j^Z) = \operatorname*{arg\,min}_{\alpha, \theta_Z, \theta_j} \sum_{i=1}^{n} (Y_i - \alpha - Z_i^\mathrm{T}\theta_Z - X_{ij}\theta_j)^2, \quad \hat{k}_n = \operatorname*{arg\,max}_{j=1,\ldots,d} (\hat{\theta}_j^Z)^2 \hat{V}_{j|Z},$$

where $\hat{V}_{j|Z} = \hat{V}_{jj} - \hat{V}_{Zj}^\mathrm{T} \hat{V}_{ZZ}^{-1} \hat{V}_{Zj}$, $\hat{V}_{ZZ} = \mathrm{cov}_n(Z)$ is the sample covariance matrix of Z, $\hat{V}_{Zj} = \mathrm{cov}_n(Z, X_j)$ is the sample covariance of Z and X_j, and $\hat{V}_{jj} = \mathrm{var}_n(X_j)$ is the sample variance of X_j. We then define the conditional marginal regression estimator of θ_0^Z as $\hat{\theta}_n^Z = \hat{\theta}_{\hat{k}_n}^Z$. For any given j, the standard error of $\hat{\theta}_j^Z$ by regressing Y on Z and X_j is $\hat{\sigma}_j/(n\hat{V}_{j|Z})^{1/2}$, where $\hat{\sigma}_j^2 = \sum_{i=1}^{n} \hat{e}_{ij}^2/(n-q-2)$ with $\hat{e}_{ij} = Y_i - \hat{\alpha}_j - Z_i^\mathrm{T}\hat{\theta}_Z^j - X_{ij}\hat{\theta}_j^Z$. Write $\hat{\sigma}_n = \hat{\sigma}_{\hat{k}_n}$ and $\hat{V}_{n|Z} = \hat{V}_{\hat{k}_n|Z}$. For testing H_0, one natural choice is the maximum-type test statistic

$$T_n = \hat{\theta}_n^Z / \{\hat{\sigma}_n/(n\hat{V}_{n|Z})^{1/2}\} = (n\hat{V}_{n|Z})^{1/2}\hat{\theta}_n^Z / \hat{\sigma}_n,$$

but its calibration is difficult. Firstly, under H_0, k_0 can be any index and so is not identifiable. Secondly, $\hat{\theta}_n^Z$ has a nonnormal limiting distribution. In particular, the limiting distribution of $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_0^Z)/\hat{\sigma}_n$ is discontinuous in the neighbourhood of $\theta_0^Z = 0$, so the limiting normal distribution that holds away from $\theta_0^Z = 0$ cannot be used to calibrate the test. Discussion of nonuniform convergence caused by nonregularity can be found in Cheng (2015) and McKeague & Qian (2015).

2·2. *Limiting distribution of T_n under the local model*

To construct a suitable test based on T_n, it is crucial to study local models. We consider the local regression model that replaces $\beta_{X,0}$ in (1) by

$$\beta_{X,n} = \beta_{X,0} + n^{-1/2}b_0. \tag{3}$$

Throughout, we assume that the distributions of ε and $(Z^{\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}$ are fixed, and only the distribution of Y depends on n through $\beta_{X,n}$; we also suppress n in the notation for Y. Under the local model, we rewrite k_0 defined in (2) as k_n and write $\theta_n^Z = \theta_{k_n}^Z$.

For any $b \in \mathbb{R}^d$, define $\bar{k}(b) = \arg\max_{j=1,\dots,d} |\mathrm{corr}(\tilde{X}_{j|Z}, X^{\mathrm{T}}b)|$. From Proposition 1 it is easy to see that under the local model (3), $k_n = \bar{k}(\beta_{X,n})$. If $\beta_{X,0} \neq 0$ and $\bar{k}(\beta_{X,0})$ is unique, then $k_n \to k_0$ and $\hat{\theta}_n^Z$ is asymptotically bounded away from zero, representing a nonlocal alternative. If $\beta_{X,0} = b_0 = 0$, then k_n is not well defined and $\theta_n^Z = 0$, representing the null model. However, if $\beta_{X,0} = 0$ and $b_0 \neq 0$, θ_n^Z lies in the neighbourhood of zero, representing a local alternative, and $\bar{k}(b_0)$ is weakly identifiable even though the local parameter b_0 is nonidentifiable. The benefit of studying the local model is that we can establish and express the nonregular asymptotic distribution of $\hat{\theta}_n^Z$ at $\beta_{X,0} = 0$ in terms of the local parameter b_0.

Define $S_j(\beta_X) = \tilde{X}_{j|Z}[\varepsilon + \{X - \tilde{X}_{j|Z}C_j/V_{j|Z} - \mathrm{cov}(X, Z^{\mathrm{T}}V_{ZZ}^{-1})Z\}^{\mathrm{T}}\beta_X]$, where $C_j = \mathrm{cov}(X, \tilde{X}_{j|Z})$. It is easy to show that $\theta_j^Z = C_j^{\mathrm{T}}\beta_{X,0}/V_{j|Z}$ and $S_j(\beta_{X,0}) = \tilde{X}_{j|Z}(Y - \alpha_0 - Z^{\mathrm{T}}\theta_Z^0 - \tilde{X}_{j|Z}\theta_j^Z)$, where $\theta_Z^0 = V_{ZZ}^{-1}\mathrm{cov}(Z, Y)$. Hence $S_j(\beta_{X,0})$ is proportional to the score function of $E\{(Y - \alpha - Z^{\mathrm{T}}\theta_Z - \tilde{X}_{j|Z}\theta_j)^2\}$ with respect to θ_j, evaluated at $(\alpha_0, \theta_Z^{0\mathrm{T}}, \theta_j^Z)^{\mathrm{T}}$. Define $W(\beta_{X,0}) = \{W_j(\beta_{X,0})\}_{j=1}^d$ as a normal random vector with mean zero and covariance matrix $\mathrm{cov}\{S(\beta_{X,0})\}$, where $S(\beta_X) = \{S_j(\beta_X)\}_{j=1}^d$. We make the following assumptions:

Assumption 1. (i) $E(Z) = 0$ and $E(X) = 0$; (ii) $E\{\|(Z^{\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}\|^4\}$ is bounded; (iii) the conditioning design matrix $(Z_1, \dots, Z_n)^{\mathrm{T}}$ is of full rank; (iv) for each $X_j \in X$ $(j = 1, \dots, d)$, X_j cannot be linearly represented by Z or, equivalently, $V_{j|Z} = V_{jj} - V_{Zj}^{\mathrm{T}}V_{ZZ}^{-1}V_{Zj} > 0$; (v) for $X_j, X_k \in X$, $|\mathrm{corr}(X_j, X_k)| < 1$ for $j \neq k$; (vi) p and q are fixed positive integers, and $q < n$;

Assumption 2. in model (1), ε has zero mean and finite variance, and is uncorrelated with $(Z^{\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}$.

