Robust testing in generalized linear models by sign-flipping score contributions[*]

Jesse Hemerik[†‡], Jelle Goeman[§] and Livio Finos[¶]

May 13, 2020

Abstract

Generalized linear models are often misspecified due to overdispersion, heteroscedasticity and ignored nuisance variables. Existing quasi-likelihood methods for testing in misspecified models often do not provide satisfactory type-I error rate control. We provide a novel semi-parametric test, based on sign-flipping individual score contributions. The tested parameter is allowed to be multi-dimensional and even high-dimensional. Our test is often robust against the mentioned forms of misspecification and provides better type-I error control than its competitors. When nuisance parameters are estimated, our basic test becomes conservative. We show how to take nuisance estimation into account to obtain an asymptotically exact test. Our proposed test is asymptotically equivalent to its parametric counterpart.

Keywords: GLM, Heteroscedasticity, High-dimensional, Permutation, Robust, Score Test, Semi-parametric, Sign-flipping.

1 Introduction

We consider the problem of testing hypotheses about parameters in potentially misspecified generalized linear models (GLMs). The types of misspecification that we consider include overdispersion and heteroscedasticity. When the model is misspecified, the traditional parametric tests tend to lose their properties, for example because they estimate the Fisher information under incorrect assumptions. By a parametric test we mean a test which fully relies on an assumed parametric model (Pesarin, 2015) to compute the null distribution of the test statistic.

---

[*]To appear in *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[†]Department of Biostatistics, Oslo Centre for Biostatistics and Epidemiology, University of Oslo

[‡]Address for correspondence: Jesse Hemerik, Oslo Centre for Biostatistics and Epidemiology, P.O.Box 1122 Blindern, 0317 Oslo, Norway. e-mail: jesse.hemerik@medisin.uio.no

[§]Biomedical Data Sciences, Leiden University Medical Center, the Netherlands

[¶]Department of Developmental Psychology and Socialization, University of Padua, Italy

When a parametric model to be tested is potentially misspecified, the most obvious approach is to extend the model with more parameters, e.g. to add an overdispersion parameter. However, such approaches still require assumptions, for example that the overdispersion is constant. Hence a fully parametric approach is not always the best option.

Another well-known approach to testing in possibly misspecified GLMs is to use a Wald-type test, where a sandwich estimate of the variance of the coefficient estimate is used. The sandwich estimate corrects for the potentially misspecified variance. As long as the linear predictor and link are correct, such a test is asymptotically exact under mild assumptions. We call a test asymptotically exact if its rejection probability is asymptotically known under the null hypothesis. For small samples, however, sandwich estimates often perform poorly and the test can be very liberal (Boos, 1992; Freedman, 2006; Maas and Hox, 2004; Kauermann and Carroll, 2000).

Recent decades have seen an increase in the use of permutation approaches for various testing problems (Tusher et al., 2001; Pesarin, 2001; Chung and Romano, 2013; Pauly et al., 2015; Winkler et al., 2016; Hemerik and Goeman, 2018a; Ganong and Jäger, 2018). These methods are useful since they require few parametric assumptions. Especially when multiple hypotheses are tested, permutation methods are often powerful since they can take into account the dependence structure in the data (Westfall and Young, 1993; Hemerik and Goeman, 2018b; Hemerik et al., 2019). In the past, permutation methods have already been used to test in linear models (Winkler et al., 2014, and references therein). Rather than permutations, sometimes other transformations are used, such as rotations (Solari et al., 2014) and sign-flipping of residuals (Winkler et al., 2014). The existing permutation tests for GLMs, however, are limited to models with identity link function.

Like some existing methods for testing in linear models, this paper presents a sign-flipping approach. Our approach is new however, since rather than flipping residuals, we flip individual score contributions (note that the score, the derivative of the log-likelihood, is a sum of n individual score contributions). Moreover, we allow testing in a wide range of models, not only regression models with identity link. Under mild assumptions, the only requirement for the test to be asymptotically exact, is that the individual score contributions have mean 0. Consequently, if the link function is correct, our method is often robust against several types of model specification, such as arbitrary overdispersion, heteroscedasticity and, in some cases, ignored nuisance parameters.

The main reason for this robustness is that we do not require to estimate the variance of the score, the Fisher information. Rather, we perform a permutation-type test based on the score contributions, where rather than permutation, we use sign-flipping. Intuitively, the advantage of this approach over explicitly estimating the variance, is the following: if the score contributions are independent and perfectly symmetric around zero under the null, then our test is exact for small n, even if the score contributions have misspecified variances and shapes (Pesarin and Salmaso, 2010b). A parametric test, on the other hand, is then usually not exact.

In case nuisance parameters are estimated, the individual score contributions become dependent and our basic sign-flipping test is no longer asymptotically exact. To

deal with this problem, we consider the *effective score*, which is less dependent on the nuisance estimate than the basic score (Hall and Mathiason, 1990; Marohn, 2002). In this case we need slightly more assumptions: the variance misspecification is not always allowed to depend on the covariates. The resulting test is asymptotically exact.

The methods in this paper have been implemented in the R package *flipscores*, available on CRAN.

In Section 2 we consider the scenario that no nuisance effects need to be estimated. In Section 3 we show how the estimation of nuisance effects can be taken into account. Section 4 provides tests of hypotheses about parameters of more than one dimension. Section 5 contains simulations and Section 6 an analysis of real data.

2 Models with known nuisance parameters

Consider random variables $\nu_1, ..., \nu_n$, which satisfy Assumption 1 below. These will often be individual score contributions (see Section 3, Rao, 1948, or Hall and Mathiason, 1990, p. 86), but the results in Section 2.1 hold for any random variables satisfying this assumption.

Assumption 1. The random variables ν_i, $i \in \mathbb{N}$, are independent of each other, have finite variances and satisfy the following. For every $\epsilon > 0$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\nu_i^2 \mathbb{1}_{\{\nu_i/\sqrt{n} > \epsilon\}}\right) = 0.$$

Further, as $n \to \infty$, $s_n^2 := \frac{1}{n} \sum_{i=1}^{n} var(\nu_i) \to s^2$ for some constant $s^2 > 0$.

Throughout Section 2, we consider any null hypothesis H_0 which implies that $\mathbb{E}\nu_i = 0$ for all $1 \le i \le n$. If $\nu_1, ..., \nu_n$ are score contributions and H_0 is a point hypothesis, then under mild assumptions, $\mathbb{E}\nu_i = 0$ is satisfied under H_0.

A key assumption throughout Section 2 is that the ν_i, $i \in \mathbb{N}$, are independent. As soon as nuisance parameters need to be estimated, however, score contributions become dependent. Section 3 is devoted to dealing with estimated nuisance.

2.1 Basic sign-flipping test

Let $\alpha \in [0, 1)$. For any $a \in \mathbb{R}$, let $\lceil a \rceil$ be the smallest integer which is larger than or equal to a and let $\lfloor a \rfloor$ be the largest integer which is at most a. Given values $T_1^n, ..., T_w^n \in \mathbb{R}$, we let $T_{(1)}^n \le ... \le T_{(w)}^n$ be the sorted values and write $T_{[1-\alpha]}^n = T_{(\lceil (1-\alpha)w \rceil)}^n$.

Throughout this paper, $w \in \{2, 3, ...\}$ denotes the number of random sign-flipping transformations to be used. Define $g_1 = (1, ..., 1) \in \mathbb{R}^n$ and for every $2 \le j \le w$ let $g_j = (g_{j1}, ..., g_{jn})$ be independent and uniformly distributed on $\{-1, 1\}^n$. Throughout the rest of Section 2, for every $1 \le j \le w$, we let

$$T_j^n = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_i.$$

We now state that the basic sign-flipping test is asymptotically exact for the point null hypothesis H_0 that implies $\mathbb{E}\nu_i = 0$, $1 \leq i \leq n$. All proofs are in the appendices.

Theorem 1. *Suppose that Assumption 1 holds. Consider the test that rejects H_0 if and only if $T_1^n > T_{\lceil 1-\alpha \rceil}^n$. Then, as $n \to \infty$, the rejection probability of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$ under H_0. Moreover, the statistics $T_1^n, ..., T_w^n$ are asymptotically normal and independent with mean 0 and common variance $\lim_{n \to \infty} s_n^2$ under H_0.*

We now provide an extension of Theorem 1 to interval hypotheses. The proof is a straightforward adaptation of the proof of Theorem 1.

Corollary 2 (Interval hypotheses)**.** *Suppose Assumption 1 holds. Consider a null hypothesis H' which implies that $\mathbb{E}\nu_i \leq 0$ (respectively $\mathbb{E}\nu_i \geq 0$) for all $1 \leq i \leq n$. Then for every $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that under H', for every $n > N$, $\mathbb{P}\{T_1^n > T_{(\lceil (1-\alpha)w \rceil)}^n\}$ (respectively $\mathbb{P}\{T_1^n < T_{(\lfloor \alpha w + 1 \rfloor)}^n\}$) is at most $\lfloor \alpha w \rfloor / w + \epsilon$.*

The following corollary extends Theorem 1 to two-sided tests. The proof is analogous to that of Theorem 1.

Corollary 3 (Two-sided test)**.** *Suppose Assumption 1 holds. Consider $\alpha_1, \alpha_2 \in \{0/w, 1/w, ..., (w-1)/w\}$. Under H_0, as $n \to \infty$,*

$$\mathbb{P}\left[\left\{T_1^n < T_{(\alpha_1 w + 1)}^n\right\} \cup \left\{T_1^n > T_{((1-\alpha_2)w)}^n\right\}\right] \to \alpha_1 + \alpha_2.$$

Note that our test does not rely on an approximate symmetry assumption (as e.g. Canay et al., 2017). Indeed, even if the scores are very skewed, asymptotically the test of Theorem 1 is exact. However, if the ν_i are symmetric, then even for small n the size is always at most α, as noted in the following proposition. A special case of this result is already discussed in Fisher (1935, §21), where every element of $\{1, -1\}^n$ is used once.

