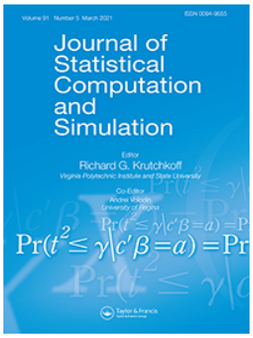

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Permutation testing in high-dimensional linear models: an empirical investigation

Jesse Hemerik, Magne Thoresen & Livio Finos

To cite this article: Jesse Hemerik, Magne Thoresen & Livio Finos (2021) Permutation testing in high-dimensional linear models: an empirical investigation, Journal of Statistical Computation and Simulation, 91:5, 897-914, DOI: 10.1080/00949655.2020.1836183

To link to this article: https://doi.org/10.1080/00949655.2020.1836183

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2021, VOL. 91, NO. 5, 897–914
https://doi.org/10.1080/00949655.2020.1836183

Permutation testing in high-dimensional linear models: an empirical investigation

Jesse Hemerik [a], Magne Thoresen [b] and Livio Finos [c]

[a]Biometris, Wageningen University & Research, Wageningen, Netherlands; [b]Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway; [c]Department of Developmental Psychology and Socialization, University of Padua, Padua, Italy

ABSTRACT

Permutation testing in linear models, where the number of nuisance coefficients is smaller than the sample size, is a well-studied topic. The common approach of such tests is to permute residuals after regressing on the nuisance covariates. Permutation-based tests are valuable in particular because they can be highly robust to violations of the standard linear model, such as non-normality and heteroscedasticity. Moreover, in some cases they can be combined with existing, powerful permutation-based multiple testing methods. Here, we propose permutation tests for models where the number of nuisance coefficients exceeds the sample size. The performance of the novel tests is investigated with simulations. In a wide range of simulation scenarios our proposed permutation methods provided appropriate type I error rate control, unlike some competing tests, while having good power.

ARTICLE HISTORY
Received 20 May 2020
Accepted 7 October 2020

KEYWORDS
Permutation test; group invariance test; high-dimensional; heteroscedasticity; semi-parametric

1. Introduction

We consider the problem of testing hypotheses about coefficients in linear models, where the outcome may be non-Gaussian and heteroscedastic, and the number of nuisance coefficients exceeds the sample size. By the nuisance coefficients we mean the coefficients that are not tested by the particular test at hand, but still need to be dealt with since they lead to confounding effects. In recent decades, the literature on permutation methods has strongly expanded [1–9]. While the permutation test dates far back [10], most of the permutation tests in the presence of nuisance were published in the last four decades. To our knowledge, the existing methods are limited to low-dimensional nuisance. For the high-dimensional case, an approach similar to a permutation test is proposed in Dezeure et al. [11].

Permutation tests for low-dimensional linear models are valuable for two main reasons. First, they are robust to violations of certain standard assumptions, such as normality and homoscedasticity [12,13]. Second, when the outcome is multidimensional, a

CONTACT Jesse Hemerik jesse.hemerik@wur.nl Biometris, Wageningen University & Research, PO Box 16, 6700 AC Wageningen, Netherlands

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2020.1836183

permutation-based test can be combined with existing permutation-based multiple testing methods, which tend to be relatively powerful, since they take into account the dependence structure of the outcomes [5–7,14]. For example, under strong positive dependence among p-values, the Bonferroni-Holm multiple testing method [15] is greatly improved by a permutation method [16].

For the low-dimensional general linear model, with identity link but not necessarily Gaussian or homoscedastic residuals, several different permutation tests have been proposed. The main approach that these methods have in common, is to permute residuals after regressing on the nuisance covariates. For overviews of the available methods, see Anderson and Legendre [17], Anderson and Robinson [18], Winkler et al. [19] and in particular Winkler et al. [13]. Among the existing permutation methods, the Freedman–Lane approach [20] is most commonly used and provides excellent power and type I error control.

Because the existing permutation tests require estimating the nuisance coefficients using maximum likelihood, these methods cannot be used when the number of covariates exceeds the sample size. In recent years, important theoretical results have been published on testing in such high-dimensional linear models. Several of these tests have proven asymptotic properties. In particular, the method in Zhang and Zhang [21] has been shown to be asymptotically optimal under certain assumptions [22]. Dezeure et al. [11] propose a bootstrap approach, which is related to the method in Zhang and Zhang [21]. Software implementations of tests for high-dimensional models include those described in Dezeure et al. [23] and Chernozhukov et al. [24].

Testing in high-dimensional linear models is very challenging, because a large number of unknown nuisance effects needs to be dealt with, using a relatively small sample size. Consequently, tests tend to sacrifice much power compared to the situation where all nuisance coefficients would be known. Further, the asymptotic properties of the mentioned methods rely on complex assumptions and sparsity. The test by Zhang and Zhang [21] can be rather anti-conservative in settings where a substantial fraction of the coefficients are non-zero. Moreover, these methods are not based on permutations. Hence they do not generally have the above-mentioned advantages, such as robustness against certain violations of the standard linear model. An exception is the bootstrap method in Dezeure et al. [11], which tends to be more robust to such violations.

We propose two novel tests, which, to our knowledge, are the first permutation tests in the presence of high-dimensional nuisance. One is an extension of the low-dimensional method in Freedman and Lane [20] and the other is somewhat related to a method by Kennedy [25,26]. Further, we allow the tested parameter to be multi-dimensional, unlike many existing methods. Using simulations we show that our methods provide appropriate type I error rate control in a wide range of situations. In particular, we illustrate empirically that our tests have the above-mentioned robustness properties. The methods in this paper have been implemented in the R package *phd*, available on CRAN.

This paper is built up as follows. In Section 2 we discuss permutation testing in settings with low-dimensional nuisance. This section contains some novel observations that will be used in Section 3. There, we propose permutation tests for high-dimensional settings. We assess the performance of our methods with simulations in Section 4. An analysis of real data is in Section 5.

2. Low-dimensional nuisance

2.1. Notation and basic ideas

We consider the general linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ is a $n \times d$ matrix of covariates of interest, $\boldsymbol{Z}$ an $n \times q$ matrix of nuisance covariates and $\boldsymbol{\epsilon}$ an n-vector of i.i.d. errors with mean 0 and non-zero variance, which are independent of the covariates. Here the rows of $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are i.i.d.. The matrix $\boldsymbol{Z}$ is assumed to have full rank with probability 1. The parameter $\boldsymbol{\beta} \in \mathbb{R}^d$ is of interest and $\boldsymbol{\gamma} \in \mathbb{R}^q$ is a nuisance parameter. We want to test the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{0} \in \mathbb{R}^d$. Here $\boldsymbol{0}$ might be replaced by another constant: the extension is straightforward.

Let w be a positive integer, which will denote the number of random permutations or other transformations. In this paper, all permutation p-values are of the form

$$p = w^{-1}\big|\{1 \leq j \leq w : T_j \geq T_1\}\big|, \tag{1}$$

or, in case of a two-sided test where both small and large values of T_1 are evidence against H_0,

$$p = 2w^{-1} \min\left\{\big|\{1 \leq j \leq w : T_j \geq T_1\}\big|, \big|\{1 \leq j \leq w : T_j \leq T_1\}\big|\right\}. \tag{2}$$

Here $T_1, \ldots, T_w \in \mathbb{R}$ are statistics whose definition depends on the particular permutation method. They are specified in the sections below. For every $2 \leq j \leq w$, the statistic T_j corresponds to the jth permutation. The statistic T_1 is based on the original, unpermuted data. All existing and novel methods in this paper only differ with respect to how $T_1, \ldots, T_w$ are computed.

