Goodness of fit tests for high-dimensional linear models

Rajen D. Shah* University of Cambridge

Peter Bühlmann ETH Zürich

April 8, 2017

Abstract

In this work 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 (RP) 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 RP 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 nonlinearity.

1 Introduction

High-dimensional data, where the number of variables may greatly exceed the number of observations, has become increasingly more prevalent across a variety of disciplines. While 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 different 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, 2014, 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 (CPSS) [Shah and Samworth, 2013] can also be used to select important variables whilst preserving error control in a wider variety of settings.

Despite these advances, something still lacking from the practitioner's toolbox is a corresponding set of diagnostic checks to help assess the validity of, for example, the high-dimensional linear

*Supported in part by the Forschungsintitut für Mathematik (FIM) at ETH Zürich.

model. For instance, there are no well-established methods for detecting heteroscedasticity in high-dimensional linear models, or whether a nonlinear 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 e.g. 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 (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 nonlinear alternatives. When restricting to sparse linear models specified by (1), the situation is different though and misspecification can happen [Bühlmann and van de Geer, 2015]; we will take a sparse Gaussian linear model as our null hypothesis (see also Theorems 3 and 4). We discuss an approach to handle a relaxation of the Gaussian error assumption in Section B of the 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 nonlinear model might be more interpretable and may generalise better, for example. Moreover, the Lasso and other sparse estimation procedures may have poor performance, undermining the various different high-dimensional inference methods 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 (RP) 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 $\sigma\boldsymbol{\varepsilon}$. On the other hand, if the signal is such that the Lasso performs poorly and instead a nonlinear model were more appropriate, for example, some of the (nonlinear) 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 nonlinear signal (an example may be Random Forest [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.

1. 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 if $\mathbf{X}$ is low-dimensional.

2. A *residual prediction method* (RP method) that is suited to predicting the particular signal expected in the residuals under the alternative(s) under consideration.

3. Some measure of the predictive capability of the RP method. Typically this would be the residual sum of squares (RSS), but in certain situations a cross-validated estimate of prediction error may be more appropriate, for example.

We will refer to the composition of a residual prediction method and an estimator of prediction error as a *residual prediction 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 nonlinearity, 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 can be predicted better than one would expect under the null hypothesis with a model as in (1), this provides evidence against the null.

Clearly in order to use RP tests to perform formal hypothesis tests, one needs knowledge of the distribution of the test statistic under the null, in order to calculate p-values. Closed form expressions are difficult if not impossible to come by, particularly when the residual prediction method is something as intractable as Random Forest.

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 in order 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. In order 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 rather 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 in order to have power against a diverse set of alternatives.

Although formally the null hypothesis tested by our approach is that of the sparse Gaussian linear model (1), an RP test geared towards nonlinearity is unlikely to reject purely due to non-Gaussianity of the errors, and so the effective null hypothesis typically allows for more general error distributions. By using the nonparametric 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 supplementary material, where we see that 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], Chatterjee and Lahiri [2011], Camponovo [2014] and Zhou [2014] who study 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 do not provide theoretical guarantees within a hypothesis testing framework as we do.

1.3 Organisation of the paper

Simulating the residuals under the null is particularly simple when rather than using the Lasso residuals, ordinary least squares residuals are used. We study this simple situation in Section 2 not only to help 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 conclude with a discussion in Section 6. The 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 of the proofs. The `R` [R Development Core Team, 2005] package `RPtests` provides an implementation of the methodology.

2 Ordinary least squares RP tests

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

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

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 let us write $\mathbf{Z} \in \mathbb{R}^{n\times q}$ for an additional group of variables. Let $\mathbf{P}_{\text{all}}$ be the orthogonal projection on to all available predictors, that is projection on to $\mathbf{X}_{\text{all}} = (\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n\times p_{\text{all}}}$, where $p_{\text{all}} = p + q$. When the RP method is OLS regression of the scaled residuals on to $\mathbf{X}_{\text{all}}$, the resulting RSS is

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

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

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

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

An alternative way to arrive at the F-test is to first residualise $\mathbf{Z}$ with respect to $\mathbf{X}$ and define new variables $\tilde{\mathbf{Z}} = (\mathbf{I}-\mathbf{P})\mathbf{Z}$. Let us write $\tilde{\mathbf{P}}$ for the orthogonal projection on to $\tilde{\mathbf{Z}}$. Now if our RP

method is OLS regression of the scaled residuals on to $\tilde{\mathbf{Z}}$, 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 $\mathbf{P}+\tilde{\mathbf{P}} = \mathbf{P}_{\text{all}}$, 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 generalisation, where rather than using OLS as a prediction method, we use other RP methods 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 i.i.d. 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 of the scaled residuals to obtain a p-value

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

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 different RP methods to gain power over a diverse set of alternatives, as we describe in the next section.

3 Aggregating RP tests

In many situations, we would like to try a variety of different RP methods, in order to have power against various different 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, nonlinearity, heteroscedasticity and correlation between the errors, simultaneously.

To motivate our approach for combining the results of multiple RP tests, we consider the famous diabetes dataset 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 one year after baseline. Given the null hypothesis of a Gaussian linear model, we wish to test for the presence of interactions and quadratic effects. In order to have power against alternatives composed of sparse coefficients for these effects, we consider as RP methods the Lasso applied to quadratic effects residualised with respect to the linear terms via OLS. We regress the OLS scaled residuals onto the transformed quadratic effects using the Lasso with tuning parameters on a grid of λ values, giving a family of RP tests.

We plot the residual sums of squares from the Lasso fits to the scaled residuals in Figure 1, as a function of λ. Also shown are the residual sums of squares from Lasso fits to scaled residuals

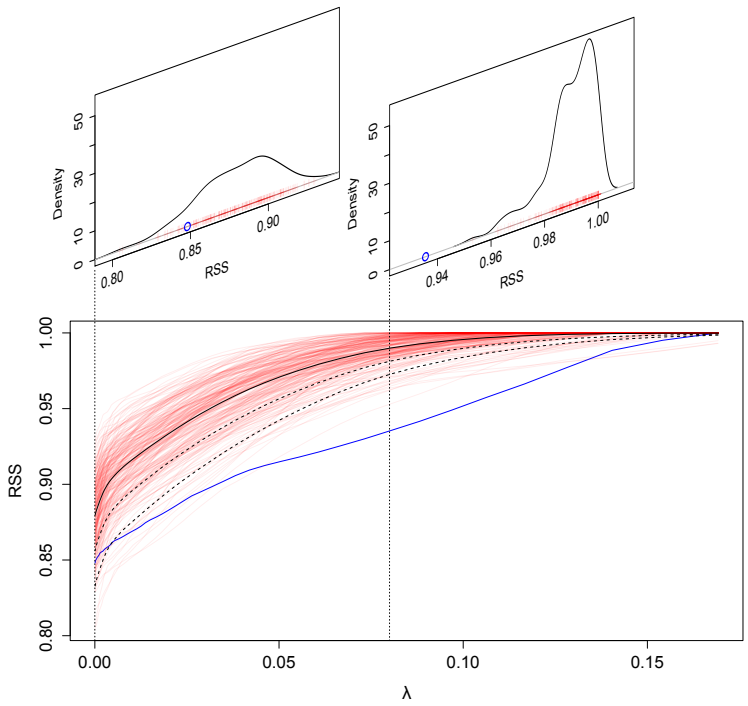

Figure 1: Bottom plot: the residual sums of squares from Lasso fits to the original scaled residuals (blue) and simulated residuals (red), as well as the mean of the latter (black) and the mean displaced by one and two standard deviations (black, dotted). Top plots: kernel density plots for the simulated residual sums of squares at $\lambda = 0$ (left) and $\lambda = 0.8$ (right) with the original residual sums of squares in blue.

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 the top left density plot shows. Indeed, were we 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 the top right density plot shows. Here the observed RSS is far to the left of the support of the simulated residual sums of squares. In order 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 null. Forming a p-value based on such a test statistic is straightforward, as we now explain.

Suppose we have residual prediction 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

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

be any measure of how extreme the curve $\mathbf{f}^{(b)}$ is compared to 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 will 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 the simulated scaled residuals are constructed as in* (2)*. Setting*

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

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

The result above is a straightforward consequence of the fact that under the null $\{\tilde{Q}_b\}_{b=0}^{B}$ form an exchangeable sequence, and standard results on Monte Carlo testing (see Davison and Hinkley [1997] Ch. 4 for example). Under an alternative, we expect $\tilde{Q}_0$ to be smaller and $\tilde{Q}_b$ for $b \geq 1$ to be larger on average, than under the null. Thus this approach will have more power than directly comparing $\tilde{Q}_0$ to 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, maximised along the curve. The intuition is that were $f_l^{(1)}$ 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 $\boldsymbol{\zeta}^{(b)}$ to be resamples from the vector of residuals (see Section B of the 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 RP 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 will see that under certain conditions, the distribution of scaled Lasso residuals are not wholly sensitive to the parameters $\boldsymbol{\beta}$ and σ in (1).

Let us write $\hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})$ for the scaled Lasso residuals when the tuning parameter is λ (in square-root parametrisation, see below):

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

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

Note that under (1) $\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) \in \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2/\sqrt{n} + \lambda\|\mathbf{b}\|_1\}$ 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 will 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 parametrisation [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 parametrised differently. For this reason, we will simply refer to (5) as the Lasso solution. Note that while 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 will 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 based on Lasso residuals using the parametric bootstrap in Algorithm 1 below. 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.

Algorithm 1 Lasso RP tests

1. Let $\check{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$, typically a Lasso estimate selected by cross-validation.
2. Set $\check{\sigma} = \|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\beta}}\|_2/\sqrt{n}$.
3. 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 i.i.d. draws from $\mathcal{N}_n(\mathbf{0}, \mathbf{I})$, and λ chosen according to the proposal of Sun and Zhang [2013].
4. Based on 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.

4.1 Justification of Lasso RP 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 the following we let $S = \{j : \beta_j \neq 0\}$ be the support set of $\boldsymbol{\beta}$.

Theorem 2. *Suppose* $\check{\boldsymbol{\beta}}$ *is such that* $\operatorname{sgn}(\check{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta})$. *For* $t \in [0, 1)$ *and* $\lambda > 0$, *consider the deterministic set*

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

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

In words, provided 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 of the quantities in the result are deterministic. Under reasonable conditions and for a sensible choice of λ (see Theorem 3), 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 2 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 parametrisation 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 2 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 RP tests

Under a so-called beta-min condition (see Theorem 3 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 based on maximising 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 we do not recommend the use of Algorithm 2 in practice; we only introduce it to facilitate theoretical analysis which sheds light on our proposed procedure Algorithm 1. Let $s = |S|$ and $s' = |\{j : \beta'_j \neq 0\}|$. The theorem below gives conditions under which with high probability, $s' \geq s$ 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 $\mathscr{C}(\xi, T) = \{\mathbf{u} : \|\mathbf{u}_{T^c}\|_1 \leq \xi\|\mathbf{u}_T\|_1, \mathbf{u} \neq \mathbf{0}\}$. The restricted eigenvalue [Bickel et al., 2009, Koltchinskii, 2009] is defined by

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

Algorithm 2 Modified Lasso RP tests (only used for Theorem 3)

1. Let $\boldsymbol{\beta}' = \hat{\boldsymbol{\beta}}_\lambda(\sigma\boldsymbol{\varepsilon})$ be the Lasso estimate of $\boldsymbol{\beta}$.
2. Let $s' = |\{j : \beta'_j \neq 0\}|$ and suppose $0 < |\beta'_{j'_s}| \leq \cdots \leq |\beta'_{j_1}|$ are the non-zero components of $\boldsymbol{\beta}'$ arranged in order of non-decreasing magnitude. Define $\hat{S}^{(k)} = \{j_1, j_2, \ldots, j_k\}$.
3. 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 $\check{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$.
4. 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 i.i.d. draws from $\mathcal{N}_n(\mathbf{0}, \mathbf{I})$. Use these to create corresponding p-values Q_k for RP tests based on (3) or the method introduced in Section 3.
5. Output $Q = \max_{k=0,\ldots,s'} Q_k$ as the final approximate p-value.

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

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

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 the design matrix $\mathbf{X}$; we will require $\kappa^(\xi), \phi(\xi) > 0$ for some $\xi > 1$. Note that in the random design setting where the rows of $\mathbf{X}$ are i.i.d. 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 3. *Suppose the data follows the Gaussian linear model* (1). *Let* $\lambda = A\sqrt{2\log(p/\eta)/n}$ *with* $A > \sqrt{2}$ *and* $pe^{-s-2} > \eta > 0$. *Suppose for* $\xi > 1$ *that*

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

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 $x \in [0, 1]$,

$$\mathbb{P}(Q \leq x) \leq x + \frac{2(1 + r_{n-s})\eta}{\sqrt{\pi\log(p/\eta)}} + e^{-n/8} \tag{10}$$

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

Although the beta-min condition, which is of the form is of the form $\min_{j \in S} |\beta_j| \geq \text{const.} \times \sqrt{s \log(p)/n}$, 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 only required under the null. 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 whilst 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 3 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 3 (see Theorems 4 and 5) which shows that neither the beta-min condition nor the maximisation over candidate p-values of Algorithm 2 is necessary for error control to hold. More generally, in our experiments we have found $Q_{s'}$ 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 maximisation (which amounts to Algorithm 1) is able to 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

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 3 suggests a fixed λ, since $\check{\boldsymbol{\beta}}$ only needs to be computed once, we can use cross-validation. This is the 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 minimises 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 8 times.

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

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 2). Though Theorem 3 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] has a more involved theory which suggests a slightly smaller λ. We therefore use their method, the default in the `R` package 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 $\beta_k \neq 0$. Note that 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 3, 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 residualised 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 will residualise $\mathbf{X}_k$ using the square-root Lasso:

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

This RP method is closely related to the pioneering idea by Zhang and Zhang [2014] and similar to that of Ning and Liu [2014], who consider using the regular Lasso (without the square-root parametrisation) at each stage. If $\mathbf{X}_k$ were not residualised 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] studies an analogous procedure in the context of the Gaussian graphical model.

Let $\mathbf{W}_k$ be the residual $\mathbf{X}_k - \mathbf{X}_{-k}\boldsymbol{\Psi}_k$. Note for each k we may write

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

where $\boldsymbol{\Theta}_k = \boldsymbol{\beta}_{-k} + \beta_k\boldsymbol{\Psi}_k \in \mathbb{R}^{p-1}$. Let $\hat{\boldsymbol{\Theta}}_k$ be the square-root Lasso regression of $\mathbf{y}$ on to $\mathbf{X}_{-k}$ with tuning parameter λ. Our RP function will be the RSS from OLS regression of the scaled Lasso residuals $(\mathbf{y} - \mathbf{X}_{-k}\hat{\boldsymbol{\Theta}}_k)/\|\mathbf{y} - \mathbf{X}_{-k}\hat{\boldsymbol{\Theta}}_k\|_2$ on to $\mathbf{W}_k$. Note this is an RP function even though it involves the residualised 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

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

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

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

where $\mathbf{y}_k^* = \mathbf{X}_{-k}\hat{\mathbf{\Theta}}_k + \check{\sigma}\boldsymbol{\varepsilon}^*$, $\boldsymbol{\varepsilon}^* \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$ and $\hat{\mathbf{\Theta}}_k^*$ is the Lasso regression of $\mathbf{y}^*$ on $\mathbf{X}_{-k}$. Here we will 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 will take to be $\{1, \ldots, s\}$, and also let $N = \{1, \ldots, p\} \setminus S$ be the set of true nulls. Assume $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. The following result shows that only a relatively mild compatibility condition is needed in order 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 the following we denote the cumulative distribution function of the standard normal by Φ.

