ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)

Goodness-of-fit tests for high dimensional linear models

Rajen D. Shah

University of Cambridge, UK

and Peter Bühlmann

Eidgenössiche Technische Hochschule Zürich, Switzerland

[Received November 2015. Final revision March 2017]

Summary. We propose a framework for constructing goodness-of-fit tests in both low and high dimensional linear models. We advocate applying regression methods to the scaled residuals following either an ordinary least squares or lasso fit to the data, and using some proxy for prediction error as the final test statistic. We call this family residual prediction tests. We show that simulation can be used to obtain the critical values for such tests in the low dimensional setting and demonstrate using both theoretical results and extensive numerical studies that some form of the parametric bootstrap can do the same when the high dimensional linear model is under consideration. We show that residual prediction tests can be used to test for significance of groups or individual variables as special cases, and here they compare favourably with state of the art methods, but we also argue that they can be designed to test for as diverse model misspecifications as heteroscedasticity and non-linearity.

Keywords: Bootstrap; Diagnostics; Goodness of fit; High dimensional models; Lasso

1. Introduction

High dimensional data, where the number of variables may greatly exceed the number of observations, have become increasingly more prevalent across a variety of disciplines. Although such data pose many challenges to statisticians, we now have a variety of methods for fitting models to high dimensional data, many of which are based on the lasso (Tibshirani, 1996); see Bühlmann and van de Geer (2011) for a review of some of the developments.

More recently, huge strides have been made in quantifying uncertainty about parameter estimates. For the important special case of the high dimensional linear model, frequentist p-values for individual parameters or groups of parameters can now be obtained through an array of techniques (Wasserman and Roeder, 2009; Meinshausen *et al*., 2009; Bühlmann, 2013; Zhang and Zhang, 2014; Lockhart *et al*., 2014; van de Geer *et al*., 2014; Javanmard and Montanari, 2014; Meinshausen, 2015; Ning and Liu, 2017; Voorman *et al*., 2014; Zhou, 2015)—see Dezeure *et al*. (2015) for an overview of some of these methods. Subsampling techniques such as stability selection (Meinshausen and Bühlmann, 2010) and its variant complementary pairs stability selection (Shah and Samworth, 2013) can also be used to select important variables while preserving error control in a wider variety of settings.

Address for correspondence: Rajen D. Shah, Statistical Laboratory, Centre for Mathematical Sciences; Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: r.shah@statslab.cam.ac.uk

 1369–7412/17/80000

Despite these advances, something still lacking from the practitioner's toolbox is a corresponding set of diagnostic checks to help to assess the validity of, for example, the high dimensional linear model. For instance, there are no well-established methods for detecting heteroscedasticity in high dimensional linear models, or whether a non-linear model may be more appropriate.

In this paper, we introduce an approach for creating diagnostic measures or goodness-of-fit tests that are sensitive to different sorts of departures from the 'standard' high dimensional linear model. As the measures are derived from examining the residuals following for example a lasso fit to the data, we use the name residual prediction (RP) tests. To the best of our knowledge, it is the first methodology for deriving confirmatory statistical conclusions, in terms of p-values, to test for a broad range of deviations from a high dimensional linear model. In Section 1.2 we give a brief overview of the idea, but first we discuss what we mean by goodness of fit in a high dimensional setting.

1.1. Model misspecification in high dimensional linear models

Consider the Gaussian linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a response vector, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the fixed design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown vector of coefficients, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$ is a vector of uncorrelated Gaussian errors and $\sigma^2 > 0$ is the variance of the noise. In the low dimensional situation where $p < n$, we may speak of model (1) being misspecified such that $\mathbb{E}(\mathbf{y}) \neq \mathbf{X}\boldsymbol{\beta}$. When $\mathbf{X}$ has full row rank, however, any vector in $\mathbb{R}^n$ can be expressed as $\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$, leaving in general no room for non-linear alternatives. When restricting to sparse linear models specified by model (1), the situation is different though and misspecification can happen (Bühlmann and van de Geer, 2015); we shall take a sparse Gaussian linear model as our null hypothesis (see also theorems 2 and 3 in Section 4). We discuss an approach to handle a relaxation of the Gaussian error assumption in section B of the on-line supplementary material.

When there is no good sparse approximation to $\mathbf{X}\boldsymbol{\beta}$, a high dimensional linear model may not be an appropriate model for the data-generating process: a sparse non-linear model might be more interpretable and may generalize better, for example. Moreover, the lasso and other sparse estimation procedures may have poor performance, undermining the various high dimensional inference methods that were mentioned above that make use of them. Our proposed RP tests investigate whether the lasso is a good estimator of the signal.

1.2. Overview of residual prediction tests and main contributions

Let $\hat{\mathbf{R}}$ be the residuals following a lasso fit to $\mathbf{X}$. If $\mathbf{X}\boldsymbol{\beta}$ is such that it can be well estimated by the lasso, then the residuals should contain very little signal and instead should behave roughly like the noise term $\sigma\boldsymbol{\varepsilon}$. In contrast, if the signal is such that the lasso performs poorly and instead a non-linear model were more appropriate, for example, some of the (non-linear) signal should be present in the residuals, as the lasso would be incapable of fitting to it.

Now if we use a regression procedure that is well suited to predicting the non-linear signal (an example may be random forests (Breiman, 2001)), applying this to the residuals and computing the resulting mean residual sum of squares (RSS) or any other proxy for prediction error will give us a test statistic that under the null hypothesis of a sparse linear model we expect to be relatively large, and under the alternative we expect to be relatively small. Different regression procedures applied to the residuals can be used to test for different sorts of departures from the Gaussian linear model. Thus RP tests consist of three components:

(a) an initial procedure that regresses $\mathbf{y}$ on $\mathbf{X}$ to give a set of residuals (this is typically the lasso if $p > n$ or could be ordinary least squares (OLS) if $\mathbf{X}$ is low dimensional);
(b) an RP method that is suited to predicting the particular signal expected in the residuals under the alternative(s) under consideration;
(c) some measure of the predictive capability of the RP method. Typically this would be the RSS, but in certain situations a cross-validated estimate of prediction error may be more appropriate, for example.

We shall refer to the composition of an RP method and an estimator of prediction error as an RP function. This must be a (measurable) function f of the residuals and all available predictors, p_{all} of them in total, to the reals $f : \mathbb{R}^n \times \mathbb{R}^{n \times p_{\text{all}}} \to \mathbb{R}$. For example, if, rather than testing for non-linearity, we wanted to ascertain whether any additional variables were significant after accounting for those in $\mathbf{X}$, we could consider the mean RSS after regressing the residuals on a matrix of predictors containing both $\mathbf{X}$ and the additional variables, using the lasso. If the residuals could be predicted better than we would expect under the null hypothesis with a model as in equation (1), this would provide evidence against the null.

Clearly to use RP tests to perform formal hypothesis tests, we need knowledge of the distribution of the test statistic under the null, to calculate p-values. Closed form expressions are difficult if not impossible to come by, particularly when the RP method is something as intractable as random forests.

In this work, we show that, under certain conditions, the parametric bootstrap (Efron and Tibshirani, 1994) can be used, with some modifications, to calibrate *any* RP test. Thus the RP method can be as exotic as needed to detect the particular departure from the null hypothesis that is of interest, and there are no restrictions requiring it to be a smooth function of the data, for example. To obtain such a general result, the conditions are necessarily strong; nevertheless, we demonstrate empirically that, for a variety of interesting RP tests, bootstrap calibration tends to be quite accurate even when the conditions cannot be expected to be met. As well as providing a way of calibrating RP tests, we also introduce a framework for combining several RP tests to have power against a diverse set of alternatives.

Although formally the null hypothesis that is tested by our approach is that of the sparse Gaussian linear model (1), an RP test geared towards non-linearity is unlikely to reject purely because of non-Gaussianity of the errors, and so the effective null hypothesis typically allows for more general error distributions. By using the non-parametric bootstrap rather than the parametric bootstrap, we can allow for non-Gaussian error distributions more explicitly. We discuss this approach in section B of the on-line supplementary material, where we see that the type I error is very well controlled even in settings with t_3 and exponential errors.

Some work related to ours here is that of Chatterjee and Lahiri (2010, 2011), Camponovo (2015) and Zhou (2014) who studied the use of the bootstrap with the (adaptive) lasso for constructing confidence sets for the regression coefficients. Work that is more closely aligned to our aim of creating diagnostic measures for high dimensional models is that of Nan and Yang (2014), though their approach is specifically geared towards variable selection and they did not provide theoretical guarantees within a hypothesis testing framework like we do.

1.3. Organization of the paper

Simulating the residuals under the null hypothesis is particularly simple when, rather than using the lasso residuals, OLS residuals are used. We study this simple situation in Section 2 not only to help to motivate our approach in the high dimensional setting, but also to present what we believe is a useful method in its own right. In Section 3 we explain how several RP tests can

be aggregated into a single test that combines the powers of each of the tests. In Section 4 we describe the use of RP tests in the high dimensional setting and prove the validity of a calibration procedure based on the parametric bootstrap. We give several applications of RP tests in Section 5 along with the results of extensive numerical experiments, and we conclude with a discussion in Section 6. The on-line supplementary material contains further discussion of the power of RP tests, a proposal for how to test null hypotheses of the form (1) allowing for more general error distributions, additional numerical results, a short comment concerning the interpretation of p-values and all the proofs. The R (R Development Core Team, 2005) package `RPtests` provides an implementation of the methodology.

2. Ordinary least squares residual predication tests

A simple but nevertheless important version of RP tests uses residuals from OLS in the first stage. For this, we require $p < n$ in the set-up of model (1). Let $\mathbf{P}$ denote the orthogonal projection onto the column space of $\mathbf{X}$. Then, under the null hypothesis that the model (1) is correct, the scaled residuals $\hat{\mathbf{R}}$ are

$$\hat{\mathbf{R}} := \frac{(\mathbf{I}-\mathbf{P})\mathbf{y}}{\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2} = \frac{(\mathbf{I}-\mathbf{P})\varepsilon}{\|(\mathbf{I}-\mathbf{P})\varepsilon\|_2},$$

and so their distribution does not depend on any unknown parameters: they form an ancillary statistic. Note that the scaling of the residuals eliminates the dependence on σ^2. It is thus simple to simulate from the distribution of any function of the scaled residuals, and this allows critical values to be calculated for tests using any RP method.