Proposition 4. *Suppose that $\nu_1, ..., \nu_n$ are independent and continuous and that under H_0, for each $1 \leq i \leq n$, $\nu_i \overset{d}{=} -\nu_i$. Then the size of the test of Theorem 1 is at most α for any $n \in \mathbb{N}$. Moreover, if $g_2, ..., g_w$ are uniformly drawn from $\{1, -1\}^n \setminus (1, ..., 1)$ without replacement (so that only g_1 takes the value $(1, ..., 1)$), then the rejection probability under H_0 is exactly $\lfloor \alpha w \rfloor / w$. (Note that w cannot exceed 2^n then.)*

If the g_j are drawn with replacement or the ν_i are discrete, then under H_0 the rejection probability of the test of Proposition 4 is (slightly) smaller than $\lfloor \alpha w \rfloor / w$ for finite n, due to the possibility of ties among the test statistics T_j^n, $1 \leq j \leq w$. Otherwise the rejection probability under H_0 is $\lfloor \alpha w \rfloor / w$.

Note that when the rejection probability under H_0 is $\lfloor \alpha w \rfloor / w$, it can be advantageous to take w such that α is a multiple of $1/w$, to exhaust the nominal level.

In Theorem 1, we did not assume continuity of the observations ν_i. There, under the mild Assumption 1, for $n \to \infty$, $\mathbb{P}(T_j^n = T_k^n) \to 0$ for any $1 \leq j < k \leq w$, regardless of the distribution of the ν_i. This allows using Theorem 1 for discrete GLMs.

2.2 Robustness

As a main example we consider the exponential family, i.e., suppose independent variables $Y_1, ..., Y_n$ have densities of the form

$$f(y_i; \eta_i) = \exp\left\{\frac{y_i\eta_i - b(\eta_i)}{a_i} + c(y_i)\right\},$$

where $\eta_i = x_i\beta + \boldsymbol{z}_i\boldsymbol{\gamma}$, $x_i, \beta \in \mathbb{R}$, $\boldsymbol{z}_i, \boldsymbol{\gamma} \in \mathbb{R}^m$ for some $m \in \mathbb{N}$. Here β is the coefficient of interest and presently we assume the other covariates $\boldsymbol{\gamma}$ to be known. The canonical link function g satisfies $\eta_i = g(\mu_i)$, where $g^{-1}(\eta_i) = \mu_i = \mathbb{E}(y_i) = b'(\eta_i)$ and $a_i = \text{var}(y_i)/b''(\eta_i)$ (Agresti, 2015). For $H_0 : \beta = \beta_0$, the score $\sum_{i=1}^n \nu_i = \sum_{i=1}^n \frac{\partial}{\partial\beta} \log\{f(y_i; \eta_i)\}|_{\beta=\beta_0}$ is

$$\sum_{i=1}^n \frac{x_i(y_i - b'(\eta_i))}{a_i}\bigg|_{\beta=\beta_0} = \sum_{i=1}^n \frac{x_i(y_i - \mathbb{E}(y_i))}{a_i}\bigg|_{\beta=\beta_0}.$$

For example, the Poisson model has link function $g = \log$, $b(\eta_i) = \exp(\eta_i)$, $a_i = 1$ and $c(y_i) = -\log(y_i!)$. Hence $\mathbb{E}(y_i) = b'(\eta_i) = \exp(\eta_i)$. Thus the score function is

$$\sum_{i=1}^n x_i(y_i - \mu_i)|_{\beta=\beta_0} = \sum_{i=1}^n x_i\{y_i - \exp(x_i\beta_0 + \boldsymbol{z}_i\boldsymbol{\gamma})\}.$$

For the normal distribution, $a_i = \sigma^2$, so that the score is

$$\sum_{i=1}^n \frac{x_i(y_i - \eta_i)}{\sigma^2}\bigg|_{\beta=\beta_0}.$$

Apart from some mild assumptions, the main assumption made in Theorem 1 is that $\mathbb{E}(\nu_i) = 0$, $i = 1, ..., n$. This is satisfied as soon as $\mu_i|_{\beta=\beta_0}$ is the true expected value of Y_i. Then the test is asymptotically exact even if the a_i are misspecified, i.e., if the variance or distributional shape of Y_i is misspecified. The a_i are even allowed to be misspecified by a factor which depends on the covariates, as long as Assumption 1 holds.

As a concrete example, consider the normal model with identity link function, which assumes that $var(Y_1) = ... = var(Y_n)$. If the real distribution is heteroscedastic, then the test will still be exact for finite n, since the ν_i are symmetric. The parametric test, however, loses its properties, for example because the estimated variance does not have the assumed chi-squared distribution. In Section 5 it is illustrated that our approach can be much more robust against heteroscedasticity than a parametric test.

Another example is the situation where the model is Poisson, i.e., $var(Y_i) = \mu_i$ is assumed, but in reality $var(Y_i) > \mu_i$, a form of overdispersion which occurs very often in practice. Then the parametric score test underestimates the Fisher information and is anti-conservative. To take the overdispersion into account it could be explicitly estimated. However, if the overdispersion factor is not constant, but depends on the

5

covariates, then again the parametric test loses its properties. Theorem 1, however, often still applies, so that an asymptotically exact test is obtained.

Further, note that if $\mathbb{E}(Y_i)$ depends on a nuisance variable Z_i^l which is latent and ignored, where Z_i^l is independent of X_i, then the test may still be valid. The reason is that marginal over Z_i^l, $\mathbb{E}(Y_i)$ may still be computed correctly (see, for example, Section 5.2). Such latent nuisance variables will increase the variance of Y_i, however, which can pose a problem for the classical parametric score test, which needs to compute the Fisher information. When the latent variable is not independent of X_i, this usually does pose a problem for our test (even as $n \to \infty$), since $\mathbb{E}(Y_i - \mu_i)$ becomes dependent on X_i under H_0.

3 Taking into account nuisance estimation

Consider independent and identically distributed pairs $(\boldsymbol{X}_i, Y_i)$, $i = 1, ..., n$, where $\boldsymbol{X}_i$ is some covariate vector and $Y_i \in \mathbb{R}$ has distribution $\mathbb{P}_{\beta, \boldsymbol{\gamma}_0, \boldsymbol{X}_i}$, which depends on parameter of interest $\beta \in \mathbb{R}$ and unknown nuisance parameter $\boldsymbol{\gamma}_0$, which lies in a set $\mathbb{G} \subseteq \mathbb{R}^{k-1}$, where k is the total number of modeled parameters. We will discuss the issues arising from estimating $\boldsymbol{\gamma}_0$ and propose a solution, which allows us to obtain an asymptotically exact test based on score flipping. Note that in this paper, the model defined above is the model considered by the user. It is the model used to compute the scores. We consider this model to be correct, unless explicitly stated otherwise, for example in Section 3.2. The parameter $\boldsymbol{\gamma}_0$ is part of the model considered by the user, so it is always modeled and estimated. For example, $\boldsymbol{\gamma}_0$ never represents ignored overdispersion.

We consider the null hypothesis $H_0 : \beta = \beta_0 \in \mathbb{R}$. Generalizations to interval hypotheses and two-sided tests can be obtained as in Corollaries 2 and 3. The case that the parameter of interest is multi-dimensional is considered in Section 4.

Suppose that for all $\boldsymbol{\gamma} \in \mathbb{G}$, $\mathbb{P}_{\beta, \boldsymbol{\gamma}, \boldsymbol{X}_i}$ has a density $f_{\beta, \boldsymbol{\gamma}, \boldsymbol{X}_i}$ around β_0 with respect to some dominating measure. For $1 \leq i \leq n$ write

$$\nu_{\boldsymbol{\gamma}, i} = \frac{\partial}{\partial \beta} \log f_{\beta, \boldsymbol{\gamma}, \boldsymbol{X}_i}(Y_i)|_{\beta = \beta_0},$$

where we assume the derivative exists. The value ν_i is the score for the i-th observation. Under H_0, $\mathbb{E}(\nu_{\boldsymbol{\gamma}_0, i}) = 0$, $i = 1, ..., n$. The score for all n observations simultaneously is $n^{1/2} S_{\boldsymbol{\gamma}}$, where

$$S_{\boldsymbol{\gamma}} = n^{-1/2} \sum_{i=1}^{n} \nu_{\boldsymbol{\gamma}, i}.$$

Assume that $\hat{\boldsymbol{\gamma}}$ is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\gamma}_0$, taking values in $\mathbb{G}$. For every $1 \leq i \leq n$, let

$$\boldsymbol{\nu}_{\hat{\boldsymbol{\gamma}}, i}^{(k-1)} = \frac{\partial}{\partial \boldsymbol{\gamma}} \log f_{\beta_0, \boldsymbol{\gamma}, \boldsymbol{X}_i}(Y_i) \Big|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}} \in \mathbb{R}^{k-1}$$

denote the $(k-1)$-vector of score contributions for the nuisance parameters, which is

assumed to exist. Let

$$S_{\hat{\gamma}}^{(k-1)} = n^{-1/2} \sum_{i=1}^{n} \nu_{\hat{\gamma},i}^{(k-1)} \in \mathbb{R}^{k-1}$$

be the vector of nuisance scores.

For $1 \leq j \leq w$, let the superscript j denote that g_j has been applied:

$$S_{\hat{\gamma}}^{j} = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_{\hat{\gamma},i},$$

$$S_{\hat{\gamma}}^{(k-1),j} = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_{\hat{\gamma},i}^{(k-1)}.$$

3.1 Asymptotically exact test

When the nuisance parameter $\boldsymbol{\gamma}_0$ is unknown, it needs to be estimated, which is typically done by maximizing the likelihood of the data under the null hypothesis. The distribution of $S_{\hat{\gamma}}$ can be substantially different from that of $S_{\boldsymbol{\gamma}_0}$, the score based on the true nuisance parameters. Indeed, under the null hypothesis, the asymptotic variance of $S_{\hat{\gamma}}$ is not the Fisher information, but the *effective Fisher information* (Rippon and Rayner, 2010; Rayner, 1997; Hall and Mathiason, 1990; Marohn, 2002; Cox and Hinkley, 1979, section 9.3), which is also the asymptotic variance of the *effective score*, defined below. The effective information is smaller than the information, given that the score for the parameter of interest and the nuisance score are correlated. Intuitively, the reason is that the nuisance variable will be used to explain part of the apparent effect of the variable of interest, also asymptotically.