Although we will often write 'permutation', sign-flipping of residuals can also be used [13]. The existing methods, as well as the novel methods in this paper, consist of the following steps.

(1) Compute a test statistic T_1 based on the original data.
(2) Compute a test statistic T_2 in a similar way, but after randomly permuting certain residuals. Repeat to obtain $T_3, \ldots, T_w$.
(3) The p-value equals (1) or (2).

Most of the existing permutation methods use residualization of $\boldsymbol{Y}$ or $\boldsymbol{X}$ with respect to the nuisance $\boldsymbol{Z}$. In the low-dimensional situation, the residual forming matrix is

$$\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'.$$

When $d = 1$ we will sometimes consider $\boldsymbol{RX} \in \mathbb{R}^n$, which is assumed to be nonzero with probability 1. In Section 2 we assume $\boldsymbol{Z}$ contains a column of 1's. This implies that the entries of $\boldsymbol{RX}$ and $\boldsymbol{RY}$ sum up to 0.

Note that if we use permutation, we can write the transformed residuals as $\boldsymbol{PRY}$, where $\boldsymbol{P}$ is an $n \times n$ matrix with exactly one 1 in every row and column and elsewhere 0's. In case of sign-flipping, $\boldsymbol{P}$ is instead an $n \times n$ diagonal matrix with diagonal elements in $\{1, -1\}$ [13]. We write $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_w$ to distinguish the w random permutation matrices. Here $\boldsymbol{P}_1$ is the identity matrix and $\boldsymbol{P}_2, \ldots, \boldsymbol{P}_w$ are random.

2.2. Choice of test statistics

Here we discuss the choice of test statistics within the permutation method of Freedman and Lane [13,20]. The purpose of this section is to discuss some existing and novel results that we will use in Section 3.

The Freedman–Lane permutation method is known to provide excellent type I error control, with both its level and power staying very close to the parametric F-test, under the Gaussian model. The test statistic T_1 is based on the unpermuted model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \epsilon$. The other statistics are obtained after randomly transforming the residuals. That is, for $2 \leq j \leq w$ the statistic T_j is based on the model $(\boldsymbol{P}_j\boldsymbol{R} + \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \epsilon$, where the same test statistic, say T, is used as for computing T_1. Thus

$$T_1 = T(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}), \tag{3}$$

$$T_j = T\big(\boldsymbol{X}, \boldsymbol{Z}, (\boldsymbol{P}_j\boldsymbol{R} + \boldsymbol{H})\boldsymbol{Y}\big), \tag{4}$$

where T is a suitable test statistic, the choice of which we now discuss.

It is usually important to take T to be an asymptotically pivotal statistic, i.e. a statistic whose asymptotic null distribution does not depend on any unknowns under H_0 ([p.926–927][26], [p.382][13], [27,28]). A pivotal statistic T will always involve estimation of the nuisance parameters. Thus, after every permutation, the nuisance parameters need to be estimated anew. Examples of pivotal test statistics are the F-statistic and Wald statistic. These are equivalent: the resulting permutation p-value (1) is the same.

In case X is one-dimensional, the F-statistic is also equivalent to the square of the *partial correlation* [29,30], which is used in Anderson and Robinson [18]. The partial correlation is the sample Pearson correlation of $\boldsymbol{RY}$ and $\boldsymbol{RX}$,

$$\rho\big(\boldsymbol{RY}, \boldsymbol{RX}\big) = \frac{(\boldsymbol{RY})'\boldsymbol{RX}}{\sqrt{\sum_i (\boldsymbol{RY})_i^2 \sum_i (\boldsymbol{RX})_i^2}}. \tag{5}$$

Here we used that the sample means of $\boldsymbol{RY}$ and $\boldsymbol{RX}$ are 0. If we use the partial correlation in the Freedman–Lane permutation test, this means that we take $T(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) = \rho(\boldsymbol{RY}, \boldsymbol{RX})$, so that (3) and (4) become

$$T_1 = \rho\big(\boldsymbol{RY}, \boldsymbol{RX}\big) \tag{6}$$

$$T_j = \rho\big(\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R} + \boldsymbol{H})\boldsymbol{Y}, \boldsymbol{RX}\big), \tag{7}$$

where $\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R} + \boldsymbol{H})$ could be simplified to $\boldsymbol{RP}_j\boldsymbol{R}$, since $\boldsymbol{RH} = \boldsymbol{0}$.

The numerator in (5) is

$$(\boldsymbol{RY})'\boldsymbol{RX} = \boldsymbol{Y}'\boldsymbol{R}'\boldsymbol{RX} = \boldsymbol{Y}'\boldsymbol{R}'\boldsymbol{X} = (\boldsymbol{RY})'\boldsymbol{X},$$

so that (5) equals

$$\frac{(\boldsymbol{RY})'\boldsymbol{X}}{\sqrt{\sum_i (\boldsymbol{RY})_i^2 \sum_i (\boldsymbol{RX})_i^2}}. \tag{8}$$

The Freedman–Lane test with T defined by (8) remains unchanged if in (8) we replace $\sum_i (\boldsymbol{RX})_i^2$ by 1 or by the constant $\sum_i \boldsymbol{X}_i^2$. Indeed, $T_1, \ldots, T_w$ will just be multiplied by the

same constant. Thus, with respect to the permutation test, the statistic (5) is equivalent to

$$\frac{(\boldsymbol{RY})'\boldsymbol{X}}{\sqrt{\sum_i(\boldsymbol{RY})_i^2\sum_i \boldsymbol{X}_i^2}}. \tag{9}$$

If $\boldsymbol{X}$ has been centred around 0, then this equals

$$\rho\big(\boldsymbol{RY},\boldsymbol{X}\big) = \frac{(\boldsymbol{RY})'(\boldsymbol{X}-\boldsymbol{\mu}_x)}{\sqrt{\sum_i(\boldsymbol{RY})_i^2\sum_i(\boldsymbol{X}_i-\boldsymbol{\mu}_x)^2}}, \tag{10}$$

where $\boldsymbol{\mu}_x$ denotes the n-vector with entries equal to the sample mean of $\boldsymbol{X}$. This is the sample correlation of $\boldsymbol{RY}$ and $\boldsymbol{X}$ and is called the *semi-partial correlation*. Thus, if $\boldsymbol{X}$ is centred, using the partial correlation is equivalent to using the semi-partial correlation.