We note that, by using OLS applied to a larger set of variables as the RP method, and the RSS from the resulting fit as the estimate of prediction error, the overall test is equivalent to a partial F-test for the significance of the additional group of variables. To see this write $\mathbf{Z} \in \mathbb{R}^{n \times q}$ for an additional group of variables. Let $\mathbf{P}_{\mathrm{all}}$ be the orthogonal projection onto all available predictors, i.e. projection onto $\mathbf{X}_{\mathrm{all}} = (\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n \times p_{\mathrm{all}}}$, where $p_{\mathrm{all}} = p + q$. When the RP method is OLS regression of the scaled residuals onto $\mathbf{X}_{\mathrm{all}}$, the resulting RSS is

$$\|(\mathbf{I}-\mathbf{P}_{\mathrm{all}})\hat{\mathbf{R}}\|_2^2 = \frac{\|(\mathbf{I}-\mathbf{P}_{\mathrm{all}})(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2}{\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2} = \frac{\|(\mathbf{I}-\mathbf{P}_{\mathrm{all}})\mathbf{y}\|_2^2}{\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2},$$

since $(\mathbf{I}-\mathbf{P}_{\mathrm{all}})\mathbf{P} = \mathbf{0}$. We reject for small values of the quantity above, or equivalently large values of

$$\frac{\|(\mathbf{P}_{\mathrm{all}}-\mathbf{P})\mathbf{y}\|_2^2}{\|(\mathbf{I}-\mathbf{P}_{\mathrm{all}})\mathbf{y}\|_2^2} \frac{n-p_{\mathrm{all}}}{p_{\mathrm{all}}-p},$$

which is precisely the F-statistic for testing the hypothesis in question.

An alternative way to arrive at the F-test is first to residualize $\mathbf{Z}$ with respect to $\mathbf{X}$ and to define new variables $\tilde{\mathbf{Z}} = (\mathbf{I}-\mathbf{P})\mathbf{Z}$. Write $\tilde{\mathbf{P}}$ for the orthogonal projection onto $\tilde{\mathbf{Z}}$. Now, if our RP method is OLS regression of the scaled residuals onto $\tilde{\mathbf{Z}}$, we may write our RSS as

$$\|(\mathbf{I}-\tilde{\mathbf{P}})\hat{\mathbf{R}}\|_2^2 = \frac{\|(\mathbf{I}-\tilde{\mathbf{P}})(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2}{\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2} = \frac{\|\{\mathbf{I}-(\mathbf{P}+\tilde{\mathbf{P}})\}\mathbf{y}\|_2^2}{\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2},$$

the final equality following from the fact that the column spaces of $\mathbf{X}$ and $\tilde{\mathbf{Z}}$ and hence $\mathbf{P}$ and $\tilde{\mathbf{P}}$ are orthogonal. It is easy to see that $\mathbf{P}+\tilde{\mathbf{P}} = \mathbf{P}_{\mathrm{all}}$, and so we arrive at the F-test once more.

We can use each of the two versions of the F-test above as starting points for generalization, where, rather than using OLS as a prediction method, we use other RP methods that are more tailored to specific alternatives of interest. The distribution of the output of an RP method under the null hypothesis of a linear model can be computed via simulation as follows. For a given $B > 1$ we generate independent n-vectors with independently and identically distributed (IID) standard normal components, $\boldsymbol{\zeta}^{(1)}, \ldots, \boldsymbol{\zeta}^{(B)}$. From these we form scaled residuals

$$\hat{\mathbf{R}}^{(b)} = \frac{(\mathbf{I} - \mathbf{P})\boldsymbol{\zeta}^{(b)}}{\|(\mathbf{I} - \mathbf{P})\boldsymbol{\zeta}^{(b)}\|_2}. \tag{2}$$

Let $\mathbf{X}_{\text{all}}$ be the full matrix of predictors. Writing the original scaled residuals as $\hat{\mathbf{R}}$ we apply our chosen RP function f to all the scaled residuals to obtain a p-value

$$\frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{f(\hat{\mathbf{R}}^{(b)}, \mathbf{X}_{\text{all}}) \leqslant f(\hat{\mathbf{R}}, \mathbf{X}_{\text{all}})\}}\right). \tag{3}$$

See also Section 4 for the extension to the case using lasso residuals.

Even in situations where the usual F-test may seem the natural choice, an RP test with a carefully chosen RP method can often be more powerful against alternatives of interest. This is particularly true when we aggregate the results of various RP methods to gain power over a diverse set of alternatives, as we describe in the next section.

3. Aggregating residual prediction tests

In many situations, we would like to try a variety of RP methods, to have power against various alternatives. A key example is when an RP method involves a tuning parameter such as the lasso. Each different value of the tuning parameter effectively gives a different RP method. One could also aim to create a generic omnibus test to test for, say, non-linearity, heteroscedasticity and correlation between the errors, simultaneously.

To motivate our approach for combining the results of multiple RP tests, we consider the famous diabetes data set of Efron *et al.* (2004). This has $p = 10$ predictors measured for $n = 442$ diabetes patients and includes a response that is a quantitative measure of disease progression 1 year after baseline. Given the null hypothesis of a Gaussian linear model, we wish to test for interactions and quadratic effects. To have power against alternatives composed of sparse coefficients for these effects, we consider as RP methods the lasso applied to quadratic effects residualized with respect to the linear terms via OLS. We regress the OLS scaled residuals onto the transformed quadratic effects by using the lasso with tuning parameters on a grid of λ-values, giving a family of RP tests.

We plot the RSSs from the lasso fits to the scaled residuals in Fig. 1, as a function of λ. Also shown are the RSSs from lasso fits to scaled residuals simulated under the null hypothesis of a Gaussian linear model, as simulation under the null hypothesis is the general principle which we use for deriving p-values.

At the point $\lambda = 0$, the observed RSS is not drastically smaller than those of the simulated residuals, as Fig. 1(a) shows. Indeed, if we were to calculate a p-value just based on the $\lambda = 0$ results corresponding to the F-test, we would obtain roughly 10%. The output at $\lambda = 0.8$, however, does provide compelling evidence against the null hypothesis, as Fig. 1(b) shows. Here the observed RSS is far to the left of the support of the simulated RSSs. To create a p-value for the presence of interactions based on all of the output, we need a measure of how 'extreme' the entire blue curve is, with respect to the red curves, in terms of carrying evidence against the

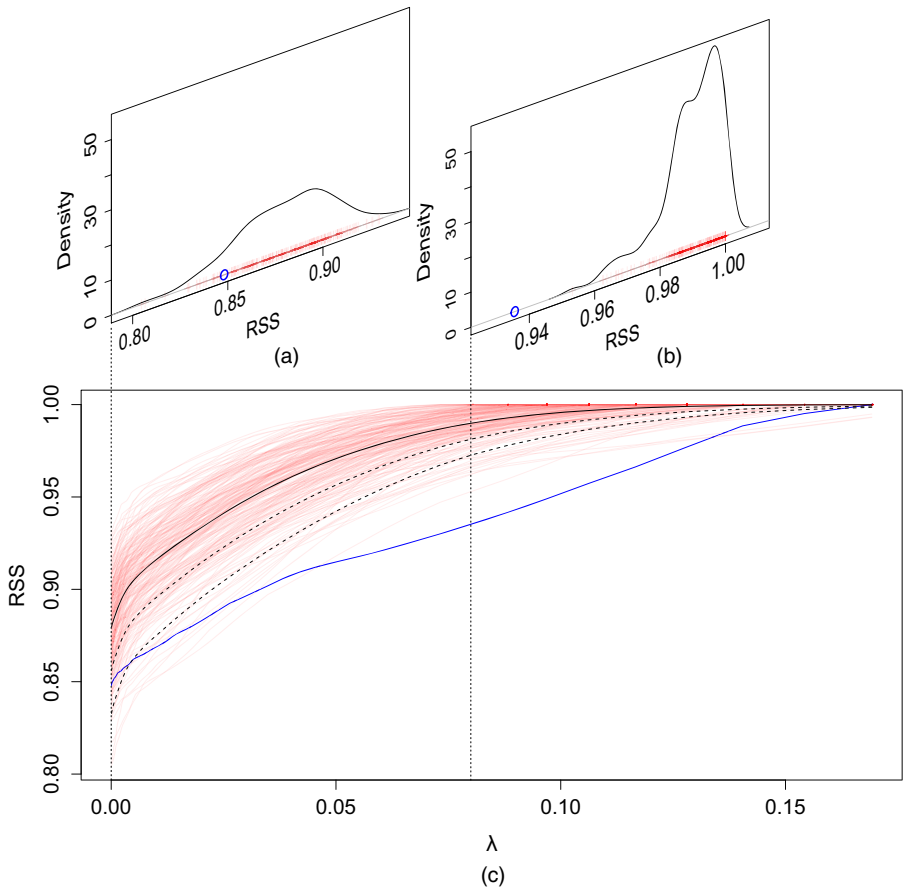

Fig. 1. Kernel density plots for the simulated RSSs at (a) $\lambda = 0$ and (b) $\lambda = 0.8$ (———, original RSSs) and (c) RSSs from lasso fits to the original scaled residuals (———) and simulated residuals (———), as well as the mean of the latter (———) and the mean displaced by 1 and 2 standard deviations (· · · · · ·)

null hypothesis. Forming a p-value based on such a test statistic is straightforward, as we now explain.

Suppose that we have RP functions f_l, $l = 1, \ldots, L$ (in our example these would be the RSS when using the lasso with tuning parameter λ_l) and their evaluations on the true scaled residuals $\hat{\mathbf{R}}^{(0)} := \hat{\mathbf{R}}$ and simulated scaled residuals $\{\hat{\mathbf{R}}^{(b)}\}_{b=1}^{B}$. Writing $f_l^{(b)} = f_l(\hat{\mathbf{R}}^{(b)}, \mathbf{X}_{\text{all}})$, let $\mathbf{f}^{(b)} = \{f_l^{(b)}\}_{l=1}^{L}$ be the curve or vector of RP function evaluations at the bth scaled residuals, and denote by $\mathbf{f}^{(-b)} = \{\mathbf{f}^{(b')}\}_{b' \neq b}$ the entire collection of curves, potentially including the curve for the true scaled residuals $\mathbf{R}^{(0)}$, but excluding the bth curve. Let

$$\tilde{Q} : \mathbb{R}^L \times \mathbb{R}^{L \times B} \to \mathbb{R},$$
$$(\mathbf{f}^{(b)}, \mathbf{f}^{(-b)}) \mapsto \tilde{Q}(\mathbf{f}^{(b)}, \mathbf{f}^{(-b)})$$

be any measure of how extreme the curve $\mathbf{f}^{(b)}$ is compared with the rest of the curves $\mathbf{f}^{(-b)}$ (larger values indicating more extreme). Here $\tilde{Q}$ can be any function such that $\tilde{Q}_b := \tilde{Q}(\mathbf{f}^{(b)}, \mathbf{f}^{(-b)})$ does not depend on the particular ordering of the curves in $\mathbf{f}^{(-b)}$; we shall give a concrete example below. We can use the $\{\tilde{Q}_b\}_{b \neq 0}$ to calibrate our test statistic $\tilde{Q}_0$ as detailed in the following proposition.