The estimation of $\boldsymbol{\gamma}_0$ makes the summands $\nu_{\hat{\gamma},1}, ..., \nu_{\hat{\gamma},n}$ underlying $S_{\hat{\gamma}}$ correlated, in such a way that $var(S_{\hat{\gamma}}) < var(S_{\boldsymbol{\gamma}_0})$ (if the score is correlated with the nuisance score). Note however that after random flipping, the summands are not correlated anymore. This means that the variance of $S_{\hat{\gamma}}$ is asymptotically smaller than the variance of $S_{\hat{\gamma}}^{j}$, $2 \leq j \leq w$ (see the proof of Theorem 5). Hence, using $\nu_{\hat{\gamma},1}, ..., \nu_{\hat{\gamma},n}$ in the test of Theorem 1 can lead to a conservative test, even as $n \to \infty$.

To make the test asymptotically exact again, we would like to adapt the individual scores such that they are less dependent on the random variation of $\hat{\boldsymbol{\gamma}}$. We do this by considering the so-called *effective score*, which "is 'less dependent' on the nuisance parameter than the usual score statistic" (Marohn, 2002, p. 344).

The effective score $S_{\hat{\gamma}}^{*}$ and the underlying summands $\nu_{\hat{\gamma},i}^{*}$, $i = 1, ..., n$ (which we assume have nonzero variance for $\hat{\gamma} = \gamma_0$) are defined as

$$S_{\hat{\gamma}}^{*} = S_{\hat{\gamma}} - \hat{\mathcal{I}}_{12}' \hat{\mathcal{I}}_{22}^{-1} S_{\hat{\gamma}}^{(k-1)},$$

$$\nu_{\hat{\gamma},i}^{*} = \nu_{\hat{\gamma},i} - \hat{\mathcal{I}}_{12}' \hat{\mathcal{I}}_{22}^{-1} \nu_{\hat{\gamma},i}^{(k-1)},$$

so that

$$S_{\hat{\gamma}}^{*} = n^{-1/2} \sum_{i=1}^{n} \nu_{\hat{\gamma},i}^{*}.$$

7

Here

$$\hat{\mathcal{I}} = \begin{bmatrix} \hat{\mathcal{I}}_{11} & \hat{\mathcal{I}}'_{12} \\ \hat{\mathcal{I}}_{12} & \hat{\mathcal{I}}_{22} \end{bmatrix},$$

with $\hat{\mathcal{I}}_{11} \in \mathbb{R}$ and the $(k-1) \times (k-1)$ matrix $\hat{\mathcal{I}}_{22}$ assumed invertible, is a consistent estimate of the population Fisher information $\mathcal{I}$, which is assumed to exist and is the variance of $(\nu_{\gamma_0,i}, \boldsymbol{\nu}_{\gamma_0,i}^{(k-1)'})'$ marginal over $\boldsymbol{X}_i$, under H_0. The matrix $\mathcal{I}$ is assumed to be continuous in the parameters. In GLMs, typically $\hat{\mathcal{I}} = n^{-1} \boldsymbol{X}' \hat{\boldsymbol{W}} \boldsymbol{X}$, where $\boldsymbol{X}$ is the design matrix and $\hat{\boldsymbol{W}}$ the estimated weight matrix (Agresti, 2015, p. 126). Further, for $1 \le j \le w$ we write

$$S_{\hat{\gamma}}^{*j} = S_{\hat{\gamma}}^j - \hat{\mathcal{I}}'_{12} \hat{\mathcal{I}}_{22}^{-1} \boldsymbol{S}_{\hat{\gamma}}^{(k-1),j}.$$

As discussed, $S_{\hat{\gamma}}$ is not generally asymptotically equivalent to S_{γ_0}. The effective score $S_{\gamma_0}^*$ (based on $\hat{\mathcal{I}} = \mathcal{I}$) however is the residual from the projection of the score S_{γ_0} on the space spanned by the nuisance scores. Hence $S_{\gamma_0}^*$ is uncorrelated with the nuisance scores $\boldsymbol{S}_{\gamma_0}^{(k-1)}$ (Marohn, 2002, p. 344). Correspondingly, as noted in the proof of Theorem 5, under mild regularity assumptions $S_{\hat{\gamma}}^* = S_{\gamma_0}^* + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$, i.e., asymptotically the effective score is not affected by the nuisance estimation.

Note that if $\hat{\gamma}$ is the maximum likelihood estimate under H_0, then $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = \boldsymbol{0}$, so that $S_{\hat{\gamma}}^* = S_{\hat{\gamma}}$. The summands $\nu_{\hat{\gamma},i}^*$ and $\nu_{\hat{\gamma},i}$ are different, however, and the key point is that $S_{\hat{\gamma}}^* = S_{\gamma_0}^* + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$.

Like Marohn (2002), we assume that if $\xi \in \mathbb{R}$ and $\beta = \beta^n = \beta_0 + n^{-1/2}\xi$, then

$$S_{\hat{\gamma}} = S_{\gamma_0} - \mathcal{I}'_{12}\sqrt{n}(\hat{\gamma} - \gamma_0) + o_{\mathbb{P}_{\beta^n,\gamma_0}}(1),$$

$$\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = \boldsymbol{S}_{\gamma_0}^{(k-1)} - \mathcal{I}_{22}\sqrt{n}(\hat{\gamma} - \gamma_0) + o_{\mathbb{P}_{\beta^n,\gamma_0}}(1),$$

which is satisfied under mild assumptions such as continuous second order derivatives.

Theorem 5. *Consider the test of Theorem 1 with $T_j^n = S_{\hat{\gamma}}^{*j}$, $1 \le j \le w$. As $n \to \infty$, under H_0 the rejection probability converges to $\lfloor \alpha w \rfloor / w \le \alpha$.*

The test of Theorem 5 has a parametric counterpart, which uses that under H_0, $S_{\hat{\gamma}}^*$ is asymptotically normal with zero mean and known variance, the effective information (Marohn, 2002, p. 341). This test is asymptotically equivalent to the test of Theorem 5, as the following proposition says.

Proposition 6. *Let $\xi \in \mathbb{R}$ and suppose the true parameter satisfies $\beta = \beta^n = \beta_0 + n^{-1/2}\xi$. As in Theorem 5, let $T_j^n = S_{\hat{\gamma}}^{*j}$, $1 \le j \le w$. Define $\phi_{n,w} = \mathbb{1}_{\{T_1^n > T_{[1-\alpha]}^n\}}$ to be the test of Theorem 5. Let ϕ'_n be the parametric test $\mathbb{1}_{\{T_1^n > \sigma_0 \Phi(1-\alpha)\}}$, where $\sigma_0^2 \in \mathbb{R}$ is the effective Fisher information and Φ the cdf of the standard normal distribution. Then $\lim_{w \to \infty} \liminf_{n \to \infty} \mathbb{P}(\phi_{n,w} = \phi'_n) = 1$.*

3.2 Robustness

In Section 2.2 it was explained that the test of Theorem 1 is often robust against misspecification of the variance of the score. The test of Theorem 5 is also robust against certain forms of variance misspecification. An example is the case that $S_{\hat{\gamma}}$ and $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)}$ are misspecified by the same factor, see Proposition 7. This happens in particular if the variance is misspecified by a factor which is independent of the covariates.

Proposition 7. *Suppose that* $\hat{\boldsymbol{\mathcal{I}}} = n^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}$, *where* $\boldsymbol{X}$ *is an* $n \times k$ *design matrix with i.i.d. rows and* $\hat{\boldsymbol{W}}$ *a weight matrix. Consider a misspecification factor* $c_1 > 0$ *and misspecified scores*

$$\tilde{\nu}_{\hat{\gamma},i} = c_1 \nu_{\hat{\gamma},i}, \quad \tilde{\boldsymbol{\nu}}_{\hat{\gamma},i}^{(k-1)} = c_1 \boldsymbol{\nu}_{\hat{\gamma},i}^{(k-1)}, \quad i = 1,...,n.$$

Further, for $c_2 > 0$ *consider the misspecified weight matrix* $\tilde{\boldsymbol{W}} = c_2\hat{\boldsymbol{W}}$. *Let* $\tilde{\boldsymbol{\mathcal{I}}} = n^{-1}\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}$ *be the misspecified average Fisher information. Let* $\tilde{\nu}_{\hat{\gamma},i}^* = \tilde{\nu}_{\hat{\gamma},i} - \tilde{\boldsymbol{\mathcal{I}}}_{12}'\tilde{\boldsymbol{\mathcal{I}}}_{22}^{-1}\tilde{\boldsymbol{\nu}}_{\hat{\gamma},i}^{(k-1)}$ *be the misspecified effective scores,* $i = 1,...,n$. *Consider the test of Theorem 5, with* $S_{\hat{\gamma}}^{*j}$, $j = 1,...,w$, *replaced by the misspecified effective score*

$$\tilde{S}_{\hat{\gamma}}^{*j} = n^{-1/2}\sum_{i=1}^{n} g_{ji}\tilde{\nu}_{\hat{\gamma},i}^*.$$

Under H_0, *as* $n \to \infty$, *the rejection probability of this test converges to* $\lfloor\alpha w\rfloor/w \leq \alpha$.