Theorem 4. *Let $\lambda = A_1\sqrt{2\log(p)/n}$ for some constant $A_1 > 1$ and suppose that $s\sqrt{\log(p)^2/n}/\kappa^2(\xi, S) \to 0$ 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*

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

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, 2014], for example.

Theorem 5. *Let λ and γ be as in Theorem 4. 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 $|T|\sqrt{\log(p)^2/n}/\kappa^2(\xi, T, \mathbf{X}_{-k}) \to 0$ and $\sqrt{\log(p)}\|\mathbf{\Theta}_{k,T^c}\|_1 \to 0$ 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*

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

If $\mathbf{\Psi}_k$ and hence $\mathbf{\Theta}_k$ were sparse, we could take T as the set of nonzeroes and the second condition involving $\|\mathbf{\Theta}_{k,T^c}\|_1$ would be vacuous. This would be the case with high probability in the random design setting where $\mathbf{X}$ has i.i.d. Gaussian rows with sparse inverse covariance matrix [van de Geer et al., 2014]. However, $\mathbf{\Theta}_{k,T^c}$ can also have many small coefficients provided they have small ℓ_1-norm. The result above shows that the power of our method is comparable to the proposals of Zhang and Zhang [2014] and van de Geer et al. [2014] based on the debaised Lasso. If $\|\mathbf{W}_k\|_2 = O(\sqrt{n})$ as would typically be the case 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 nonlinearities are given in Section A of the 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 to a standard normal distribution. However our experience has been that this can be slightly anti-conservative 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 occurring. 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 tests 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 required to compute $\mathbf{W}_k$ can be chosen in the same way, and we also note that these nodewise regressions only need to be done once rather than for each bootstrap sample. Great computational savings can be realised 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 substantially reduce the computational burden; not just in the case of testing significance of individual predictors but for RP tests in general.

5 Applications

5.1 Low-dimensional nulls

Here we return to the problem of testing for quadratic effects in the diabetes dataset used in the example of Figure 1. In order to further investigate 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 using i.i.d. 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 i.i.d. 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 $s \in \{1, 4, 10, 20, 35, 54\}$. Note that 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 to the *global test* procedure of Goeman et al. [2006]. The results, shown in Figure 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.

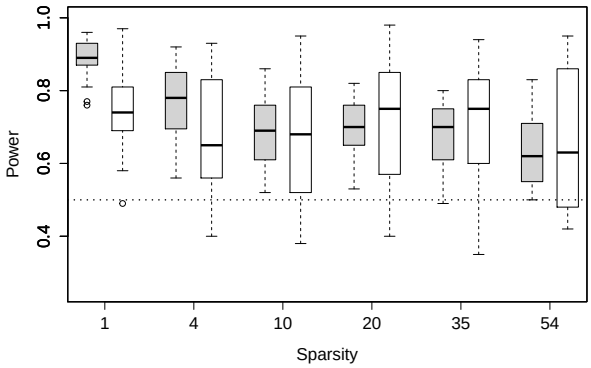

Figure 2: Boxplots of the power of RP tests (grey) and the global test (white) across the 25 signals estimated through 100 repetitions, for each of the sparsity levels s; the power of the F-test is fixed at 0.5 and shown as a dotted line.

5.2 High-dimensional nulls

In this section we report the results of using RP tests (Algorithm 1) 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 in order to have reasonable power against these alternatives. The rows of the matrices are distributed as $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ given by the three types described in Table 1.

Table 1: Generation of Σ.

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

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], in order 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 $\sqrt{n}$.

In order 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 considered in Dezeure et al. [2015] and also resembles the estimated signal from regression of the true response associated with the gene expression data on to the predictors using the Lasso or MCP [Zhang, 2010]. Other constructions for generating the non-zero regression coefficients are considered in Section C in the 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 used was 100 when testing for significance of individual predictors and fixed at 249 in all other settings.

We now explain interpretation of the plots in Figures 3–5; a description of Figure 6 is given in Section 5.2.4. The top and bottom rows of each of Figures 3–5 concern settings under null and alternative hypotheses respectively. Thin red curves trace the empirical cumulative distribution functions (CDFs) of the p-values obtained using RP tests, whilst 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 these are averages over different simulation settings.

The black dashed 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 rise above it in the bottom rows indicating good power. Of course, even if all of the p-value distributions were stochastically larger than uniform so the type I error was always controlled, we would not expect their estimated distributions i.e. the empirical CDFs to always lie below the dashed line. The black dotted curve allows us to assess type I error control across the simulation settings more easily. It is constructed such that in each of the plots, were the type I error to be controlled exactly, we would expect on average 1 out of the 25 empirical CDFs for RP tests to escape above the region the line encloses. Thus several curves not completely enclosed under the dotted line 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 $\text{Bin}(B+1, x)/(B+1)$ 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

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

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

We now describe the particular RP tests used in Figures 3–5, and the alternatives investigated, as well as the results shown in Figure 6 concerning testing for the significance of individual predictors

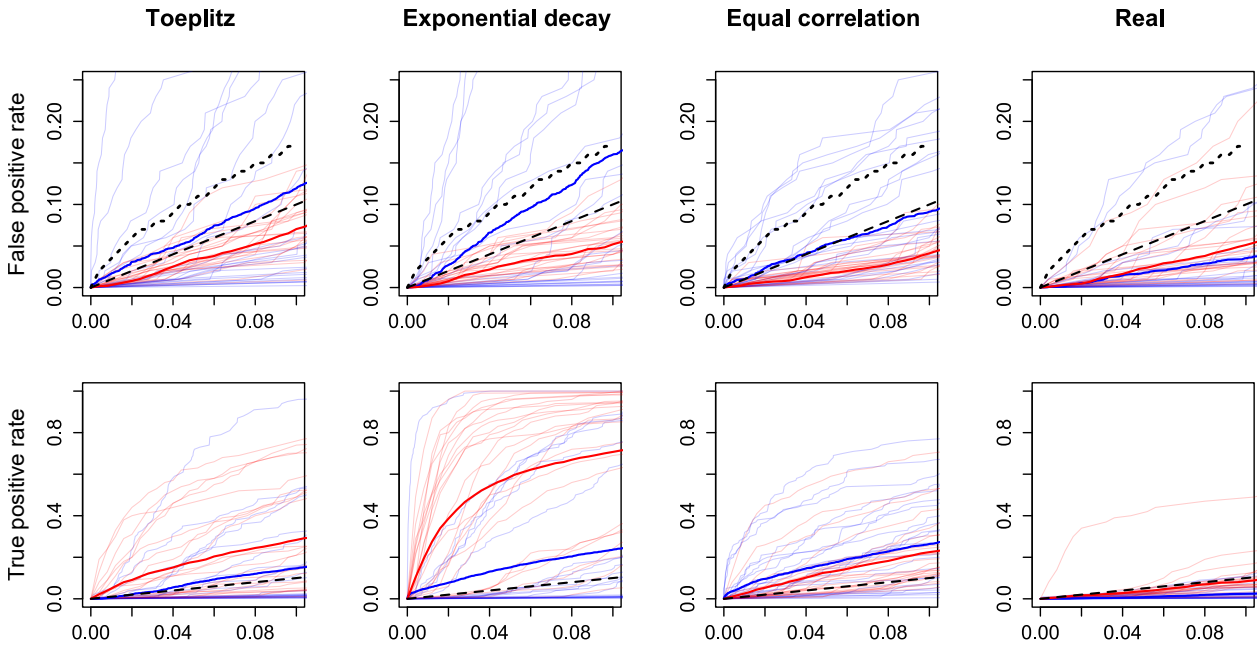

Figure 3: Testing significance of groups: the empirical distribution functions of the p-values from RP tests (red) and the debiased Lasso (blue) under the null (top row) and alternative (bottom row) respectively. The dashed line equals the 45 degree line corresponding to the Unif$[0, 1]$ distribution function, and the dotted curve is explained in the main text.

as detailed in Section 4.2.

5.2.1 Groups

We consider the problem of testing the null hypothesis $\boldsymbol{\beta}_G = \mathbf{0}$ within linear model (1). One approach is to regress each column of $\mathbf{X}_G$ on to $\mathbf{X}_{G^c}$ in turn using the square-root Lasso (c.f. Section 4.2), and consider a matrix of residuals $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times |G|}$. We may then use Lasso regression on to $\tilde{\mathbf{X}}$ 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 Figure 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, $\boldsymbol{\beta}_{G^c}$ had 12 non-zero components whilst $\boldsymbol{\beta}_G = \mathbf{0}$. The alternatives, corresponding to the bottom row, also modify the signal such that $\boldsymbol{\beta}_A$ is non-zero (in addition to $\boldsymbol{\beta}_S$ being non-zero as was the case under the null) with coefficients generated in exactly the same way as for $\boldsymbol{\beta}_S$ and A being a randomly selected set of 12 variables chosen from G.

The blue lines trace the empirical CDFs of p-values constructed using the debiased Lasso proposal of van de Geer et al. [2014] and implemented in the `hdi` package Dezeure et al. [2015] for `R`. More specifically, we use the minimum of the p-values 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 using the Westfall–Young procedure [Westfall and Young, 1993] as explained in Bühlmann [2013]. This ensures that no power is lost due to correlations among 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.

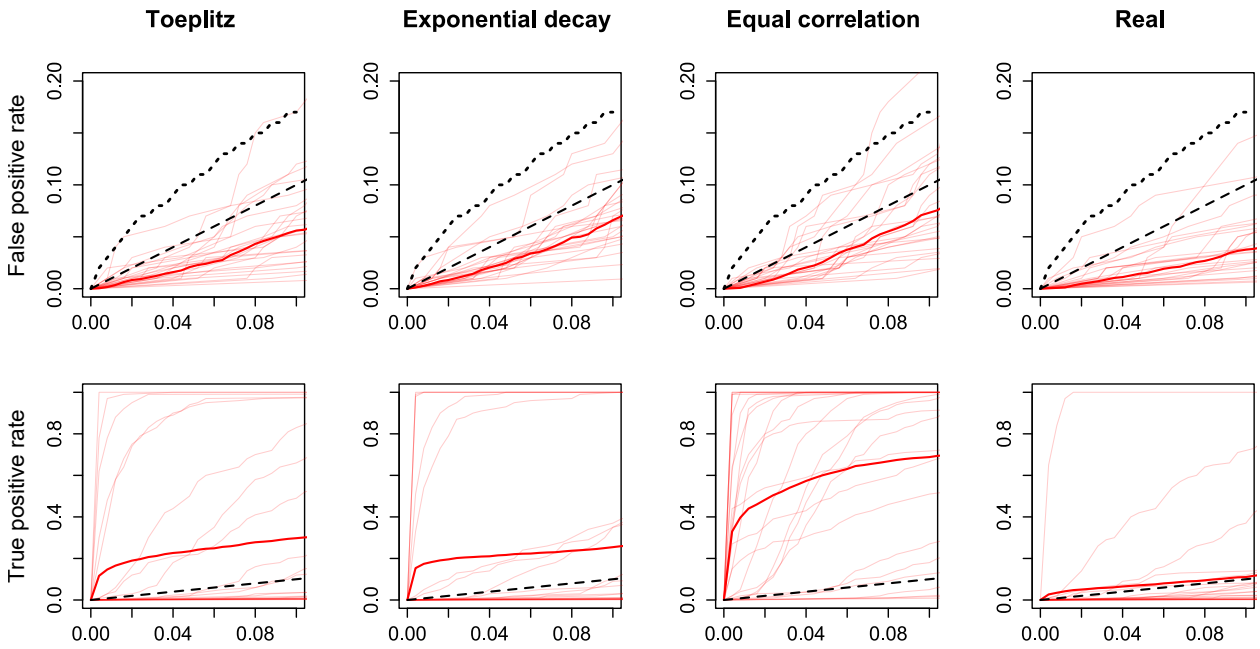

Figure 4: Testing for nonlinearity; the interpretation is similar to that of Figure 3.

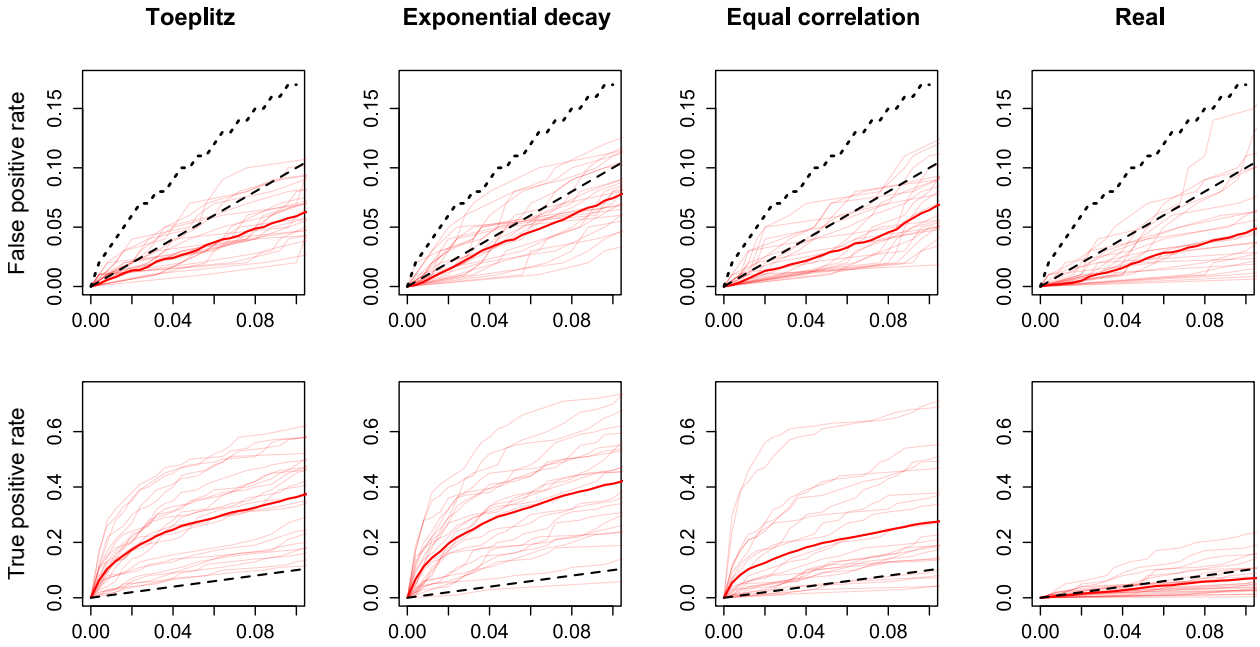

Figure 5: Testing for heteroscedasticity; the interpretation is similar to that of Figure 3.