Proposition 1. Suppose that the simulated scaled residuals are constructed as in equation (2). Setting

$$Q = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{\tilde{Q}_b \geqslant \tilde{Q}_0\}}\right),$$

we have that, under the null hypothesis (1), $\mathbb{P}(Q \leqslant x) \leqslant x$ for all $x \in [0, 1]$, so Q constitutes a valid p-value.

This result is a straightforward consequence of the fact that under the null hypothesis $\{\tilde{Q}_b\}_{b=0}^{B}$ form an exchangeable sequence, and standard results on Monte Carlo testing (see Davison and Hinkley (1997), chapter 4, for example). Under an alternative, we expect $\tilde{Q}_0$ to be smaller and $\tilde{Q}_b$ for $b \geqslant 1$ to be larger on average than under the null. Thus this approach will have more power than directly comparing $\tilde{Q}_0$ with a sample from its null distribution.

We recommend constructing $\tilde{Q}$ as follows. Let $\hat{\mu}_l^{(-b)}$ and $\hat{\sigma}_l^{(-b)}$ respectively be the empirical mean and standard deviation of $\{f_l^{(b')}\}_{b' \neq b}$. We then set

$$\tilde{Q}_b = \max_l \{(\hat{\mu}_l^{(-b)} - f_l^{(b)})/\hat{\sigma}_l^{(-b)}\}, \tag{4}$$

the number of standard deviations by which the bth curve lies below the rest of the curves, maximized along the curve. The intuition is that, if $f_l^{(1)}$ were to have a Gaussian distribution under the null for each l, $\Phi\{(f_l^{(0)} - \hat{\mu}_l^{(-0)})/\hat{\sigma}^{(-0)}\}$ would be an approximate p-value based on the lth RP function, whence $\Phi(\tilde{Q}_0)$ would be the minimum of these p-values. Though it would be impossible to match the power of the most powerful test for the alternative in question (perhaps that corresponding to $\lambda = 0.8$ in our diabetes example) among the L tests considered, one would hope to come close. We stress, however, that this choice of $\tilde{Q}$ (4) yields valid p-values regardless of the distribution of $f_l^{(1)}$ under the null.

Using this approach with a grid of $L = 100$ λ-values, we obtain a p-value of under 1% for the diabetes example. As discussed in Section 1.2, this low p-value is unlikely to be due to a deviation from Gaussian errors, and indeed, when we take our simulated errors $\zeta^{(b)}$ to be resamples from the vector of residuals (see section B of the on-line supplementary material), we also obtain a p-value under 1%; clear evidence that a model including only main effects is inappropriate for the data. Further simulations demonstrating the power of this approach are presented in Section 5.

4. Lasso residual prediction tests

When the null hypothesis is itself high dimensional, we can use lasso residuals in the first stage of the RP testing procedure. Although, unlike scaled OLS residuals, scaled lasso residuals are not ancillary, we shall see that, under certain conditions, the distribution of scaled lasso residuals is not wholly sensitive to the parameters $\boldsymbol{\beta}$ and σ in model (1).

Write $\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})$ for the scaled lasso residuals when the tuning parameter is λ (in a square-root parameterization, see below):

$$\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) \in \underset{\mathbf{b} \in \mathbb{R}^p}{\arg\min}\{\|\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) + \sigma\boldsymbol{\varepsilon}\|_2/\sqrt{n} + \lambda\|\mathbf{b}\|_1\}, \tag{5}$$

$$\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) = \frac{\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})\} + \sigma\boldsymbol{\varepsilon}}{\|\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})\} + \sigma\boldsymbol{\varepsilon}\|_2}.$$

Note that, under model (1),

Table 1. Algorithm 1: lasso RP tests

(a) Let $\check{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$, typically a lasso estimate selected by cross-validation
(b) Set $\check{\sigma} = \|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\beta}}\|_2/\sqrt{n}$
(c) Form B scaled simulated residuals $\{\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta}^{(b)})\}_{b=1}^B$ where the $\boldsymbol{\zeta}^{(b)}$ are IID draws from $\mathcal{N}_n(\mathbf{0}, \mathbf{I})$, and λ is chosen according to the proposal of Sun and Zhang (2013)
(d) On the basis of the scaled simulated residuals $\{\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta}^{(b)})\}_{b=1}^B$, compute a p-value (3) or use these to form an aggregated p-value as described in Section 3

$$\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) \in \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p}\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2/\sqrt{n} + \lambda\|\mathbf{b}\|_1\}$$

and

$$\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) = \{\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})\}/\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})\|_2.$$

Sometimes we shall omit the first argument of $\hat{\boldsymbol{\beta}}$ for convenience, in which case it will always be the true parameter value under the null, $\boldsymbol{\beta}$. Here we are using the lasso in the square-root parameterization (Belloni *et al.*, 2011; Sun and Zhang, 2012) rather than the conventional version where the term in the objective assessing the model fit would be $\|\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) + \sigma\boldsymbol{\varepsilon}\|_2^2$. We note that the two versions of the lasso have identical solution paths but these will simply be parameterized differently. For this reason, we shall simply refer to expression (5) as the lasso solution. Note that, although the lasso solution may potentially be non-unique, the residuals are always uniquely defined, as the fitted values from a lasso fit are unique (see Tibshirani (2013), for example). Throughout we shall assume that the columns of $\mathbf{X}$ have been scaled to have l_2-norm $\sqrt{n}$.

We set out our proposal for calibrating RP tests on the basis of lasso residuals by using the parametric bootstrap in algorithm 1 in Table 1. In the following section we aim to justify the use of the parametric bootstrap from a theoretical perspective and also discuss the particular choices $\check{\boldsymbol{\beta}}$, $\check{\sigma}$ and λ used above.

4.1. Justification of lasso residual prediction tests

Given $\mathbf{b} \in \mathbb{R}^p$ and a set $A \subseteq \{1, \ldots, p\}$, let $\mathbf{b}_A$ be the subvector of $\mathbf{b}$ with components consisting of those indexed by A. Also, for a matrix $\mathbf{M}$, let $\mathbf{M}_A$ be the submatrix of $\mathbf{M}$ containing those columns indexed by A, and let $\mathbf{M}_k = \mathbf{M}_{\{k\}}$, the kth column. The following result shows that if $\operatorname{sgn}(\check{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta})$, with the sign function understood as being applied componentwise, we have partial ancillarity of the scaled residuals. In what follows we let $S = \{j : \beta_j \neq 0\}$ be the support set of $\boldsymbol{\beta}$.

Theorem 1. Suppose that $\check{\boldsymbol{\beta}}$ is such that $\operatorname{sgn}(\check{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta})$. For $t \in [0, 1)$ and $\lambda > 0$, consider the deterministic set

$$\Lambda_{\lambda,t} = \{\boldsymbol{\zeta} \in \mathbb{R}^n : \operatorname{sgn}\{\hat{\boldsymbol{\beta}}_{\lambda,S}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} = \operatorname{sgn}(\boldsymbol{\beta}_S) \text{ and } \min_{j\in S} \hat{\beta}_j(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})/\beta_j > t\}.$$

Then we have that, for all $\boldsymbol{\zeta} \in \Lambda_{\lambda,t}$, $\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}) = \hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta})$ provided that

$$0 < \check{\sigma}/\sigma < \min_{j\in S} \check{\beta}_j/\{(1-t)\beta_j\}.$$

In words, provided that the error $\boldsymbol{\zeta}$ is in the set $\Lambda_{\lambda,t}$ and conditions for $\check{\boldsymbol{\beta}}$ and $\check{\sigma}$ are met, the scaled residuals from a lasso fit to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\zeta}$ are precisely equal to the scaled residuals from a

lasso fit to $\mathbf{X}\check{\boldsymbol{\beta}} + \check{\sigma}\boldsymbol{\zeta}$. Note that all the quantities in the result are deterministic. Under reasonable conditions and for a sensible choice of λ (see theorem 2), when $\boldsymbol{\zeta} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$, we can expect the event $\boldsymbol{\zeta} \in \Lambda_{\lambda,t}$ to have large probability. Thus theorem 1 shows that the scaled residuals are not very sensitive to the parameter σ or to the magnitudes of the components of $\boldsymbol{\beta}$ but instead depend largely on the signs of the latter. It is the square-root parameterization that allows the result to hold for a large range of values of $\check{\sigma}$, and in particular for all $\check{\sigma}$ sufficiently small.

Theorem 1 does not directly justify a way to simulate from the distribution of the scaled lasso residuals as in algorithm 1 since the sign pattern of $\check{\boldsymbol{\beta}}$ must equal that of $\boldsymbol{\beta}$. Accurate estimation of the sign pattern of $\boldsymbol{\beta}$ using the lasso requires a strong irrepresentable or neighbourhood stability condition (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). Nevertheless, we now show that we can modify algorithm 1 to yield provable error control under more reasonable conditions. In Section 4.1.2 we argue heuristically that the same error control should hold for algorithm 1 in a wide range of settings.

4.1.1. Modified lasso residual prediction tests

Under a so-called beta-min condition (see theorem 2 below), with high probability we can arrive at an initial estimate of $\boldsymbol{\beta}$, $\boldsymbol{\beta}'$ via the lasso for which $\operatorname{sgn}(\boldsymbol{\beta}'_S) = \operatorname{sgn}(\boldsymbol{\beta}_S)$, and where $\min_{j \in S} |\beta'_j| > \max_{j \in S^c} |\beta'_j|$. With such a $\boldsymbol{\beta}'$, we can aim to seek a threshold τ for which the lasso applied only on the subset of variables $\mathbf{X}_{S_\tau}$ where $S_\tau := \{j : |\beta'_j| > \tau\}$ yields an estimate that has the necessary sign agreement with $\boldsymbol{\beta}$. This then motivates algorithm 2 (Table 2) on the basis of maximizing over the candidate p-values obtained through different $\boldsymbol{\beta}$-estimates derived from applying the lasso to different subsets of the initial active set (see also Chatterjee and Lahiri (2011) which introduces a related scheme).

Note that we do not recommend the use of algorithm 2 in practice; we introduce it only to facilitate theoretical analysis which sheds light on our proposed procedure algorithm 1. Let $s = |S|$ and $s' = |\{j : \beta'_j \neq 0\}|$. Theorem 2 below gives conditions under which, with high probability, $s' \geqslant s$ and residuals from responses generated around $\check{\boldsymbol{\beta}}^{(s)}$ will equal the true residuals. This then shows that the maximum p-value Q will in general be a conservative p-value as it will always be at least as large as Q_s, on an event with high probability.