Proposition 7 is useful, since it tells us that if in a GLM $var(Y_i)$ is misspecified by a constant, such that $\hat{\boldsymbol{W}}$ and the scores are misspecified by a constant, the resulting test is still asymptotically exact. In Proposition 7 we assume that the misspecification factors of the weights and the scores are the same for all observations. This is satisfied for example when the model is binomial or Poisson, but the true distribution is respectively quasi-binomial or quasi-Poisson. Moreover, in practice the test can be very robust against heteroscedasticity (see Section 5). The variance misspecification is not generally allowed to depend on the covariates, since then $S_{\hat{\gamma}}$ and $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)}$ can be misspecified by different factors asymptotically. There are exceptions however, see Sections 3.3 and 5.

When there are estimated nuisance parameters, one can sometimes nevertheless decide to use the test of Theorem 1 with the basic scores $\nu_{\hat{\gamma},i}$ plugged in (rather than using effective scores). Indeed, this test has been shown to be very robust to misspecification, as long as $\mathbb{E}\nu_{\hat{\gamma},i} = 0$, $1 \leq i \leq n$. It is asymptotically conservative if the score S_{γ_0} is correlated with the nuisance scores $\boldsymbol{S}_{\gamma_0}^{(k-1)}$, i.e., when $\boldsymbol{\mathcal{I}}_{12} \neq 0$. Hence, when using this test, it can be useful to redefine the covariates such that $\boldsymbol{\mathcal{I}}_{12} = 0$ (as in Cox and Reid, 1987). When $\hat{\boldsymbol{W}} = b\boldsymbol{I}$, $b > 0$, this means ensuring that the nuisance covariates are orthogonal to the covariate of interest. If the model is potentially misspecified, then the weights and hence $\boldsymbol{\mathcal{I}}_{12}$ are not asymptotically known, but the user could substitute a best guess for the weights.

3.3 An example

As discussed, the test of Theorem 5 is not generally asymptotically exact if the variance misspecification depends on the covariates. An important exception is the case where the model is

$$Y_i \sim N(\gamma_0 + \beta X_i, \sigma^2) \quad i = 1, ..., n, \tag{1}$$

where γ_0 is the unknown intercept and $X_i \in \mathbb{R}$. If the null hypothesis is $H_0 : \beta = \beta_0$, then γ_0 is a nuisance parameter that needs to be estimated. (We do not need to know σ and can simply substitute 1 for it.) Hence, we compute the effective score. Note that for $1 \le i \le n$,

$$\nu_{\hat{\gamma},i} = x_i(y_i - \hat{\mu}_i)/\sigma^2,$$
$$\boldsymbol{\nu}_{\hat{\gamma},i}^{(k-1)} = (y_i - \hat{\mu}_i)/\sigma^2.$$

Note that we can consistently estimate $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}$ by $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, so that the effective score contributions are

$$\nu_{\hat{\gamma},i}^* = (x_i - \overline{x})(y_i - \hat{\mu}_i)/\sigma^2.$$

Thus, the effective score contributions are exactly the basic score contributions after centering $x_1, ..., x_n$ around 0. Similarly, if $x_1, ..., x_n$ are already centered, then $\nu_{\hat{\gamma},i}$ and $\nu_{\hat{\gamma},i}^*$ coincide, since then $\hat{\boldsymbol{\mathcal{I}}}_{12} = 0$.

The test of Theorem 5 is not always asymptotically exact if $S_{\hat{\gamma}}$ and $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)}$ are misspecified by different factors. However, if $\hat{\boldsymbol{\mathcal{I}}}_{12} = 0$, then this does not apply anymore. The test of Theorem 5 then remains asymptotically exact and reduces to the test based on the basic score. For the model (1), this means that even if the misspecification of $var(Y_i)$ depends on X_i, we obtain an asymptotically exact test.

A particular case where this principle applies is the generalized Behrens-Fisher problem, where the aim is to test equality of the means μ^1 and μ^2 of two populations (or to test if $\mu^1 \le \mu^2$ or $\mu^1 \ge \mu^2$). In this problem, it is only assumed that two independent samples from these populations are available, without making other assumptions such as equal variances. It is well-known that this problem has no exact solution under normality (Pesarin and Salmaso, 2010b; Lehmann and Romano, 2005). Under mild assumptions, we obtain an asymptotically exact test for this problem. Pesarin and Salmaso (2010b) already suggested sign-flipping residuals to solve this problem. This is equivalent to flipping scores in our linear model (1) if $|x_1| = ... = |x_n|$.

4 Multi-dimensional parameter of interest

Until now we have considered hypotheses about a one-dimensional parameter $\beta \in \mathbb{R}$. Here we extend our results to hypotheses about a multi-dimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, $d \in \mathbb{N}$. Our tests are defined even if $d > n$, but in the theoretical results that follow we consider d fixed and let n increase to infinity. The extension to multi-dimensional $\boldsymbol{\beta}$ shares important characteristics with the test for a one-dimensional parameter, such as robustness and asymptotic equivalence with the parametric score test.

4.1 Asymptotically exact test

Our tests below are related to the existing nonparametric combination (NPC) methodology (Pesarin, 2001; Pesarin and Salmaso, 2010b,a). This is a very general permutation-based methodology that allows combining test statistics for many hypotheses into a single test of the intersection hypothesis. NPC can be extended to the score-flipping framework. Our tests below could be considered a special case of such an extension of the NPC methodology. This special case has certain power-optimality properties, discussed below.

The parametric score test has a classical extension to a hypothesis on a multi-dimensional parameter, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathbb{R}^d$ (Rao, 1948). We will extend our test in an analogous way. We first assume the nuisance $\boldsymbol{\gamma}_0 \in \mathbb{R}^{k-d}$ to be known. Since $\boldsymbol{\beta} \in \mathbb{R}^d$, the score is $\boldsymbol{S}_{\boldsymbol{\gamma}_0} = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nu}_{\boldsymbol{\gamma}_0,i} \in \mathbb{R}^d$, where

$$\boldsymbol{\nu}_{\boldsymbol{\gamma}_0,i} = \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{X}_i}(Y_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \in \mathbb{R}^d,$$

$1 \leq i \leq n$, which are now d-vectors. We assume the derivatives exist. About the elements of $\boldsymbol{\nu}_{\boldsymbol{\gamma},i}$ (and the nuisance scores considered later) we make the assumptions which are analogous to the earlier assumptions about $\nu_{\boldsymbol{\gamma},i}$.

Let $\hat{\boldsymbol{\mathcal{I}}}_{11}$ to be a consistent estimate of $\boldsymbol{\mathcal{I}}_{11}$, the $d \times d$ Fisher information for $\boldsymbol{\beta} \in \mathbb{R}^d$. Rao's classical statistic for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathbb{R}^d$ is

$$\boldsymbol{S}'_{\boldsymbol{\gamma}_0} \hat{\boldsymbol{\mathcal{I}}}_{11}^{-1} \boldsymbol{S}_{\boldsymbol{\gamma}_0} = \Big(n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nu}'_{\boldsymbol{\gamma}_0,i} \Big) \hat{\boldsymbol{\mathcal{I}}}_{11}^{-1} \Big(n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nu}_{\boldsymbol{\gamma}_0,i} \Big),$$

which asymptotically has a χ_d^2 distribution under H_0.

Instead of requiring a matrix $\hat{\boldsymbol{\mathcal{I}}}^{-1}$ which converges to the inverse of the Fisher Information, in our test that follows we allow to replace the Fisher information by any random matrix $\hat{\boldsymbol{V}}$ converging to some non-zero matrix $\boldsymbol{V}$. That is, we do not require the Fisher Information to be asymptotically known, just like in the one-dimensional case. The matrix $\boldsymbol{V}$ can be any matrix of preference, including $\boldsymbol{\mathcal{I}}_{11}^{-1}$ (if $\boldsymbol{\mathcal{I}}_{11}$ is invertible), or we can take $\hat{\boldsymbol{V}} = \boldsymbol{V} = \boldsymbol{I}$. We will discuss different choices of $\boldsymbol{V}$ shortly.

Theorem 8. *The result of Theorem 1 still applies if for $1 \leq j \leq w$ we define*

$$T_j^n = \Big(n^{-1/2} \sum_{i=1}^{n} g_{ji} \boldsymbol{\nu}'_{\boldsymbol{\gamma}_0,i} \Big) \hat{\boldsymbol{V}} \Big(n^{-1/2} \sum_{i=1}^{n} g_{ji} \boldsymbol{\nu}_{\boldsymbol{\gamma}_0,i} \Big).$$

In case the nuisance parameter $\boldsymbol{\gamma}_0$ is unknown and we have a $\sqrt{n}$-consistent estimate $\hat{\boldsymbol{\gamma}}$, we can use the same test, but with effective scores instead of basic scores plugged in. See Theorem 9. For multi-dimensional $\boldsymbol{\beta}$, the effective score contributions are

$$\boldsymbol{\nu}^*_{\hat{\boldsymbol{\gamma}},i} = \boldsymbol{\nu}_{\hat{\boldsymbol{\gamma}},i} - \hat{\boldsymbol{\mathcal{I}}}'_{12} \hat{\boldsymbol{\mathcal{I}}}_{22}^{-1} \boldsymbol{\nu}^{(k-d)}_{\hat{\boldsymbol{\gamma}},i} \in \mathbb{R}^d,$$

$1 \leq i \leq n$, where

$$\boldsymbol{\nu}^{(k-d)}_{\hat{\boldsymbol{\gamma}},i} = \frac{\partial}{\partial \boldsymbol{\gamma}} \log f_{\boldsymbol{\beta}_0,\boldsymbol{\gamma},\boldsymbol{X}_i}(Y_i)\Big|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} \in \mathbb{R}^{k-d}.$$

Here $\hat{\boldsymbol{\mathcal{I}}}_{12}$ and $\hat{\boldsymbol{\mathcal{I}}}_{22}$ are $(k-d) \times d$ and $(k-d) \times (k-d)$ matrices, respectively.