If we take T to be the semi-partial correlation, then (3) and (4) become $T_1 = \rho(\boldsymbol{RY},\boldsymbol{X})$ and

$$T_j = \rho\big(\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R}+\boldsymbol{H})\boldsymbol{Y},\boldsymbol{X}\big) = \frac{\big(\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R}+\boldsymbol{H})\boldsymbol{Y}\big)'(\boldsymbol{X}-\boldsymbol{\mu}_x)}{\sqrt{\sum_i\big(\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R}+\boldsymbol{H})\boldsymbol{Y}\big)_i^2\sum_i(\boldsymbol{X}_i-\boldsymbol{\mu}_x)^2}}, \tag{11}$$

where $\boldsymbol{R}(\boldsymbol{P}_j\boldsymbol{R}+\boldsymbol{H})$ could be simplified to $\boldsymbol{RP}_j\boldsymbol{R}$. Note that we could simply leave the constant $\sum_i(\boldsymbol{X}_i-\boldsymbol{\mu}_x)^2$ out without changing the result of the permutation test. Although for centred $\boldsymbol{X}$ the statistics (5) and (10) are equivalent, their counterparts in the high-dimensional setting are not, as will be discussed in Section 3.1.

3. High-dimensional nuisance

When the nuisance parameter $\boldsymbol{\gamma}$ has dimension $q \geq n$, the existing permutation methods cannot be used. Here, these approaches are adapted to obtain tests which can account for high-dimensional nuisance. We first consider the case that X is one-dimensional, i.e.$d = 1$. The case that $d > 1$ is discussed in Section 3.3. We assume that the entries of $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ have expected value 0. Consequently, the intercept is 0.

All existing tests rely on residualization steps, where $\boldsymbol{Y}$ or $\boldsymbol{X}$ is regressed on $\boldsymbol{Z}$. A natural way to adapt this step to the high-dimensional setting, is to instead estimate the residuals using some type of elastic net regularization. We will consider ridge regression. For minimizing prediction error, ridge regression is often preferrable to Lasso, principal components regression, variable subset selection and partial least squares [31,32].

Compared to the existing methods, including the Freedman–Lane approach discussed in Section 2.2, using ridge regression comes down to replacing the projections $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ and $\hat{\boldsymbol{X}} = \boldsymbol{HX}$ by ridge estimates $\tilde{\boldsymbol{H}}_\lambda\boldsymbol{Y}$ and $\tilde{\boldsymbol{H}}_{\lambda_X}\boldsymbol{X}$, with $\lambda, \lambda_X > 0$. Here, for $\lambda' > 0$,

$$\tilde{\boldsymbol{H}}_{\lambda'} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\lambda'\boldsymbol{I}_q)^{-1}\boldsymbol{Z}', \tag{12}$$

which satisfies

$$\tilde{\boldsymbol{H}}_{\lambda'}\boldsymbol{Y} = \boldsymbol{Z}\,\mathrm{argmin}_{\boldsymbol{\gamma}}\Big(\|\boldsymbol{Y}-\boldsymbol{Z\gamma}\|_2^2+\lambda'\|\boldsymbol{\gamma}\|_2^2\Big)$$

and similarly for $\boldsymbol{X}$. The values λ, λ_X are the regularization parameters, whose selection will be discussed. Using ridge regression, the residuals become $\tilde{\boldsymbol{R}}_\lambda\boldsymbol{Y}$ and $\tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}$, where $\tilde{\boldsymbol{R}}_\lambda = (\boldsymbol{I}-\tilde{\boldsymbol{H}}_\lambda)$ and $\tilde{\boldsymbol{R}}_{\lambda_X} = (\boldsymbol{I}-\tilde{\boldsymbol{H}}_{\lambda_X})$.

Table 1. Permutation schemes for four different methods.

Method	Model after permutation
Freedman–Lane	$(\boldsymbol{PR}+\boldsymbol{H})\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{Z\gamma}+\boldsymbol{\epsilon}$
Kennedy	$\boldsymbol{PRY}=\boldsymbol{RX\beta}+\boldsymbol{\epsilon}$
Freedman–Lane HD	$(\boldsymbol{P}\tilde{\boldsymbol{R}}_\lambda+\tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{Z\gamma}+\boldsymbol{\epsilon}$
Double residualization	$(\boldsymbol{P}\tilde{\boldsymbol{R}}_\lambda+\tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}=\tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X\beta}+\boldsymbol{\epsilon}$

Note: The last two methods are novel and can account for high-dimensional nuisance.

The last two rows of Table 1 outline the permutation schemes that we will consider in Sections 3.1 and 3.2. The first two rows summarize the Freedman–Lane method discussed in Section 2.2 and the Kennedy method [13,25,26]. This table is analogous to Table 2 in Winkler et al. [13] and allows easy comparison of the new methods with the existing methods discussed in Winkler et al. [13].

Although Table 1 outlines the permutation schemes that we will use, several crucial specifics remain to be filled in. For example, several choices of the regularization parameters λ and λ_X can be considered. Moreover, the computational challenge of performing nuisance estimation in every step needs to be addressed. Finally and importantly, we must determine what test statistics are suitable to use within our permutation tests.

3.1. Freedman–Lane HD

As discussed in Section 2.2, the low-dimensional Freedman–Lane method is known to provide excellent type I error control and power. Here we will provide an extension to the case of high-dimensional nuisance. We will refer to this test as *Freedman–Lane HD*. The permutation scheme that we use is analogous to that of Freedman–Lane and is shown in the third row of Table 1.

As in the Freedman–Lane method, after every permutation, we will require nuisance estimation to compute T_j. We will choose ridge regression to do this. Note however that when many permutations are used, performing a ridge regression after every permutation can be a large computational burden. We will therefore compute λ only once, for the unpermuted model. We take λ to be the value that gives the minimal mean cross-validated error; see Section 4.1 for more details. After each permutation, we then use the same parameter λ in the ridge regression. Thus, after the jth permutation, to compute the new ridge residuals, we will only need to pre-multiply the transformed outcome $(\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda+\tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}$ by $\tilde{\boldsymbol{R}}_\lambda$. We only need to compute $\tilde{\boldsymbol{R}}_\lambda$ once. Owing to this approach, essentially we need to perform ridge regression only once.

An important consideration is the test statistic T used within the permutation test. The usual F-statistic and Wald statistic are only defined when the nuisance is low-dimensional. Extending these definitions to the high-dimensional setting with $q \geq n$ is problematic. For example, a Wald-type statistic would require an unbiased estimate of β and a variance estimate. The partial correlation (5), however, is more naturally generalized to the $q \geq n$ setting: we can replace the residuals $\boldsymbol{RY}$ and $\boldsymbol{RX}$ by the ridge residuals $\tilde{\boldsymbol{R}}_\lambda\boldsymbol{Y}$ and $\tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}$. Similarly we can generalize the semi-partial correlation (10), by replacing $\boldsymbol{RY}$ by $\tilde{\boldsymbol{R}}_\lambda\boldsymbol{Y}$. This

gives the following test statistics, which generalize the partial correlation (5) and the semi-partial correlation (10), respectively:

$$\rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X}\big) = \frac{(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y} - \boldsymbol{\mu}_1)'(\tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X} - \boldsymbol{\mu}_2)}{\sqrt{\sum_i (\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y} - \boldsymbol{\mu}_1)_i^2 \sum_i (\tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X} - \boldsymbol{\mu}_2)_i^2}}, \tag{13}$$

$$\rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \boldsymbol{X}\big) = \frac{(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y} - \boldsymbol{\mu}_1)'(\boldsymbol{X} - \boldsymbol{\mu}_x)}{\sqrt{\sum_i (\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y} - \boldsymbol{\mu}_1)_i^2 \sum_i (\boldsymbol{X} - \boldsymbol{\mu}_x)_i^2}}. \tag{14}$$

Here, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_x$ are n-vectors whose entries are the sample means of $\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}$, $\tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X}$ and $\boldsymbol{X}$, respectively. Zhu and Bradic [33] also use a type of generalized partial correlation as the test statistic.