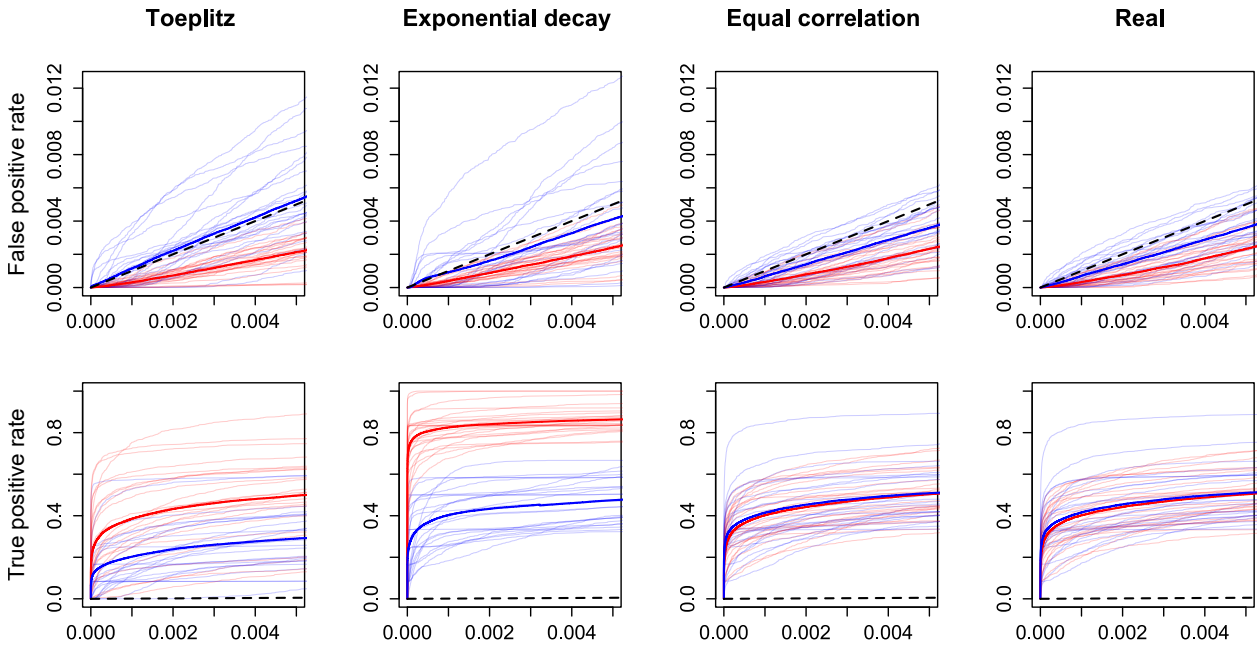

Figure 6: Testing individual variables: the plots give the proportion of $|S^c| = 488$ null (top row) and $|S| = 12$ true variables (bottom row) selected at various threshold levels with RP tests (red) and the debiased Lasso (blue).

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. On the other hand, 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 Nonlinearity

In order to test for nonlinearity, we consider an RP method based on Random Forest [Breiman, 2001]. We used the default settings for Random Forest 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 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 nonlinear signal also feature in sparse linear approximations to the truth. Applying a Random Forest to the entire set of variables may have slightly greater power when this is not the case, but would have greatly diminished power in the more natural situations where this 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 nonlinear 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 nonlinear component of the signal. This nonlinear 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 displayed in Figure 4 show that RP tests are able to deliver reasonable power in many of the settings 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 $n = 300$ in order to have reasonable power against the alternative. The data-generation procedure under the null was left unchanged. In order to generate vectors of variances for the alternative settings, we randomly selected 3 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 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 Figure 5. RP tests are able to deliver reasonable power in the the simulated design settings, but do 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

Figure 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 positives (bottom row) that would be selected given p-value thresholds varying along the x-axis. Thus for example in order to obtain the expected number of false positives selected at a given threshold, the y values should be multiplied by $p - |S| = 488$. The blue curves display the same results for the debaised Lasso as implemented in the `hdi` package. The dashed 45 degree line gives the expected proportion of false positives that would be incurred by an exact test.

We see that even at the low p-value thresholds particularly relevant for multiple testing correction, RP tests give consistent error control whilst 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 is not always able to 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 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 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 generalised linear models, for example.

References

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous Analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242, 2013.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional data: Methods, Theory and Applications*. Springer, 2011.

P. Bühlmann and S. van de Geer. High-dimensional inference in misspecified linear models. *Electron. J. Statist.*, 9:1449–1473, 2015.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

L. Camponovo. On the validity of the pairs bootstrap for lasso estimators. *Biometrika, to appear*, 2014.

A. Chatterjee and S. Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proc. Am. Math. Soc.*, 138(12):4497–4509, 2010.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *J. Am. Statist. Ass.*, 106(494): 608–625, 2011.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional Inference: Confidence intervals, p-values and R-Software hdi. *Statistical Science*, 30:533–558, 2015.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Ann. Statist.*, 32: 407–451, 2004.

J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen. Testing against a high-dimensional alternative. *J. R. Statist. Soc. B*, 68:477–493, 2006.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42:413–468, 2014.

N. Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Statist. Soc. B*, 77(5):923–945, 2015.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34:1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection (with discussion). *J. R. Statist. Soc. B*, 72: 417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Am. Statist. Ass.*, 104:1671–1681, 2009.

Y. Nan and Y. Yang. Variable selection diagnostics measures for high-dimensional regression. *J. Computnl Graph. Statist.*, 23(3):636–656, 2014.

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765*, 2014.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.r-project.org`.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica, to appear*, 2016.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.*, 43:991–1026, 2015.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *J. R. Statist. Soc. B*, 75(1):55–80, 2013.

T. Sun. *scalreg: Scaled sparse linear regression*, 2013. URL `https://CRAN.R-project.org/package=scalreg`. R package version 1.0.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *The J. Mach. Learn. Res.*, 14 (1):3385–3418, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

A. Voorman, A. Shojaie, and D. Witten. Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*, 2014.

L. Wasserman and K. Roeder. High dimensional variable selection. *Ann. Statist.*, 37:2178–2201, 2009.

P. Westfall and S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38: 894–942, 2009.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76:217–242, 2014.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

Q. Zhou. Monte carlo simulation for lasso-type problems by estimator augmentation. *J. Am. Statist. Ass.*, 109(508):1495–1516, 2014.

Q. Zhou. Uncertainty quantification under group sparsity. *arXiv preprint arXiv:1507.01296*, 2015.

Supplementary material

This supplementary material is organised as follows. Section A contains results on the power of the RP tests approach for detecting nonlinearity.

In Section B we discuss how the RP tests methodology can be extended to test for null hypotheses of linear models with non-Gaussian errors, and present numerical results in support of our proposed scheme. Additional numerical results to complement those of Section 5 in the main paper are presented in Section C. In Section D we provide some brief comments on the interpretation of p-values derived from RP tests. Finally the proofs of all of the results in the main paper, we well as those stated in Section A, are collected in Section E. Note that all equations numbered 1–10 are in the main paper.

A Power of Lasso RP tests

In this section we briefly discuss the power of RP tests for detecting nonlinearity. Suppose the response is generated according to

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

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a sparse vector with $S = \{j : \beta_j \neq 0\}$, $s = |S|$ and as before $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. The nonlinear term $\mathbf{f}$ is to be thought of as a vector of function evaluations of some nonlinear function: $f_i = f(\mathbf{x}_{i,S})$ where $f : \mathbb{R}^{|S|} \to \mathbb{R}$, though this is not assumed in the sequel.

As in Section 4 of the main paper, here we require that the columns of $\mathbf{X}$ have been scaled to have ℓ_2-norm $\sqrt{n}$. To facilitate theoretical analysis, we will assume $\check{\boldsymbol{\beta}}$ is a Lasso estimate with fixed $\lambda = A_1\sqrt{2\log(p)/n}$ and $A_1 > 1$, rather than with the tuning parameter selected by cross-validation as in Algorithm 1. Furthermore, we will also take this to be the tuning parameter used in the construction of the Lasso scaled residuals $\hat{\mathbf{R}} = \hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \mathbf{f} + \sigma\boldsymbol{\varepsilon})$. Let the bootstrap scaled residuals be $\hat{\mathbf{R}}^* = \hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta})$. We will also take this λ to be the tuning parameter used in the construction of the Lasso scaled residuals $\hat{\mathbf{R}} = \hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \mathbf{f} + \sigma\boldsymbol{\varepsilon})$. Let the bootstrap scaled residuals derived from $\check{\boldsymbol{\beta}}$ and $\check{\sigma} := \|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\beta}}\|_2/\sqrt{n}$ be $\hat{\mathbf{R}}^* := \hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta})$ where $\boldsymbol{\zeta} \in \mathcal{N}_n(\mathbf{0}, \mathbf{I})$.

To quantify the potential power of RP tests, we define

$$\boldsymbol{\psi}_\gamma = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \{\|\mathbf{f} - \mathbf{X}\mathbf{b}\|_2/\sqrt{n} + \gamma\|\mathbf{b}\|_1\}$$

and let $\mathbf{w}_\gamma = \mathbf{f} - \mathbf{X}\boldsymbol{\psi}_\gamma$ be the nonlinear signal $\mathbf{f}$ residualised with respect to $\mathbf{X}$. As in Section 4.2 we consider an asymptotic regime where p, $\mathbf{X}$, $\boldsymbol{\beta}$, S and $\mathbf{f}$ can all change as $n \to \infty$, though we suppress this in the notation. Also, as in Theorem 4, let $\mathscr{B} = \{\mathbf{b} \in \mathbb{R}^p : \mathbf{b}_{S^c} = 0\}$.

The result, which follows from Theorem 5 and its proof, shows that whilst the true residuals are positively correlated with the residualised signal $\mathbf{w}_\gamma$, the bootstrap residuals are not.

Corollary 6. *Suppose* $\|\mathbf{f}\|_2/\sqrt{n} \to 0$ *and for some* γ *we have* $\sqrt{n}\gamma\|\boldsymbol{\psi}_{\gamma,S^c}\|_1 \to 0$ *and* $\|\mathbf{f}\|_2\gamma = o(\sqrt{\log(p)})$. *Assume there exists* $\xi > (A_1+1)/(A_1-1)$ *with* $s\gamma\sqrt{\log(p)}/\kappa^2(\xi, S) \to 0$. *We have*

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} \left|\mathbb{P}(\mathbf{w}_\gamma^T\hat{\mathbf{R}}/\|\mathbf{w}_\gamma\|_2 \leq x) - \Phi\Big(x - \|\mathbf{w}_\gamma\|_2/\sqrt{\sigma^2 + \|\mathbf{w}_\gamma\|_2^2/n}\Big)\right| \to 0,$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} |\mathbb{P}(\mathbf{w}_\gamma^T\hat{\mathbf{R}}^*/\|\mathbf{w}_\gamma\|_2 \leq x|\boldsymbol{\varepsilon}) - \Phi(x)| \xrightarrow{p} 0.$$

An interesting application of the result above is quantification of the power to detect interactions, as we now discuss. Consider a random design setting where $\mathbf{X}$ is a scaled version of a matrix $\mathbf{Z}$ whose rows $\mathbf{z}_i$ are independent with $\mathbf{z}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ and $\Sigma_{jj} = 1$ for all j. That is we have $\mathbf{X}_k = \sqrt{n}\mathbf{Z}_k/\|\mathbf{Z}_k\|_2$. Let $f_i = \mathbf{z}_{i,S}^T \mathbf{\Theta} \mathbf{z}_{i,S}$ where without loss of generality, $\mathbf{\Theta} \in \mathbb{R}^{s\times s}$ is a symmetric matrix. Thus the nonlinear component of the signal is a quadratic function of the variables in S. As before, we will consider asymptotics where $n, p \to \infty$ and now also $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ will change as $p \to \infty$, though we suppress this in the notation. We will assume a restricted eigenvalue-type condition on the sequence of covariance matrices $\mathbf{\Sigma}$: let

$$\phi_0(\xi) = \inf\left\{ \frac{\|\mathbf{\Sigma}^{1/2}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} : \mathbf{u} \in \mathscr{C}(\xi, S) \right\}$$

and assume that for $\xi > (A_1 + 1)/(A_1 - 1)$, we have $\phi_0(\xi) > \phi > 0$ as $n \to \infty$. Note that this is weaker than assuming the minimum eigenvalue of $\mathbf{\Sigma}$ is bounded away from zero, for example.

Theorem 7. *Suppose* $n^{1/3}\mathbb{E}(f_1^2) \to 0$ *and* $s\log(p)/n^{1/3} \to 0$. *We have*

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} \left| \mathbb{P}(\mathbf{f}^T\hat{\mathbf{R}}/\|\mathbf{f}\|_2 \le x|\mathbf{Z}) - \Phi\big(x - \|\mathbf{f}\|_2/\sqrt{\sigma^2 + \|\mathbf{f}\|_2^2/n}\big) \right| \xrightarrow{p} 0,$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, x\in\mathbb{R}} |\mathbb{P}(\mathbf{f}^T\hat{\mathbf{R}}^*/\|\mathbf{f}\|_2 \le x|\boldsymbol{\varepsilon}, \mathbf{Z}) - \Phi(x)| \xrightarrow{p} 0.$$

Note that the theorem allows for $\mathbb{E}(\|\mathbf{f}\|_2^2) = n\mathbb{E}(f_1^2) \to \infty$, though we do need $\mathbb{E}(f_1^2) \to 0$. We see that the nonlinear signal is positively correlated with the true Lasso residuals, but not with the bootstrap residuals. Thus the nonlinear signal is present in the true Lasso residuals, and in principle can be detected by a suitable RP method.

B Non-Gaussian errors

Although the null hypothesis that the Gaussian linear model (1) is correct is often of interest, one may wish to consider a larger null hypothesis that allows for non-Gaussian errors. Theorem 3 cannot easily be extended to this setting as it allows for arbitrary (collections of) RP functions to be used, including those that might directly test for normality. We do not pursue this further here but note that knowing errors are non-Gaussian can be helpful for designing a different objective function to use with ℓ_1 penalisation that may be more efficient for estimation. We also note that one could in principle extend the results of Theorems 4 and 5 to allow for non-Gaussian error distributions under the null. The result for a single variable follows via the central limit theorem, but the uniformity of variables in S^c requires the deep results of Chernozhukov et al. [2014].

Nevertheless, it is desirable that a test for e.g. nonlinearity should not reject more often when a sparse linear model with non-normal errors holds. When non-Gaussian errors must be included in the null hypothesis, we recommend taking the simulated errors $\boldsymbol{\zeta}^{(b)}$ to be a sample with replacement from the original scaled residuals $\hat{\mathbf{R}}$.

Figures 7 and 8 are identical to Figures 3 and 4 but with exponential errors rather than Gaussian errors used in all simulations. Similarly Figures 9 and 10 use t-distributed errors with 3 degrees of freedom scaled to have variance 1. We use the nonparametric bootstrap approach described above. We see that type I error is very well controlled for RP tests across all the settings with the power also competitive.