Assumption 1(i) is imposed for simplicity, and can be satisfied by centring the covariates prior to data analysis. Assumption 1(ii) requires the tails of the predictor vectors to be well behaved, and (iii) requires a nonsingular conditioning design matrix. Assumption 1(iv) and (v) are needed for model identifiability, and neither requires X to be of full rank; (iv) rules out the case where the remaining predictors are linearly dependent on Z, while (v) excludes the case where two components in X are linearly dependent and is imposed to ensure the uniqueness of $K(b_0)$ in Theorem 1. Assumption 1(vi) specifies a fixed dimension. However, the proposed method can be used when $p > n$ as long as $q < n$, though no theory is provided in this paper concerning the asymptotic properties of the resulting procedure. Assumption 2 only requires ε and the covariates to be uncorrelated, and can accommodate some heteroscedasticity in the error.

THEOREM 1. *Suppose that Assumptions* 1 *and* 2 *hold, that* $k_0 = \bar{k}(\beta_{X,0})$ *is unique when* $\beta_{X,0} \neq 0$, *and that* $\bar{k}(b_0)$ *is unique when* $\beta_{X,0} = 0$ *and* $b_0 \neq 0$. *Under the local model* (3),

$(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$ *converges in distribution to*

$$
\begin{cases}
W_{k_0}(\beta_{X,0})/(V_{k_0|Z}^{1/2}\sigma_{k_0}), & \beta_{X,0} \neq 0, \\[2mm]
\dfrac{W_{K(b_0)}(0)}{V_{K(b_0)|Z}^{1/2}\sigma_{K(b_0)}} + \dfrac{V_{K(b_0)|Z}^{1/2}}{\sigma_{K(b_0)}}\left\{\dfrac{C_{K(b_0)}}{V_{K(b_0)|Z}} - \dfrac{C_{\bar{k}(b_0)}}{V_{\bar{k}(b_0)|Z}}\right\}^{\mathrm{T}} b_0, & \beta_{X,0} = 0,
\end{cases}
$$

where $\sigma_j^2 = E\{(Y - \alpha_0 - Z^{\mathrm{T}}\theta_Z^j - X_j\theta_j^Z)^2\}$ *and* $K(b_0) = \arg\max_{j=1,\ldots,d}\{W_j(0) + C_j^{\mathrm{T}}b_0\}^2/V_{j|Z}$.

Theorem 1 indicates that the limiting distribution takes a different form in each of two cases and is discontinuous at $\beta_{X,0} = 0$. Under the null hypothesis $\beta_{X,0} = b_0 = 0$, $\sigma_{K(0)} = \sigma_\varepsilon$ and $T_n = (n\hat{V}_{n|Z})^{1/2}\hat{\theta}_n^Z/\hat{\sigma}_n \to W_{K(0)}(0)/\{V_{K(0)|Z}^{1/2}\sigma_\varepsilon\}$ in distribution. Let $Q_{\gamma,0}$ be the upper γth quantile of $W_{K(0)}(0)$ or, equivalently, $\varepsilon\tilde{X}_{K(0)|Z}$; then $\mathrm{pr}[|T_n| > Q_{\gamma/2,0}/\{V_{K(0)|Z}^{1/2}\sigma_\varepsilon\} \mid H_0] = \gamma$. Under the local model (3), assuming that $\bar{k}(b_0)$ is unique when $\beta_{X,0} = 0$ and $b_0 \neq 0$, we have

$$
T_n = (n\hat{V}_{n|Z})^{1/2}\hat{\theta}_n^Z/\hat{\sigma}_n \to \{W_{K(b_0)}(0) + C_{K(b_0)}^{\mathrm{T}}b_0\}/\{V_{K(b_0)|Z}^{1/2}\sigma_{K(b_0)}\}, \quad \text{as } n \to \infty,
$$

in distribution. The following corollary establishes the local power of the test based on T_n.

COROLLARY 1. *Suppose that Assumptions* 1 *and* 2 *hold, and* $\bar{k}(b_0)$ *is unique when* $\beta_{X,0} = 0$ *and* $b_0 \neq 0$. *Then under the local model* (3),

$$
\begin{aligned}
\mathrm{pr}(\text{reject } H_0 \mid H_a) &= \mathrm{pr}\big[|T_n| > Q_{\gamma/2,0}/\{V_{K(0)|Z}^{1/2}\sigma_\varepsilon\} \mid H_a\big] \\
&\to \mathrm{pr}\{\varepsilon\tilde{X}_{K(b_0)|Z} \leqslant -A - C_{K(b_0)}^{\mathrm{T}}b_0\} + \mathrm{pr}\{\varepsilon\tilde{X}_{K(b_0)|Z} \geqslant A - C_{K(b_0)}^{\mathrm{T}}b_0\},
\end{aligned}
$$

as $n \to \infty$, *where* $A = Q_{\gamma/2,0}V_{K(b_0)|Z}^{1/2}\sigma_{K(b_0)}/\{V_{K(0)|Z}^{1/2}\sigma_\varepsilon\}$. *Consequently,* $\mathrm{pr}(\text{reject } H_0 \mid H_a) \to 1$ *if* $C_{K(b_0)}^{\mathrm{T}}b_0 \to \infty$, *as* $n \to \infty$.

2·3. *Conditional adaptive resampling for test calibration*

In this subsection, we present a bootstrap method for calibrating the null distribution of T_n. Because of the nonregularity of the limiting distribution in Theorem 1 at $\beta_{X,0} = 0$, the conventional bootstrap (Efron & Tibshirani, 1993) is not consistent for estimating the distribution of $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$. We propose a conditional adaptive resampling approach that takes into account the asymptotic behaviour of T_n under the local model. The main idea is to decompose $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$ into two components, corresponding to the nonlocal case with $\beta_{X,0} \neq 0$ and the local case with $\beta_{X,0} = 0$. We separate the two cases by comparing $|T_n|$ with some positive threshold λ_n that satisfies $\lambda_n \to \infty$ and $\lambda_n = o(n^{1/2})$.