In Section 2.2 we reasoned that if $\boldsymbol{X}$ has been centred, (5) and (10) are equivalent with respect to the permutation test. This does not apply to (13) and (14). In simulations, using the statistic (14) tended to result in somewhat higher power than using the statistic (13). In Section 4 we consider both methods.

In case the generalization of the partial correlation is used, the test statistics $T_1, \ldots, T_w$ on which Freedman–Lane HD is based are

$$T_1 = \rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X}\big), \tag{15}$$

$$\begin{aligned} T_j &= \rho\big(\tilde{\boldsymbol{R}}_\lambda \big(\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda\big)\boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X}\big) \\ &= \frac{\big(\tilde{\boldsymbol{R}}_\lambda (\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y} - \boldsymbol{\mu}^j\big)'(\tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X} - \boldsymbol{\mu}_2)}{\sqrt{\sum_i \big(\tilde{\boldsymbol{R}}_\lambda (\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y} - \boldsymbol{\mu}^j\big)_i^2 \sum_i (\tilde{\boldsymbol{R}}_{\lambda_X} \boldsymbol{X} - \boldsymbol{\mu}_2)_i^2}}, \end{aligned} \tag{16}$$

where $2 \le j \le w$. Here $\boldsymbol{\mu}^j$ is an n-vector whose entries are the sample mean of $\tilde{\boldsymbol{R}}_\lambda (\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}$. For the version based on the generalization of the semi-partial correlation, the statistics are

$$T_1 = \rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \boldsymbol{X}\big), \tag{17}$$

$$T_j = \rho\big(\tilde{\boldsymbol{R}}_\lambda (\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \boldsymbol{X}\big). \tag{18}$$

As usual, T_1 is just T_j with $\boldsymbol{P}_j = \boldsymbol{I}_n$. The pseudo-code for the version based on semi-partial correlations is in Algorithm 1.

If $q < n$, as $\lambda \downarrow 0$, the test converges to the test for $\lambda = 0$, which is the classical Freedman–Lane method. In the wide range of simulation settings considered in Section 4, the Freedman–Lane HD method stayed on the conservative side, in the sense that the size was less than α. This may due to the fact that if $\lambda > 0$ and $2 \le j < k \le w$, the correlation between T_1 and T_j tended to be larger than the correlation between T_j and T_k in simulations. This may be related to the fact that the correlation between $\boldsymbol{Y}$ and $\boldsymbol{Y}^{*j}$ is strictly larger than the correlation between $\boldsymbol{Y}^{*j}$ and $\boldsymbol{Y}^{*k}$, where $\boldsymbol{Y}^{*j} := (\boldsymbol{P}_j \tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}$. This inequality is proved in the Supplementary Material.

As discussed, to perform the test, λ and hence $\tilde{\boldsymbol{R}}_\lambda$ need to be computed only once. Thus, like the low-dimensional Freedman–Lane procedure, the test requires nuisance estimation after every permutation, but this is not a large computational burden. The method is often

computationally feasible even when many millions of permutations are used; see Section 4. It is also worth mentioning that there exist approximate methods for reducing the number of permutations while still allowing for very small, accurate p-values [19,34].

Algorithm 1 Freedman- -Lane HD (version based on semi-partial correlations)

1: Compute $\tilde{\boldsymbol{H}}_\lambda = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}_q)^{-1}\boldsymbol{Z}'$ and the residual forming matrix $\tilde{\boldsymbol{R}}_\lambda = \boldsymbol{I} - \tilde{\boldsymbol{H}}_\lambda$. Here λ is taken to give the minimal mean cross-validated error (see main text).
2: Let $T_1 = \rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \boldsymbol{X}\big)$, the sample Pearson correlation of the $\boldsymbol{Y}$-residuals with $\boldsymbol{X}$.
3: **for** $2 \leq j \leq w$ **do**
4: Let $T_j = \rho\big(\tilde{\boldsymbol{R}}_\lambda(\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \boldsymbol{X}\big)$, where the random matrix $\boldsymbol{P}_j$ encodes random permutation or sign-flipping.
5: **end for**
6: The two-sided p-value p equals (2).
7: **return** p

3.2. Double residualization

Here we propose a test that we refer to as the *Double Residualization* method. The method is somewhat related to the Kennedy procedure [13,25,26], but not analogous. The Kennedy method residualizes both $\boldsymbol{Y}$ and $\boldsymbol{X}$ and proceeds to permute the $\boldsymbol{Y}$-residuals. Here we replace the least squares regression by ridge regression. Moreover, unlike Kennedy's permutation scheme, we keep $\tilde{\boldsymbol{H}}_\lambda \boldsymbol{Y}$ in the model; see Table 1. The test statistic that we use within the permutation test is the sample correlation. Thus, the test is based on the statistics

$$\begin{aligned} T_1 &= \rho\big(\boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}\big), \\ T_j &= \rho\big((\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}\big), \end{aligned} \tag{19}$$

where $2 \leq j \leq w$. The difference between (19) and (16) is that (16) contains an additional $\tilde{\boldsymbol{R}}_\lambda$. The pseudo-code for the Double Residualization method is in Algorithm 2. We take λ and λ_X to be the values that give the minimal mean cross-validated error; see Section 4.1 for more details. For fixed q, as $n \to \infty$, the Double Residualization method becomes equivalent to the Kennedy method and the Freedman–Lane method if the penalty is $o_{\mathbb{P}}(n^{1/2})$, as shown in the Supplementary Material. The case that $q > n$ is investigated in Section 4.

3.3. Multi-dimensional parameter of interest

In the above we considered the case that the tested parameter β has dimension $d = 1$. Our tests can be extended to the case $d > 1$ by using Pesarin's Non-Parametric Combination (NPC) approach [35, ch. 4]. This is a general method for combining permutation tests of different hypotheses into a test for the intersection hypothesis. The NPC principle can be applied in a wide range of scenarios. In simpler settings with no nuisance, NPC has important proven properties, such as asymptotically optimal power. Here, we will explain how NPC can be applied in our setting. For convenience, we will focus on the application

Algorithm 2 Double Residualization

1: Compute $\tilde{\boldsymbol{H}}_\lambda = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}_q)^{-1}\boldsymbol{Z}'$ and, analogously, $\tilde{\boldsymbol{H}}_{\lambda_X}$. Here λ and λ_X are determined through cross-validation (see main text). Let $\tilde{\boldsymbol{R}}_\lambda = \boldsymbol{I} - \tilde{\boldsymbol{H}}_\lambda$ and $\tilde{\boldsymbol{R}}_{\lambda_X} = \boldsymbol{I} - \tilde{\boldsymbol{H}}_{\lambda_X}$.
2: Let $T_1 = \rho\big(\boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}\big)$, the sample Pearson correlation of $\boldsymbol{Y}$ and $\tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}$.
3: **for** $2 \le j \le w$ **do**
4: Let $T_j = \rho\big((\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \tilde{\boldsymbol{R}}_{\lambda_X}\boldsymbol{X}\big)$, where the random matrix $\boldsymbol{P}_j$ encodes random permutation or sign-flipping.
5: **end for**
6: The two-sided p-value p equals (2).
7: **return** p

of NPC to our test of Algorithm 1, i.e. Freedman–Lane HD based on the generalized semi-partial correlation. Combining NPC with our other tests can be done similarly, but can be computationally much less efficient for large d, as will be explained below.