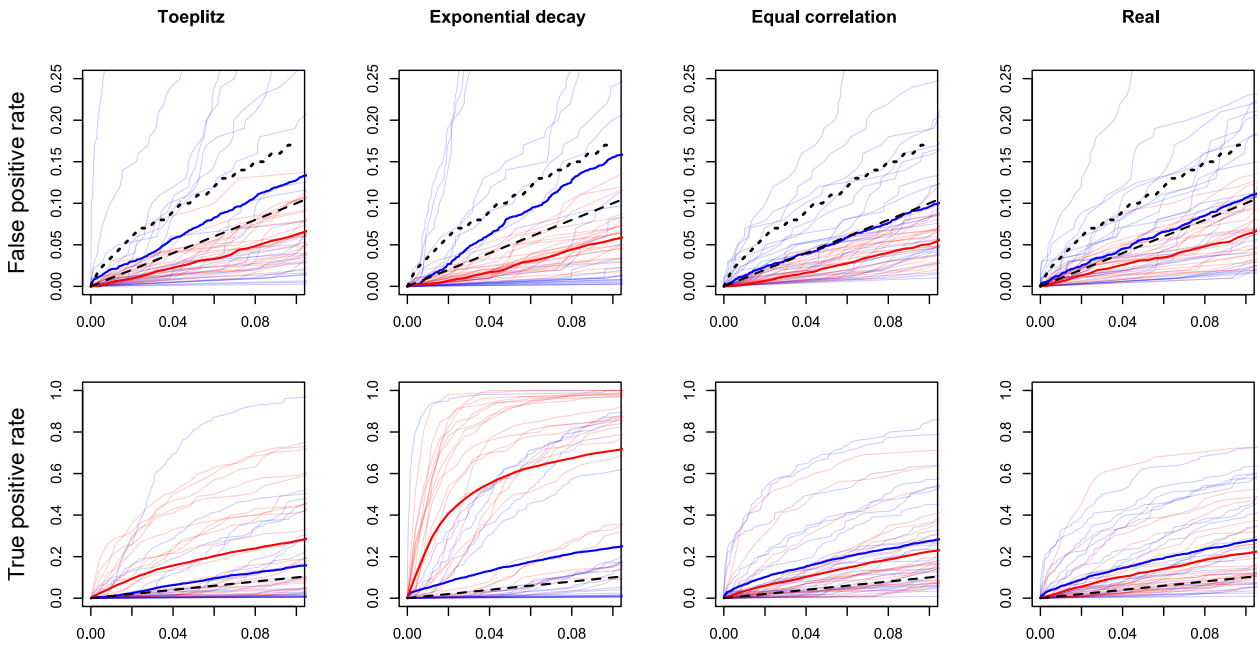

Figure 7: Testing significance of groups with exponential errors; the interpretation is similar to that of Figure 3.

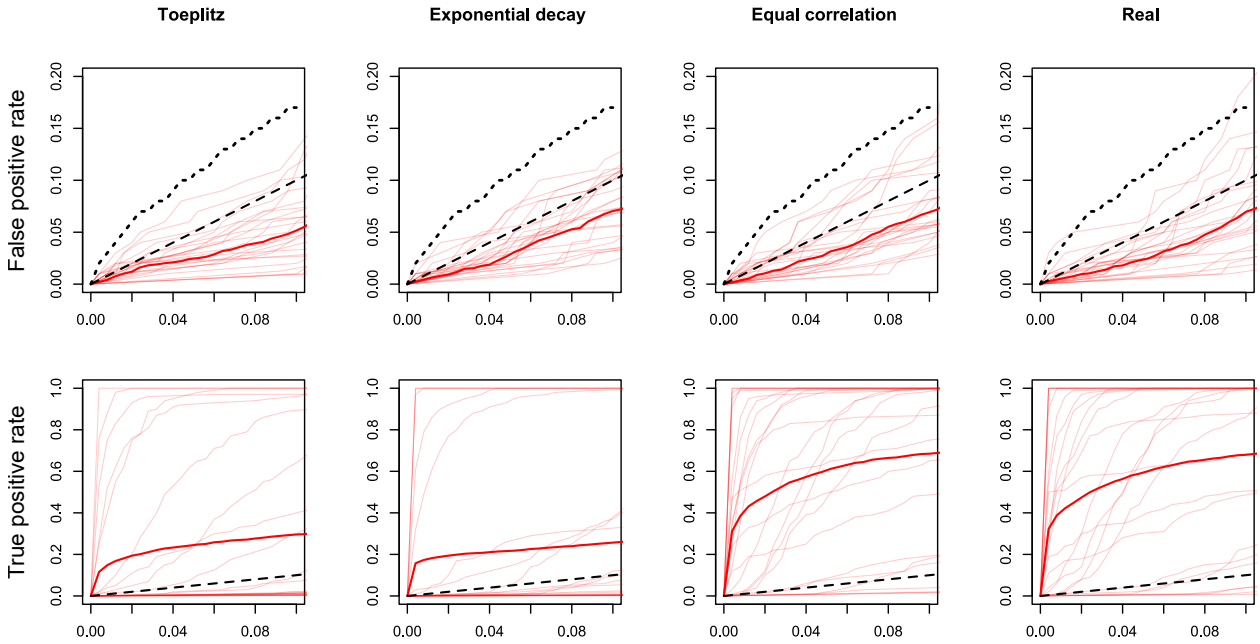

Figure 8: Testing for nonlinearity with exponential errors; the interpretation is similar to that of Figure 4.

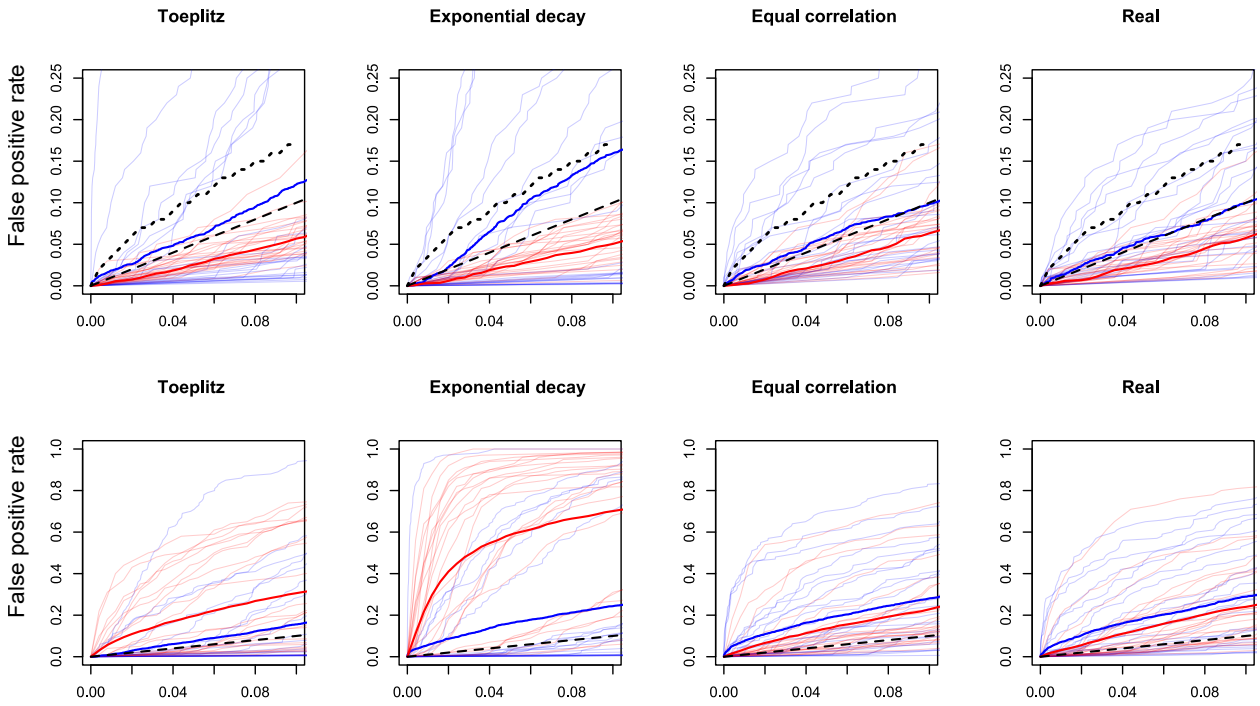

Figure 9: Testing significance of groups with t-distributed errors with 3 degrees of freedom; the interpretation is similar to that of Figure 3.

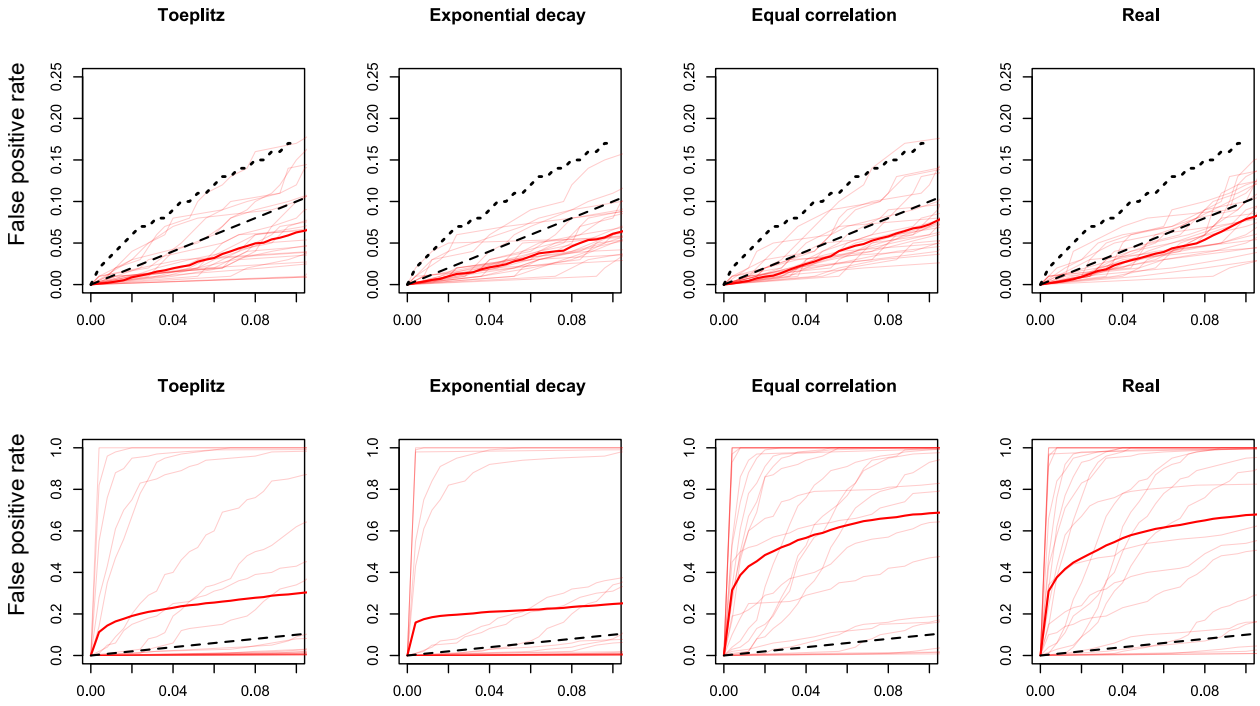

Figure 10: Testing for nonlinearity with t-distributed errors with 3 degrees of freedom; the interpretation is similar to that of Figure 4.

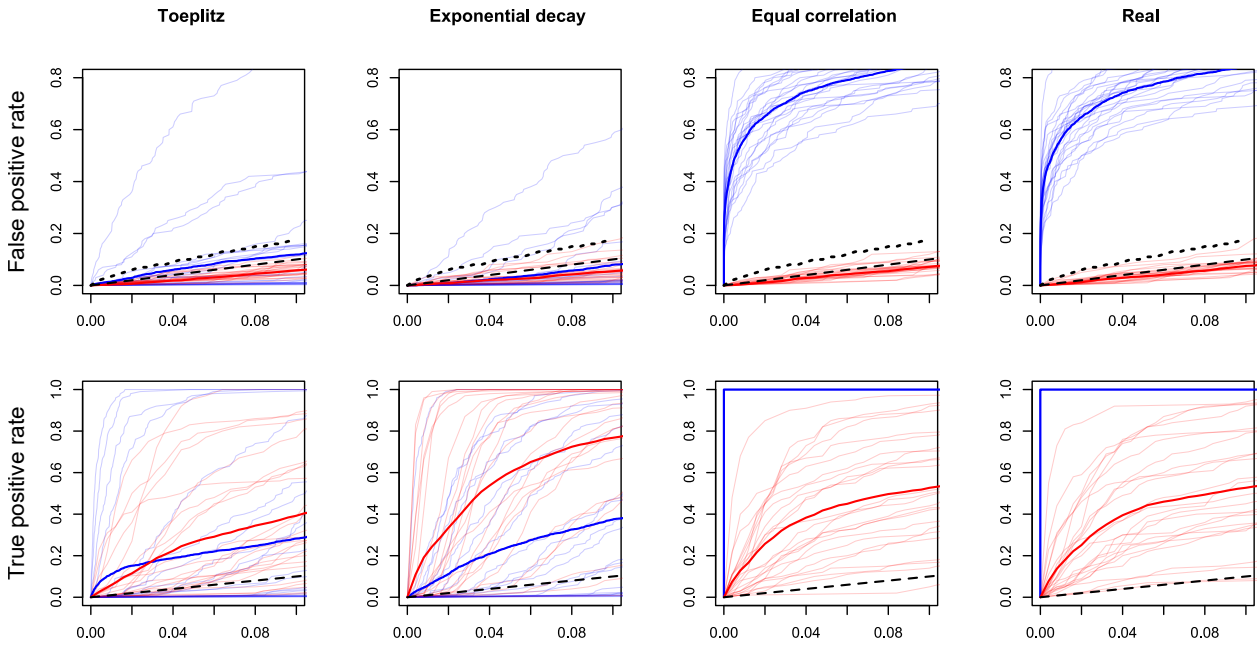

Figure 11: Testing significance of groups when coefficients are chosen according to (11); the interpretation is similar to that of Figure 3.

C Additional numerical results

In this section we present additional numerical results of the same format as those in Section 5 in the main paper, but where the nonzero coefficients given active variables $S = \{j_1, \ldots, j_{12}\}$ are chosen as follows:

$$\beta_{j_k} \propto \frac{1}{\sqrt{k}}. \tag{11}$$

The coefficients are then scaled to have an ℓ_1-norm of 12. Modulo these modifications, Figures 11 and 12 are exactly analogous to Figures 3 and 4 respectively.

We see the results are very much in line with those of Section 5. Note that the reduced power of RP tests compared to the debiased Lasso in the equal correlation and real design settings of Figure 11 is due to the poor calibration of the debiased approach, which here tends to greatly exceeds its nominal level.

D Uses of RP tests and interpretation of p-values

One use of RP tests is as a form of reassurance that Lasso-based inference is safe to use on the data at hand: a large p-value indicates a lack of evidence for the Lasso having poor performance. A small value on the other hand suggests a sparse linear model is inappropriate.

As we explain in this work however, the RP testing framework is general enough to include tests for significance of groups or individual predictors. It is important to note though that as with all p-values, low values in these cases can only indicate inadequacy of the null models considered; extra assumptions are required to draw conclusions, such as that a variable is significant, on the basis of having observed a low p-value.