Let $\hat{X}_{j|Z} = X_j - \hat{V}_{Zj}^{\mathrm{T}}\hat{V}_{ZZ}^{-1}Z$, where $\hat{V}_{Zj}$ and $\hat{V}_{ZZ}$ are sample estimates of V_{Zj} and V_{ZZ}. It is easy to see that the sample covariance $\mathrm{cov}_n(\hat{X}_{j|Z}, Z)$ is equal to zero for $j = 1, \ldots, d$. Under the local model (3) with $\beta_{X,0} = 0$, and by Assumption 2, we have

$$
n^{1/2}\theta_n^Z = n^{1/2}\mathrm{cov}(\tilde{X}_{k_n|Z}, X^{\mathrm{T}}n^{-1/2}b_0 + \varepsilon)/V_{k_n|Z} = \mathrm{cov}(\tilde{X}_{k_n|Z}, X^{\mathrm{T}}b_0)/V_{k_n|Z}, \tag{4}
$$

$$
n^{1/2}\hat{\theta}_n^Z = n^{1/2}\mathrm{cov}_n(\hat{X}_{\hat{k}_n|Z}, X^{\mathrm{T}}n^{-1/2}b_0 + \varepsilon)/\hat{V}_{\hat{k}_n|Z} = \{\mathbb{Z}_{n,\hat{k}_n} + \mathrm{cov}_n(\hat{X}_{\hat{k}_n|Z}, X^{\mathrm{T}}b_0)\}/\hat{V}_{\hat{k}_n|Z}, \tag{5}
$$

where $\mathbb{Z}_{n,j} = n^{1/2}\mathrm{cov}_n(\hat{X}_{j|Z}, \varepsilon) = n^{-1/2}\sum_{i=1}^{n}\varepsilon_i(\hat{X}_{ij|Z} - n^{-1}\sum_{i=1}^{n}\hat{X}_{ij|Z})$. Therefore, $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$ equals

$$(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n I(|T_n| > \lambda_n \text{ or } \beta_{X,0} \neq 0)$$

$$+ \left[\frac{\hat{V}_{n|Z}^{1/2}}{\hat{\sigma}_n}\left\{\frac{\mathbb{Z}_{n,\hat{k}_n} + \mathrm{cov}_n(\hat{X}_{\hat{k}_n|Z}, X^{\mathrm{T}}b_0)}{\hat{V}_{\hat{k}_n|Z}} - \frac{\mathrm{cov}(\tilde{X}_{k_n|Z}, X^{\mathrm{T}}b_0)}{V_{k_n|Z}}\right\}\right]I(|T_n| \leqslant \lambda_n, \beta_{X,0} = 0). \quad (6)$$

When $\lambda_n \to \infty$ and $\lambda_n = o(n^{1/2})$, we can show that $\mathrm{pr}(|T_n| > \lambda_n) \to I(\beta_{X,0} \neq 0)$. Thus the first term in (6) can be consistently bootstrapped. We can express the second term in the brackets as $\mathbb{V}_n(b)$, a process indexed by $b \in \mathbb{R}^d$, and then bootstrap $\mathbb{V}_n$. Define $\mathbb{K}_n(b) = \arg\max_{j=1,\ldots,d}\{\mathbb{Z}_{n,j} + \mathrm{cov}_n(\hat{X}_{j|Z}, X^{\mathrm{T}}b)\}^2/\hat{V}_{j|Z}$. By (5) and the definition of $\mathbb{Z}_{n,j}$, we have $\mathbb{K}_n(b_0) = \arg\max_j(\hat{\theta}_j^Z)^2\hat{V}_{j|Z} = \hat{k}_n$ when $\beta_{X,0} = 0$. By Proposition 1, $k_n = \bar{k}(b_0)$. This, combined with (4), motivates us to take

$$\mathbb{V}_n(b) = \frac{\hat{V}_{\mathbb{K}_n(b)|Z}^{1/2}}{\hat{\sigma}_{\mathbb{K}_n(b)}}\left[\frac{\mathbb{Z}_{n,\mathbb{K}_n(b)} + \mathrm{cov}_n\{\hat{X}_{\mathbb{K}_n(b)|Z}, X^{\mathrm{T}}b\}}{\hat{V}_{\mathbb{K}_n(b)|Z}} - \frac{\mathrm{cov}\{\tilde{X}_{\bar{k}(b)|Z}, X^{\mathrm{T}}b\}}{V_{\bar{k}(b)|Z}}\right].$$

Therefore, to approximate the asymptotic distribution of $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$, it suffices to bootstrap $(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n I(|T_n| > \lambda_n \text{ or } \beta_{X,0} \neq 0) + \mathbb{V}_n(b_0)I(|T_n| \leqslant \lambda_n, \beta_{X,0} = 0)$.

Hereafter, we use a superscript $*$ in any statistic to denote its bootstrap version based on the bootstrap sample, obtained by resampling $\{(Y_i, Z_i, X_i), i = 1, \ldots, n\}$ with replacement. Define

$$R_{n1}^* = (n\hat{V}_{n|Z}^*)^{1/2}(\hat{\theta}_n^{Z*} - \hat{\theta}_n^Z)/\hat{\sigma}_n^* I(|T_n^*| > \lambda_n \text{ or } |T_n| > \lambda_n)$$

$$+ \mathbb{V}_n^*(b_0)I(|T_n^*| \leqslant \lambda_n, |T_n| \leqslant \lambda_n). \quad (7)$$

The following theorem shows that the distribution of the maximum-type test statistic can be bootstrapped consistently under the local model (3).

THEOREM 2. *Assume the conditions of Theorem 1 and that $\lambda_n = o(n^{1/2})$ and $\lambda_n \to \infty$ as $n \to \infty$. Then under the local model* (3), $R_{n1}^* \to (n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n$ *in distribution, conditional on the data.*

To test the significance of X conditional on Z, we only need to obtain the sampling distribution of T_n under the null hypothesis $\theta_n^Z = 0$; that is, we only need to apply the bootstrap procedure with $b_0 = 0$. For any given significance level γ, let $c_l(\lambda_n)$ and $c_u(\lambda_n)$ be the lower and upper $(\gamma/2)$th quantiles of R_{n1}^* with $b_0 = 0$. If T_n falls outside the interval $[c_l(\lambda_n), c_u(\lambda_n)]$, then we reject H_0 and conclude that there exists at least one predictor in X that has significant correlation with Y conditional on Z. Hereafter, we refer to the bootstrap procedure for testing the presence of a conditional effect of X on Y given Z as the conditional adaptive resampling test.

2·4. *The hybrid test procedure*

Maximum-type statistics are known to be better at capturing sparse and strong signals, but less powerful in regimes with dense and weak signals. To adapt the test to accommodate signals of unknown sparsity, we propose a hybrid test statistic $\tilde{T}_n(\omega) = \omega T_n^2 + (1 - \omega)S_n$, where ω

is a tuning parameter between 0 and 1, and $S_n = d^{-1} \sum_{j=1}^{d} T_{nj}^2$ is the average of the squared conditional marginal t-statistics $T_{nj} = (n\hat{V}_{j|Z})^{1/2}\hat{\theta}_j^Z/\hat{\sigma}_j$.