Suppose $d > 1$. We are interested in $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, where we assume $\boldsymbol{\beta}_0 = \boldsymbol{0}$ again for notational convenience. For every $1 \le l \le d$, let β_l be the l-th entry of $\boldsymbol{\beta}$. The hypothesis of interest H_0 is the intersection of $H^1, \ldots, H^d$, where H^l is the hypothesis that β_l equals 0. To test $H_0 = H^1 \cap \cdots \cap H^d$, we proceed as follows. As usual, sample random matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_w$ that encode permutation (or sign-flipping). For every $1 \le l \le d$ and $1 \le j \le w$, define

$$T_j^l = \rho\big(\tilde{\boldsymbol{R}}_\lambda(\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \boldsymbol{X}_{\cdot l}\big),$$

where $\boldsymbol{X}_{\cdot l}$ is the lth column of $\boldsymbol{X}$. A key point here is that the same permutation matrix $\boldsymbol{P}_j$ is used to compute each of the statistics $T_j^1, \ldots., T_j^d$. Due to this manner of simultaneous permutation, the dependence structure of $(T_j^1, \ldots., T_j^d)$ mimics that of $(T_1^1, \ldots., T_1^d)$. Indeed, if $\boldsymbol{\gamma}$ were exactly known so that we could replace $\tilde{\boldsymbol{R}}_\lambda\boldsymbol{Y}$ and $\tilde{\boldsymbol{H}}_\lambda\boldsymbol{Y}$ by $\boldsymbol{\epsilon}$ and $\boldsymbol{Z}\boldsymbol{\gamma}$, then $(T_j^1, \ldots., T_j^d)$ and $(T_1^1, \ldots., T_1^d)$ would have exactly the same dependence structure under H_0.

Consider a function $\Psi : \mathbb{R}^d \to \mathbb{R}$, which will be used to compute a combination statistic [35, ch. 4]. For every $1 \le j \le w$ define $\Psi_j = \Psi(T_j^1, \ldots., T_j^d)$. Note that if $\tilde{\boldsymbol{R}}_\lambda\boldsymbol{Y}$ and $\tilde{\boldsymbol{H}}_\lambda\boldsymbol{Y}$ would be the exact errors and expected values, then under H_0, $\Psi_1, \ldots, \Psi_w$ would be identically distributed and exchangeable. The p-value for testing H_0 is now computed as in (1) but with T_j replaced by the combination statistic Ψ_j. The pseudo-code for this test is in Algorithm 3. Note that if $d = 1$ and Ψ is the identity and a two-sided p-value is computed, then this method reduces to the test of Algorithm 1.

The function Ψ should be chosen such that high values of Ψ_1 indicate evidence against H_0. The choice of Ψ influences power. Examples of functions Ψ are $\Psi(t_1, \ldots, t_d) = \max(|t_1|, \ldots, |t_d|)$ and $\Psi(t_1, \ldots, t_d) = d^{-1}\sum_{l=1}^d |t_l|$. The former choice of Ψ if often used when one or few of the coefficients $\beta_1, \ldots, \beta_d$ are expected to be nonzero under the alternative. Otherwise, the latter choice of Ψ is often used. Other examples of combining functions Ψ are in Pesarin and Salmaso [35, ch. 4].

Applying NPC to the other tests of Sections 3.1 and 3.2 tends to be computationally less efficient than the method of Algorithm 3. For example, applying NPC to our Double

Algorithm 3 Extension of the test of Algorithm 1 to the case that $d > 1$.

Compute $\tilde{\boldsymbol{H}}_\lambda = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}_q)^{-1}\boldsymbol{Z}'$ and the residual forming matrix $\tilde{\boldsymbol{R}}_\lambda = \boldsymbol{I} - \tilde{\boldsymbol{H}}_\lambda$. Here λ is taken to give the minimal mean cross-validated error.
for $1 \leq l \leq d$ **do**
  Let $T_1^l = \rho\big(\tilde{\boldsymbol{R}}_\lambda \boldsymbol{Y}, \boldsymbol{X}_{\cdot l}\big)$, where $\boldsymbol{X}_{\cdot l}$ is the l-th column of $\boldsymbol{X}$.
end for
for $2 \leq j \leq w$ **do**
  Consider a random $n \times n$ matrix $\boldsymbol{P}_j$ encoding random permutation or sign-flipping.
  Compute $\tilde{\boldsymbol{R}}_\lambda(\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}$.
  for $1 \leq l \leq d$ **do**
    Let $T_j^l = \rho\big(\tilde{\boldsymbol{R}}_\lambda(\boldsymbol{P}_j\tilde{\boldsymbol{R}}_\lambda + \tilde{\boldsymbol{H}}_\lambda)\boldsymbol{Y}, \boldsymbol{X}_{\cdot l}\big)$.
  end for
end for
for $1 \leq j \leq w$ **do**
  Compute $\Psi_j = \Psi(T_j^1, ..., T_j^d)$, where Ψ is the combining function.
end for
The p-value p equals $w^{-1}\big|\{1 \leq j \leq w : \Psi_j \geq \Psi_1\}\big|$.
return p

Residualization method would require ridge-regressing each of the d variables of interest (corresponding to $\beta_1, \ldots, \beta_d$) on the nuisance variables.

4. Simulations

We used simulations to gain additional insight into the performance of the new tests, as well as existing tests. The simulations were performed with *R* version 3.6.0 on a server with 40 cores and 1TB RAM. In Section 4.2 we consider scenarios where the outcome Y follows a standard Gaussian high-dimensional linear model. In Section 4.3 we consider non-standard settings with non-normality and heteroscedasticity. We consider simulated datasets where the covariates have equal variances. It is well-known that when the data are not standardized, this can affect the accuracy of the model obtained with ridge regression [36, p.257].

4.1. Simulation settings and tests

We considered the model in Section 2.1, where the variable of interest was one-dimensional, i.e. $\beta \in \mathbb{R}$. The case $d > 1$ is considered in Section 4.4. In every simulation, the covariates had mean 0 and variance 1. They were sampled from a multivariate normal distribution with homogenous correlation ρ', unless stated otherwise. The errors $\boldsymbol{\epsilon}$ had variance 1, unless stated otherwise. The intercept was $\gamma_1 = 0$, i.e. Y had mean 0. The tested hypothesis was $H_0 : \beta = 0$. The sample size in the reported simulations was $n = 30$, unless stated otherwise. We obtained comparable results for other sample sizes. The esti-

mated probabilities in the tables are based on 10^4 repeated simulations, unless stated otherwise.

In the power simulations we usually took $|\beta|$ to be relatively large compared to most of the nuisance coefficients. The reason is that testing in high-dimensional models is very challenging. For example, in settings with $|\beta| = |\gamma_1| = \cdots = |\gamma_q| > 0$ the power of all the tests considered (including the competitors) usually barely exceeds the type I error rate.