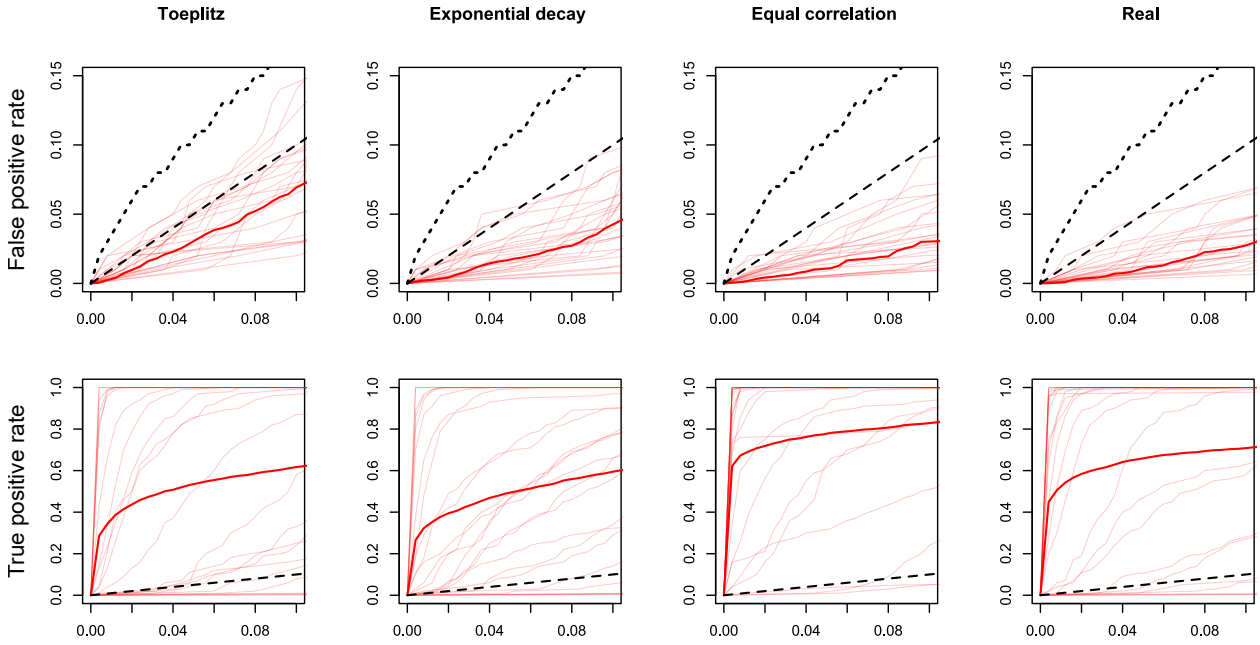

Figure 12: Testing for nonlinearity; the interpretation is similar to that of Figure 4.

E Proofs

In the proofs that follow, we will let $c_1, \ldots, c_4 \geq 0$ denote constants, which may change from line to line.

E.1 Proof of Theorem 2

In the following we suppress the dependence of $\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ on λ for notational simplicity. We know that every Lasso solution $\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ is characterised by the KKT conditions

$$\frac{1}{\sqrt{n}} \frac{\mathbf{X}^T[\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} + \sigma\boldsymbol{\zeta}]}{\|\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} + \sigma\boldsymbol{\zeta}\|_2} = \lambda\hat{\boldsymbol{\nu}},$$

where $\|\hat{\boldsymbol{\nu}}\|_\infty \leq 1$ and $\hat{\boldsymbol{\nu}}_{\hat{S}} = \operatorname{sgn}(\hat{\boldsymbol{\beta}}_{\hat{S}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}))$ with $\hat{S} = \{j : \hat{\beta}_j(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}) \neq 0\}$.

Now picking a particular Lasso solution $\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ in the case where it is not unique, let

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta}) = \check{\boldsymbol{\beta}} + \frac{\check{\sigma}}{\sigma}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}) - \boldsymbol{\beta}\}.$$

Note that when $\boldsymbol{\zeta} \in \Lambda_{\lambda,t}$, the upper bound on $\check{\sigma}$ ensures that we have $\operatorname{sgn}(\tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta})) = \operatorname{sgn}(\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}))$. Next observe that

$$\mathbf{X}\{\check{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta})\} + \check{\sigma}\boldsymbol{\zeta} = -\frac{\check{\sigma}}{\sigma}\mathbf{X}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta}) - \boldsymbol{\beta}\} + \check{\sigma}\boldsymbol{\zeta} = \frac{\check{\sigma}}{\sigma}[\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} + \sigma\boldsymbol{\zeta}], \tag{12}$$

so

$$\frac{1}{\sqrt{n}} \frac{\mathbf{X}^T[\mathbf{X}\{\check{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta})\} + \check{\sigma}\boldsymbol{\zeta}]}{\|\mathbf{X}\{\check{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta})\} + \check{\sigma}\boldsymbol{\zeta}\|_2} = \frac{1}{\sqrt{n}} \frac{\mathbf{X}^T[\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} + \sigma\boldsymbol{\zeta}]}{\|\mathbf{X}\{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})\} + \sigma\boldsymbol{\zeta}\|_2} = \lambda\hat{\boldsymbol{\nu}}(\boldsymbol{\beta}, \boldsymbol{\zeta}).$$

But this shows that $\tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta})$ satisfies the KKT conditions for $\hat{\boldsymbol{\beta}}(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta})$. Since Lasso fitted values are unique, we must have $\mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\zeta}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta})$. Now substituting into (12) finally shows that $\hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta}) = \hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ as required.

E.2 Results from Sun and Zhang [2012]

The proofs of Theorems 3–7 make use of Theorem 2 and Corollary 1 in Sun and Zhang [2012]. We re-state a subset of these results here for convenience. We have modified the notation in Sun and Zhang [2012] in order to avoid clashes with our own notation. Furthermore, we have replaced the sign-restricted cone invertibility factor $F_2(\xi, S)$ (equation 21 in Sun and Zhang [2012]) with its lower bound $F_2(\xi, S) \geq \phi^2(\xi)/(1+\xi)$ [Zhang and Zhang, 2012].

Consider the linear model setup of (1) though without any assumptions on the distribution of $\boldsymbol{\varepsilon}$ initially. For $\xi > 1$ and $\lambda > 0$, define

$$\mu(\lambda, \xi, \boldsymbol{\beta}, \mathbf{X}) = (\xi+1) \min_T \inf_{0<\nu<1} \max \left[\frac{\|\boldsymbol{\beta}_{T^c}\|_1}{\nu}, \frac{\lambda|T|/\{2(1-\nu)\}}{\kappa^2\{(\xi+\nu)/(1-\nu), T, \mathbf{X}\}} \right], \tag{13}$$

where the minimum is over $T \subset \{1, \ldots, p\}$ and $T^c = \{1, \ldots, p\} \setminus T$. Further let $\tilde{\sigma} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2/\sqrt{n}$ and $\tau^2 = \tau^2(\tilde{\sigma}, \lambda, \xi, \boldsymbol{\beta}, \mathbf{X}) = \lambda\mu(\tilde{\sigma}\lambda, \xi, \boldsymbol{\beta}, \mathbf{X})/\tilde{\sigma}$. Writing $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})$, define $\hat{\sigma} = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2/\sqrt{n}$.

Theorem 8 (Theorem 2 and Corollary 1 in Sun and Zhang [2012])**.** *Let*

$$\Lambda_1 = \left\{ \boldsymbol{\zeta} \in \mathbb{R}^n : \frac{\|\mathbf{X}^T\boldsymbol{\zeta}\|_\infty}{\sqrt{n}\|\boldsymbol{\zeta}\|_2} \leq (1-\tau^2)\lambda\frac{\xi-1}{\xi+1} \right\}. \tag{14}$$

When $\boldsymbol{\varepsilon} \in \Lambda_1$,

$$\begin{aligned} \max(1 - \hat{\sigma}/\tilde{\sigma}, 1 - \tilde{\sigma}/\hat{\sigma}) &\leq \tau^2, \\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\leq \mu(\tilde{\sigma}\lambda, \xi, \boldsymbol{\beta}, \mathbf{X})/(1-\tau^2) \\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 &\leq \frac{2\tilde{\sigma}\xi\sqrt{s}\lambda}{(1-\tau^2)\phi^2(\xi)}. \end{aligned}$$

Lemma 9. *Let* $n, p, m \in \mathbb{N}$ *with* $n \geq m \geq 3$. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I})$. *Let* $\mathbf{a} \in \mathbb{R}^m$ *and suppose* $0 < \eta < pe^{-(n-m)-2}$. *Then*

$$\mathbb{P}(\mathbf{a}^T\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2 > \sqrt{2\log(p/\eta)/n}\,\|\mathbf{a}\|_2) \leq \frac{1}{p}\frac{(1+r_m)\eta}{\sqrt{\pi\log(p/\eta)}}$$

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

Proof. We follow the proof of part (ii) of Theorem 2 of Sun and Zhang [2012]. Let $z = \mathbf{a}^T\boldsymbol{\varepsilon}/(\|\mathbf{a}\|_2\|\boldsymbol{\varepsilon}\|_2)$. Then $z/\sqrt{(1-z^2)/(m-1)}$ follows a t-distribution with $m-1$ degrees of freedom. The only change we need to make in the aforementioned proof is to note that

$$\eta < pe^{-(n-m)-2} \Rightarrow m - 2 - \log(p/\eta) \geq n - 2\log(p/\eta).$$

and modify (A8) in Sun and Zhang [2012] appropriately. □

E.3 Proof of Theorem 3

Without loss of generality assume $S = \{1, \ldots, s\}$. Let $\lambda_0 = 2\sqrt{\log(p/\eta)/n}$ and let $\mathbf{P}$ denote the orthogonal projection on to $\mathbf{X}_S$. Define the following subsets of $\mathbb{R}^n$:

$$\Lambda_1 = \left\{\boldsymbol{\zeta} : \max_j \frac{|\mathbf{X}_j^T(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}|}{\|(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}\|_2} \leq \lambda_0 \|(\mathbf{I}-\mathbf{P})\mathbf{X}_j\|_2/\sqrt{2}\right\}$$
$$\Lambda_2 = \left\{\boldsymbol{\zeta} : \max_j \frac{|\mathbf{X}_j^T\mathbf{P}\boldsymbol{\zeta}|}{\|\boldsymbol{\zeta}\|_2} \leq \lambda_0 \|\mathbf{P}\mathbf{X}_j\|_2/\sqrt{2}\right\}$$
$$\Lambda_3 = \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\|_2/\sqrt{n} \leq \sqrt{2}\}.$$

Let $\Lambda = \Lambda_1 \cap \Lambda_2 \cap \Lambda_3$ and let $\Omega = \{\boldsymbol{\varepsilon} \in \Lambda\}$. Lemma 9 shows that on the event Ω we have the following properties.

(i) $\mathrm{sgn}(\boldsymbol{\beta}'_S) = \mathrm{sgn}(\boldsymbol{\beta}_S)$ and $\min_{j\in S} \beta'_j/\beta_j > 1 - 1/(2\sqrt{2})$.

(ii) $\check{S}^{(s)} = S$ and $\mathrm{sgn}(\check{\boldsymbol{\beta}}^{(s)}) = \mathrm{sgn}(\boldsymbol{\beta}_S)$.

(iii) $2\sqrt{2}\sigma \min_{j\in S} \check{\beta}_j^{(s)}/\beta_j > \check{\sigma}^{(s)}$.

In the following, we suppress dependence on λ. Let $\check{\boldsymbol{\beta}} \in \mathbb{R}^p$ be $\check{\boldsymbol{\beta}}^{(s)}$ with $p-s$ zeroes added: $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}^{(s)}, 0, \ldots, 0)$ and let $\check{\sigma} = \check{\sigma}^{(s)}$. Note that by Theorem 2, on Ω we know that $\hat{\mathbf{R}}(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta}) = \hat{\mathbf{R}}(\boldsymbol{\beta}, \sigma\boldsymbol{\zeta})$ for all $\boldsymbol{\zeta} \in \Lambda$. Moreover, Lemma 11 shows that conditional on Ω, $\hat{\mathbf{R}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})$ and $\mathbf{X}\check{\boldsymbol{\beta}}/\check{\sigma}$ are independent. Write $\hat{\mathbf{R}}^{(b)} = \hat{\mathbf{R}}_\lambda(\check{\boldsymbol{\beta}}, \check{\sigma}\boldsymbol{\zeta}^{(b)})$ for $b = 1, \ldots, B$ and let $\hat{\mathbf{R}}^{(0)} = \hat{\mathbf{R}}_\lambda(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon})$. We see that conditional on Ω, $\{\hat{\mathbf{R}}^{(b)}\}_{b=0}^B$ are independent. Also $\hat{\mathbf{R}}^{(0)}\mathbb{1}_{\{\boldsymbol{\varepsilon}\in\Lambda\}}, \{\hat{\mathbf{R}}^{(b)}\mathbb{1}_{\{\boldsymbol{\zeta}^{(b)}\in\Lambda\}}\}_{b=1}^B$ are independent and identically distributed.

Let $\tilde{Q}_b$, $b = 0, \ldots, B$ be derived as in Section 3 by applying the function $\tilde{Q}$ to appropriate functions of scaled residuals $\{\hat{\mathbf{R}}^{(b)}\}_{b=0}^B$. Recall that $Q = \sum_{b=0}^B \mathbb{1}_{\{\tilde{Q}_b \geq \tilde{Q}_0\}}/(B+1)$. Let $R = \{b : \boldsymbol{\zeta}^{(b)} \in \Lambda\}$ and note that letting $\delta = \mathbb{P}(\Omega^c)$ we have $|R| \sim \mathrm{Bin}(1-\delta, B)$. We have

$$\begin{aligned}\mathbb{P}(Q \leq x) &\leq \mathbb{P}(Q \leq x, \Omega) + \delta \\ &= (1-\delta)\mathbb{E}\{\mathbb{P}(Q \leq x | R, \Omega)\} + \delta.\end{aligned}$$

Lemma 10 gives the required bound on δ; it only remains to show the first term on the RHS is at most x. From the above, conditional on R and the event Ω, $\{\tilde{Q}_0, \{\tilde{Q}_b\}_{b\in R}\}$ are exchangeable. Now there can be at most $\lfloor x(B+1)\rfloor$ values $b \in \{0, \ldots, B\}$ with $\sum_{b'\neq b} \mathbb{1}_{\{\tilde{Q}_{b'} \geq \tilde{Q}_b\}}/(B+1) \leq x$. This entails that

$$\begin{aligned}(|R|+1)\mathbb{P}(Q \leq x | R, \Omega) &\leq \lfloor x(B+1)\rfloor \\ \mathbb{P}(Q \leq x | R, \Omega) &\leq \frac{x(B+1)}{|R|+1}.\end{aligned}$$

Therefore

$$\begin{aligned}(1-\delta)\mathbb{E}\{\mathbb{P}(Q \le x|R,\Omega)\} &\le x(1-\delta)\mathbb{E}\left(\frac{B+1}{|R|+1}\right)\\ &= x\sum_{r=0}^{B}(1-\delta)^{r+1}\delta^{B-r}\binom{B}{r}\frac{B+1}{r+1}\\ &= x\sum_{r=1}^{B+1}(1-\delta)^{r}\delta^{B+1-r}\binom{B+1}{r}\\ &= x(1-\delta^{B+1}) \le x.\end{aligned}$$

E.4 Proofs of Theorems 4 and 5

The proofs of Theorems 4 and 5 rest on the following decomposition of T_k and an analogous one for T_k^*:

$$T_k = \frac{1}{\hat{\sigma}_k}\|\mathbf{W}_k\|_2\beta_k + \frac{\sigma}{\hat{\sigma}_k}Z_k + \frac{1}{\hat{\sigma}_k}\delta_k \tag{15}$$

where

$$\hat{\sigma}_k = \|\mathbf{y} - \mathbf{X}_{-k}\hat{\boldsymbol{\Theta}}_k\|_2/\sqrt{n}.$$

The first term in (15) is zero for $k \in N$ and

$$\delta_k = \frac{\mathbf{W}_k^T}{\|\mathbf{W}_k\|_2}\mathbf{X}_{-k}(\boldsymbol{\Theta}_k - \hat{\boldsymbol{\Theta}}_k), \tag{16}$$

$$Z_k = \frac{\mathbf{W}_k^T\boldsymbol{\varepsilon}}{\|\mathbf{W}_k\|_2} \sim \mathcal{N}(0,1). \tag{17}$$

The main term we have to control is δ_k. The result of Sun and Zhang [2012] shows that $\|\boldsymbol{\Theta}_k - \hat{\boldsymbol{\Theta}}_k\|_1$ is small with high probability. Next appealing to the KKT conditions for the square-root Lasso, we obtain

$$\|\mathbf{X}_{-k}^T\mathbf{W}_k\|_\infty/\|\mathbf{W}_k\|_2 \le \sqrt{n}\gamma. \tag{18}$$

Hölder's inequality then gives

$$|\delta_k| \le \sqrt{n}\gamma\|\boldsymbol{\Theta}_k - \hat{\boldsymbol{\Theta}}_k\|_1, \tag{19}$$

which forms the basis of the proofs.