The sum-type statistic S_n is regular and its limiting distribution has no discontinuity, so we can approximate the critical value for $\tilde{T}_n(\omega)$ by using a modified bootstrap version $\tilde{T}_n^*(\omega) = \omega R_{n1}^{*2} + (1-\omega)R_{n2}^*$, where R_{n1}^* is defined in (7) with $b_0 = 0$ and $R_{n2}^* = d^{-1}\sum_{j=1}^{d}\{(n\hat{V}_{j|Z}^*)^{1/2}(\hat{\theta}_j^{Z*} - \hat{\theta}_j^Z)/\hat{\sigma}_j^*\}^2$ is the standard bootstrap version of S_n under H_0. Specifically, at significance level γ, the hybrid test rejects H_0 if $\tilde{T}_n(\omega) > \tilde{c}_\gamma(\omega, \lambda_n)$, where $\tilde{c}_\gamma(\omega, \lambda_n)$ is the upper γth quantile of $\tilde{T}_n^*(\omega)$. The validity of this modified bootstrap for the hybrid test statistic is established in Theorem 3.

THEOREM 3. *Suppose that the assumptions in Theorem 1 hold and that* $\lambda_n = o(n^{1/2})$ *and* $\lambda_n \to \infty$ *as* $n \to \infty$. *Then for a given* $\omega \in [0, 1]$, *under the local model* (3),

$$\tilde{T}_n^*(\omega) \to \omega\big\{(n\hat{V}_{n|Z})^{1/2}(\hat{\theta}_n^Z - \theta_n^Z)/\hat{\sigma}_n\big\}^2 + (1-\omega)d^{-1}\sum_{j=1}^{d}\big\{(n\hat{V}_{j|Z})^{1/2}(\hat{\theta}_j^Z - \theta_j^Z)/\hat{\sigma}_j\big\}^2$$

in distribution, conditional on the data.

The proposed hybrid test has connections with the method of Fan et al. (2015), which improves the power of the sum-type test against sparse signals through an added enhancement term that is negligible under the null but large under the alternative. Consequently, the test in Fan et al. (2015) may have inflated Type I error if the enhancement component is not tuned properly. In contrast, our proposed hybrid test can control the Type I error rate for any $\omega \in [0, 1]$, though it may lose power with a poor choice of ω. Ideally, a larger ω is preferred for sparse and large alternatives, and vice versa. We suggest a data-driven procedure for choosing ω. Let $\chi_j = T_{nj}^2$ and let $\chi_{(j)}$ be the jth order statistic ($j = 1, \ldots, d$). Define $\chi_n = \{\chi_{(d)} - \bar{\chi}_{-(d)} - \mu_\chi\}/\sigma_\chi$, where $\bar{\chi}_{-(d)} = (d-1)^{-1}\sum_{j=1}^{d-1}\chi_{(j)}$ and μ_χ and σ_χ are the expectation and standard deviation of $\chi_{(d)} - \bar{\chi}_{-(d)}$, which are approximated through Monte Carlo simulation of independent $\chi^2(1)$ random variables. A large χ_n value suggests that the maximum t-statistic is far away from the rest and therefore indicates large and sparse signals. We set the data-adaptive weight ω^* to be a continuous function of χ_n, where ω^* equals 0 if $\chi_n \leqslant 3$, equals 1 if $\chi_n \geqslant 5$, and is linear in χ_n if $3 < \chi_n < 5$; the test based on this weight exhibits competitive power for both sparse and dense alternatives.

2·5. *Some computational issues*

The adaptive resampling procedure depends on the tuning parameter λ_n. The conditions $\lambda_n = o(n^{1/2})$ and $\lambda_n \to \infty$ ensure that the pretest is consistent with asymptotically negligible Type I error, since $\lim_{n\to\infty}\text{pr}(|T_n| > \lambda_n \mid \theta_n^Z = 0) = 0$. In practice, the test is conservative if λ_n is chosen too large, and is liberal otherwise. When $\lambda_n = 0$, the test is equivalent to the centred percentile bootstrap, which estimates the critical values by the upper and lower quantiles of $(n\hat{V}_{n|Z}^*)^{1/2}(\hat{\theta}_n^{Z*} - \hat{\theta}_n^Z)/\hat{\sigma}_n^*$; see Efron & Tibshirani (1993). We show in § 3·1 that the centred percentile bootstrap often fails to control the Type I error rate. Our numerical investigation shows that the rule-of-thumb suggested by McKeague & Qian (2015), $\lambda_n = \max[(a\log n)^{1/2}, \Phi^{-1}\{1 - \gamma/(2d)\}]$ with $a \in [5, 25]$, is a good choice for large samples, where Φ is the distribution function of $N(0, 1)$ and γ is the significance level. In our implementation, we choose the constant a from a grid $a_1 < \cdots < a_m \in [0, 25]$ by adopting the following double bootstrap procedure. We describe the procedure only for the maximum-type statistic, as that for the hybrid statistic is similar.

Step 1. Generate a bootstrap sample $\{(Y_i^*, Z_i^*, X_i^*), i = 1, \ldots, n\}$ by resampling $\{(Y_i, Z_i, X_i), i = 1, \ldots, n\}$ with replacement, and obtain the bootstrap estimates $\hat{\theta}_n^{Z*}$, $\hat{V}_{n|Z}^*$ and $\hat{\sigma}_n^*$.

Step 2. Generate B_2 double bootstrap samples by resampling $\{(Y_i^*, Z_i^*, X_i^*), i = 1, \ldots, n\}$ with replacement. Let $\lambda_{n,k} = \max[(a_k \log n)^{1/2}, \Phi^{-1}\{1 - \gamma/(2d)\}]$ $(k = 1, \ldots, m)$, and write $R_{n1,k}^{**}$ for the double bootstrap statistic of R_{n1}^* with $b_0 = 0$ and $\lambda_n = \lambda_{n,k}$. Then denote the lower and upper $(\gamma/2)$th quantiles of $R_{n1,k}^{**}$ by $c_l^*(\lambda_{n,k})$ and $c_u^*(\lambda_{n,k})$.

Step 3. Repeat Steps 1 and 2 B_1 times, calculate the proportion of times that the test statistic $(n\hat{V}_{n|Z}^*)^{1/2}(\hat{\theta}_n^{Z*} - \hat{\theta}_n^Z)/\hat{\sigma}_n^*$ falls outside the interval $[c_l^*(\lambda_{n,k}), c_u^*(\lambda_{n,k})]$ for each $\lambda_{n,k}$, and choose the tuning parameter λ_n^* to be the one that gives the rejection proportion closest to γ.

The double bootstrap can be computationally costly. However, when all the a_k in the grid lead to the same rejection conclusion, we can avoid the double bootstrap by taking any constant from the grid to perform the test and thus reduce the computational burden.