As well as a beta-min condition, the result requires some relatively mild assumptions on the design matrix. Let $\mathcal{C}(\xi, T) = \{\mathbf{u} : \|\mathbf{u}_{T^c}\|_1 \leqslant \xi \|\mathbf{u}_T\|_1, \mathbf{u} \neq \mathbf{0}\}$. The restricted eigenvalue (Bickel *et al.*, 2009; Koltchinskii, 2009) is defined by

$$\phi(\xi) = \inf\left\{ \frac{\|\mathbf{X}\mathbf{u}\|_2/\sqrt{n}}{\|\mathbf{u}\|_2} : \mathbf{u} \in \mathcal{C}(\xi, S) \right\}. \tag{6}$$

For a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$, $T \subset \{1, \ldots, p\}$ and $\xi > 1$, the compatibility factor $\kappa(\xi, T, \mathbf{M})$ (van de Geer and Bühlmann, 2009) is given by

Table 2. Algorithm 2: modified lasso RP tests (only used for theorem 2)

(a) Let $\boldsymbol{\beta}' = \hat{\boldsymbol{\beta}}_\lambda(\sigma\varepsilon)$ be the lasso estimate of $\boldsymbol{\beta}$
(b) Let $s' = |\{j : \beta'_j \neq 0\}|$ and suppose that $0 < |\beta'_{j_s}| \leqslant \ldots \leqslant |\beta'_{j_1}|$ are the non-zero components of $\boldsymbol{\beta}'$ arranged in order of non-decreasing magnitude; define $\hat{S}^{(k)} = \{j_1, j_2, \ldots, j_k\}$
(c) For $k = 1, \ldots, s'$ let $\check{\boldsymbol{\beta}}^{(k)}$ be the lasso estimate from regressing $\mathbf{y}$ on $\mathbf{X}_{\hat{S}^{(k)}}$; further set $\check{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$.
(d) Using each of the $\check{\boldsymbol{\beta}}^{(k)}$ in turn and $\check{\sigma}^{(k)} = \|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\beta}}^{(k)}\|_2/\sqrt{n}$, generate sets of residuals $\{\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}^{(k)}, \check{\sigma}^{(k)}\boldsymbol{\zeta}^{(b)})\}_{b=1}^B$ where the $\boldsymbol{\zeta}^{(b)}$ are IID draws from $\mathcal{N}_n(\mathbf{0}, \mathbf{I})$; use these to create corresponding p-values Q_k for RP tests based on equation (3) or the method introduced in Section 3
(e) Output $Q = \max_{k=0,\ldots,s'} Q_k$ as the final approximate p-value

$$\kappa(\xi, T, \mathbf{M}) = \inf\left\{ \frac{\|\mathbf{M}\mathbf{u}\|_2/\sqrt{n}}{\|\mathbf{u}_T\|_1/|T|} : \mathbf{u} \in \mathscr{C}(\xi, T) \right\}. \tag{7}$$

When either of the final two arguments are omitted, we shall take them to be S and $\mathbf{X}$ respectively; the more general form is required in Section 4.2. The sizes of $\kappa(\xi)$ and $\phi(\xi)$ quantify the ill-posedness of the design matrix $\mathbf{X}$; we shall require $\kappa(\xi), \phi(\xi) > 0$ for some $\xi > 1$. Note that, in the random-design setting where the rows of $\mathbf{X}$ are IID multivariate normal with the minimum eigenvalue of the covariance matrix bounded away from zero, the factors (6) and (7) can be thought of as positive constants in asymptotic regimes where $s\log(p)/n \to 0$. We refer the reader to van de Geer and Bühlmann (2009) and Zhang and Zhang (2012) for further details.

Theorem 2. Suppose that the data follow the Gaussian linear model (1). Let

$$\lambda = A\sqrt{\left\{ \frac{2\log(p/\eta)}{n} \right\}}$$

with $A > \sqrt{2}$ and $p\exp(-s-2) > \eta > 0$. Suppose for $\xi > 1$ that

$$\frac{s\log(p/\eta)}{n\kappa^2(\xi)} \leqslant \frac{1}{A^2(\xi+1)} \min\left\{ 1 - \frac{(\xi+1)\sqrt{2}}{A(\xi-1)}, \frac{1}{5} \right\}. \tag{8}$$

Assume a beta-min condition

$$\min_{j \in S} |\beta_j| > 10\sqrt{2}A\xi \frac{\sigma\sqrt{\{s\log(p/\eta)\}}}{\phi^2(\xi)\sqrt{n}}. \tag{9}$$

Then, for all $x \in [0, 1]$,

$$\mathbb{P}(Q \leqslant x) \leqslant x + \frac{2(1 + r_{n-s})\eta}{\sqrt{\{\pi\log(p/\eta)\}}} + \exp\left(-\frac{n}{8}\right) \tag{10}$$

where $r_m \to 0$ as $m \to \infty$.

Although the beta-min condition, which is of the form $\min_{j \in S} |\beta_j| \geqslant \text{constant} \times \sqrt{\{s\log(p)/n\}}$, may be regarded as somewhat strong, the conclusion is correspondingly strong: any RP method or collection of RP methods with arbitrary $\tilde{Q}$ for combining tests can be applied to the residuals and the result remains valid. It is also worth noting, however, that the conditions are required only under the null hypothesis. For example, if the alternative of interest was that an additional variable $\mathbf{z}$ was related to the response after accounting for those in the original design matrix $\mathbf{X}$, no conditions on the relationship between $\mathbf{X}$ and $\mathbf{z}$ are required for the test to be valid.

More importantly, though, the conditions are certainly not necessary for the conclusion to hold. The scaled residuals are a function of the fitted values and the response, and do not involve lasso parameter estimates directly. Thus although duplicated columns in $\mathbf{X}$ could be problematic for inferential procedures relying directly on lasso estimates such as the debiased lasso (Zhang and Zhang, 2014), they pose no problem for RP tests. In addition, given a particular RP method, exact equality of the residuals would not be needed to guarantee a result of the form (10).

4.1.2. Relevance of theorem 2 to algorithm 1

In the special case of testing for the significance of a single predictor described above, we have a much stronger result than theorem 2 (see theorems 3 and 4) which shows that neither the beta-min condition nor the maximization over candidate p-values of algorithm 2 is necessary for error control to hold. More generally, in our experiments we have found that $Q_{s'}$ is usually equal to or close to the maximum Q for large B across a variety of settings. Thus selecting $Q_{s'}$ rather than

performing the maximization (which amounts to algorithm 1) can deliver conservative error control as evidenced by the simulations in Section 5.

A heuristic explanation for why the error is controlled is that typically the amount of signal remaining in $\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}^{(k)}, \check{\sigma}\boldsymbol{\zeta})$ increases with k, simply because typically $\|\mathbf{X}\check{\boldsymbol{\beta}}^{(k)}\|_2$ also increases with k. This can result in the prediction error of a procedure applied to the various residuals decreasing with k because the signal-to-noise ratios tend to be increasing; thus the p-values tend to increase with k.

In addition, when the lasso performs well, we would expect residuals to contain very little signal, and any differences in the signals contained in $\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ and $\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\varepsilon})$ to be smaller still, particularly when $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}\check{\boldsymbol{\beta}}$ are close. Typically the RP function will be insensitive to such small differences since they are unlikely to be too close to directions against which power is desired. We now discuss the choices of $\check{\boldsymbol{\beta}}$, λ and $\check{\sigma}$ in algorithm 1.

4.1.3. Practical considerations

4.1.3.1. Choice of $\check{\boldsymbol{\beta}}$. In view of the preceding discussion, it suffices for $\check{\boldsymbol{\beta}}$ to satisfy a screening-type property: we would like the support of $\check{\boldsymbol{\beta}}$ to contain that of $\boldsymbol{\beta}$. Though theorem 2 suggests a fixed λ, since $\check{\boldsymbol{\beta}}$ needs to be computed only once, we can use cross-validation. This is perhaps the most standard way of producing an estimate that performs well for screening (see for example section 2.5.1 of Bühlmann and van de Geer (2011)).

If the folds for cross-validation are chosen at random, the estimate will have undesirable randomness beyond that of the data. We thus suggest taking many random partitions into folds and using an estimate based on a λ that minimizes the cross-validation error curve based on all of the folds used. In our simulations in Section 5 we partition the observations into 10 random folds a total of eight times.

4.1.3.2. Choice of $\check{\sigma}$. The normalized RSS is perhaps the most natural choice for $\check{\sigma}^2$ (see also Reid *et al.* (2016)), though, as theorem 1 suggests, the results are essentially unchanged when this is doubled or halved, for example.

4.1.3.3. Choice of λ for the lasso residuals. The choice of λ should be such that, with high probability, the resulting estimate contains the support of $\boldsymbol{\beta}$ (see theorem 1). Though theorem 2 suggests taking $\lambda = A\sqrt{\{2\log(p)/n\}}$ for $A > \sqrt{2}$, the restriction on A is an artefact of basing our result on oracle inequalities from Sun and Zhang (2012), which place relatively simple conditions on the design. Sun and Zhang (2013) used a more involved theory which suggests a slightly smaller λ. We therefore use their method, the default in the R package of Sun (2013), as a convenient fixed choice of λ.

4.2. Testing the significance of individual predictors

Here we consider the collection of null hypotheses $H_k : \beta_k = 0$ and their corresponding alternatives that $\beta_k \neq 0$. For this setting there are many other approaches that can perform the required tests. Our aim here is to show that RP tests can be valid under weaker assumptions than those laid out in theorem 2, and moreover that the simpler approach of algorithm 1 can control type I error.

We begin with some notation. For $A_k := \{1, \ldots, p\} \setminus \{k\}$ and $\mathbf{b} \in \mathbb{R}^p$, let $\mathbf{b}_{-k} = \mathbf{b}_{A_k}$ and $\mathbf{X}_{-k} = \mathbf{X}_{A_k}$. For each variable k, our RP method will be a least squares regression onto a version of $\mathbf{X}_k$ that has been residualized with respect to $\mathbf{X}_{-k}$. Since in the high dimensional setting $\mathbf{X}_{-k}$ will typically have full row rank, an OLS regression of $\mathbf{X}_k$ on $\mathbf{X}_{-k}$ will return the $\mathbf{0}$-vector as residuals. Hence we shall residualize $\mathbf{X}_k$ by using the square-root lasso:

$$\Psi_k = \underset{\mathbf{b}\in\mathbb{R}^{p-1}}{\arg\min}\{\|\mathbf{X}_k - \mathbf{X}_{-k}\mathbf{b}\|_2/\sqrt{n} + \gamma\|\mathbf{b}\|_1\}.$$

This RP method is closely related to the pioneering idea by Zhang and Zhang (2014) and similar to that of Ning and Liu (2017), who considered using the regular lasso (without the square-root parameterization) at each stage. If $\mathbf{X}_k$ were not residualized with respect to $\mathbf{X}_{-k}$, and the regular lasso were used, the resulting RP method would be similar to that of Voorman *et al.* (2014). The work of Ren *et al.* (2015) studied an analogous procedure in the context of the Gaussian graphical model.