The penalty λ was chosen to give the minimal mean error, based on 10-fold cross validation. The penalty λ_X was chosen analogously. To compute the penalties, we used the *cv.glmnet()* function in the R package *glmnet*. We used $[10^{-5}, 10^5]$ as the range of candidate values for the penalty. The penalty obtained with *cv.glmnet()* is scaled by a factor n, so we multiplied this penalty by n to obtain λ. We included an intercept in the ridge regressions, but excluding the intercept gave very similar results.

All tests used were two-sided. The tests corresponding to the columns of the tables in this section are the following.

'FLH1' is the Freedman–Lane HD test defined in Section 3.1, with test statistics $T_1, \ldots, T_w$ based on the generalized partial correlation as in (16). 'FLH2' is the same, except that $T_1, .., T_w$ are based on the generalized *semi*-partial correlation as in (18). 'DR' is the Double Residualization method of Section 3.2. Each of these tests used $w = 2 \cdot 10^4$ permutations.

'BM' is a high-dimensional test based on ridge projections, proposed in Bühlmann [37]. This test is based on a bias-corrected estimate $|\hat{\beta}_{\text{corr}}|$ of $|\beta| \in \mathbb{R}$ and an asymptotic upper bound of its distribution. We used the implementation in the R package *hdi* [23].

'ZZ' is a high-dimensional test based on Lasso projections, proposed in Zhang and Zhang [21]. This method constructs a different bias-corrected estimate $\hat{b}$ of β, which has an asymptotically known normal distribution under certain assumptions, such as sparsity. For this test we also used the *hdi* package. We could not include this test in the simulations with a very high number of nuisance parameters, since it is computationally very time-consuming when q is large, as also noted in Dezeure et al. [23]. We expect the test to have good power in these settings.

'BO' is the bootstrap approach in Dezeure et al. [11], which is also implemented in the *hdi* package. We set the number of bootstrap samples per test to 1000 and considered the robust version of the method. We used the shortcut, which avoids repeated tuning of the penalty. Still, the method was very slow, so that we used 10^3 instead of 10^4 repeated simulations of this method per setting. Also, we did not include the test in the simulations with very large q.

4.2. Gaussian, homoscedastic outcome

We first consider some settings with a moderately large number of nuisance coefficients, $q = 60$. We first simulated a setting with $\gamma_2 = \cdots = \gamma_{60} = 0.05$, i.e, $\boldsymbol{\gamma}$ was dense. We took $\rho' = 0.5$. The estimated level and power of the tests described above, for different p-value cut-offs α, are shown in Table 2. The tests rejected H_0 if the p-value was smaller than α. The level of a test should be at most α.

Table 2. Dense setting with $\rho' = 0.5, n = 30, q = 60$.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0281	.0333	.0219	.0087	.0666	.063
	0.01	.0042	.0063	.0021	.0024	.0311	.023
	0.001	.0003	.0006	0001	.0005	.0121	.009
power	0.05	.9062	.9273	.9616	.8901	.9934	.982
	0.01	.8373	.8819	.7984	.7679	.9799	.939
	0.001	.6716	.7996	.3263	.5795	.9441	.857

Note: Power is shown for $\beta = 1.5$.

Table 3. Sparse setting with $\rho' = 0.9, n = 30, q = 60$.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0302	.0270	.0348	.0106	.0358	.051
	0.01	.0050	.0035	.0044	.0013	.0104	.012
	0.001	.0003	.0001	.0001	.0000	.0022	.002
power	0.05	.4494	.5426	.4804	.3234	.6050	.554
	0.01	.2283	.3379	.2135	.1506	.4154	.346
	0.001	.0685	.1195	.0445	.0501	.2296	.206

Note: Power is shown for $\beta = 1.5$.

Table 2 shows that the test ZZ by Zhang and Zhang [21] was rather anti-conservative. Especially for small α, its level was many times larger than α. This is partly due to the anti-sparsity. Indeed, ZZ only has proven asymptotic properties under a sparsity assumption. The bootstrap approach BO of Dezeure et al. [11] was much less liberal, but still seemed to be somewhat anti-conservative for small α. Of the other tests, Freedman–Lane HD 2 (FLH2) often had the most power. The Double Residualization method had relatively low power when α was small, e.g. 0.001.

We also considered a setting with very high correlation $\rho' = 0.9$, see Table 3. We took $\gamma_2 = \gamma_3 = 1$ and $\gamma_4 = \cdots . = \gamma_{60} = 0$. The first 4 methods provided appropriate type I error control. For small cut-offs α, the method ZZ by Zhang and Zhang [21] was relatively powerful, but also seemed to be somewhat anti-conservative. This method seems more suitable for settings where q is many times larger than n. Among our permutation methods, Freedman–Lane HD 2 had the best power, while incurring few type I errors. The method BM by Bühlmann [37] was relatively conservative.

We repeated the same simulation scenario, but with $n = 15$ instead of $n = 30$. The results are in Table 4. The methods ZZ of Zhang and Zhang [21] and BO of Dezeure et al. [11] were very anti-conservative for $\alpha = 0.01$ and $\alpha = 0.001$. Our methods provided appropriate type I error control.

Further, we considered a simulation where there were clusters of correlated covariates. The setting was as before, except that there were three independent clusters of size 20. Each cluster had a multivariate normal distribution with all correlations equal to 0.9. We took $\gamma_2 = \cdots = \gamma_{60} = 0.05$. The results are in Table 5. As before, the tests ZZ of Zhang and Zhang [21] and BO of Dezeure et al. [11] had good power, but were anti-conservative.

Table 4. Sparse setting with $\rho' = 0.9, n = 15, q = 60$.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0268	.0244	.0294	.0030	.0392	.050
	0.01	.0048	.0030	.0028	.0004	.0124	.026
	0.001	.0008	.0000	.0000	.0002	.0032	.020
power	0.05	.5020	.6034	.5090	.4038	.7586	.692
	0.01	.2822	.4558	.2094	.2384	.6248	.552
	0.001	.0730	.1982	.0438	.1244	.4614	.386

Note: Power is shown for $\beta = 3$.

Table 5. Dense setting with $n = 30, q = 60$ and three clusters of dependent covariates.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0356	.0224	.0344	.0130	.0520	.073
	0.01	.0059	.0025	.0048	.0022	.0248	.023
	0.001	.0010	.0002	.0002	.0007	.0087	.008
power	0.05	.4892	.5706	.5043	.4188	7382	.620
	0.01	.2672	.3393	.2226	.2399	.6199	.454
	0.001	.0814	.1007	.0382	.0977	.4741	.322

Note: Power is shown for $\beta = 1.5$.

Table 6. Sparse setting with a large number ($q = 1000$) of nuisance variables.

		Method			
	α	FLH1	FLH2	DR	BM
level	0.05	.0068	.0065	.0145	.0001
	0.01	.0013	.0011	.0011	.0000
	0.001	.0002	.0001	.0000	.0000
power	0.05	.5577	.5469	.9613	.7820
	0.01	.5060	.5043	.8007	.6510
	0.001	.3752	.4049	.3463	.4851

Notes: Here $\rho' = 0.5, n = 30$. Power is shown for $\beta = 2$.