When ε is independent of $(Z^T, X^T)^T$, we have $\sigma_{K(0)} = \sigma_\varepsilon$ and $W_{K(0)}(0) \sim N(0, \Sigma_W)$ under H_0, where $\Sigma_W = \sigma_\varepsilon^2 \Sigma_{X|Z}$ and $\Sigma_{X|Z} = \text{cov}(\tilde{X}_{1|Z}, \ldots, \tilde{X}_{d|Z})$. By Theorem 1 we can obtain the critical value for T_n by constructing $W_{K(0)}(0)/V_{K(0)|Z}^{1/2}$ with random draws from a normal distribution with mean zero and covariance $\hat{\Sigma}_{X|Z}$, an estimator of $\Sigma_{X|Z}$. This test calibration approach is computationally convenient, but our empirical studies show that it often leads to distorted Type I error in small samples. In addition, the method relies on a stronger independence assumption between ε and X. When ε and X are dependent, we would need to estimate moments of the form $E(\varepsilon^2 \tilde{X}_{j|Z} \tilde{X}_{k|Z})$ to simulate draws from the asymptotic null distribution of T_n, and this is not feasible without further distributional information.

2·6. *Forward regression via a sequential conditional adaptive resampling test*

The proposed conditional adaptive resampling test can be used sequentially as a stopping rule in forward regression to select important predictors. To account for multiple testing in the sequential procedure, we consider a two-stage forward regression based on the multiple test adjustment idea in Holm (1979). Our numerical studies show that this two-stage procedure controls the familywise error rate well for modest and large samples.

In the first stage, initial forward regression is carried out in the following steps.

Step 1. Test for the presence of significant predictors by applying the adaptive resampling test of McKeague & Qian (2015), which is a special case of the conditional adaptive resampling test with $Z = \emptyset$. Let p_1 be the associated p-value in this step. If $p_1 > \gamma$, we stop the procedure and declare that there is no significant predictor; otherwise we let $Z^{(1)} = \{X_{\hat{k}_1}\}$, where $\hat{k}_1$ is the index of the predictor that has the strongest marginal dependence with Y.

Step 2. Conditional on $Z^{(1)}$, perform the conditional adaptive resampling test with $X^{(1)} = X \backslash Z^{(1)}$. Let p_2 be the associated p-value. If $p_2 > \gamma$, we stop; otherwise we repeat the proposed test with updated covariate vectors $Z^{(2)} = \{X_{\hat{k}_1}, X_{\hat{k}_2}\}$ and $X^{(2)} = X \backslash Z^{(2)}$, where $X_{\hat{k}_2}$ is the index of the predictor that is most correlated with Y conditioning on $Z^{(1)}$.

Step 3. Repeat Step 2 until no more significant predictors are detected. Assume that the procedure stops after N steps, and denote the selected covariate set by $Z^{(N)} = \{X_{\hat{k}_1}, \ldots, X_{\hat{k}_N}\}$ and the associated p-values by $\{p_1, \ldots, p_N\}$, where $p_j \leqslant \gamma$ $(j = 1, \ldots, N)$.

In the second stage, multiple test adjustment is performed as follows. Suppose that $N \geqslant 1$. Define $\tilde{N} = 1$ if $p_1 > \gamma/N$, or $\tilde{N} = \max_{1 \leqslant j \leqslant N}\{j : p_l \leqslant \gamma/(N - l + 1), \, l = 1, \ldots, j\}$ otherwise; the finally selected covariate set is chosen as $Z^{(\tilde{N})} = \{X_{\hat{k}_1}, \ldots, X_{\hat{k}_{\tilde{N}}}\}$.

The above forward regression procedure differs from the stepwise procedure in McKeague & Qian (2015), which successively applies the adaptive resampling test method by performing marginal regression on the remaining predictors, treating residuals from the previous stage as the new response. As demonstrated in Barut & Wang (2015), the performance of the stepwise adaptive resampling test procedure deteriorates with higher correlation between covariates. In contrast, our sequential procedure re-estimates the coefficients of variables already included at each step, and thus is less susceptible to problems due to high correlation. Furthermore, we control the familywise error rate for the sequential tests. Our numerical studies in § 3 show that this procedure outperforms the stepwise procedure of McKeague & Qian (2015) in cases with highly correlated covariates.

3. SIMULATION STUDY

3·1. *Size and power*

We test the significance of a subset of predictors after conditioning on Z, a known subset of conditioning covariates, using the conditional adaptive resampling test. Because the distribution of T_n is symmetric, the test based on T_n is equivalent to T_n^2, i.e., $\tilde{T}_n(\omega)$ with $\omega = 1$. We compare: (i) the adaptive resampling test based on $\tilde{T}_n(\omega)$ with ω equal to 1, 0·5, ω^* and 0, where ω^* is the weight selected by the data-driven rule-of-thumb; (ii) the test based on direct estimation of the asymptotic critical value of the maximum-type statistic T_n, assuming independence of the noise and the covariates; (iii) the centred percentile bootstrap that constructs the critical value of T_n with standard bootstrap quantiles; (iv) the partial covariance-based test proposed by Lan et al. (2014); (v) the permutation test based on the maximal absolute partial correlation; and (vi) marginal t-tests with Bonferroni correction. The test of Lan et al. (2014) is based on the partial covariance of the response and the predictors after controlling for Z, and is applicable when $d > n$. The permutation test is based on the test statistic $\max_{j=1,\ldots,d} |\mathrm{corr}(Y, X_j \mid Z)|$, where the critical value is constructed by the upper γth sample quantile of $\max_{j=1,\ldots,d} |\mathrm{corr}(Y, X_j^* \mid Z)|$, with X_j^* being a row-permuted version of X_j. For the centred percentile bootstrap, permutation and conditional adaptive resampling tests, we use 500 bootstrap samples or permutations. For the proposed test based on $\tilde{T}_n(\omega)$ with $\omega > 0$, we select the tuning parameter λ_n by a double bootstrap of 100 samples.

We consider the standard regression model given by

$$Y_i = Z_i^{\mathrm{T}}\alpha_Z + X_i^{\mathrm{T}}\beta_X + \varepsilon_i \quad (i = 1, \ldots, n),$$

where $(Z_i^{\mathrm{T}}, X_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p \sim N(0, \Sigma)$, with Σ having an exchangeable structure with ones on the diagonal and ρ on the off-diagonals, $Z_i \in \mathbb{R}^5$, $\alpha_Z = (1, 0·8, 0·6, 0·4, 0·2)^{\mathrm{T}}$, and $\varepsilon_i \sim N(0, 1)$ is independent of Z_i and X_i. We consider different scenarios with $n \in \{100, 400\}$, $\rho \in \{0·5, 0·8\}$ and $p \in \{100, 400\}$. For each scenario, the simulation is repeated 1000 times. We let $\beta_X = 0$ for size assessment, and consider two types of β_X for power analysis: (i) sparse signals, $\beta_X = (\beta_1, 0^{\mathrm{T}})^{\mathrm{T}}$ where β_1 ranges from 0 to $10/n^{1/2}$; (ii) dense signals, $\beta_X = \kappa(\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ where $\beta_1, \ldots, \beta_d$ are simulated from a standard normal distribution and κ is selected so that the signal strength $\delta = \beta_X^{\mathrm{T}}\Sigma_{X|Z}\beta_X$ ranges from 0 to $20/n^{1/2}$, with $\Sigma_{X|Z}$ being the conditional covariance matrix of X given Z. The nominal level is set at $\gamma = 0·05$.