Let $\mathbf{W}_k$ be the residual $\mathbf{X}_k - \mathbf{X}_{-k}\Psi_k$. For each k we may write

$$\mathbf{y} = \mathbf{X}_{-k}\Theta_k + \beta_k\mathbf{W}_k + \sigma\varepsilon$$

where $\Theta_k = \boldsymbol{\beta}_{-k} + \beta_k\Psi_k \in \mathbb{R}^{p-1}$. Let $\hat{\Theta}_k$ be the square-root lasso regression of $\mathbf{y}$ onto $\mathbf{X}_{-k}$ with tuning parameter λ. Our RP function will be the RSS from OLS regression of the scaled lasso residuals $(\mathbf{y} - \mathbf{X}_{-k}\hat{\Theta}_k)/\|\mathbf{y} - \mathbf{X}_{-k}\hat{\Theta}_k\|_2$ onto $\mathbf{W}_k$. Note that this is an RP function even though it involves the residualized version of $\mathbf{X}_k$, $\mathbf{W}_k$; the latter is simply a function of $\mathbf{X}$. Equivalently, we can consider the test statistic T_k^2 with T_k defined by

$$T_k = \frac{\mathbf{W}_k^{\mathrm{T}}(\mathbf{y} - \mathbf{X}_{-k}\hat{\Theta}_k)}{\|\mathbf{W}_k\|_2\|\mathbf{y} - \mathbf{X}_{-k}\hat{\Theta}_k\|_2/\sqrt{n}}.$$

Note that T_k is simply a regularized partial correlation between $\mathbf{y}$ and $\mathbf{X}_k$ given $\mathbf{X}_{-k}$. The bootstrap version is

$$T_k^* = \frac{\mathbf{W}_k^{\mathrm{T}}(\mathbf{y}_k^* - \mathbf{X}_{-k}\hat{\Theta}_k^*)}{\|\mathbf{W}_k\|_2\|\mathbf{y}_k^* - \mathbf{X}_{-k}\hat{\Theta}_k^*\|_2/\sqrt{n}},$$

where $\mathbf{y}_k^* = \mathbf{X}_{-k}\hat{\Theta}_k + \check{\sigma}\varepsilon^*$, $\varepsilon^* \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$, and $\hat{\Theta}_k^*$ is the lasso regression of $\mathbf{y}^*$ on $\mathbf{X}_{-k}$. Here we shall consider taking $\check{\sigma} = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2/\sqrt{n}$ where $\hat{\boldsymbol{\beta}}$ is the square-root lasso regression of $\mathbf{y}$ on the full design matrix $\mathbf{X}$.

As before, let S be the support of $\boldsymbol{\beta}$, which without loss of generality we shall take to be $\{1, \ldots, s\}$, and also let $N = \{1, \ldots, p\} \setminus S$ be the set of true nulls. Assume that $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. The following result shows that only a relatively mild compatibility condition is needed to ensure that the type I error is controlled. We consider an asymptotic regime with $n \to \infty$ where $\boldsymbol{\beta}$, $\mathbf{X}$ and p are all allowed to vary with n though we suppress this in the notation. In what follows we denote the cumulative distribution function of the standard normal by Φ.

Theorem 3. Let $\lambda = A_1\sqrt{\{2\log(p)/n\}}$ for some constant $A_1 > 1$ and suppose that

$$\frac{s\sqrt{\{\log(p)^2/n\}}}{\kappa^2(\xi, S)} \to 0$$

for some $\xi > (A_1 + 1)/(A_1 - 1)$. Let $\gamma = A_2\sqrt{\{2\log(p)/n\}}$ for some constant $A_2 > 0$. Define $\mathscr{B} = \{\mathbf{b} \in \mathbb{R}^p : \mathbf{b}_N = 0\}$. Then

$$\sup_{k\in N, \boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} |\mathbb{P}(T_k \leqslant x) - \Phi(x)| \to 0,$$

$$\sup_{k\in N, \boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} |\mathbb{P}(T_k^* \leqslant x|\varepsilon) - \Phi(x)| \xrightarrow{\mathrm{p}} 0.$$

We see that a bootstrap approach can control the type I error uniformly across the noise variables and $\boldsymbol{\beta} \in \mathscr{B}$. We note that this result is for a fixed design $\mathbf{X}$ and does not require any sparsity

assumptions on the inverse covariance matrix of a distribution that could have generated the rows of $\mathbf{X}$ (Ning and Liu, 2017), for example.

Theorem 4. Let λ and γ be as in theorem 3. Assume that for some ξ with $\xi > (A_1+1)/(A_1-1)$ there is a sequence of sets $\{1,\ldots,s-1\} \subseteq T \subset \{1,\ldots,p-1\}$ such that

$$\frac{|T|\sqrt{\{\log(p)^2/n\}}}{\kappa^2(\xi, T, \mathbf{X}_{-k})} \to 0$$

and $\sqrt{\log(p)}\|\boldsymbol{\Theta}_{k,T^c}\|_1 \to 0$ where $T^c = \{1,\ldots,p-1\} \setminus T$. Further assume that $\beta_k\|\mathbf{W}_k\|_2/\sqrt{n} \to 0$. Define $\mathscr{B}_k = \{\mathbf{b} \in \mathscr{B} : b_k = \beta_k\}$. Then

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}_k, x\in\mathbb{R}} \left|\mathbb{P}(T_k \leqslant x) - \Phi\left\{x - \frac{\beta_k\|\mathbf{W}_k\|_2}{\sqrt{(\sigma^2 + \beta_k^2\|\mathbf{W}_k\|_2^2/n)}}\right\}\right| \to 0,$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}_k, x\in\mathbb{R}} |\mathbb{P}(T_k^* \leqslant x|\boldsymbol{\varepsilon}) - \Phi(x)| \xrightarrow{\mathrm{p}} 0.$$

If $\boldsymbol{\Psi}_k$ and hence $\boldsymbol{\Theta}_k$ were sparse, we could take T as the set of non-zeros and the second condition involving $\|\boldsymbol{\Theta}_{k,T^c}\|_1$ would be vacuous. This would be so with high probability in the random-design setting where $\mathbf{X}$ has IID Gaussian rows with sparse inverse covariance matrix (van de Geer *et al.*, 2014). However, $\boldsymbol{\Theta}_{k,T^c}$ can also have many small coefficients provided that they have small l_1-norm. The result above shows that the power of our method is comparable with the proposals of Zhang and Zhang (2014) and van de Geer *et al.* (2014) based on the debiased lasso. If $\|\mathbf{W}_k\|_2 = O(\sqrt{n})$ as would typically be so in the random-design setting discussed above, we would have power tending to 1 if $\beta_k \to 0$ but $\sqrt{n}|\beta_k| \to \infty$. Further results on power to detect non-linearities are given in section A of the on-line supplementary material.

The theoretical results do not suggest any real benefit from using the bootstrap, as to test hypotheses we can simply compare T_k with a standard normal distribution. However, our experience has been that this can be slightly anticonservative in certain settings. Instead, we propose to use the bootstrap to estimate the mean and standard deviation of the null distribution of the T_k by computing the empirical mean $\hat{m}_k$ and standard deviation $\hat{v}_k$ of B samples of T_k^*. Then we take as our p-values $2\{1 - \Phi(|T_k - \hat{m}_k/\hat{v}_k|)\}$.

This construction of p-values appears to yield tests that very rarely have size exceeding their nominal level. Indeed in all our numerical experiments we found no evidence of this violation. An additional advantage is that only a modest number of bootstrap samples is needed to yield the sort of low p-values that could fall below the threshold of a typical multiple-testing procedure. We recommend choosing B between 50 and 100.

4.2.1. Computational considerations

Using our bootstrap approach for calibration presents a significant computational burden when it is applied to test for the significance of each of a large number of variables in turn. Some modifications to algorithm 1 can help to overcome this issue and allow this form of RP test to be applied to typical high dimensional data with large p.

Firstly, rather than using cross-validation to choose λ for computation of $\hat{\Theta}_k$, we recommend using the fixed λ of Sun and Zhang (2013) (see also Section 4.1.3). The tuning parameter γ that is required to compute $\mathbf{W}_k$ can be chosen in the same way, and we also note that these nodewise regressions need to be done only once rather than for each bootstrap sample. Great computational savings can be realized by first regressing $\mathbf{y}$ on $\mathbf{X}$ to yield coefficients $\hat{\boldsymbol{\beta}}$. Writing $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$, we know that, for each $k \notin \hat{S}$, $\hat{\boldsymbol{\Theta}}_k = \hat{\boldsymbol{\beta}}_{-k}$, so we only need to compute $\hat{\boldsymbol{\Theta}}_k$ for those k in $\hat{S}$. The same logic can be applied to computation of $\hat{\boldsymbol{\Theta}}_k^*$ for the bootstrap replicates.

We also remark that approaches for directly simulating lasso estimates (Zhou, 2014) may be used to produce simulated residuals. These have the potential to reduce the computational burden substantially; not just in the case of testing significance of individual predictors but for RP tests in general.

5. Applications

5.1. Low dimensional null hypotheses

Here we return to the problem of testing for quadratic effects in the diabetes data set that was used in the example of Fig. 1. To investigate further the power of the aggregate RP test constructed through lasso regressions on a grid of 100 λ-values as described in Section 3, we created artificial signals from which we simulated responses. The signals (mean responses) were constructed by selecting at random s of the quadratic terms and giving these coefficients generated by using IID Unif$[-1, 1]$ random variables. The remaining coefficients for the variables were set to 0, so s determined the sparsity level of the signal. Responses were generated by adding IID Gaussian noise to the signals, with variance chosen such that the F-test for the presence of quadratic effects has power 0.5 when the size is fixed at 0.05. We created 25 artificial signals at each sparsity level $s \in \{1, 4, 10, 20, 35, 54\}$. The total number of possible quadratic effects was 54 (as one of the variables was binary), so the final sparsity level represents fully dense alternatives where we might expect the F-test to have good power. We note, however, that the average power of the F-test in the dense case rests critically on the form of the covariance between the generated quadratic coefficients, with optimality guarantees only in special circumstances (see section 8 of Goeman *et al.* (2006)). For the RP tests, we set the number of bootstrap samples B to be 249.

We also compare the power of RP tests with the *global test* procedure of Goeman *et al.* (2006). The results, which are shown in Fig. 2, suggest that RP tests can outperform the F-test in a variety of settings, most notably when the alternative is sparse, but also in dense settings. When there are small effects spread out across many variables ($s \in \{35, 54\}$), the global test tends to do best; indeed in such settings it is optimal. In the sparser settings, RP tests perform better.