E.4.1 Proof of Theorem 4

In view of Lemma 14 applied with $\mathcal{F}_n$ simply a constant, for the first part we need only show that $\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N}\delta_k \xrightarrow{p} 0$ and $\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N}|\hat{\sigma}_k - \sigma| \xrightarrow{p} 0$. First note that

$$\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N}\mu(\lambda,\xi,\boldsymbol{\beta}_{-k},\mathbf{X}_{-k}) \le \sup_{\boldsymbol{\beta}\in\mathscr{B}}\mu(\lambda,\xi,\boldsymbol{\beta},\mathbf{X}) \to 0$$

in view of Lemma 12 and consequently (as clearly $\lambda \to 0$)

$$\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N}\tau^2(\tilde{\sigma},\lambda,\xi,\boldsymbol{\beta}_{-k},\mathbf{X}_{-k}) \le \sup_{\boldsymbol{\beta}\in\mathscr{B}}\tau^2(\tilde{\sigma},\lambda,\xi,\boldsymbol{\beta},\mathbf{X}) \to 0.$$

Now writing $\tau^2 = \sup_{\boldsymbol{\beta}\in\mathscr{B}} \tau^2(\tilde{\sigma}, \lambda, \xi, \boldsymbol{\beta}, \mathbf{X})$ for convenience, we see that for n sufficiently large it must be the case that $(1-\tau^2)\lambda(\xi-1)/(\xi+1) > \sqrt{2\log(p)/n}$. By Theorem 8 there is a sequence of events with probability tending to 1 on which we have

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \max\left(1 - \frac{\hat{\sigma}_k}{\tilde{\sigma}}, 1 - \frac{\tilde{\sigma}}{\hat{\sigma}_k}\right) \leq \tau^2, \tag{20}$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_1 \leq \frac{(\xi+1)\tilde{\sigma}\lambda s}{(1-\tau^2)\kappa^2(\xi, S)}, \tag{21}$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}} \max\left(1 - \frac{\check{\sigma}}{\tilde{\sigma}}, 1 - \frac{\tilde{\sigma}}{\check{\sigma}}\right) \leq \tau^2, \tag{22}$$

where $\tilde{\sigma} = \sigma\|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$. From Lemma 13 we know that $\tilde{\sigma} \xrightarrow{p} \sigma$, so in particular we have $\sup_{\boldsymbol{\beta}\in\mathscr{B},,k\in N} |\hat{\sigma}_k - \sigma| \xrightarrow{p} 0$. Thus also, on a sequence of events with probability tending to 1, applying (19) to (21) we have

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} |\delta_k| \leq c_1\sigma\frac{\sqrt{\log(p)}\lambda s}{\kappa^2(\xi, S)} \to 0,$$

which completes the proof of the first part.

Turning to the bootstrap results, we know that that on a sequence of events of the form $\Omega_n = \{\boldsymbol{\varepsilon} \in \Delta_n\}$ with probability tending to 1, we have

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \sqrt{n}\gamma\|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_1 \geq \sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \sqrt{n}\gamma\|\hat{\boldsymbol{\Theta}}_{k,N_k}\|_1 \to 0 \text{ and } \sup_{\boldsymbol{\beta}\in\mathscr{B}} |\check{\sigma} - \sigma| \to 0.$$

Here N_k corresponds to the set of noise components of $\boldsymbol{\Theta}_k$. Thus on Ω_n, by Lemma 12, we have

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \sqrt{n}\gamma\mu(c_1\lambda, \xi^*, \hat{\boldsymbol{\Theta}}_k, \mathbf{X}_{-k}) \to 0 \tag{23}$$

for any fixed $c_1 > 0$ and some $\xi^* > (A_1+1)/(A_1-1)$. Let δ_k^* and $\hat{\sigma}_k^*$ be the bootstrap equivalents of δ_k and $\hat{\sigma}_k$ respectively. By applying Lemma 14 now with $\mathcal{F}_n = \boldsymbol{\varepsilon}$, we see that it is enough to show that on Ω_n, for all $\eta > 0$ we have

$$\mathbb{P}(\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} |\delta_k^*| > \eta|\boldsymbol{\varepsilon}) \to 0 \text{ and } \mathbb{P}(\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} |\hat{\sigma}_k^* - \sigma| > \eta|\boldsymbol{\varepsilon}) \to 0.$$

For this it is sufficient to exhibit a sequence $\Omega_n^* = \{\boldsymbol{\varepsilon}^* \in \Delta_n^*\}$ whose probability tends to 1 such that on $\Omega_n \cap \Omega_n^*$ we have

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} |\delta_k^*| \to 0 \tag{24}$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} |\hat{\sigma}_k^* - \sigma| \to 0. \tag{25}$$

Let $\tilde{\sigma}^* = \check{\sigma}\|\boldsymbol{\varepsilon}^*\|_n/\sqrt{n}$. Define

$$\tau_*^2 = \sup_{\boldsymbol{\beta}\in\mathscr{B}, k\in N} \tau^2(\tilde{\sigma}^*, \lambda, \xi^*, \hat{\boldsymbol{\Theta}}_k, \mathbf{X}_{-k}).$$

Provided $\tau_*^2 \to 0$ we have $(1-\tau_*^2)\lambda(\xi^*-1)/(\xi^*+1) > \sqrt{2\log(p)/n}$ for n sufficiently large. This gives us the equivalent of (20) and (21) for n sufficiently large:

$$\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N} \max\left(1-\frac{\hat{\sigma}_k^*}{\tilde{\sigma}^*}, 1-\frac{\tilde{\sigma}^*}{\hat{\sigma}_k^*}\right) \leq \tau_*^2, \tag{26}$$

$$\sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N} \|\hat{\boldsymbol{\Theta}}_k^* - \boldsymbol{\Theta}_k^*\|_1 \leq c_2 \sup_{\boldsymbol{\beta}\in\mathscr{B},k\in N} \mu(\tilde{\sigma}^*\lambda, \xi^*, \hat{\boldsymbol{\Theta}}_k, \mathbf{X}_{-k}). \tag{27}$$

By Lemma 13, conditional on $\boldsymbol{\varepsilon}$, $\sup_{\boldsymbol{\beta}\in\mathscr{B}} |\tilde{\sigma}^* - \check{\sigma}| \xrightarrow{p} 0$. Thus τ_*^2 tending to 0 and therefore also (26) and (27) occur on a sequence of events with probability tending to 1, $\Omega_n^* \cap \Omega_n$, where Ω_n^* is of the form $\{\boldsymbol{\varepsilon}^* \in \Delta_n^*\}$. As on $\Omega_n^* \cap \Omega_n$, $\sup_{\boldsymbol{\beta}\in\mathscr{B}} |\check{\sigma} - \sigma| \to 0$, (22) gives (25). Applying (19) to (27) and (23) then gives (24), which completes the proof.

E.4.2 Proof of Theorem 5

The proof proceeds similarly to that of Theorem 4. For the first result, we use Lemma 14 with $\mathcal{F}_n$ simply constant. Thus it suffices to show $\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} |\hat{\sigma}_k - \sigma| \xrightarrow{p} 0$, $\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} |\delta_k| \xrightarrow{p} 0$ and $\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} |\beta_k| \|\mathbf{W}_k\|_2 |{\sigma'_k}^{-1} - \hat{\sigma}_k^{-1}| \xrightarrow{p} 0$ where $\sigma'_k = \sqrt{\sigma^2 + \|\mathbf{W}_k\|_2^2\beta_k^2/n}$. To this end, first note that by Lemma 12, for some $\xi' > (A_1+1)/(A_1-1)$ we have $\log(p)\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} \mu(\lambda, \xi', \boldsymbol{\Theta}_k, \mathbf{X}_{-k}) := \log(p)\mu_k \to 0$. Let

$$\tilde{\sigma}_k = \frac{1}{\sqrt{n}}\|y - \mathbf{X}_{-k}\boldsymbol{\Theta}_k\|_2 = \frac{1}{\sqrt{n}}\|\sigma\boldsymbol{\varepsilon} + \beta_k\mathbf{W}_k\|_2.$$

Then

$$\tilde{\sigma}_k^2 = \frac{1}{n}\left(\sigma^2\|\boldsymbol{\varepsilon}\|_2^2 + \beta_k^2\|\mathbf{W}_k\|_2^2 + 2\sigma\beta_k\|\mathbf{W}_k\|_2 Z_k\right)$$

where Z_k is defined as in (17). Since $\beta_k\|\mathbf{W}_k\|_2/\sqrt{n} \to 0$, we have that $\tilde{\sigma}_k \xrightarrow{p} \sigma$ by Lemma 13. For later use we also note that

$$\sqrt{n}(\tilde{\sigma}_k^2 - {\sigma'_k}^2) = \sqrt{n}(\sigma^2\|\boldsymbol{\varepsilon}\|_2^2/n - 1) + 2\sigma\beta_k\|\mathbf{W}_k\|_2 Z_k/\sqrt{n} = O_P(1). \tag{28}$$

by the central limit theorem and as $\beta_k\|\mathbf{W}_k\|_2/\sqrt{n} \to 0$. Thus we have $\tau_k^2 := \sup_{\boldsymbol{\beta}\in\mathscr{B}_k} \tau^2(\tilde{\sigma}_k, \lambda, \xi', \boldsymbol{\Theta}_k, \mathbf{X}_{-k}) \xrightarrow{p} 0$. Note that by (18),

$$\frac{1}{n}\beta_k\|\mathbf{X}_{-k}^T\mathbf{W}_k\|_\infty \leq \frac{1}{\sqrt{n}}\beta_k\gamma\|\mathbf{W}_k\|_2 = o\sqrt{\frac{\log(p)}{n}}.$$

Therefore

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\frac{\|\mathbf{X}_{-k}^T(\sigma\boldsymbol{\varepsilon} + \beta_k\mathbf{W}_k)\|_\infty/n}{\tilde{\sigma}_k} \geq (1-\tau_k^2)\lambda\frac{\xi-1}{\xi+1}\right) \\
\geq &\lim_{n\to\infty} \mathbb{P}\left(\frac{\|\mathbf{X}_{-k}^T\boldsymbol{\varepsilon}\|_\infty/n}{\|\boldsymbol{\varepsilon}\|_2/\sqrt{n}} \leq \sqrt{2\log(p)/n}\right) = 1.
\end{aligned}$$

Thus by Theorem 8 we have that on a sequence of events $\Omega_n = \{\boldsymbol{\varepsilon} \in \Delta_n\}$ with probability tending to 1, $|\tilde{\sigma}_k - \sigma| \to 0$, $\sqrt{n}\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} |\hat{\sigma}_k - \tilde{\sigma}_k| \to 0$, and $\log(p)\sup_{\boldsymbol{\beta}\in\mathscr{B}_k} \|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_1 \to 0$. The latter in

conjunction with (19) shows $\sup_{\boldsymbol{\beta}\in\mathscr{B}_k}|\delta_k| \xrightarrow{p} 0$. That $\sup_{\boldsymbol{\beta}\in\mathscr{B}_k}|\beta_k|\|\mathbf{W}_k\|_2|\sigma_k'^{-1} - \hat{\sigma}_k^{-1}| \xrightarrow{p} 0$ follows from (28).

Now we derive the result concerning the bootstrap test statistic. Note that on Ω_n,

$$\sqrt{n}\gamma \sup_{\boldsymbol{\beta}\in\mathscr{B}_k}\|\hat{\boldsymbol{\Theta}}_{k,T^c}\|_1 \leq \sqrt{n}\gamma\{\|\boldsymbol{\Theta}_{k,T^c}\|_1 + \sup_{\boldsymbol{\beta}\in\mathscr{B}_k}\|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_1\} \to 0, \tag{29}$$

and so also $\sqrt{n}\gamma\sup_{\boldsymbol{\beta}\in\mathscr{B}_k}\mu(c_1\lambda, \xi^*, \hat{\boldsymbol{\Theta}}_k, \mathbf{X}_{-k}) \to 0$ for any fixed $c_1 > 0$ and some $\xi^* > (A_1 + 1)/(A_1 - 1)$. The rest of the proof then proceeds exactly as the proof for the bootstrap statistic in Theorem 4.

E.5 Proof of Corollary 6

The proof is essentially identical to that of Theorem 5, but with $\beta_k\mathbf{W}_k$ replaced by $\mathbf{w}_\gamma$ and $\boldsymbol{\Theta}_k$ replaced by $\boldsymbol{\beta} + \boldsymbol{\psi}_\gamma$.

E.6 Proof of Theorem 7

By Corollary 1 of Raskutti et al. [2010], with probability tending to 1 we have $\kappa(\xi) \geq \phi(\xi) > \phi/8 > 0$. By Lemma 15 and the fact that $\sqrt{\mathbb{E}(f_1^2)} = o(n^{-1/6})$ we see that $\|\mathbf{f}\|_2 = o_P(n^{1/3})$. Indeed, we have $\mathbb{P}(\|\mathbf{f}\|_2/n^{1/3} \leq n^{1/6}\sqrt{2E(f_1^2)}) \to 1$ as $n \to \infty$. Also, with probability tending to 1,

$$\frac{1}{n}\|\mathbf{X}^T\mathbf{f}\|_\infty \leq c_1\frac{\sqrt{\log(p)}}{n^{1/3}}\cdot\frac{1}{\sqrt{n}}\|\mathbf{f}\|_2 \leq c_1\frac{\sqrt{\log(p)}}{n^{1/3}}\cdot o(n^{-1/6}) = o(\sqrt{\log(p)/n}),$$

i.e. $\|\mathbf{X}^T\mathbf{f}\|_\infty/n = o_P(\sqrt{\log(p)/n})$. We now temporarily make the dependence of $\mathbf{Z}$ and p on n explicit by writing $\mathbf{Z}^{(n)}$ and p_n respectively, in order to explain the structure of the argument to follow. From the above, we have a sequence of sets $\Lambda_n \subseteq \mathbb{R}^{n\times p_n}$ for which $\mathbb{P}(\mathbf{Z}^{(n)} \in \Lambda_n) \to 1$, and on the respective events $\|\mathbf{f}\|_2 = o(n^{1/3})$ and $\frac{1}{n}\|\mathbf{X}^T\mathbf{f}\|_\infty = o(\sqrt{\log(p)/n})$ (uniformly). Let $\hat{\sigma} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}'\|_2/\sqrt{n}$. In relation to Corollary 6, we will take $\gamma = c_1\sqrt{\log(p)}/n^{1/3}$ sufficiently large such that the Lasso regression of $\mathbf{f}$ on $\mathbf{X}$ produces the zero vector. By Lemma 14, it suffices to show that for each $\eta > 0$,

$$\begin{aligned}
\sup_{\mathbf{Z}^{(n)}\in\Lambda_n}\mathbb{P}\big(\sup_{\boldsymbol{\beta}\in\mathscr{B}}|\hat{\sigma} - \sigma| > \eta \,|\, \mathbf{Z}^{(n)}\big) &\to 0 \\
\sup_{\mathbf{Z}^{(n)}\in\Lambda_n}\mathbb{P}\big(\sup_{\boldsymbol{\beta}\in\mathscr{B}}\|\mathbf{X}^T\mathbf{f}\|_\infty\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1/\|\mathbf{f}\|_2 > \eta \,|\, \mathbf{Z}^{(n)}\big) &\to 0 \\
\sup_{\mathbf{Z}^{(n)}\in\Lambda_n}\mathbb{P}\big(\sup_{\boldsymbol{\beta}\in\mathscr{B}}\|\mathbf{f}\|_2|\sigma'^{-1} - \hat{\sigma}^{-1}| > \eta \,|\, \mathbf{Z}^{(n)}\big) &\to 0.
\end{aligned} \tag{30}$$

Here $\sigma' := \sqrt{\sigma^2 + \|\mathbf{f}\|_2^2/n}$. The remainder of the argument to arrive at the first result is essentially identical to that in Theorem 5, with $\mathbf{f}$ playing the role of $\beta_k\mathbf{W}_k$, and with the probabilities being conditional on $\mathbf{Z}^{(n)}$. The only difference is that in place of (28) (which leads to the equivalent of (30)), we have

$$n^{1/3}|\tilde{\sigma}^2 - \sigma'^2| = |n^{1/3}\sigma^2(\|\boldsymbol{\varepsilon}\|_2^2/n - 1) + 2\sigma\mathbf{f}^T\boldsymbol{\varepsilon}/n^{2/3}|,$$

where $\tilde{\sigma} = \|\sigma\boldsymbol{\varepsilon} + \mathbf{f}\|_2/\sqrt{n}$. Since for all $\mathbf{Z}^{(n)} \in \Lambda_n$, $\|\mathbf{f}\|_2 = o(n^{1/3})$ uniformly, it is straightforward to show that

$$\sup_{\mathbf{Z}^{(n)} \in \Lambda_n} \mathbb{P}\big(\sup_{\boldsymbol{\beta} \in \mathscr{B}} \|\mathbf{f}\|_2 |\sigma' - \tilde{\sigma}| > \eta \mid \mathbf{Z}^{(n)} \big) \to 0.$$

for any $\eta > 0$, which then leads to (30).