Theorem 9 (Unknown nuisance). *The result of Theorem 1 still applies if for $1 \leq j \leq w$ we define*

$$T_j^n = \left(n^{-1/2} \sum_{i=1}^n g_{ji} \boldsymbol{\nu}_{\hat{\gamma},i}^{*'}\right) \hat{\boldsymbol{V}} \left(n^{-1/2} \sum_{i=1}^n g_{ji} \boldsymbol{\nu}_{\hat{\gamma},i}^{*}\right).$$

The test of Theorem 9 is asymptotically equivalent to its parametric counterpart, as Proposition 10 states. In particular, if we take $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{\mathcal{I}}}^*)^{-1}$, where $(\hat{\boldsymbol{\mathcal{I}}}^*)^{-1}$ is a consistent estimate of the inverse of the effective Fisher information, then the test of Theorem 9 is asymptotically equivalent to the parametric score test (Hall and Mathiason, 1990, p. 86).

Proposition 10 (Equivalence with parametric counterpart). *Define T_j^n as in Theorem 9, $1 \leq j \leq w$. Let $\boldsymbol{\xi} \in \mathbb{R}^d$ and suppose the true value of the parameter of interest is $\boldsymbol{\beta} = \boldsymbol{\beta}^n = \boldsymbol{\beta}_0 + n^{-1/2} \boldsymbol{\xi}$. Let $\phi_{n,w} = \mathbb{1}_{\{T_1^n > T_{[1-\alpha]}^n\}}$. This is the test of Theorem 9. Let ϕ_n' be the parametric test $\mathbb{1}_{\{T_1^n > q_\alpha\}}$, where q_α is the $(1-\alpha)$-quantile of the distribution to which T_1^n converges as $n \to \infty$ under $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. (This is the χ_d^2 distribution if $\boldsymbol{V}$ is the inverse of the effective information matrix $\boldsymbol{\mathcal{I}}^*$). Then $\lim_{w\to\infty} \liminf_{n\to\infty} \mathbb{P}(\phi_{n,w} = \phi_n') = 1$.*

We have seen that the test of Theorem 1 is often robust against overdispersion and heteroscedasticity: as long as the score contributions have mean 0, the test is asymptotically exact, under very mild assumptions. Moreover, it is not required to estimate the Fisher information. The same applies to the multi-dimensional extension in Theorem 8.

The test that takes into account nuisance estimation (Theorem 9) uses effective scores, so that it does require estimating the information. However, as in the one-dimensional case, it can be seen that the test remains valid if the information matrix is asymptotically misspecified by a constant (as in Proposition 7). Additional robustness is illustrated with simulations in Section 5.5.

4.2 Connection with the global test

The test of Theorem 8 is related to the global test, which was developed in Goeman et al. (2004, 2006, 2011). We can combine the global test with the score-flipping approach. In certain cases, the resulting test coincides with the test of Theorem 8.

The global test is a parametric test of H_0. For the test to be defined, it is not required that $d \leq n$. For GLMs with canonical link function, the test statistic of the global test is

$$\boldsymbol{S}_{\gamma_0}' \boldsymbol{\Sigma} \boldsymbol{S}_{\gamma_0}, \tag{2}$$

with $\boldsymbol{\Sigma}$ a freely chosen positive (semi)definite $d \times d$ matrix (Goeman et al., 2006, 2011). The choice of $\boldsymbol{\Sigma}$ influences the power properties.

Note that when $\hat{\boldsymbol{V}} = \boldsymbol{\Sigma}$, the statistic (2) coincides with the statistic of Theorem 8. Thus, it immediately follows from our results that the global test can be combined with

our sign-flipping approach, leading to a test which becomes asymptotically exact as $n \to \infty$ and asymptotically equivalent to its parametric counterpart, the original global test (by Proposition 10). Combining the global test with sign-flipping is useful in the light of our robustness results. Moreover, the sign-flipping variant can be combined with existing permutation-based multiple testing methodology (Westfall and Young, 1993; Hemerik and Goeman, 2018b; Hemerik et al., 2019).

Goeman et al. (2006) provide results on the power properties of the global test as depending on the choice of $\mathbf{\Sigma}$. Since the global test is asymptotically equivalent to its sign-flipping counterpart, these results can be used as recommendations on the choice of $\hat{\mathbf{V}}$ in Theorem 8. In particular, according to Goeman et al. (2006, Section 8), taking $\hat{\mathbf{V}} = \mathbf{I}$ leads to good power if one expects that relatively much of the variance of $\mathbf{Y}$ is explained by the large variance principal components of the design matrix. If this is not the case, taking $\hat{\mathbf{V}}$ to be an estimate of the inverse of the Fisher information (if invertible) can provide better power. In general, the global test has optimal power on average (over $\boldsymbol{\beta}$) in a neighbourhood of $\boldsymbol{\beta}_0$ that depends on $\mathbf{\Sigma}$ (Goeman et al., 2006). Hence the same holds asymptotically for the test of Theorem 8, for GLMs with canonical link.

5 Simulations

To compare the tests in this paper with each other and existing tests, we applied them to simulated data. In particular we considered scenarios where the model was misspecified. Simulations with a multi-dimensional parameter of interest are in Section 5.5.

5.1 Overdispersion, heteroscedasticity and estimated nuisance

In Sections 5.1 and 5.2 the assumed model was Poisson, but in fact $Y_1, ..., Y_n$ were drawn from a negative binomial distribution.

The covariates $X, Z, Z^l \in \mathbb{R}$ were drawn from a multivariate normal distribution with zero mean and $var(X) = var(Z) = var(Z^l) = 1$. (For nonzero means, similar simulation results were obtained as below.) The response satisfied $\log\{\mathbb{E}(Y_i)\} = \log(\mu_i) = \eta_i =$

$$0 + \beta \cdot X_i + \gamma_0 \cdot Z_i + \gamma_0^l \cdot Z_i^l.$$

The null hypothesis was $H_0 : \beta = 0$. In Section 5.1 we took $\gamma_0^l = 0$. The coefficient γ_0 and the intercept 0 were nuisance parameters that were estimated by maximum likelihood under H_0. We took $\gamma_0 = 1$ and $\rho(X_i, Z_i) = 0.5$, $\rho(Z_i^l, Z_i) = 0$, $\rho(Z_i^l, X_i) = 0$. We took the dispersion parameter of the negative binomial distribution to be 1, so that $var(Y_i) = \mu_i + \mu_i^2$.

The assumed model, however, was Poisson, i.e., $var(Y_i) = \mu_i$ was assumed. Thus the true variance was larger than the assumed variance and the variance misspecification factor depended on μ_i, i.e., on the covariate Z_i. The assumed log link function was correct and in Section 5.1 the linear predictor was correct as well.
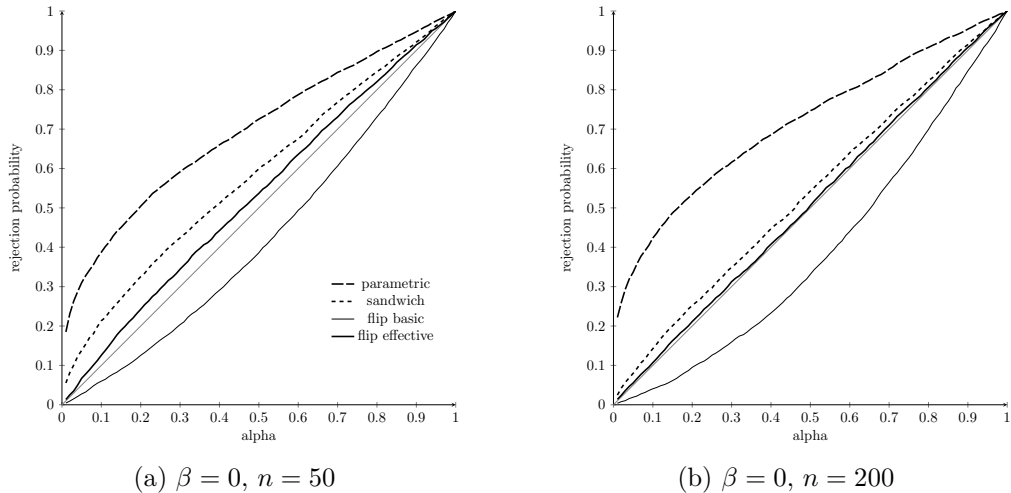
Figure 1: Estimated rejection probabilities for four tests under misspecified variance and estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

In Figure 1 the estimated rejection probabilites of four tests under $H_0 : \beta = 0$ are compared, based on 5000 repeated simulations. In all simulations the tests were two-sided.

One of the tests considered was the parametric score test. Since the assumed model was Poisson, the computed Fisher information was too small and the test was anti-conservative.

We also applied a Wald test, where we used a sandwich estimate (Agresti, 2015, p. 280) of the variance of $\hat{\beta}$, to correct for the misspecified variance function. We used the R package *gee* for this (available on CRAN), specifying blocks of size 1. As can be seen in Figure 1, this test was rather anti-conservative (especially for small α, e.g. $\alpha = 0.01$). This was in particular due to the estimation error of the sandwich (Boos, 1992; Freedman, 2006; Maas and Hox, 2004; Kauermann and Carroll, 2000).

Further, we applied the sign-flipping test based on the basic scores $\nu_{\hat{\gamma},i}$. Due to the estimation of γ_0, the variance of the score was shrunk and the test was conservative, as explained in Section 3.1. In the simulations under H_0 we took $w = 200$. Taking w larger led to a very similar level (see also Marriott, 1979). In the power simulations we took $w = 1000$.