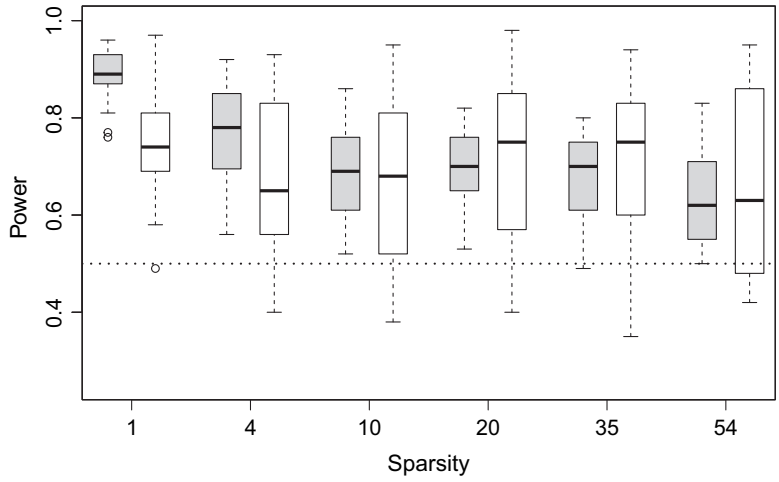

Fig. 2. Boxplots of the power of RP tests (■) and the global test (□) across the 25 signals estimated through 100 repetitions, for each of the sparsity levels s; the power of the F-test is fixed at 0.5 (· · · · · ·)

Table 3. Generation of Σ

Toeplitz: $\Sigma_{jk} = 0.9^{\lvert j-k \rvert}$
Exponential decay: $(\boldsymbol{\Sigma}^{-1})_{jk} = 0.4^{\lvert j-k \rvert/5}$
Equal correlation: $\Sigma_{jk} = 0.8$ if $j \neq k$ and $\Sigma_{jk} = 1$ otherwise

5.2. High dimensional nulls

In this section we report the results of using RP tests (algorithm 1) that are tailored to detect particular alternatives on a variety of simulated examples where the null hypothesis is high dimensional. We investigate both control of the type I error and the powers of the procedures.

Our examples are inspired by Dezeure *et al.* (2015). We use $n \times p$ simulated design matrices with $p = 500$ and $n = 100$ except for the setting where we test for heteroscedasticity in which we increase n to 300 to have reasonable power against these alternatives. The rows of the matrices are distributed as $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ given by the three types that are described in Table 3.

In addition to the randomly generated design matrices, we also used a publicly available real design matrix from gene expression data of *Bacillus subtilis* with $n = 71$ observations and $p = 4088$ predictors (Bühlmann *et al.*, 2014). Similarly to Dezeure *et al.* (2015), to keep the computational burden of the simulations manageable, we reduced the number of variables to $p = 500$ by selecting only those with the highest empirical variance. For each of the four design settings, we generated 25 design matrices (those from the real data were all the same). The columns of the design matrices were mean centred and scaled to have l_2-norm $\sqrt{n}$.

To create responses under the null hypothesis, for each of these 100 design matrices, we randomly generated a vector of coefficients $\boldsymbol{\beta}$ as follows. We selected a set S of 12 variables from $\{1, \ldots, p\}$. We then assigned $\boldsymbol{\beta}_{S^c} = \mathbf{0}$ and each β_k with $k \in S$ was generated according to Unif$[-2, 2]$ independently of other coefficients. This form of signal is similar to the most challenging signal settings that were considered in Dezeure *et al.* (2015) and also resembles the estimated signal from regression of the true response associated with the gene expression data onto the predictors by using the lasso or the minimax concave penalty (Zhang, 2010). Other constructions for generating the non-zero regression coefficients are considered in section C in the on-line supplementary material. Given $\mathbf{X}$ and $\boldsymbol{\beta}$, we generated $r = 100$ responses according to the linear model (1) with $\sigma = 1$. Thus, in total, here we evaluate the type I error control of our procedures on over 100 data-generating processes. The number of bootstrap samples B that were used was 100 when testing for significance of individual predictors and was fixed at 249 in all other settings.

We now explain interpretation of the plots in Figs 3–5; a description of Fig. 6 is given in Section 5.2.4. The top and bottom rows of each of Figs 3–5 concern settings under null and alternative hypotheses respectively. Thin red curves trace the empirical cumulative distribution functions (CDFs) of the p-values that were obtained by using RP tests, whereas thin blue curves, where shown, represent the same for debiased lasso-based approaches. In all plots, thickened coloured curves are averages of their respective thin coloured curves; note that these are averages over different simulation settings.

The black broken line is the CDF of the uniform distribution; thus we would hope for the empirical CDFs to be close to this in the null settings (top rows), and to rise above it in the bottom rows, indicating good power. Of course, even if all the p-value distributions were stochastically larger than uniform so that the type I error was always controlled, we would not expect their estimated distributions, i.e. the empirical CDFs, always to lie below the broken line. The black dotted curve allows us to assess type I error control across the simulation settings more easily.

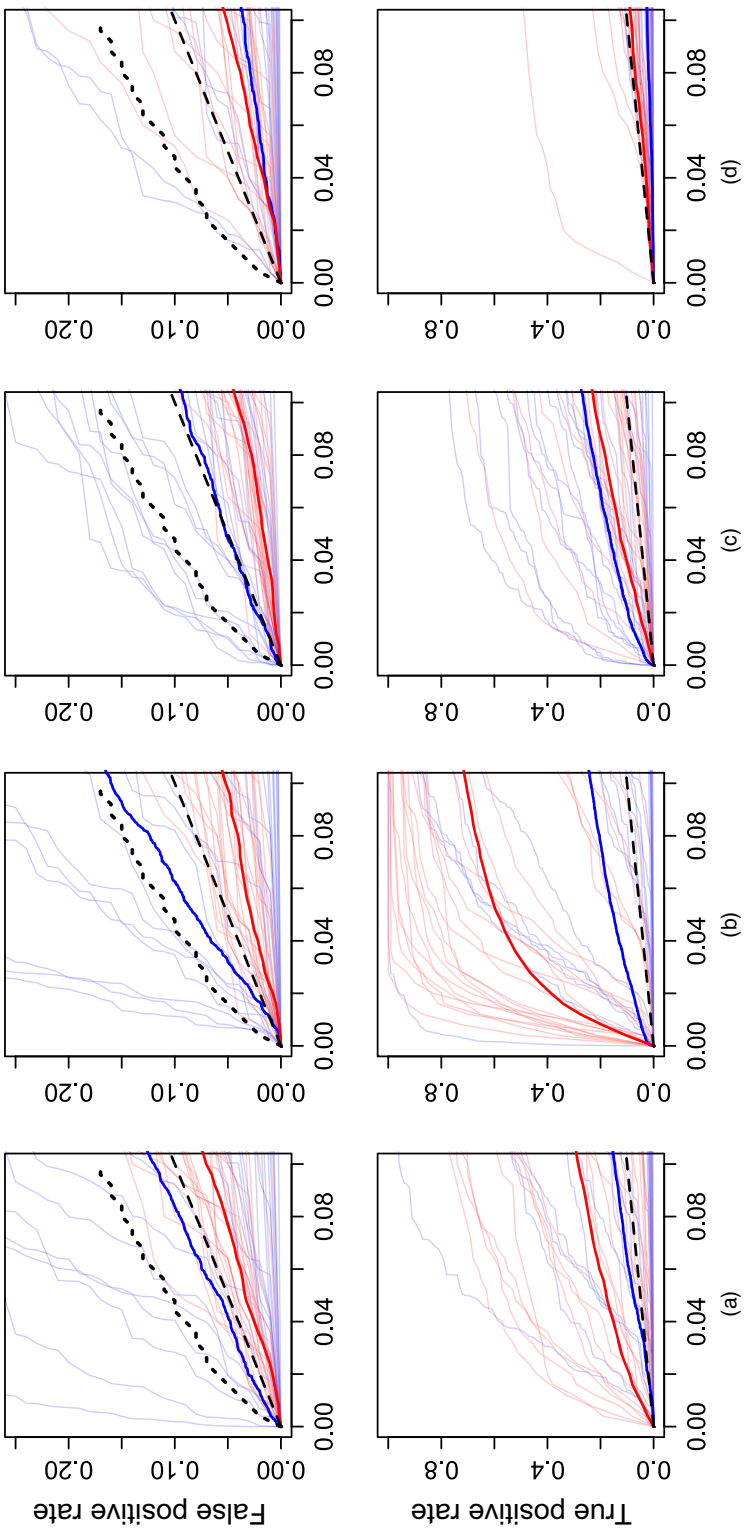

Fig. 3. Testing significance of groups—the empirical distribution functions of the p-values from RP tests (——) and the debiased lasso (——) under the null (top row) and alternative (bottom row) respectively (– – –, 45° line corresponding to the Unif[0, 1] distribution function; - - - - -, aid to assess type I error control): (a) Toeplitz; (b) exponential decay; (c) equal correlation; (d) real