Table 1. *Simulation results under the null hypothesis: the reported values are average rejection rates* (%) *for different methods with a nominal level of* $\gamma = 0.05$

ρ	n	p	$\tilde{T}_n(1)$	$\tilde{T}_n(0.5)$	$\tilde{T}_n(\omega^*)$	$\tilde{T}_n(0)$	ASYM	PCBT	BON	PERM	CPB
0.5	100	100	6.1	6.1	4.9	3.5	8.0	4.5	7.3	9.4	72.5
		400	7.4	7.1	4.9	3.5	9.5	5.0	8.3	12.3	89.1
	400	100	4.3	4.3	4.8	4.5	5.0	4.5	4.2	7.4	73.9
		400	7.1	6.8	5.9	4.5	7.8	4.7	6.7	14.1	93.8
0.8	100	100	4.5	4.6	3.2	2.6	7.5	4.2	6.2	25.7	70.8
		400	6.7	6.0	4.2	3.0	8.7	4.8	7.8	44.6	87.2
	400	100	5.5	6.0	4.1	3.8	6.3	3.8	4.9	25.6	74.7
		400	5.5	6.3	4.4	3.9	6.4	4.7	5.5	46.4	92.3

$\tilde{T}_n(\omega)$, the proposed conditional adaptive resampling test with $\omega = 1, 0.5, \omega^*, 0$; ASYM, the test based on direct estimation of the asymptotic critical value of T_n; PCBT, the partial covariance-based test of Lan et al. (2014); BON, the t-test with Bonferroni correction; PERM, the permutation test; CPB, the standard centred percentile bootstrap.

The computation was done using R (R Development Core Team, 2018) on a server comparable to a MacBook Pro, with 2.5 GHz Intel Core i5 and 8 GB 1600 MHz DDR3. The computational complexity for the test based on $\tilde{T}_n(\omega)$ is mainly due to the bootstrap resampling of R^*_{n1} and R^*_{n2}. Thus, in terms of computing time, there are no differences between performing a test with a single ω or with multiple ones. Among 1000 replicates, the average computing times for the adaptive resampling test with 500 bootstrap repetitions and a given λ_n were 3.3, 7.0, 18.0 and 54.1 seconds for $(n, p) = (100, 100), (400, 100), (100, 400)$ and $(400, 400)$, respectively, and the corresponding standard errors were 0.09, 0.11, 0.17 and 0.33 seconds.

Table 1 summarizes the empirical sizes of all methods for testing the effect of X_i conditional on Z_i. Both the proposed resampling test and the test of Lan et al. (2014) control the sizes reasonably well, except in cases with large d and small n. The centred percentile bootstrap gives highly inflated Type I error, and the asymptotic-based test and the t-test with Bonferroni correction give inflated Type I error for the small sample size of $n = 100$. When X and Z are uncorrelated, the permutation test performs well in terms of both Type I error and power; see the Supplementary Material. However, the permutation test gives inflated Type I error when X and Z are correlated, and the inflation is more severe for larger ρ.

In Fig. 1 we plot the power curves for $\tilde{T}_n(\omega)$ with ω equal to 1, 0.5, the data-driven weight ω^* and 0 as well as for the test of Lan et al. (2014), in the case where $\rho = 0.5$, $p = 100$ and $n = 400$; the results for $n = 100$ can be found in the Supplementary Material, and the results for $\rho = 0.8$ are similar and hence omitted. We omit other methods from the power analysis because of their inflated Type I errors. When the signal is sparse and large, the hybrid test using the selected weight ω^* is competitive with those using $\omega = 1$ and 0.5, and it appears to be more powerful than the test with $\omega = 0$ and the test of Lan et al. (2014). When the signals are dense and weak, the hybrid test with selected weight ω^* results in powers almost identical to those of the sum-type test with $\omega = 0$ and the test of Lan et al. (2014).

As suggested by a referee, we also tried the case with larger noise, where $\varepsilon_i \sim N(0, 4)$. The empirical sizes are almost the same as when $\varepsilon_i \sim N(0, 1)$, but the signals must be doubled to achieve the same power.

3.2. *Forward selection*

We compare the proposed sequential forward regression procedure with (i) the stepwise method in McKeague & Qian (2015), (ii) the crossvalidation-based method, and (iii) the
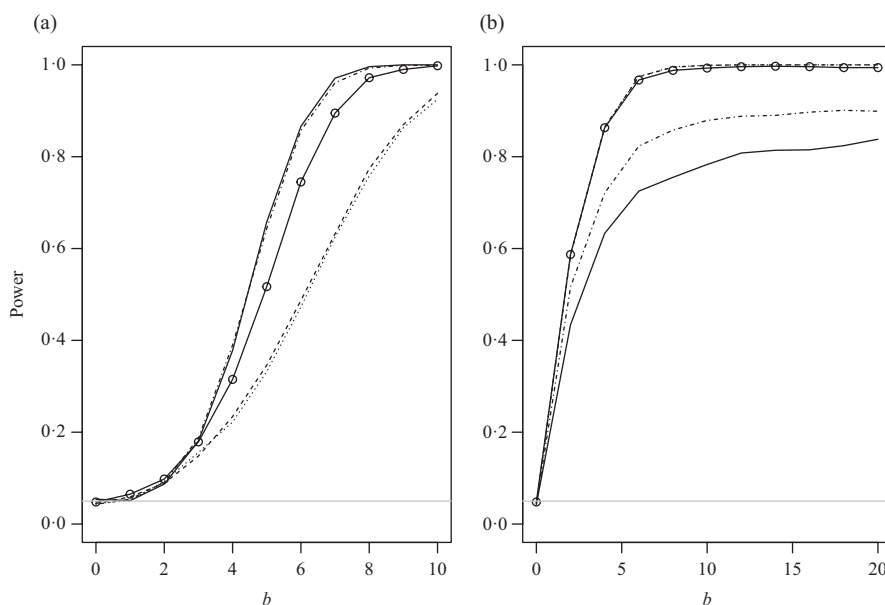
Fig. 1. Power curves of different methods under (a) sparse and (b) dense alternatives: conditional adaptive resampling test based on $\omega = 1$ (solid), 0·5 (dash-dot), ω^* (solid with circles) or 0 (dashed) and the test of Lan et al. (2014) (dotted), for the case with $\rho = 0.5$, $p = 100$ and $n = 400$. In each panel the horizontal axis measures the signal strength: $\beta_1 = b/n^{1/2}$ for the sparse alternative and $\delta = b/n^{1/2}$ for the dense alternative; the grey horizontal line indicates the nominal level of 0·05.