Finally, we used the sign-flipping test of Theorem 5, which is based on the effective scores $\nu_{\hat{\gamma},i}^*$. In Section 3.2 it was already shown that this test is asymptotically exact under constant variance misspecification. Here, however, the variance misspecification factor was $1 + \mu_i$ (i.e., it depended on Z_i). Nevertheless the rejection probability under H_0 was approximately α. This illustrates that the test has some additional robustness, which we have not theoretically shown.
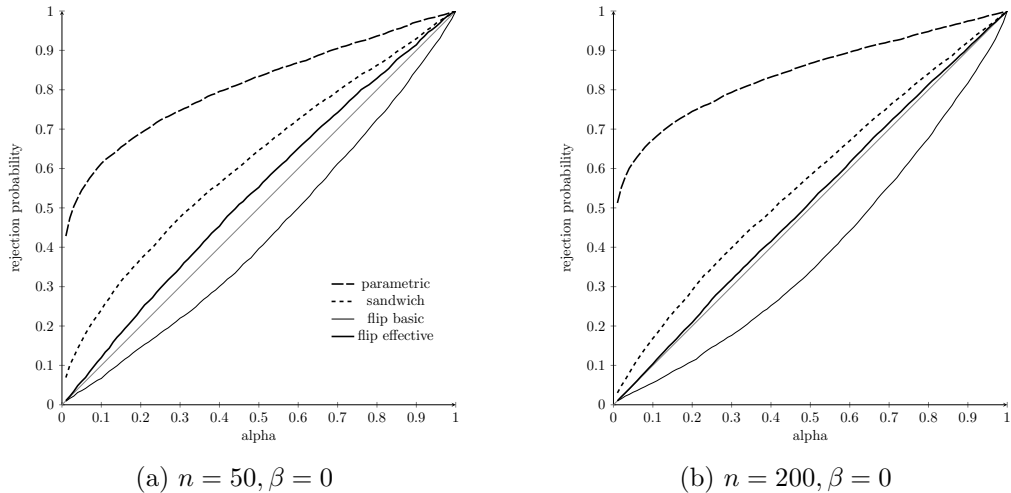
Figure 2: Estimated rejection probabilities for four tests under misspecified variance, estimated nuisance and ignored nuisance. The null hypothesis was $H_0 : \beta = 0$.

5.2 Ignored nuisance

The same simulations were performed as in Section 5.1, but with $\gamma_0^l = 1$. Since $\gamma_0^l = 0$ was assumed, Z_i^l represented an ignored, latent variable. Figure 2 shows similar results as Figure 1. The parametric test was even more anti-conservative than in Section 5.1. The reason is that the introduction of Z_i^l increased the variance Y_i, so that the variance of the score was even more misspecified than in Section 5.1.

The test of Theorem 5 was still nearly exact for $n = 200$, even though μ_i was misspecified. (Even marginally over Z_i^l, μ_i was misspecified. Possibly the estimation of the intercept corrected for the misspecification.)

A conclusion from the simulations of Sections 5.1 and 5.2, is that the sandwich-based approach should not always be seen as the most reliable way of testing models with misspecified variance functions. Indeed, in our simulations the test of Theorem 5 was substantially less anti-conservative (while having similar power, see Section 5.3).

5.3 Power

For a meaningful power comparison of the four tests, we considered the scenario where the assumed model was correct, i.e., the data distribution was Poisson and γ_0^l was 0. See figure 3. The estimated probabilities are based on $2 \cdot 10^4$ simulation loops.

Since the model was correct, asymptotically there was no better choice than the parametric test. The sign-flipping test of Theorem 5 had very similar power. The basic sign-flipping test was again conservative due to the estimation of γ_0. The sandwich-based test had the most power, but was anti-conservative (null behavior not shown).

15

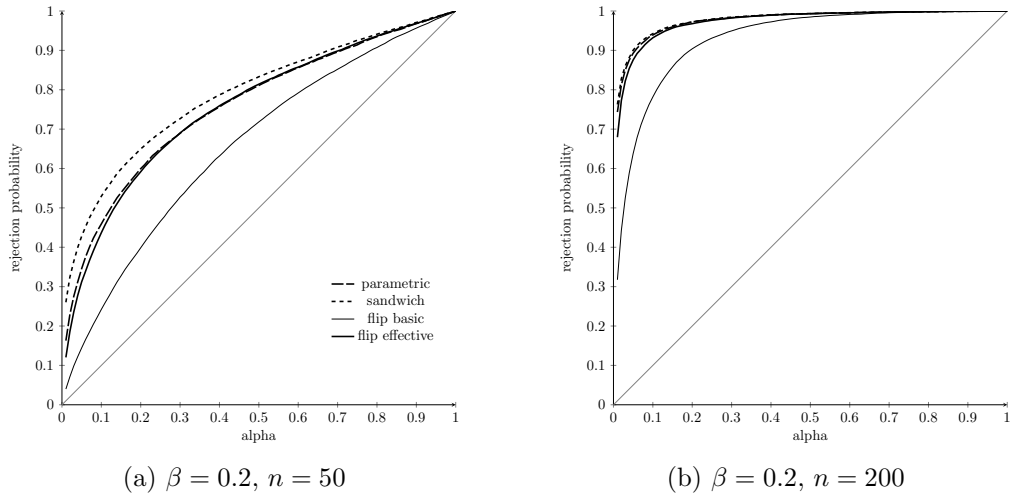(a) $\beta = 0.2$, $n = 50$ (b) $\beta = 0.2$, $n = 200$

Figure 3: Power comparison of four two-sided tests under the correct model, with estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

5.4 Strong heteroscedasticity

When a Gaussian linear model is considered with $Y_i \sim N(\beta x_i, \sigma^2)$, $x_1 = ... = x_n = 1$ and $H_0 : \beta = 0$, the score contributions are $\nu_i = X_i(Y_i - 0)/\sigma^2 = Y_i/\sigma^2$. Thus the test of Theorem 1 simply flips the observations Y_i, $1 \leq i \leq n$. The parametric counterpart of this test is the one-sample t-test. The t-test needs to explicitly estimate the nuisance parameter σ^2; the sign-flipping test does not (simply substitute $\sigma = 1$).

We simulated strongly heteroscedastic data: we took $Y_i \sim N(\beta x_i, \sigma_i^2)$, with $\sigma_i = \exp(i)$, $1 \leq i \leq n = 10$. Consequently the t-statistic did not have the assumed distribution and under H_0 the rejection probability of the t-test was far from the nominal level for most α, see Figure 4a. The sign-flipping test did not need to estimate the variance. In this setting the test has rejection probability $\lfloor \alpha w \rfloor / w$ exactly if the transformations $g_1, ..., g_w$ are drawn without replacement, since the observations are symmetric, see Proposition 4. (We drew $g_1, ..., g_w$ with replacement for convenience, but this gives almost the same test as drawing without replacement, due to the small probability of ties.)

For a meaningful power comparison, we considered the correct, homoscedastic model with $\sigma_1 = ... = \sigma_{10} = 1$. Figure 4b, based on 10^5 repeated simulations, shows that the tests had virtually the same power.

5.5 Multi-dimensional parameter of interest

We considered the same setting as in Section 5.2, except that $\boldsymbol{\beta}$ and the estimated nuisance parameter $\boldsymbol{\gamma}_0 = (0.5, 0.2, 0, 0, 0)$ were 5-dimensional (so $\boldsymbol{X}_i$, $\boldsymbol{Z}_i \in \mathbb{R}^5$). All corresponding covariates were correlated ($\rho = 0.5$). There was an ignored nuisance covariate as before ($\gamma_0^l = 0.5$), which was uncorrelated with the other covariates. Thus

16

(a) $\mu = 0$, strong heteroscedasticity

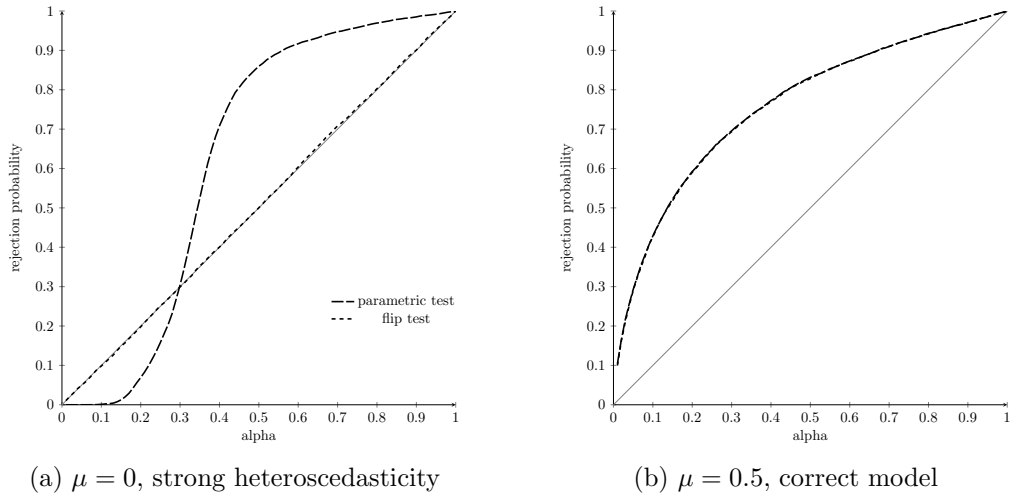(b) $\mu = 0.5$, correct model

Figure 4: Comparison of the one-sample t-test and the sign-flipping test. The null hypothesis was $H_0 : \mu = 0$.

there were in total 11 covariates. We took the overdispersion such that $var(Y_i) = \mu_i + 0.5\mu_i^2$, i.e., the overdispersion again depended on the covariates (heteroscedasticity).

Instead of the basic score test for a one-dimensional parameter we now used the multi-dimensional extension in Theorem 8. Similarly, instead of the test of Theorem 5 based on effective scores, we used the multi-dimensional extension in Theorem 9. We took $\hat{V} = V$ to be the identity matrix.

In Sections 5.1 and 5.2 we compared our tests with a Wald test based on a sandwich estimate of $Var(\hat{\beta})$. Here we proceeded analogously, using a sandwich estimate of the 5×5 matrix $Var(\hat{\beta})$ in the multi-dimensional Wald test. This test uses that $\hat{\beta}' Var(\hat{\beta})^{-1} \hat{\beta}$ asymptotically has a χ_d^2 distribution under the null hypothesis $H_0 : \beta = \mathbf{0}$.