We also performed simulations with a very large number of nuisance variables ($q = 1000$). We first took $\gamma_2 = \gamma_3 = 1$, $\gamma_4 = \cdots = \gamma_{10} = 0.2$, $\gamma_{11} = \cdots = \gamma_{1000} = 0$. See Table 6 for simulations with $\rho' = 0.5$ and Table 7 for simulations with $\rho' = 0.9$. All permutation methods provided appropriate type I error control. Double Residualization (DR) had relatively high power for large cut-offs α, but not for small cut-offs. The method BM by Bühlmann [37] had relatively good power for $\rho' = 0.5$ but low power for $\rho' = 0.9$.

We also performed simulations where γ was very anti-sparse, e.g. with $\gamma_2 = 1$, $\gamma_3 = \cdots = \gamma_{800} = 0.002$ and $\rho' = 0.9$. We also considered negative coefficients and we varied the magnitude of the coefficients and the errors $\boldsymbol{\epsilon}$ and the sample size. We also considered more settings where there were multiple independent clusters of correlated covariates. Also in these settings, the type I error rate was controlled.

Table 7. Sparse setting with a large number ($q = 1000$) of nuisance variables and high correlation $\rho' = 0.9$.

		Method			
	α	FLH1	FLH2	DR	BM
level	0.05	.0236	.0319	.0358	.0006
	0.01	.0040	.0074	.0057	.0000
	0.001	.0003	.0006	.0001	.0000
power	0.05	.4766	.5317	.7127	.2115
	0.01	.3106	.4254	.4137	.1042
	0.001	.1303	.2500	.1344	.0407

Note: Power is shown for $\beta = 2$.

Table 8. Same sparse setting as at Table 3 but with very heavy-tailed errors.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0345	.0313	.0336	.0034	.0215	.022
	0.01	.0059	.0051	.0053	.0001	.0043	.004
	0.001	.0005	.0002	.0002	.0000	.0006	.002
power	0.05	.4498	.5493	.4593	.2173	.5433	.566
	0.01	.2295	.3353	.2016	.0730	.3173	.390
	0.001	.0780	.1309	.0492	.0151	.1374	.215

4.3. Violations of the Gaussian model

Permutation tests can be robust to violations of the standard linear model, such as non-normality and heteroscedasticity [12,13]. The power of parametric methods is often substantially decreased when the residuals have heavy tails. The power of the permutation tests is more robust to such deviations from normality. This is illustrated in Table 8. Here, the data distribution was the same as in the setting corresponding to Table 3, except that the errors $\boldsymbol{\epsilon}$ were not standard normally distributed, but had very heavy (cubed exponential) tails, scaled such that the errors had standard deviation 1. Note in Table 8 that the permutation and bootstrap methods still had roughly the same power as at Table 3, while the power of BM and ZZ was strongly reduced compared to Table 3.

As a second type of violation of the standard linear model, we considered heteroscedasticity. We simulated errors ϵ_i which were normally distributed, but with standard deviation proportional to the absolute value covariate of interest, $|X_i|$. We again took $\gamma_2 = \gamma_3 = 1$, $\gamma_4 = \cdots = \gamma_{60} = 0$. We took $\rho' = 0$ for illustration, since in that case the method ZZ by Zhang and Zhang [21] turned out to be very anti-conservative under heteroscedasticity. Otherwise, the simulated data were again as those used for Table 3. The results are in Table 9. Note that despite the heteroscedasticity, the permutation-based tests provided appropriate type I error control. The bootstrap approach BO of Dezeure et al. [11] seemed to be anti-conservative for small α. The test BM from Bühlmann [37] had higher power than the permutation methods in this specific setting, but was anti-conservative for small α.

Table 9. Sparse setting with heteroscedastic errors, $\rho' = 0$, $n = 30, q = 60$.

		Method					
	α	FLH1	FLH2	DR	BM	ZZ	BO
level	0.05	.0352	.0354	.0271	.0338	.1490	.077
	0.01	.0065	.0069	.0050	.0109	.0648	.028
	0.001	.0010	.0009	.0008	.0029	.0280	.011
power	0.05	.7901	.8060	.7855	.9403	.9902	.982
	0.01	.6787	.6861	.6454	.8534	.9741	.936
	0.001	.4910	.4909	.4498	.6903	.9332	.830

Note: Power is shown for $\beta = 1.5$.

Table 10. Multi-dimensional $\boldsymbol{\beta} \in \mathbb{R}^{10}$.

		Simulation setting		
	α	Setting 1	Setting 2	Setting 3
level	0.05	.0174	.0197	.0330
	0.01	.0023	.0024	.0055
	0.001	.0004	.0002	.0002
power	0.05	.4443	.5098	.6286
	0.01	.3740	.4552	.5731
	0.001	.2503	.3788	.4736

Note: Power is shown for $\boldsymbol{\beta} = (3, 2, 1, 0, \ldots, 0)$.

In the simulations underlying Table 9, we did not use sign-flipping, which is known to be robust to heteroscedasticity [12,13]. Surprisingly, our tests nevertheless provided appropriate type I control. We also performed these simulations with sign-flipping instead of permutation (results not shown), which further reduced the level of our tests, but also somewhat reduced the power.

4.4. Multi-dimensional parameter of interest

We simulated the test of Section 3.3 for multi-dimensional $\boldsymbol{\beta}$. As the combination statistic we used $\Psi(t_1, \ldots, t_d) = \max(t_1, \ldots, t_d)$. The parameter of interest $\boldsymbol{\beta}$ had dimension 10 and there were 490 nuisance variables, i.e. $\dim(\boldsymbol{\gamma}) = 491$, since γ_1 is the intercept. The outcome Y followed a Gaussian model, as in Section 4.2. We considered three simulation settings. The nuisance parameters were $\gamma_2 = 3, \gamma_3 = 2, \gamma_4 = 1, \gamma_5 = \cdots = \gamma_{491} = 0$ in the first two settings and $\gamma_2 = \cdots . = \gamma_{101} = 0.03, \gamma_{102} = \cdots = \gamma_{491} = 0$ in the third setting. The covariates had a multinormal distribution with homogeneous correlation $\rho' = 0.5$ in the first setting and $\rho' = 0.9$ in the last two settings. The results are in Table 10. The test provided appropriate type I error control.

We conclude from the simulations of Section 4 that our tests provide good type I error control and are rather robust to several types of model misspecification. The method ZZ from Zhang and Zhang [21] was often relatively powerful, but was quite anti-conservative in several scenarios. The bootstrap approach BO of Dezeure et al. [11] was also anti-conservative in several scenarios, but less so. The method BM from Bühlmann et al. [37] tended to be relatively conservative.

Table 11. Real data analysis.

	Fraction of rejected hypotheses					
α	FLH1	FLH2	DR	BM	ZZ	BO
0.05	.0005	.0259	.0428	0	.0135	.0272
0.01	0	.0071	.0066	0	.0022	.0051
0.001	0	.0002	.0012	0	.0007	.0024
0.0001	0	0	0	0	0	0

Note: For different p-value cut-offs α, the fraction of rejected hypotheses is shown.

5. Data analysis

We analyse a dataset about riboflavin (vitamin B2) production with *B. subtilis*. This dataset is called *riboflavin* and is publicly available [36]. It contains normalized measurements of expression rates of 4088 genes from $n = 71$ samples. We use these as input variables. Further, for each sample the dataset contains the logarithm of the riboflavin production rate, which is our one-dimensional outcome of interest. We (further) standardized the expression levels by subtracting the means and dividing by the standard deviations. We also shifted the outcome values to have mean zero.