The bootstrap result is simpler. We know there is a sequence of events depending only on $(\mathbf{Z}, \boldsymbol{\varepsilon})$ on which

$$\sup_{\boldsymbol{\beta} \in \mathscr{B}} \|\boldsymbol{\beta} - \check{\boldsymbol{\beta}}\|_1 \leq c_2 \sigma s \sqrt{\log(p)/n}/\phi^2(\xi).$$

Thus on the same sequence of events

$$\sup_{\boldsymbol{\beta} \in \mathscr{B}} \sqrt{n} \cdot c_1 \sqrt{\log(p)}/n^{1/3} \cdot \mu(\lambda, \xi, \check{\boldsymbol{\beta}}, \mathbf{X}) \to 0,$$

from which the result follows by arguing along the lines of the second part of the proof of Theorem 4.

E.7 Technical lemmas

Lemma 10. *Consider the setup of Theorem 3 and its proof. Recall that*

$$\begin{aligned}
\Lambda_1 &= \bigg\{ \boldsymbol{\zeta} : \max_j \frac{|\mathbf{X}_j^T (\mathbf{I} - \mathbf{P})\boldsymbol{\zeta}|}{\|(\mathbf{I} - \mathbf{P})\boldsymbol{\zeta}\|_2} \leq \lambda_0 \|(\mathbf{I} - \mathbf{P})\mathbf{X}_j\|_2/\sqrt{2} \bigg\}, \\
\Lambda_2 &= \bigg\{ \boldsymbol{\zeta} : \max_j \frac{|\mathbf{X}_j^T \mathbf{P}\boldsymbol{\zeta}|}{\|\boldsymbol{\zeta}\|_2} \leq \lambda_0 \|\mathbf{P}\mathbf{X}_j\|_2/\sqrt{2} \bigg\}, \\
\Lambda_3 &= \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\|_2/\sqrt{n} \leq \sqrt{2}\}, \\
\Lambda &= \Lambda_1 \cap \Lambda_2 \cap \Lambda_3,
\end{aligned}$$

with $\lambda_0 = 2\sqrt{\log(p/\eta)/n}$. *On the event* $\Omega = \{\boldsymbol{\varepsilon} \in \Lambda\}$ *we have the following properties.*

(i) $\operatorname{sgn}(\boldsymbol{\beta}'_S) = \operatorname{sgn}(\boldsymbol{\beta}_S)$ *and* $\min_{j \in S} \beta'_j/\beta_j > 1 - 1/(2\sqrt{2})$.

(ii) $\hat{S}^{(s)} = S$ *and* $\operatorname{sgn}(\check{\boldsymbol{\beta}}^{(s)}) = \operatorname{sgn}(\boldsymbol{\beta}_S)$.

(iii) $2\sqrt{2}\sigma \min_{j \in S} \check{\beta}_j^{(s)}/\beta_j > \check{\sigma}^{(s)}$.

Furthermore

$$\mathbb{P}(\Omega) \geq 1 - \frac{2(1 + r_{n-s})\eta}{\sqrt{\pi \log(p/\eta)}} - e^{-n/8}. \tag{31}$$

Proof. First we bound $\mathbb{P}(\Omega)$. From Lemma 9 and the union bound we have $\mathbb{P}(\boldsymbol{\varepsilon} \in \Lambda_2) \leq (1 + r_n)\eta/\sqrt{\pi \log(p/\eta)}$. Next note that

$$\frac{\mathbf{X}_j^T (\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon}}{\|(\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon}\|_2 \|(\mathbf{I} - \mathbf{P})\mathbf{X}_j\|_2} \overset{d}{=} \frac{\mathbf{a}^T \boldsymbol{\zeta}}{\|\boldsymbol{\zeta}\|_2 \|\mathbf{a}\|_2}$$

where $\mathbf{a} \in \mathbb{R}^{n-s}$ and $\overset{d}{=}$ denotes equality in distribution. Thus Lemma 9 gives $\mathbb{P}(\boldsymbol{\varepsilon} \in \Lambda_1) \leq (1 + r_{n-s})\eta/\sqrt{\pi \log(p/\eta)}$. Finally, Lemma 13 gives $\mathbb{P}(\boldsymbol{\varepsilon} \in \Lambda_3) \geq 1 - e^{-n/8}$.

Now turning to (i), observe that as $\|(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}\|_2 \le \|\boldsymbol{\zeta}\|_2$ and

$$(\|\mathbf{P}\mathbf{X}_j\|_2 + \|(\mathbf{I}-\mathbf{P})\mathbf{X}_j\|_2)/\sqrt{2} \le \|\mathbf{X}_j\|_2 = \sqrt{n},$$

we have

$$\Omega \subseteq \left\{ \frac{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty}{\sqrt{n}\|\boldsymbol{\varepsilon}\|_2} \le \lambda_0 \right\}.$$

Since

$$\frac{1}{2}\frac{(\xi+1)\lambda s}{\kappa^2(\xi,S)} \ge \mu(\lambda,\xi,\boldsymbol{\beta},\mathbf{X}),$$

(8) ensures that $(1-\tau^2)\lambda(\xi-1)/(\xi+1) \ge \lambda_0$. Note also that (8) gives $\tau^2 \le 1/5$. Thus by Theorem 8, the fact that $\boldsymbol{\varepsilon} \in \Lambda_3$ and (9) we have

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2 \le \frac{2\tilde{\sigma}\xi\sqrt{s}\lambda}{(1-\tau^2)\phi^2(\xi)} \le \frac{\tilde{\sigma}}{4\sigma}\min_{j\in S}|\beta_j| \le \frac{1}{2\sqrt{2}}\min_{j\in S}|\beta_j|, \tag{32}$$

where $\tilde{\sigma} = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$. This shows $\mathrm{sgn}(\boldsymbol{\beta}'_S) = \mathrm{sgn}(\boldsymbol{\beta}_S)$. Next

$$\min_{j\in S}\frac{\beta'_j}{\beta_j} \ge 1 - \frac{\|\boldsymbol{\beta}'-\boldsymbol{\beta}\|_2}{\min_{j\in S}|\beta_j|} \ge 1 - \frac{1}{2\sqrt{2}},$$

which shows (i). Now

$$\begin{aligned}\min_{j\in S}|\beta'_j| - \max_{j\in S^c}|\beta'_j| &\ge \min_{j\in S}|\beta_j| - \max_{j\in S}|\beta_j-\beta'_j| - \max_{j\in S^c}|\beta'_j| \\ &> \min_{j\in S}|\beta_j| - \sqrt{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 > 0.\end{aligned}$$

Thus $\hat{S}^{(s)} = S$. By Theorem 8, the ℓ_2 bound (32) is also satisfied by $\check{\boldsymbol{\beta}}^{(s)}$, which then shows (ii). Also note that $\check{\sigma}^{(s)} \le \tilde{\sigma}/(1-\tau^2)$ by Theorem 8. From (32) we have

$$\begin{aligned}2\sqrt{2}\frac{\sigma}{\check{\sigma}^{(s)}}\min_{j\in S}\frac{\check{\beta}_j^{(s)}}{\beta_j} &\ge 2\sqrt{2}\frac{\sigma}{\check{\sigma}^{(s)}}\left(1-\frac{\tilde{\sigma}}{4\sigma}\right) \\ &\ge 2\frac{\tilde{\sigma}}{\check{\sigma}^{(s)}} - \frac{\tilde{\sigma}}{\sqrt{2}\check{\sigma}^{(s)}}.\end{aligned}$$

We see that the RHS is at least 1 when $\tau^2 \le (\sqrt{2}-1)/(2\sqrt{2}-1)$, which then shows (iii) as $(\sqrt{2}-1)/(2\sqrt{2}-1) < 1/5$. □

Lemma 11. *Consider the setup of Theorem 3 and its proof. Conditional on the event Ω, $\hat{\mathbf{R}}(\boldsymbol{\beta},\sigma\boldsymbol{\varepsilon})$ and $(\check{\boldsymbol{\beta}}^{(s)},\check{\sigma}^{(s)})$ are independent.*

Proof. Write $v_N(\boldsymbol{\zeta}) = N\|(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}\|_2$. First we claim that on Ω, $\hat{\mathbf{R}}(\boldsymbol{\beta},\sigma\boldsymbol{\varepsilon})$ depends only on $(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/v_1(\boldsymbol{\varepsilon})$. To this end, we argue that for $\boldsymbol{\zeta} \in \Lambda$ with $v_N(\boldsymbol{\zeta}) > 1/(2\sqrt{2})$,

$$\begin{aligned}\hat{\mathbf{R}}(\boldsymbol{\beta},\sigma\boldsymbol{\zeta}) &= \hat{\mathbf{R}}(\boldsymbol{\beta}+\sigma\boldsymbol{\delta}(\boldsymbol{\zeta}),\sigma(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}) \\ &= \hat{\mathbf{R}}(\boldsymbol{\beta},\sigma(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}) \qquad (33) \\ &= \hat{\mathbf{R}}(\boldsymbol{\beta},\sigma(\mathbf{I}-\mathbf{P})\boldsymbol{\zeta}/v_N(\boldsymbol{\zeta})), \qquad (34)\end{aligned}$$

where $\boldsymbol{\delta}(\boldsymbol{\zeta}) = ((\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\boldsymbol{\zeta}, 0, \ldots, 0) \in \mathbb{R}^p$. Note the validity of the above inequalities would prove the initial claim for all $\boldsymbol{\zeta} \in \Lambda$ with $v_N(\boldsymbol{\zeta}) > 1/(2\sqrt{2})$ for each fixed N. However, since $\cup_{N>0}\{\boldsymbol{\zeta} : v_N(\boldsymbol{\zeta}) > 1/(2\sqrt{2})\} = \mathbb{R}^n$ the initial claim would then have to be true for all $\boldsymbol{\zeta} \in \Lambda$. We now set about proving these equalities. The first equality is clear by definition of $\hat{\mathbf{R}}$. Turning to the second (33), let $\boldsymbol{\zeta} \in \Lambda$ and denote by $\|\cdot\|$ the operator norm. From Lemma 10 we have

$$\begin{aligned}\sigma\|\boldsymbol{\delta}(\boldsymbol{\zeta})\|_\infty &= \sigma\|(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\boldsymbol{\zeta}\|_\infty \\ &\leq \sigma(\|\boldsymbol{\zeta}\|_2/\sqrt{n})\|(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\|\ \sqrt{s}\|\mathbf{X}_S^T\mathbf{P}\boldsymbol{\zeta}\|_\infty/(\|\boldsymbol{\zeta}\|_2/\sqrt{n}) \\ &\leq \frac{2\sigma\sqrt{s\log(p/\eta)/n}}{\phi^2(\xi)} \leq \min_{j\in S}|\beta_j|/10,\end{aligned}$$

using the facts that $\xi > 1$ and $A > \sqrt{2}$ and (9) in the final line. Note that by Lemma 10, $\Lambda \subseteq \Lambda_{\lambda,t}$ with the latter defined in Theorem 2 and $t = 1 - 1/(2\sqrt{2})$. Now clearly if $\boldsymbol{\zeta} \in \Lambda$ then also $(\mathbf{I} - \mathbf{P})\boldsymbol{\zeta} \in \Lambda$. An application of Theorem 2 then gives the desired equality (33). Equality (34) then follows from a further application of Theorem 2 noting that $v_N(\boldsymbol{\zeta}) < 1/(1-t) = 2\sqrt{2}$ by assumption.

Next we examine $\check{\boldsymbol{\beta}}^{(s)}$ and $\check{\sigma}^{(s)}$. Note that on Ω, the former is simply the Lasso estimate from regressing on $\mathbf{X}_S$ and the latter is the resulting normalised root-RSS. Write $\check{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}^{(s)}$ and $\check{\sigma} = \check{\sigma}^{(s)}$. The least squares part of the Lasso objective decomposes as

$$\|\mathbf{X}_S\boldsymbol{\beta}_S + \boldsymbol{\varepsilon} - \mathbf{X}_S\mathbf{b}\|_2 = \{\|\mathbf{X}_S\boldsymbol{\beta}_S + \mathbf{P}\boldsymbol{\varepsilon} - \mathbf{X}_S\mathbf{b}\|_2^2 + \|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2^2\}^{1/2}.$$

Thus it is clear that the fitted values $\mathbf{X}_S\check{\boldsymbol{\beta}}$ do not depend on $(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/v_1(\boldsymbol{\varepsilon})$. This then implies that $\check{\sigma}$ does not depend on $(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/v_1(\boldsymbol{\varepsilon})$ since it is determined by $\|\boldsymbol{\varepsilon}\|_2^2 = \|\mathbf{P}\boldsymbol{\varepsilon}\|_2^2 + \|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2^2$ and $\mathbf{P}\boldsymbol{\varepsilon}$.