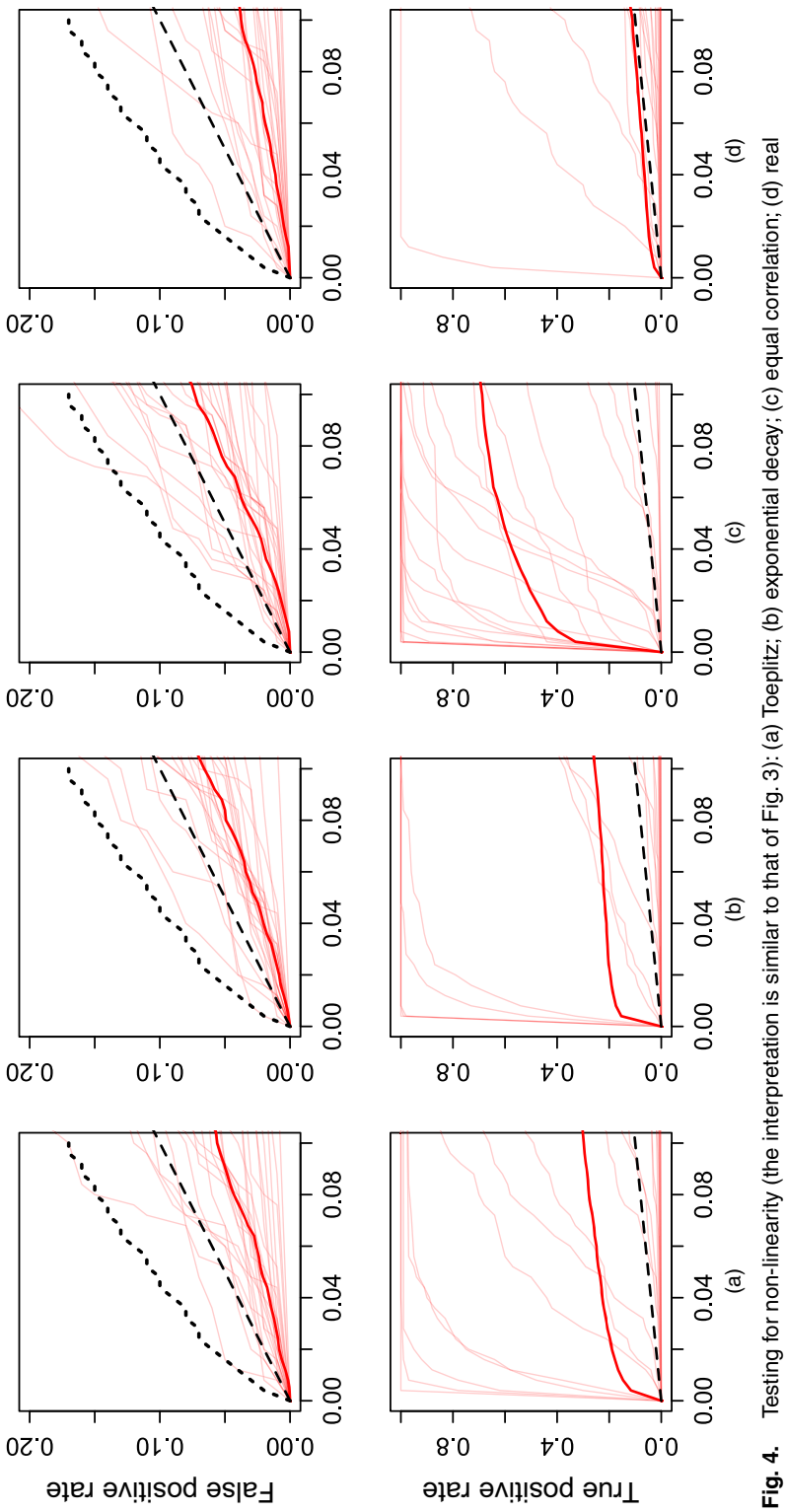

Fig. 4. Testing for non-linearity (the interpretation is similar to that of Fig. 3): (a) Toeplitz; (b) exponential decay; (c) equal correlation; (d) real

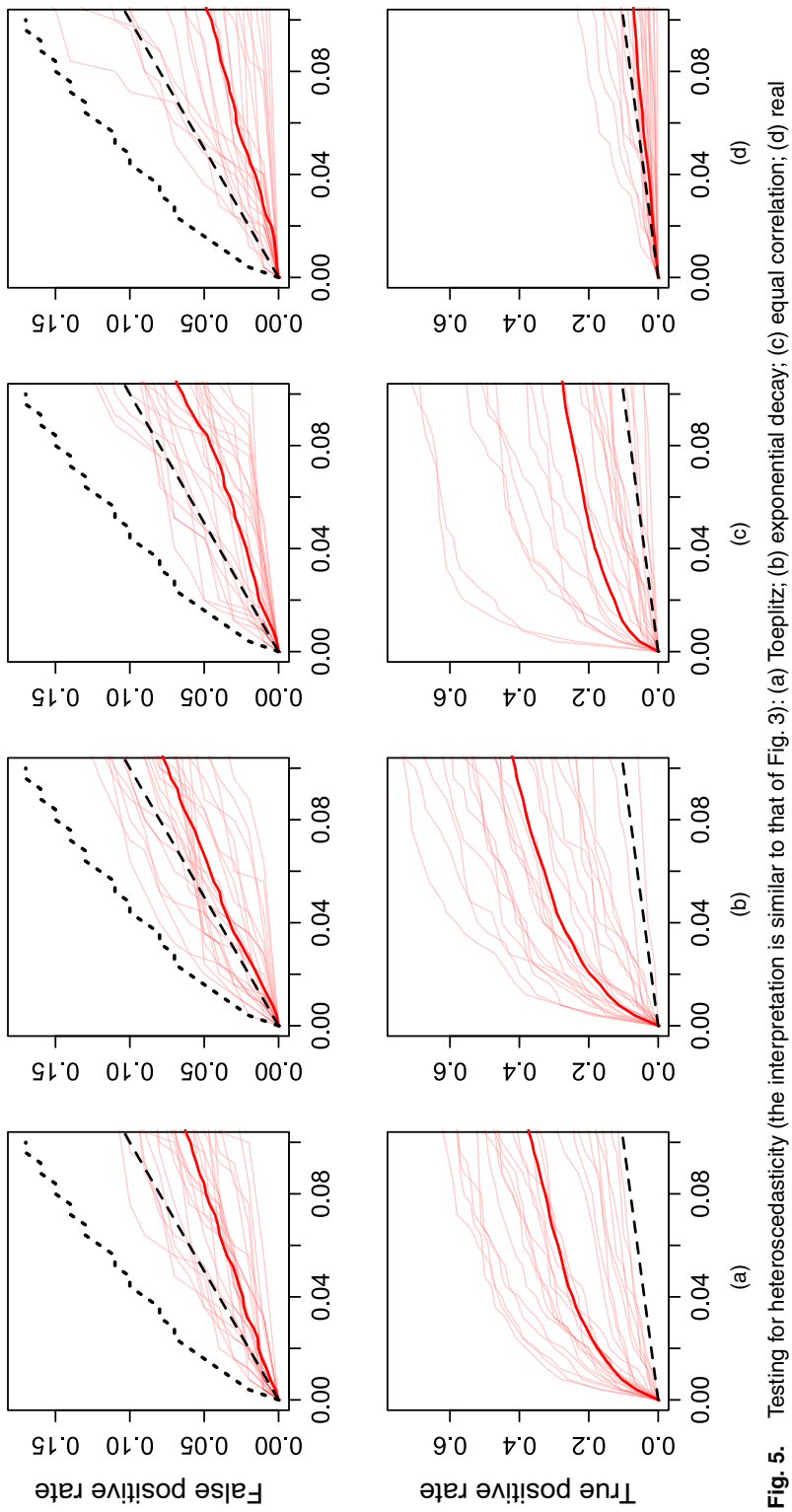

Fig. 5. Testing for heteroscedasticity (the interpretation is similar to that of Fig. 3): (a) Toeplitz; (b) exponential decay; (c) equal correlation; (d) real

It is constructed such that, in each of the plots, if the type I error were to be controlled exactly, we would expect on average one out of the 25 empirical CDFs for RP tests to escape above the region that the line encloses. Thus several curves not completely enclosed under the dotted curve in a given plot would indicate poor control of type I error. More precisely, the line is computed as follows. Let $q_\alpha(x)$ be the upper α-quantile of a $\mathrm{Bin}(B+1, x)/(B+1)$ distribution. Note this is the marginal distribution of $\hat{U}(x)$ where $\hat{U}$ is the empirical CDF of B samples from the uniform distribution on $\{1/(B+1), 2/(B+1), \ldots, 1\}$. The curve then traces $q_\alpha(x)$ with α chosen such that

$$\mathbb{P}[\max_{x \in [0,0.1]} \{\hat{U}(x) - q_\alpha(x)\} > 0] = 1/25.$$

We see that, across all of the data-generating processes and for each of the three RP testing methods, it appears that the size never exceeds the nominal level by a significant amount. Moreover the same holds for the additional 100 data-generating processes whose results are presented in the on-line supplementary material: the type I error is controlled well uniformly across all the settings that were considered.

We now describe the particular RP tests that were used in Figs 3–5, and the alternatives investigated, as well as the results shown in Fig. 6 concerning testing for the significance of individual predictors as detailed in Section 4.2.

5.2.1. Groups

We consider the problem of testing the null hypothesis $\beta_G = \mathbf{0}$ within linear model (1). One approach is to regress each column of $\mathbf{X}_G$ onto $\mathbf{X}_{G^c}$ in turn by using the square-root lasso (see Section 4.2), and to consider a matrix of residuals $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times |G|}$. We may then use lasso regression onto $\tilde{\mathbf{X}}$ as our family of RP methods and combine the resulting test statistics as in Section 3.

We use this approach on our simulated data and the results are displayed in red in Fig. 3. For the null settings (top row) we took G^c to be a randomly selected set of size $p/2$ containing S. Thus, under the null hypothesis, β_{G^c} had 12 non-zero components whereas $\beta_G = \mathbf{0}$. The alternatives, corresponding to the bottom row, also modify the signal such that β_A is non-zero (in addition to β_S being non-zero as was the case under the null) with coefficients generated in exactly the same way as for β_S and A being a randomly selected set of 12 variables chosen from G.

The blue curves trace the empirical CDFs of p-values constructed by using the debiased lasso proposal of van de Geer *et al.* (2014) and implemented in the `hdi` package of Dezeure *et al.* (2015) for R. More specifically, we use the minimum of the p-values that are associated with each of the coefficients in G (see section 2.3 of van de Geer *et al.* (2014)) as our test statistic and calibrate this by using the Westfall–Young procedure (Westfall and Young, 1993) as explained in Bühlmann (2013). This ensures that no power is lost because of correlations between the individual p-values, as would be the case with Bonferroni correction, for example. Remaining parameters were set to the defaults in the `hdi` package.

Although the sizes of the debiased lasso-based tests averaged over the equal correlation design examples are very close to the nominal level, this is due to the several settings where the size exceeds the desired level being compensated for by other examples where the tests are more conservative. In contrast, RP tests have slightly conservative type I error control across all the examples, and greater power among the Toeplitz and exponential decay settings.

5.2.2. Non-linearity

To test for non-linearity, we consider an RP method based on random forests (Breiman, 2001).

We used the default settings for random forests as implemented by Liaw and Wiener (2002), but rather than using a direct application to the residuals we apply it to the equicorrelation set: the set of variables with maximum absolute correlation with the residuals. This is invariably the set of variables that are selected in the initial lasso fit, though in situations where the lasso solution is not unique this will in general be a superset of the support of any lasso solution. Using this smaller set of variables reduces the computational burden of a random-forest fit and also gives the test greater power in situations where the variables contributing to the non-linear signal also feature in sparse linear approximations to the truth. Applying a random-forest fit to the entire set of variables may have slightly improved power when this is not so but would have greatly diminished power in the more natural situations where this holds. Rather than using the RSS from the random-forest fits as our proxy for prediction error, we use the out-of-bag error. This has the advantage of being more insensitive to the size of the equicorrelation set and tends to result in greater power.