For every $1 \leq i \leq 4088$, we tested the hypothesis H_i that the outcome was independent of the expression level of gene i, conditional on the other expression levels. We used the same tests as considered in the simulations. This time we used $w = 2 \cdot 10^5$ permutations per test.

The results of the analysis are summarized in Table 11. The columns correspond to the same methods as considered in Section 4. For every method, the fraction of rejections is shown for different p-value cut-offs α. The fraction of rejections is the number of rejected hypotheses divided by 4088, the total number of hypotheses. The hypotheses that were rejected, were those with p-values smaller than or equal to the cut-off α.

With most methods we obtain many p-values smaller than 0.05. This is not the case for the test BM by Bühlmann [37], which is known to be relatively conservative. After Bonferroni's multiple testing correction, we reject no hypotheses with any method, suggesting there is no strong signal in the data. Van de Geer et al. [22] also obtained such a result with this dataset.

6. Discussion

We have proposed novel permutation methods for testing in linear models, where the number of nuisance variables may be much larger than the sample size. Advantages of permutation approaches include robustness to certain violations of the standard linear model and compatibility with powerful permutation-based multiple testing methods.

We have proposed two novel permutation approaches, Freedman–Lane HD and Double Residualization. Within these approaches some variations are possible, with respect to how the regularization parameters are chosen and which test statistics are used. Our methods provided excellent type I error rate control in a wide range of simulation settings. In particular we considered settings with anti-sparsity, high correlations among the covariates,

clustered covariates, fat-tailedness of the outcome variable and heteroscedasticity. The simulation study was limited to settings with multivariate normal covariates. Future research may address more scenarios.

We compared our methods to the parametric tests in Bühlmann [37] and Zhang and Zhang [21] and to the bootstrap approach in Dezeure et al. [11]. One advantage of our methods compared to those in Bühlmann [37] and Zhang and Zhang [21], is that they are defined in the case that the parameter of interest is multi-dimensional. Further, our tests tended to have higher power than the method by Bühlmann [37]. The test by Zhang and Zhang [21] had relatively good power, but was rather anti-conservative in several scenarios, for example under anti-sparsity and heteroscedasticity. The bootstrap approach of Dezeure et al. [11] was also anti-conservative in some scenarios, but less so. Our permutation tests provided appropriate type I error control in all scenarios. Moreover, our permutation tests were computationally much faster than the bootstrap method.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jesse Hemerik http://orcid.org/0000-0002-9811-1336

References

[1] Albajes-Eizagirre A, Solanes A, Vieta E, et al. Voxel-based meta-analysis via permutation of subject images (psi): theory and implementation for sdm. NeuroImage. 2019;186:174–184.
[2] Berrett TB, Wang Y, Barber RF, et al. The conditional permutation test. arXiv preprint arXiv:1807.05405; 2018.
[3] Ganong P, Jäger S. A permutation test for the regression kink design. J Amer Statist Assoc. 2018;113(522):494–504.
[4] He HY, Basu K, Zhao Q, et al. Permutation p-value approximation via generalized stolarsky invariance. Ann Stat. 2019;47(1):583–611.
[5] Hemerik J, Goeman JJ. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. J R Stat Soc Ser B (Stat Methodol). 2018;80(1):137–155.
[6] Hemerik J, Solari A, Goeman JJ. Permutation-based simultaneous confidence bounds for the false discovery proportion. Biometrika. 2019;106(3):635–649.
[7] Meinshausen N, Maathuis MH, Bühlmann P. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. Ann Stat. 2011;39(6):3369–3391.
[8] Rao K, Drikvandi R, Saville B. Permutation and bayesian tests for testing random effects in linear mixed-effects models. Stat Med. 2019;38:5034–5047.
[9] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Nat Acad Sci. 2001;98(9):5116–5121.
[10] Fisher RA. 'The coefficient of racial likeness' and the future of craniometry. J Anthropol Inst Great Britain Ireland. 1936;66:57–63.
[11] Dezeure R, Bühlmann P, Zhang C-H. High-dimensional simultaneous inference with the bootstrap. Test. 2017;26(4):685–719.
[12] Hemerik J, Goeman JJ, Finos L. Robust testing in generalized linear models by sign-flipping score contributions. J R Stat Soc Ser B (Methodol). 2020;82(3):841–864.
[13] Winkler AM, Ridgway GR, Webster MA, et al. Permutation inference for the general linear model. Neuroimage. 2014;92:381–397.
[14] Meinshausen N. False discovery control for multiple tests of association under general dependence. Scand J Stat. 2006;33(2):227–237.

[15] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.
[16] Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. Vol. 279. New York: John Wiley & Sons; 1993.
[17] Anderson MJ, Legendre P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Statistical Computation and Simulation¡/DIFdel¿J Stat Comput Simul. 1999;62(3):271–303.
[18] Anderson MJ, Robinson J. Permutation tests for linear models. Aust N Z J Stat. 2001;43(1):75–88.
[19] Winkler AM, Ridgway GR, Douaud G, et al. Faster permutation inference in brain imaging. NeuroImage. 2016;141:502–516.
[20] Freedman D, Lane D. A nonstochastic interpretation of reported significance levels. J Bus Econ Stat. 1983;1(4):292–298.
[21] Zhang C-H, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B (Stat Methodol). 2014;76(1):217–242.
[22] Van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat. 2014;42(3):1166–1202.
[23] Dezeure R, Bühlmann P, Meier L. High-dimensional inference: confidence intervals, *p*-values and R-software hdi. Sci Stat Sci. 2015;30(4):533–558.
[24] Chernozhukov V, Hansen C, Spindler M. hdm: high-dimensional metrics. R Journal. 2016;8(2):185–199.
[25] Kennedy FE. Randomization tests in econometrics. J Bus Econ Stat. 1995;13(1):85–94.
[26] Kennedy PE, Cade BS. Randomization tests for multiple regression. Commun Statist Simul Comput. 1996;25(4):923–936.
[27] Hall P, Titterington D. The effect of simulation order on level accuracy and power of Monte Carlo tests. J R Stat Soc Ser B (Methodol). 1989;51(3):459–467.
[28] Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. Biometrics. 1991;47(2):757–762.
[29] Agresti A. Foundations of linear and generalized linear models. Hoboken (NJ): John Wiley & Sons; 2015.
[30] Fisher RA. The distribution of the partial correlation coefficient. Metron. 1924;3:329–332.
[31] Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993;35(2):109–135.
[32] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics. New York; 2009.
[33] Zhu Y, Bradic J. Significance testing in non-sparse high-dimensional linear models. Electron J Stat. 2018;12(2):3312–3364.
[34] Knijnenburg TA, Wessels LF, Reinders MJ, et al. Fewer permutations, more accurate *p*-values. Bioinformatics. 2009;25(12):i161–i168.
[35] Pesarin F, Salmaso L. Permutation tests for complex data: theory, applications and software. Chichester: John Wiley & Sons; 2010.
[36] Bühlmann P, Kalisch M, Meier L. High-dimensional statistics with a view toward applications in biology. Annu Rev Stat Appl. 2014;1:255–278.
[37] Bühlmann P. Statistical significance in high-dimensional linear models. Bernoulli. 2013;19(4):1212–1242.