Now observe that

$$\begin{aligned}\mathbf{P}\boldsymbol{\varepsilon} &\perp\!\!\!\perp \big(\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2,\ (\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2\big) \\ \|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2 &\perp\!\!\!\perp (\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2,\end{aligned}$$

so $\mathbf{P}\boldsymbol{\varepsilon}$, $\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2$, $(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}/\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2$ are jointly independent. Let $E_1 = \{\hat{\mathbf{R}}(\boldsymbol{\beta}, \sigma\boldsymbol{\varepsilon}) \in B_1\}$, $E_2 = \{(\mathbf{X}_S\check{\boldsymbol{\beta}}, \check{\sigma}) \in B_2\}$, where $B_1 \subseteq \mathbb{R}^n$ and $B_2 \subseteq \mathbb{R}^{n+1}$ are arbitrary Borel sets. Let $\Omega_k = \{\boldsymbol{\varepsilon} \in \Lambda_k\}$, $k = 1, 2, 3$. From the above

$$\mathbb{P}(E_1, E_2|\Omega_1, \Omega_2, \Omega_3) = \frac{\mathbb{P}(E_1, \Omega_1)}{\mathbb{P}(\Omega_1)}\frac{\mathbb{P}(E_2, \Omega_2, \Omega_3)}{\mathbb{P}(\Omega_2, \Omega_3)} = \mathbb{P}(E_1|\Omega_1)\mathbb{P}(E_2|\Omega_2, \Omega_3).$$

The conditional probabilities on the RHS remain unchanged if we modify the conditioning event to be $(\Omega_1, \Omega_2, \Omega_3)$ since $E_1 \perp\!\!\!\perp (\Omega_2, \Omega_3)$ and $\mathbb{E}_2 \perp\!\!\!\perp \Omega_1$. This completes the proof. □

Lemma 12. *Given a sequence of collections of matrices $\mathbf{M}_{k,n} \in \mathbb{R}^{n\times p_n}$, $k \in N_n$, $n = 1, 2, \ldots$, suppose there exists a sequence of collections of sets $S_{k,n} \subseteq \{1, \ldots, p_n\}$, tuning parameters $\lambda_n = A\sqrt{\log(p)/n}$ and $\xi > c$ for constants $A, c > 0$ such that the follow holds:*

$$\sup_{k\in N_n} \frac{|S_{k,n}|\sqrt{\log(p_n)^2/n}}{\kappa^2(\xi, S_{k,n}, \mathbf{M}_{k,n})} \to 0.$$

Moreover, suppose the collection of sequences $\boldsymbol{\beta}_{k,n} \in \mathbb{R}^{p_n}$ is such that

$$\sup_{k \in N_n} \sqrt{\log(p_n)} \|\boldsymbol{\beta}_{n, S_{k,n}^c}\|_1 \to 0.$$

Then there exists a $\xi' > c$ such that $\sqrt{\log(p_n)}\mu(c_1\lambda_n, \xi', \boldsymbol{\beta}_{k,n}, \mathbf{M}_{k,n}) \to 0$ for any $c_1 > 0$ (where the function μ is defined in (13)).

Proof. Let $\xi' = (\xi + c)/2 > c$ and let $\nu' > 0$ be given by $(\xi' + \nu')/(1 - \nu') = \xi$. Then we have

$$\begin{aligned}
&\sup_{k \in N_n} \sqrt{\log(p_n)}\mu(c_1\lambda_n, \xi', \boldsymbol{\beta}_{k,n}, \mathbf{M}_{k,n}) \\
\leq &(\xi' + 1)\sqrt{\log(p_n)} \max_{k \in N_n} \max \left[\frac{\|\boldsymbol{\beta}_{k,n,S_{k,n}^c}\|_1}{\nu'}, \frac{c_1\lambda_n |S_{k,n}|/\{2(1-\nu')\}}{\kappa^2(\xi, S_{k,n}, \mathbf{M}_{k,n})} \right] \to 0.
\end{aligned}$$

□

Lemma 13. *Let $Z_n \sim \chi_n^2$. We have the following tail bounds [Boucheron et al., 2013, pg. 29]:*

$$\mathbb{P}(Z_n > n + 2\sqrt{n\gamma} + 2\gamma) \leq e^{-\gamma},$$

whence taking $\gamma = n/8$,

$$\mathbb{P}(\sqrt{Z_n/n} > \sqrt{2}) < \mathbb{P}[Z_n > n\{1 + 2(1/\sqrt{8} + 1/8)\}] \leq e^{-n/8}.$$

Lemma 14. *Let $T_{k,n} \in \mathbb{R}$, $k \in N_n$, $n = 1, 2, \ldots$ be a collection of random variables which we may decompose as*

$$T_{k,n} = a_{k,n} Z_{k,n} + b_{k,n}$$

where each $Z_{k,n}$ is identically distributed with continuous distribution function F. Suppose further that each $Z_{k,n}$ is independent of the random elements $\mathcal{F}_n$, and for all $\delta > 0$, $\sup_{k \in N_n} \mathbb{P}(|b_{k,n} - d_{k,n}| > \delta | \mathcal{F}_n) \xrightarrow{p} 0$ for some random variables $d_{k,n} \in \mathbb{R}$ that are functions of $\mathcal{F}_n$, and $\sup_{k \in N_n} \mathbb{P}(|a_{k,n} - 1| > \delta | \mathcal{F}_n) \xrightarrow{p} 0$. Then

$$\sup_{k \in N_n} \sup_{x \in \mathbb{R}} |\mathbb{P}(T_{k,n} \leq x | \mathcal{F}_n) - F(x - d_{k,n})| \xrightarrow{p} 0.$$

Proof. First note that for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $c_1, c_2 \in [-\delta, \delta]$,

$$\sup_{x \in \mathbb{R}} \left| F\left(\frac{x + c_1}{1 + c_2} \right) - F(x) \right| < \epsilon. \tag{35}$$

Indeed, the function $G(x, c_1, c_2) := F\{(x + c_1)/(1 + c_2)\}$ is uniformly continuous on $\mathbb{R} \times [-\delta', \delta']^2$ for $0 < \delta' < 1$ sufficiently small (since F is uniformly continuous and compositions of uniformly continuous functions are continuous) and the LHS of (35) is $\sup_{x \in \mathbb{R}} |G(x, c_1, c_2) - G(x, 0, 0)|$.

Hence given $\epsilon > 0$, let $\delta > 0$ be such that the LHS of (35) is at most $\epsilon/2$. Then

$$\begin{aligned}
\sup_{x \in \mathbb{R}} |\mathbb{P}(T_{k,n} \leq x | \mathcal{F}_n) - F(x - d_{k,n})| &= \sup_{x \in \mathbb{R}} |\mathbb{P}(Z_{k,n} \leq (x + d_{k,n} - b_{k,n})/a_{k,n} | \mathcal{F}_n) - F(x)| \\
&\leq \sup_{x \in \mathbb{R}} \sup_{c_1, c_2 \in [-\delta, \delta]} |\mathbb{P}(Z_{k,n} \leq (x + c_1)/(1 + c_2) | \mathcal{F}_n) - F(x)| \\
&\quad + \mathbb{P}(|b_{k,n} - d_{k,n}| > \delta | \mathcal{F}_n) + \mathbb{P}(|a_{k,n} - 1| > \delta | \mathcal{F}_n) \\
&\leq \epsilon/2 + \mathbb{P}(|b_{k,n} - d_{k,n}| > \delta | \mathcal{F}_n) + \mathbb{P}(|a_{k,n} - 1| > \delta | \mathcal{F}_n).
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{P}\{\sup_{k\in N_n}\sup_{x\in\mathbb{R}}|\mathbb{P}(T_{k,n}\le x|\mathcal{F}_n)-F(x-d_{k,n})|>\epsilon\}\\
&\le \mathbb{P}\{\sup_{k\in N_n}\mathbb{P}(|b_{k,n}-d_{k,n}|>\delta|\mathcal{F}_n)>\epsilon/4\}+\mathbb{P}\{\sup_{k\in N_n}\mathbb{P}(|a_{k,n}-1|>\delta|\mathcal{F}_n)>\epsilon/4\}\\
&\to 0 \qquad \text{as } n\to\infty.
\end{aligned}$$

□

Lemma 15. *Consider the setup of Theorem 7. There exists constants $c_1, c_2 > 0$ such that for all $t < n^{1/4}$*

$$\mathbb{P}(1-tn^{-1/4}\le \|\mathbf{f}\|_2^2/\mathbb{E}(\|\mathbf{f}\|_2^2)\le 1+tn^{-1/4})\le c_1e^{-c_2t^2}, \tag{36}$$

$$\mathbb{P}\bigg(\frac{1}{\sqrt{n}}\frac{\|\mathbf{X}^T\mathbf{f}\|_\infty}{\|\mathbf{f}\|_2}\le c_1\frac{\sqrt{\log(p)}}{n^{1/3}}\bigg)\to 0. \tag{37}$$

Proof. Although in Theorem 7 certain assumptions are placed on $\mathbb{E}(f_1^2)$, since the probabilities above do not depend on $\mathbb{E}(f_1^2)$ here we may assume $\mathbb{E}(f_1^2)=1$. Fix $j\in\{1,\dots,p\}$. Using the eigendecomposition $\mathbf{PDP}^T=\boldsymbol{\Sigma}_{S,S}^{1/2}\boldsymbol{S},\boldsymbol{S}^{1/2}$ (where $\mathbf{D}$ is diagonal and $\mathbf{P}$ is orthogonal) and writing $\overset{d}{=}$ for equality in distribution, we have

$$(f_1,Z_{1j})=(\mathbf{z}_{1,S}^T\mathbf{B}\mathbf{z}_{1,S},\,Z_{1j})=(\mathbf{z}_{1,S}^T\Sigma_{S,S}^{-1/2}\mathbf{PDP}^T\Sigma_{S,S}^{-1/2}\mathbf{z}_{1,S},\,Z_{1j})\overset{d}{=}(\mathbf{u}^T\mathbf{D}\mathbf{u},v)$$

where $(\mathbf{u},v)$ is multivariate Gaussian with $\mathbf{u}\sim\mathcal{N}_s(\mathbf{0},\mathbf{I})$ and $\mathrm{Var}(v)=1$. Let the diagonal entries of $\mathbf{D}$ be $\boldsymbol{\theta}$.

First we show (36). Note that

$$\begin{aligned}
\mathbb{E}(f_1^2)=\mathbb{E}\Big\{\Big(\sum_j u_j^2\theta\Big)^2\Big\}&=\sum_j\theta_j^2\mathbb{E}(u_j^4)+\sum_{j\ne k}\theta_j\theta_k\mathbb{E}(u_j^2)\mathbb{E}(u_k^2)\\
&=2\|\boldsymbol{\theta}\|_2^2+\Big(\sum_j\theta_j\Big)^2=1.
\end{aligned}\tag{38}$$

Thus in particular, $\|\boldsymbol{\theta}\|_\infty\le 1/\sqrt{2}$, $\|\boldsymbol{\theta}\|_2^2\le 1/2$ and $\big|\sum_j\theta_j\big|\le 1$.

Now with a view to applying Lemma 16 below, consider

$$\begin{aligned}
\mathbb{E}\exp\Big\{\frac{1}{4}\Big|\Big(\sum_j u_j^2\theta_j\Big)^2\Big|^{1/2}\Big\}&\le\mathbb{E}\exp\Big(\sum_j u_j^2\theta_j/4\Big)+\mathbb{E}\exp\Big(-\sum_j u_j^2\theta_j/4\Big)\\
&=\prod_j(1-\theta_j/2)^{-1/2}+\prod_j(1+\theta_j/2)^{-1/2}
\end{aligned}$$

using the fact that $u_j^2\sim\chi_1^2$. Note that

$$(1-\theta_j/2)^{-1}=1+\frac{\theta_j}{2}+\frac{\theta_j^2}{4(1-\theta_j/2)}\le 1+\frac{\theta_j}{2}+\theta_j^2.$$

Thus, by the AM–GM inequality we have

$$\begin{aligned}\prod_j (1-\theta_j/2)^{-1} &\le \prod_j (1+\theta_j/2+\theta_j^2) \\ &\le \left(\frac{1}{s}\sum_j (1+\theta_j/2+\theta_j^2)\right)^s \\ &\le (1+1/s)^s < e,\end{aligned}$$

using (38). Similarly, $\prod_j (1+\theta_j/2)^{-1} < e$. Putting things together, we see that

$$\mathbb{E}\exp(|f_1^2 - \mathbb{E}(f_1^2)|^{1/2}/4) \le e^{1/4}\mathbb{E}\exp(|f_1|/4) < e^{1/4}\sqrt{2e}. \tag{39}$$

Lemma 16 then immediately gives (36).

To show (37), we first obtain a tail bound for $|\mathbf{f}^T\mathbf{Z}_j|/n$. Observe that if (w_1, w_2, w_3) is multivariate Gaussian with zero-mean, then since $(w_1, w_2, w_3) \stackrel{d}{=} -(w_1, w_2, w_3)$, we have

$$\mathbb{E}(w_1 w_2 w_3) = \mathbb{E}\{(-w_1)(-w_2)(-w_3)\} = -\mathbb{E}(w_1 w_2 w_3) = 0.$$

Then, by Hölder's inequality

$$\begin{aligned}\mathbb{E}\exp(|f_1 Z_{1j}|^{2/3}/4) &= \mathbb{E}\exp\Big(\Big|\sum_j v u_j^2\theta_j\Big|^{2/3}/4\Big) \\ &\le \sum_{r=0}^{\infty} \mathbb{E}\Big(\Big|\sum_j u_j^2\theta_j/4\Big|^{2r/3}|v/2|^{2r/3}\Big)\Big/r! \\ &\le \sum_{r=0}^{\infty} [\mathbb{E}\{(v^2/4)^r\}]^{1/3}\Big\{\mathbb{E}\Big(\Big|\sum_j u_j^2\theta_j/4\Big|^r\Big)\Big\}^{2/3}\Big/r! \\ &\le \sum_{r=0}^{\infty} \mathbb{E}\Big(\Big|\sum_j u_j^2\theta_j/4\Big|^r\Big)\Big/r! + \sum_{r=0}^{\infty}\mathbb{E}\{(v^2/4)^r\}/r! \\ &= \mathbb{E}\exp(|f_1|/4) + \mathbb{E}(v^2/4) \le \sqrt{2e} + \sqrt{2},\end{aligned}$$

using (39) in the final line. Note $\|\mathbf{f}\|_2/\sqrt{n} \xrightarrow{p} 1$ by (36), and $\|\mathbf{Z}_j\|/\sqrt{n} \xrightarrow{p} 1$ by Lemma 13. Thus by Lemma 16, $\mathbb{P}(|\mathbf{f}^T\mathbf{X}_j|/(\sqrt{n}\|\mathbf{f}\|_2) \ge t) \le c_1\exp(-c_2 n^{2/3}t^2)$ for $t \in [0,1]$ and some constants $c_1, c_2 > 0$. Thus for c_3 with $c_2 c_3^2 - 1 > 0$,

$$\mathbb{P}(\|\mathbf{Z}^T\mathbf{f}\|_\infty/(\sqrt{n}\|\mathbf{f}\|_2) \ge c_3\sqrt{\log(p)}/n^{1/3}) \le c_1 p\exp(-c_2 c_3^2\log(p)) = c_1 p^{-(c_2c_3^2-1)} \to 0$$

as $p \to \infty$ and $\log(p)/n^{2/3} \to 0$. □

Lemma 16 (Lemma B.4 of Hao and Zhang [2014])**.** *Let $W_1, \ldots, W_n$ be independent random variables with zero mean such that $\mathbb{E}(\exp(c_1|W_i|^\alpha)) \le c_2$ for constants $c_1, c_2 > 0$ and $\alpha \in (0,1]$. Then there exist constants $c_3, c_4 > 0$ such that for $t \in [0,1]$,*

$$\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n W_i\Big| \ge t\Big) \le c_3\exp(-c_4 n^\alpha t^2).$$

References

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *arXiv preprint arXiv:1412.3661*, 2014.

N. Hao and H. H. Zhang. Interaction screening for ultrahigh-dimensional data. *J. Am. Statist. Ass.*, 109(507):1285–1301, 2014.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *The J. Mach. Learn. Res.*, 11:2241–2259, 2010.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.