To create the non-linear signal for the alternative settings, we randomly divided S into four groups of three. Each variable x was transformed via a sigmoid composed with a random affine mapping as below:

$$x \mapsto [1 + \exp\{-5(a + bx)\}]^{-1}.$$

Here $a, b \in \mathcal{N}(0, 1)$ independently. The transformed variables in each group were multiplied together, and a linear combination of these resulting products with Unif$[-1, 1]$ generated coefficients formed the non-linear component of the signal. This non-linear signal was then scaled such that the residuals from an OLS fit to the variables in S had an empirical variance of 2 and finally added to the linear signal.

The results that are displayed in Fig. 4 show that RP tests can deliver reasonable power in many of the settings that were considered, though the real design examples appear to be particularly challenging.

5.2.3. *Heteroscedasticity*

As testing for heteroscedasticity in a high dimensional setting is rather challenging, here we increase the number of observations for the simulated design settings to $n = 300$ to have reasonable power against the alternative. The data generation procedure under the null hypothesis was left unchanged. To generate vectors of variances for the alternative settings, we randomly selected three variables from S and formed a linear combination of these variables with Unif$[-2, 2]$ coefficients. A constant was then added, so the minimum component was 0.01, and finally the vector was scaled so that the average of its components was 1. This vector then determined the variance of normal errors added to the signal.

To detect this heteroscedasticity, we used a family of RP methods given by lasso regression of the absolute values of the residuals onto the equicorrelation set. The results are shown in Fig. 5. RP tests can deliver reasonable power in the simulated design settings, but they struggle to detect the heteroscedasticity with the real design which has a lower number of observations ($n = 71$).

5.2.4. *Testing significance of individual predictors*

Fig. 6 shows the results of using RP tests as described in Section 4.2 to test hypotheses $H_k : \beta_k = 0$. The red curves give the average proportions of false (top row) and true positive results (bottom row) that would be selected given p-value thresholds varying along the x-axis. Thus, for example, to obtain the expected number of false positive results selected at a given threshold, the y-values

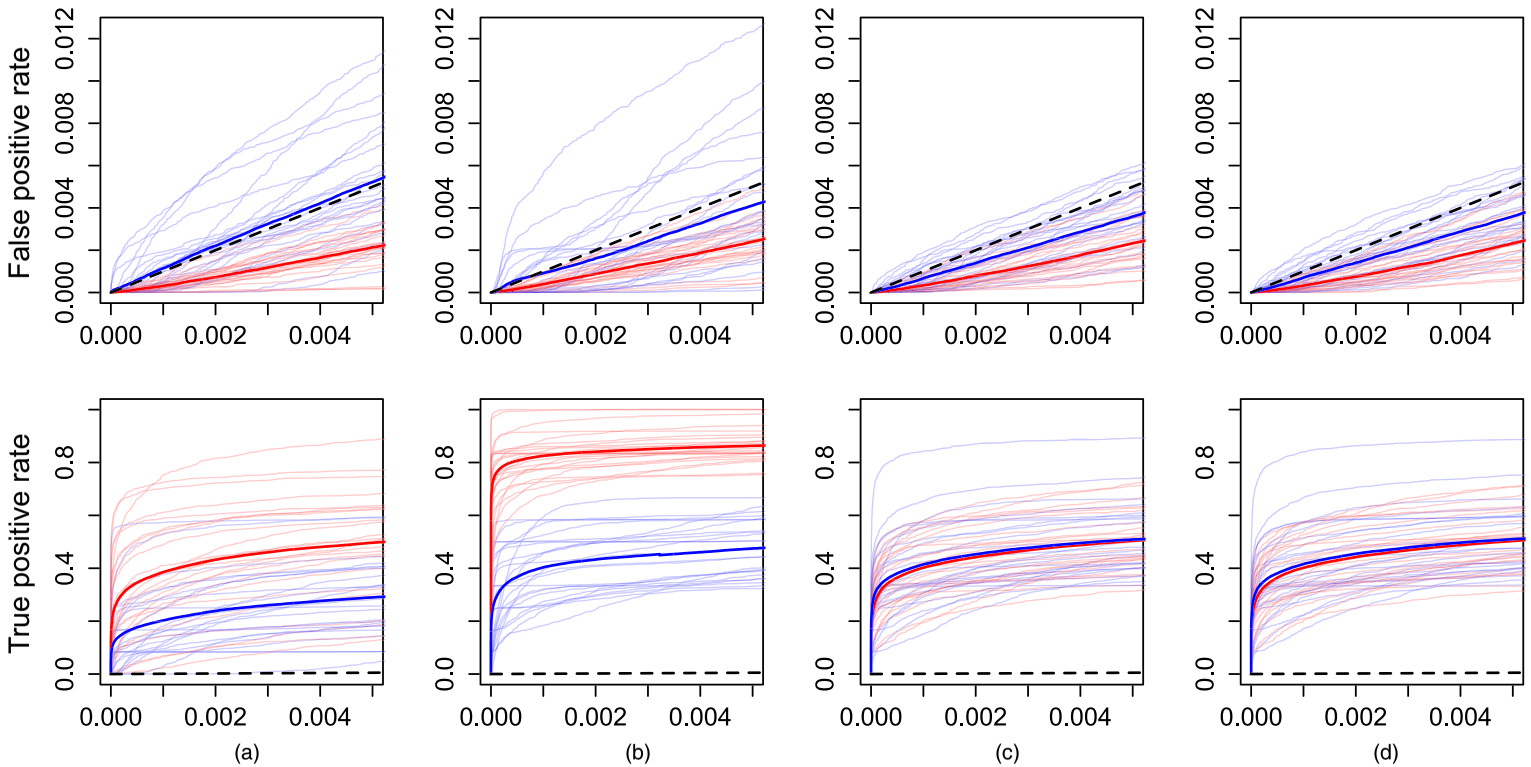

Fig. 6. Testing individual variables—the plots give the proportion of $|S^c| = 488$ null (top row) and $|S| = 12$ true variables (bottom row) selected at various threshold levels with RP tests (——) and the debiased lasso (——): (a) Toeplitz; (b) exponential decay; (c) equal correlation; (d) real

should be multiplied by $p - |S| = 488$. The blue curves display the same results for the debiased lasso as implemented in the `hdi` package. The broken 45° line gives the expected proportion of false positive results that would be incurred by an exact test.

We see that, even at the low p-value thresholds that are particularly relevant for multiple-testing correction, RP tests give consistent error control while also delivering superior or equal power. Such error control effectively requires accurate knowledge of the extreme tails of the null distribution of the test statistics. We see here that the debiased lasso approach cannot always achieve this in the Toeplitz and exponential decay settings, and indeed error control for multiple testing is rare among the currently available methods (Dezeure *et al.*, 2015).

6. Discussion

The RP testing methodology that was introduced in this work allows us to treat model checking as a prediction problem: that of fitting any (prediction) function to the scaled residuals from OLS or the lasso. This makes the problem of testing goodness of fit amenable to the entire range of prediction methods that have been developed across statistics and machine learning. We have investigated here RP tests for detecting significant single or groups of variables, heteroscedasticity or deviations from linearity, and we expect that effective RP methods can also be found for testing for correlated errors, heterogeneity and other sorts of departures from the standard Gaussian linear model. Related ideas should be applicable to test for model misspecification in high dimensional generalized linear models, for example.

Acknowledgements

Rajen Shah was supported in part by the Forschungsinstitut für Mathematik at the Eidgenössiche Technische Hochschule Zürich.

References

Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.
Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
Bühlmann, P. (2013) Statistical significance in high-dimensional linear models. *Bernoulli*, **19**, 1212–1242.
Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin: Springer.
Bühlmann, P. and van de Geer, S. (2015) High-dimensional inference in misspecified linear models. *Electron. J. Statist.*, **9**, 1449–1473.
Bühlmann, P., Kalisch, M. and Meier, L. (2014) High-dimensional statistics with a view toward applications in biology. *A. Rev. Statist. Appl.*, **1**, 255–278.
Camponovo, L. (2015) On the validity of the pairs bootstrap for lasso estimators. *Biometrika*, **102**, 981–987.
Chatterjee, A. and Lahiri, S. (2010) Asymptotic properties of the residual bootstrap for lasso estimators. *Proc. Am. Math. Soc.*, **138**, 4497–4509.
Chatterjee, A. and Lahiri, S. N. (2011) Bootstrapping lasso estimators. *J. Am. Statist. Ass.*, **106**, 608–625.
Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*, vol. 1. Cambridge: Cambridge University Press.
Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015) High-dimensional inference: confidence intervals, p-values and R-Software hdi. *Statist. Sci.*, **30**, 533–558.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–451.
Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. Boca Raton: CRC Press.
van de Geer, S. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, **3**, 1360–1392.
van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.

Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. (2006) Testing against a high dimensional alternative. *J. R. Statist. Soc.* B, **68**, 477–493.
Javanmard, A. and Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–2909.
Koltchinskii, V. (2009) The dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15**, 799–828.
Liaw, A. and Wiener, M. (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.
Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014) A significance test for the lasso. *Ann. Statist.*, **42**, 413–468.
Meinshausen, N. (2015) Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Statist. Soc.* B, **77**, 923–945.
Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.
Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.
Meinshausen, N., Meier, L. and Bühlmann, P. (2009) P-values for high-dimensional regression. *J. Am. Statist. Ass.*, **104**, 1671–1681.
Nan, Y. and Yang, Y. (2014) Variable selection diagnostics measures for high-dimensional regression. *J. Computnl Graph. Statist.*, **23**, 636–656.
Ning, Y. and Liu, H. (2017) A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, **45**, 158–195.
R Development Core Team (2005) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reid, S., Tibshirani, R. and Friedman, J. (2016) A study of error variance estimation in lasso regression. *Statist. Sin.*, to be published.
Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2015) Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.*, **43**, 991–1026.
Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: another look at stability selection. *J. R. Statist. Soc.* B, **75**, 55–80.
Sun, T. (2013) scalreg: scaled sparse linear regression. *R Package Version 1.0*. (Available from `https://CRAN.R-project.org/package=scalreg`.)
Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, **99**, 879–898.
Sun, T. and Zhang, C.-H. (2013) Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, **14**, 3385–3418.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R. J. (2013) The lasso problem and uniqueness. *Electron. J. Statist.*, **7**, 1456–1490.
Voorman, A., Shojaie, A. and Witten, D. (2014) Inference in high dimensions with the penalized score test. *Preprint arXiv:1401.2678*. University of Washington, Seattle.
Wasserman, L. and Roeder, K. (2009) High dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201.
Westfall, P. and Young, S. (1993) *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. New York: Wiley.
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.
Zhang, C.-H. and Zhang, T. (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, **27**, 576–593.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zhou, Q. (2014) Monte Carlo simulation for lasso-type problems by estimator augmentation. *J. Am. Statist. Ass.*, **109**, 1495–1516.
Zhou, Q. (2015) Uncertainty quantification under group sparsity. *Preprint arXiv:1507.01296*.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supporting materials for "Goodness of fit tests for high-dimensional linear models"'.