The results under H_0 are shown in Figure 5, where each plot is based on 10^4 simulation loops. They are comparable to those in Section 5.2, except that the sandwich-based method is now even more anti-conservative. This is because $Var(\hat{\beta})$ is now a 5×5 matrix, which is difficult to estimate accurately. For $n = 50$ and $\alpha = 0.01$, the rejection probability of the sandwich-based method was 0.27 instead of the required 0.01.

For a meaningful power comparison of the four tests, we again considered the scenario where the assumed model was correct, i.e., the data distribution was Poisson and γ_0^l was 0. See Figure 6, where each plot is based on 10^4 simulation loops. As usual, the sign-flipping test based on basic scores had low power due to nuisance estimation. The power of the sign-flipping test based on effective scores was comparable to that of the parametric score test. As in Section 5.3, the test based on a sandwich estimate was the most powerful, but this has limited meaning, since it was also rather anti-conservative under the correct model (plot not shown).

To conclude, sign-flipping provided much more reliable type-I error control than the sandwich approach, while giving satisfactory power (comparable to that of the parametric test, under the correct model).

17

(a) $\boldsymbol{\beta} = \mathbf{0}$, $n = 50$ (b) $\boldsymbol{\beta} = \mathbf{0}$, $n = 200$
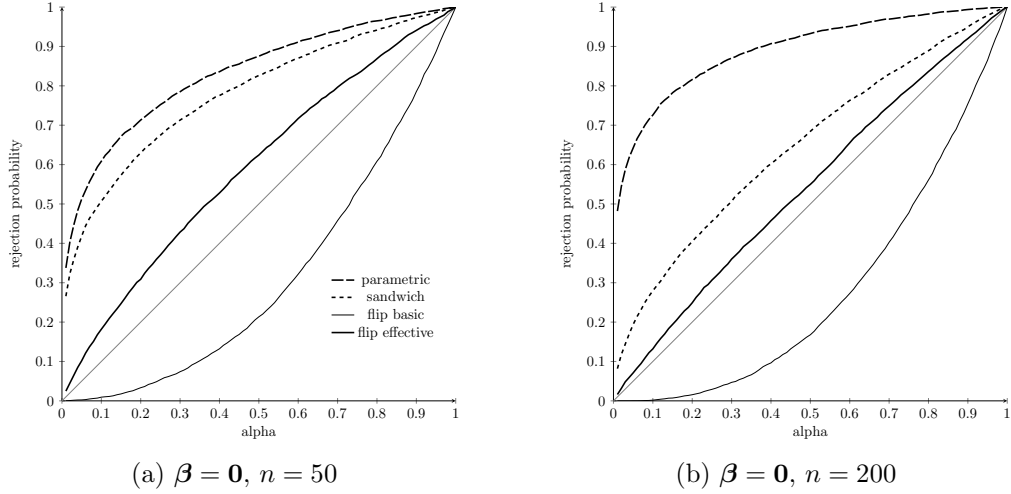
Figure 5: Estimated rejection probabilities under the null hypothesis. The model was misspecified due to overdispersion, heteroscedasticity and ignored nuisance.
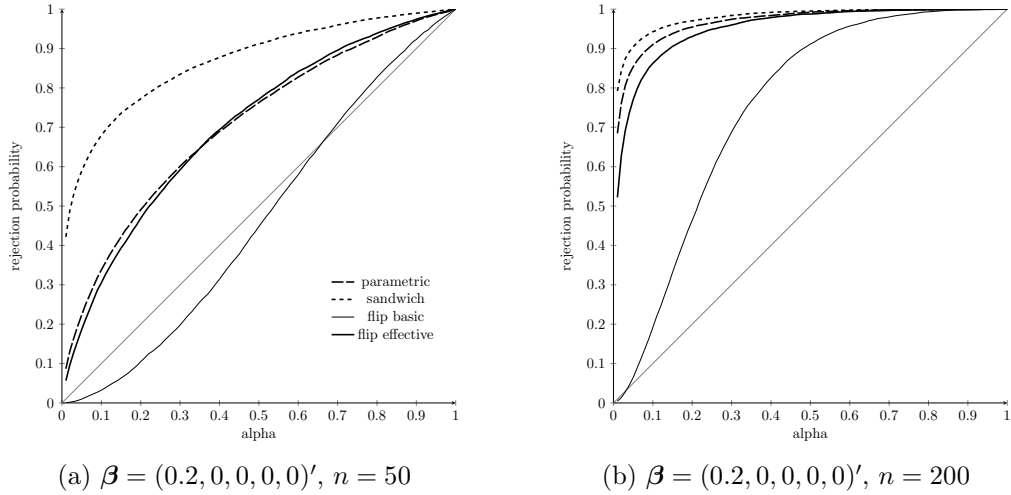

(a) $\boldsymbol{\beta} = (0.2, 0, 0, 0, 0)'$, $n = 50$ (b) $\boldsymbol{\beta} = (0.2, 0, 0, 0, 0)'$, $n = 200$

Figure 6: Power comparison under the correct model. The null hypothesis was $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

6   Data analysis

We analyzed the dataset *warpbreaks*. These data are used in the example code of the *gee* R package, available on CRAN. The dataset gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn. There are 54 observations of 3 variables: the number of breaks, the type of wool (A or B) and the tension (low, medium or high). For each of the 6 possible combinations of wool and tension, there are 9 observations. Using various methods, we tested whether the number of breaks depends on the type of wool.

We first considered a basic Poisson model with

$$\log(\mu_i) = \gamma_1 + \beta \mathbb{1}_{\{\text{wool}=B\}} + \gamma_2 \mathbb{1}_{\{\text{tension}=M\}} + \gamma_3 \mathbb{1}_{\{\text{tension}=H\}}.$$

The γ_i, $1 \leq i \leq 3$, were nuisance parameters that were estimated using maximum likelihood. We first tested $H_0 : \beta = 0$ using the parametric score test, obtaining a p-value of $6.29 \cdot 10^{-5}$. (All tests performed were two-sided.)

However, the data were clearly overdispersed: for each combination of wool and tension, the empirical variance of the 9 observations was substantially larger than the empirical mean. Thus the p-value based on the parametric test had limited meaning. Fitting a quasi-Poisson model, which assumes constant overdispersion, gave a p-value of 0.059.

As in Section 5, we also applied a Wald test, where we used a sandwich estimate (Agresti, 2015, p. 280) of the variance of $\hat{\beta}$, to correct for the misspecified variance function. This resulted in a p-value of 0.048.

Further, we used the sign-flipping test based on the basic scores $\nu_{\hat{\gamma},i}$, $i = 1,...,54$ (still using the basic Poisson model). We took $w = 10^6$. This resulted in a p-value of 0.113. This test is rather robust to model misspecification, but we know that it tends to be conservative when the score is correlated with the nuisance scores, as was the case here.

Finally, we performed the test of Theorem 5 based on the effective score. This test is asymptotically exact under the correct model and has been shown to be robust against several forms of variance misspecification. It provided a p-value of 0.065.

Based on this evidence, when maintaining a confidence level of 0.05, it seems that we cannot reject H_0. Indeed, only the sandwich-based test provided a p-value below 0.05, but this test is often anti-conservative, as discussed in Section 5.1.

Discussion

We have proposed a test which relies on the assumption that individual score distributions are independent and have mean 0 (in case of a point hypothesis) under the null. If the score contributions are misspecified due to overdispersion, heteroscedasticity or ignored nuisance covariates, then the traditional parametric tests lose their properties. The sign-flipping test is often robust to these types of misspecification and can still be asymptotically exact.

When nuisance parameters are estimated, the basic score contributions become dependent. If a nuisance score is correlated with the score of the parameter of interest, the estimation reduces the variance of the score, so that the sign-flipping test becomes conservative. As a solution we propose to use the effective score, which is asymptotically the part of the score that is orthogonal to the nuisance score. The effective score is asymptotically unaffected by the nuisance estimation, so that we again obtain an asymptotically exact test. We have proven that this is still the case when the scores and the Fisher information are misspecified by a constant, and simulations illustrate additional robustness.

When the parameter of interest is multi-dimensional, our test statistic involves a freely chosen matrix, which influences the power properties. If this matrix is taken to be the inverse of the effective Fisher information and the assumed model is correct, then our test is asymptotically equivalent to the parametric score test. Under the correct model, in certain situations our test is asymptotically equivalent to the global test (Goeman et al., 2006), which is popular for testing hypotheses about high-dimensional parameters.

A   A lemma

Lemma 11. *Suppose that for $n \to \infty$, a vector $\boldsymbol{T}^n = (T_1^n, ..., T_w^n)$ converges in distribution to a vector $\boldsymbol{T}$ of i.i.d. continuous variables. Then $\mathbb{P}(T_1^n > T_{[1-\alpha]}^n) \to \lfloor \alpha w \rfloor / w$.*

Proof. Note that $\mathbb{P}(T_1^n > T_{[1-\alpha]}^n) = \mathbb{P}(\boldsymbol{T}^n \in A)$, where

$$A = \{(t_1, ..., t_w) \in \mathbb{R}^w : |\{2 \le j \le w : t_j < t_1\}| \ge \lceil (1-\alpha)w \rceil \}.$$

Let ∂A be the boundary of A, i.e., the set of discontinuity points of $\mathbb{1}_A$. Note that if $t \in \partial A$, then $t_i = t_j$ for some $1 \le i < j \le w$. It follows that $\mathbb{P}(\boldsymbol{T} \in \partial A) = 0$. Since $\mathbb{1}_A$ is continuous on $(\partial A)^c$, it follows from the continuous mapping theorem (Van der Vaart, 1998, Theorem 2.3) that $\mathbb{1}_A(\boldsymbol{T}^n) \xrightarrow{d} \mathbb{1}_A(\boldsymbol{T})$.