Benjamini–Hochberg method that controls the false discovery rate at 0·05. In the crossvalidation-based procedure, we randomly perform a five-fold split of the data, calculate the sum of the squared crossvalidation errors, and stop the procedure if the sum of the errors is not further reduced. The Benjamini–Hochberg method is based on the adjustment in Benjamini & Hochberg (1995) to p-values obtained from the marginal t-test. The data are generated from $Y = X^{\mathrm{T}}\beta + \varepsilon$ with $\varepsilon \sim N(0,1)$, and we consider (I) $\beta = (1,1,1,0.8,0.8,0_{p-5}^{\mathrm{T}})^{\mathrm{T}}$ and (II) $\beta = (1,1,1,0.8,0.8,0.3,0.3,0_{p-7}^{\mathrm{T}})^{\mathrm{T}}$. We report only the results for the hybrid test with $\omega = 1$, as the test based on the data-driven weight ω^* performs similarly.

Table 2 summarizes the selection results for case (I) with $n = 400$, including: the average number of true positives; the average number of false positives; the proportion of times that the exact true model was selected; the proportion of false negatives, i.e., the proportion of cases missing at least one relevant predictor; and the familywise error rate, i.e., the proportion of times that at least one irrelevant predictor was selected. The selection results shown in Table 2 suggest that when covariates are independent, i.e., $\rho = 0$, the stepwise method of McKeague & Qian (2015) performs similarly to our sequential method, with both giving accurate model selection and controlling the familywise error rate well. However, the performance of the stepwise method deteriorates with higher correlation between covariates, missing on average about one relevant variable, and is even worse when $n = 100$. Our sequential procedure is less susceptible to high correlation; it gives accurate variable selection in almost all scenarios considered, and controls the familywise error rate well for both $\rho = 0$ and $\rho = 0.5$, while the stepwise method of McKeague & Qian (2015) gives a highly inflated familywise error rate for $\rho = 0.5$. The crossvalidation-based method always selects a larger model, and its familywise error rate is out of control. When $\rho = 0$, the Benjamini–Hochberg method controls the false discovery rates at 0·08 and 0·06 for $n = 100$

Table 2. *Simulation results for forward selection: comparison of
the proposed sequential forward regression procedure with other
approaches in case* (I)*, the sparse and large signal setting, for
sample size n = 400*

ρ	p	Method	NoTP	NoFP	OracleP	FN	FWER
0·0	100	ART	5·00	0·05	0·95	0·00	0·05
		CART	5·00	0·03	0·97	0·00	0·03
		CV	5·00	6·23	0·03	0·00	0·97
		BH	5·00	0·31	0·73	0·00	0·27
	400	ART	5·00	0·06	0·94	0·00	0·06
		CART	5·00	0·06	0·94	0·00	0·06
		CV	5·00	14·71	0·00	0·00	1·00
		BH	5·00	0·33	0·73	0·00	0·27
0·5	100	ART	4·01	1·46	0·00	0·62	0·64
		CART	5·00	0·06	0·94	0·00	0·06
		CV	5·00	6·05	0·04	0·00	0·96
		BH	5·00	95·00	0·00	0·00	1·00
	400	ART	3·91	1·45	0·00	0·66	0·60
		CART	5·00	0·06	0·94	0·00	0·06
		CV	5·00	14·64	0·01	0·00	0·99
		BH	5·00	395·00	0·00	0·00	1·00

NoTP, the average number of relevant predictors selected across simulation;
NoFP, the average number of irrelevant predictors selected across simulation;
OracleP, the proportion of replicates in which the true model is selected; FN,
the proportion of replicates in which at least one relevant predictor is missed;
FWER, familywise error rate, the proportion of replicates in which at least one
irrelevant predictor is selected; ART, the stepwise selection method of McKeague
& Qian (2015); CART, the proposed forward selection via sequential conditional
adaptive resampling test with multiple test adjustment; CV, the forward selection
based on five-fold crossvalidation; BH, the Benjamini–Hochberg method.

and 400, respectively, but the familywise error rate is too large. For $\rho = 0.5$, the Benjamini–
Hochberg method always selects all of the predictors, possibly because of the exchangeable
correlation among the predictors. Further results can be found in the Supplementary Material.

4. Real-data analysis

Studies of expression quantitative trait locus are often used to link variations of genotype to
deviations in gene expression levels. The dataset we analyse, available at http://www.braineac.org/
(Trabzuni et al., 2011), contains gene expression levels and genotype sequences from 134 individ-
uals with no known neurodegenerative disorders. The gene expressions were collected from up
to 12 different brain regions. Comparison of gene expressions for a specific gene from different
brain regions can be used to understand how gene expression is regulated in different areas of the
human brain. Previous analyses by the data providers showed that gene expression levels vary
across different brain regions (Ramasamy et al., 2014).

We study the expression of the ATP5G2 gene at transcript level t3456313 Chromosome 12.
The gene ATP5G2 regulates adenosine triphosphate synthase, which is an enzyme involved in
cellular energy transfer. ATP5G2 is also hypothesized to be a tumour suppressor for renal cell

Table 3. *Estimated coefficients and their* 95% *confidence intervals (in brackets) of the chosen single nucleotide polymorphisms in the expression quantitative trait locus analysis for ATP5G2 in different brain regions; the confidence intervals are calculated using conditional bootstrap*

Region	rs11170631	rs11170639	rs12818213	rs112183557	rs188609099
CC	$-0.39\ [-0.44, -0.33]$				
ME	$-0.28\ [-0.35, -0.17]$				$-0.29\ [-0.86, -0.19]$
OC	$-0.38\ [-0.44, -0.33]$				
TC	$-0.36\ [-0.43, -0.27]$				
FC		$-0.32\ [-0.37, -0.26]$			
HI		$-0.29\ [-0.35, -0.19]$			
SN		$-0.31\ [-0.38, -0.19]$			
TH			$-0.28\ [-0.33, -0.16]$		
PU				$-0.33\ [-0.39, -0.22]$	

CC, cerebellar cortex; ME, medulla; OC, occipital cortex; TC, temporal cortex; FC, frontal cortex; HI, hippocampus; SN, substantia nigra; TH, thalamus; PU, putamen.

carcinoma (Morris et al., 2011). In the original study by Ramasamy et al. (2014), the authors performed a single-variable study and showed that expression of t3456313 in all brain regions can be consistently estimated by the single nucleotide polymorphism rs12818213. In this study we use a multivariate estimator and check the consistency of the above statement when other single nucleotide polymorphisms are considered.

The single nucleotide polymorphisms under investigation are located one megabase upstream and downstream of the transcription start site of ATP5G2. Variables and individuals with missing data are removed, which reduces the size of the dataset to 3853 variables and 124 samples. Following the suggestion of the data providers, we treat the marker values as numerical; that is, we do not treat allele type 1 or 2 separately, and instead code all of the covariates as numbers ranging from 0 to 2. We fit a different regression for each brain region, where the gene expression levels of ATP5G2 in that region are treated as the response, using the same 3853 covariates, including rs12818213, in each regression. Such an analysis is useful for comparing the variation of gene dynamics in different parts of the brain.

We use the conditional adaptive resampling test with forward regression to select among the 3853 variables that are significant at the 5% nominal level. We display the results for $\omega = 1$, but the data-driven choice of ω gave similar outputs and only one extra covariate was recruited for one of the regions, namely the thalamus. The fitted coefficients of chosen single nucleotide polymorphisms are presented in Table 3, along with their 95% confidence intervals. The confidence intervals are estimated through conditional bootstrap, in which we vary the b term in $\mathbb{V}_n(b)$ and look for the widest bootstrap quantiles of $(n\hat{V}^*_{n|Z})^{-1/2}\hat{\sigma}^*_n R^*_{n1}$ over all $b_0 \in \mathbb{R}^d$, where R^*_{n1} is given in (7). This approach produces robust confidence intervals (McKeague & Qian, 2015).

Table 3 shows that the single nucleotide polymorphisms recruited by the proposed test change with respect to brain region. This difference is in line with the original study of Ramasamy et al. (2014); however, in their analysis of the ATP5G2 gene, the authors concluded that rs12818213 had a consistent signal in all brain regions. We reach a different conclusion: the proposed test finds that the variations in the ATP5G2 gene can be explained by rs12818213 only in the thalamus. In other regions, completely separate single nucleotide polymorphisms appear to be significant predictors.

We plot the correlation matrix for the displayed single nucleotide polymorphisms in Fig. 2. The correlation matrix shows that the first to the fourth single nucleotide polymorphisms in Table 3 are heavily correlated. In fact, rs11170631, rs11170639, rs12818213 and rs112183557, which

	rs11170631	rs11170639	rs12818213	rs112183557	rs188609099
rs11170631	1	0·98	0·99	0·95	0·13
rs11170639	0·98	1	0·99	0·97	0·14
rs12818213	0·99	0·99	1	0·96	0·14
rs112183557	0·95	0·97	0·96	1	0·14
rs188609099	0·13	0·14	0·14	0·14	1

Fig. 2. Correlation matrix of the chosen single nucleotide polymorphisms.

are related to all of the regions, have intercorrelations of higher than 95%. These findings are complementary to the analysis of ATP5G2 in the original paper. One surprising result from the correlation matrix is the weak correlation of rs188609099 with the others. Since rs188609099 was found to be active only in the medulla region, this indicates that different mechanisms are at work for expression of the ATP5G2 gene in the medulla and other regions of the brain.

5. Discussion

If one is interested in inference regarding the effect of some pre-chosen covariate such as the treatment, the uncertainty involved in the selection of nuisance confounding variables may have potentially devastating effects and must be accounted for appropriately. We refer to Leeb & Pötscher (2005, 2008) for further discussion of post-selection inference issues, and to Meinshausen et al. (2009), Chatterjee & Lahiri (2011), Berk et al. (2013), Lockhart et al. (2014), Zhang & Zhang (2014), Javanmard & Montanari (2014), Belloni et al. (2015), Lee et al. (2016) and Candès et al. (2016) for developments relating to post-selection inference methods.

Supplementary material

Supplementary material available at *Biometrika* online contains additional simulation results and proofs of the theoretical results.

REFERENCES

BARUT, E., FAN, J. & VERHASSELT, A. (2016). Conditional sure independence screening. *J. Am. Statist. Assoc.* **111**, 1266–77.

BARUT, E. & WANG, H. (2015). Comments on "An adaptive resampling test for detecting the presence of significant predictors" by I. McKeague and M. Qian. *J. Am. Statist. Assoc.* **110**, 1442–5.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

BELLONI, A., CHERNOZHUKOV, V. & KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other *Z*-estimation problems. *Biometrika* **102**, 77–94.

BERK, R., BROWN, L., BUJA, A., ZHANG, K. & ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–37.

CANDÈS, E., FAN, Y., JANSON, L. & LV, J. (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv:* 1610.02351.

CHATTERJEE, A. & LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Am. Statist. Assoc.* **106**, 608–25.

CHENG, X. (2015). Robust inference in nonlinear models with mixed identification strength. *J. Economet.* **189**, 207–28.

EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall/CRC.

FAN, J., LIAO, Y. & YAO, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83**, 1497–541.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

JAVANMARD, A. & MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–909.

LAN, W., WANG, H. & TSAI, C. (2014). Testing covariates in high-dimensional regression. *Ann. Inst. Statist. Math.* **66**, 279–301.

LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.

LEEB, H. & PÖTSCHER, B. M. (2005). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Economet.* **142**, 201–11.

LEEB, H. & PÖTSCHER, B. M. (2008). Model selection and inference: Facts and fiction. *Economet. Theory* **21**, 21–59.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. & TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–68.

MCKEAGUE, I. & QIAN, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *J. Am. Statist. Assoc.* **110**, 1422–33.

MEINSHAUSEN, N., MEIER, L. & BUHLMANN, P. (2009). *P*-values for high-dimensional regression. *J. Am. Statist. Assoc.* **104**, 1671–81.

MORRIS, M. R., RICKETTS, C., GENTLE, D., MCRONALD, F., CARLI, N., KHALILI, H., BROWN, M., KISHIDA, T., YAO, M., BANKS, R. E. et al. (2011). Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. *Oncogene* **30**, 1390–401.

R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

RAMASAMY, A., TRABZUNI, D., GUELFI, S., VARGHESE, V., SMITH, C., WALKER, R., DE, T., COIN, L., DE SILVA, R., COOKSON, M. R. et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neurosci.* **17**, 1418–28.

TRABZUNI, D., RYTEN, M., WALKER, R., SMITH, C., IMRAN, S., RAMASAMY, A., WEALE, M. E. & HARDY, J. (2011). Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J. Neurochem.* **119**, 275–82.

ZHANG, C. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.