The elements of $\boldsymbol{T}$ are i.i.d. draws from the same distribution. Hence it follows from the Monte Carlo testing principle (Lehmann and Romano, 2005) that under H_0, $\mathbb{P}(\boldsymbol{T} \in A) = \lfloor \alpha w \rfloor / w$. Thus $\mathbb{P}(\boldsymbol{T}^n \in A) \to \lfloor \alpha w \rfloor / w$. $\square$

B   Proofs of the results

Proof of Theorem 1. Suppose H_0 holds. We will show that $\boldsymbol{T}^n = (T_1^n, ..., T_w^n)$ converges in distribution to a multivariate normal distribution with mean $\boldsymbol{0}$ and variance $\lim_{n \to \infty} s_n^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the $w \times w$ identity matrix. It then follows from Lemma 11 that $\mathbb{P}(T_1^n > T_{[1-\alpha]}^n) \to \lfloor \alpha w \rfloor / w$.

Under H_0, for each $1 \le j \le w$, $\mathbb{E}(T_j^n) = 0$. For every $1 \le j \le w$, $var(T_j^n) = n^{-1} \sum_{i=1}^n var(\nu_i) = s_n^2$. Let $\boldsymbol{Q}_n$ be the covariance matrix of $\boldsymbol{T}^n$. $\boldsymbol{Q}_n$ has zeroes off the

diagonal. Indeed, for $1 \le j < k \le w$,

$$cov(T_j^n, T_k^n) = cov(n^{-1/2} \sum_{i=1}^{n} g_{ji}\nu_i, n^{-1/2} \sum_{i=1}^{n} g_{ki}\nu_i) = 0,$$

since the g_{ki}, $2 \le k \le w$, are independent with mean 0. Hence $\boldsymbol{Q}_n$ converges to $\lim_{n\to\infty} s_n \boldsymbol{I}$. Note that $\boldsymbol{T}^n$ is a sum of n vectors. By the multivariate Lindeberg-Feller central limit theorem (Van der Vaart, 1998) $\boldsymbol{T}^n$ converges in distribution to a multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\lim_{n\to\infty} s_n^2 \boldsymbol{I}$.

We have shown that $\boldsymbol{T}^n$ converges in distribution to a vector $\boldsymbol{T}$, say, of i.i.d. normal random variables. It now follows from Lemma 11 that $\mathbb{P}(T_1^n > T_{[1-\alpha]}^n) \to \lfloor \alpha w \rfloor / w$. $\square$

Proof of Proposition 4. Note that $(\nu_1, ..., \nu_n) \overset{d}{=} (g_{j1}\nu_1, ..., g_{jn}\nu_n)$ for every $1 \le j \le w$. This means that the test becomes a basic random transformation test and the results follow from the proof of Theorem 2 in Hemerik and Goeman (2018b).

Proof of Theorem 5. Suppose that H_0 holds. Note that

$$S_{\hat{\gamma}}^* = S_{\hat{\gamma}} - \hat{\boldsymbol{\mathcal{I}}}_{12}' \hat{\boldsymbol{\mathcal{I}}}_{22}^{-1} \boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = S_{\hat{\gamma}} - \boldsymbol{\mathcal{I}}_{12}' \boldsymbol{\mathcal{I}}_{22}^{-1} \boldsymbol{S}_{\hat{\gamma}}^{(k-1)} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1) = $$

$$S_{\gamma_0} - \boldsymbol{\mathcal{I}}_{12}' \sqrt{n}(\hat{\gamma} - \gamma_0) - \boldsymbol{\mathcal{I}}_{12}' \boldsymbol{\mathcal{I}}_{22}^{-1} \Big\{ \boldsymbol{S}_{\gamma_0}^{(k-1)} - \boldsymbol{\mathcal{I}}_{22} \sqrt{n}(\hat{\gamma} - \gamma_0) \Big\} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1) = $$

$$S_{\gamma_0}^* + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1).$$

Let $2 \le j \le w$ and

$$S_{\gamma}^{j+} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=1}\}} \nu_{\gamma, i}, \quad S_{\gamma}^{j-} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=-1}\}} \nu_{\gamma, i}.$$

Note that

$$S_{\hat{\gamma}}^j = S_{\hat{\gamma}}^{j+} - S_{\hat{\gamma}}^{j-} = \Big\{ S_{\gamma_0}^{j+} - \frac{1}{2}\sqrt{n}\boldsymbol{\mathcal{I}}_{12}'(\hat{\gamma} - \gamma_0) \Big\} - \Big\{ S_{\gamma_0}^{j-} - \frac{1}{2}\sqrt{n}\boldsymbol{\mathcal{I}}_{12}'(\hat{\gamma} - \gamma_0) \Big\} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1) = $$

$$S_{\gamma_0}^{j+} - S_{\hat{\gamma}}^{j-} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1) = S_{\gamma_0}^j + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1).$$

The intuitive reason why $S_{\hat{\gamma}}^j = S_{\gamma_0}^j + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1)$, is that the estimation of $\hat{\gamma}$ does not cause the summands underlying $S_{\hat{\gamma}}^j$ to be correlated. Similarly we find that $\boldsymbol{S}_{\hat{\gamma}}^{(k-1),j} = \boldsymbol{S}_{\gamma_0}^{(k-1),j} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1)$ and conclude that $S_{\hat{\gamma}}^{*j} = S_{\gamma_0}^{*j} + o_{\mathbb{P}_{\beta_0, \gamma_0}}(1)$.

Let $\boldsymbol{T}^n$ be as in the proof of Theorem 1, with ν_i replaced by $\nu_{\gamma_0, i}^*$. Suppose H_0 holds and $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{I}}$, so that the summands underlying T_j^n are independent. For every $1 \le i \le n$, $\mathbb{E}(\nu_{\gamma_0, i}^*) = 0$. The elements of $\boldsymbol{T}^n$ are uncorrelated and have common variance $var(\nu_{\gamma_0, 1}^*)$. By the multivariate central limit theorem (Van der Vaart, 1998; Greene, 2012), $\boldsymbol{T}^n$ converges in distribution to $N(\boldsymbol{0}, var(\nu_{\gamma_0, 1}^*)\boldsymbol{I})$. We supposed that $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{I}}$ to use the central limit theorem, but the asymptotic distribution of $\boldsymbol{T}^n$ is the same if $\hat{\boldsymbol{\mathcal{I}}}$ is any consistent estimator of $\boldsymbol{\mathcal{I}}$.

Let $\hat{\boldsymbol{T}}^n$ be as in the proof of Theorem 1, with ν_i replaced by $\nu^*_{\hat{\gamma},i}$. For every $1 \leq j \leq w$, $S^{*j}_{\hat{\gamma}} = S^{*j}_{\gamma_0} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$. Thus $\hat{\boldsymbol{T}}^n$ and $\boldsymbol{T}^n$ are asymptotically equivalent. The result now follows from Lemma 11. $\qquad\square$

Proof of Proposition 6. For $2 \leq j \leq w$ consider

$$S^{*j+}_{\hat{\gamma}} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=1}\}} \nu^*_{\hat{\gamma},i}, \quad S^{*j-}_{\hat{\gamma}} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=-1}\}} \nu^*_{\hat{\gamma},i}$$

$$S^{(k-1),j+}_{\hat{\gamma}} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=1}\}} \nu^{(k-1)}_{\hat{\gamma},i}, \quad S^{(k-1),j-}_{\hat{\gamma}} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=-1}\}} \nu^{(k-1)}_{\hat{\gamma},i}.$$

We have

$$S^{*j+}_{\hat{\gamma}} = S^{j+}_{\hat{\gamma}} - \hat{\mathcal{I}}'_{12} \hat{\mathcal{I}}^{-1}_{22} S^{(k-1),j+}_{\hat{\gamma}} =$$

$$S^{j+}_{\gamma_0} - \frac{1}{2}\sqrt{n} \mathcal{I}'_{12}(\hat{\gamma} - \gamma_0) - \mathcal{I}'_{12} \mathcal{I}^{-1}_{22} \{ S^{(k-1),j+}_{\gamma_0} - \frac{1}{2}\sqrt{n} \mathcal{I}_{22}(\hat{\gamma} - \gamma_0) \} + o_{\mathbb{P}_{\beta^n,\gamma_0}}(1) =$$

$$S^{*j+}_{\gamma_0} + o_{\mathbb{P}_{\beta^n,\gamma_0}}(1)$$

and analogously $S^{*j-}_{\hat{\gamma}} = S^{*j-}_{\gamma_0} + o_{\mathbb{P}_{\beta^n,\gamma_0}}(1)$. By Marohn (2002, p. 341), for $2 \leq j \leq w$, $S^{*j+}_{\gamma_0}$ and $S^{*j-}_{\gamma_0}$ have an asymptotic $N(\frac{1}{2}\xi\sigma^2_0, \frac{1}{2}\sigma^2_0)$ distribution. Since they are independent, it follows that $T^n_j = S^{*j+}_{\hat{\gamma}} - S^{*j-}_{\hat{\gamma}}$ has an asymptotic $N(0, \sigma^2_0)$ distribution, $2 \leq j \leq w$. With the multivariate central limit theorem we find that $(T^n_2, ..., T^n_w)$ converges in distribution to a vector of $w-1$ i.i.d. $N(0, \sigma^2_0)$ variables as $n \to \infty$.

Let $\epsilon, \epsilon' > 0$. Let $(T'_1, ..., T'_w)$ have the asymptotic distribution of $(T^n_1, ..., T^n_w)$. Let $(T''_1, ..., T''_w)$ be a vector of w i.i.d. $N(0, \sigma^2_0)$ variables. Apart from the first element, these two vectors have the same distribution. For $w \in \{2, 3, ...\}$, define $T^{[w]}_{[1-\alpha]}$ like $T^n_{[1-\alpha]}$, but based on the values $T'_1, ..., T'_w$ instead of $T^n_1, ..., T^n_w$. Also define $T^{[[w]]}_{[1-\alpha]}$ like $T^n_{[1-\alpha]}$, but based on the values $T''_1, ..., T''_w$. Note that as $w \to \infty$, the empirical quantile $T^{[[w]]}_{[1-\alpha]}$ converges in distribution to the constant $\sigma_0\Phi(1-\alpha)$. Further note that for $w \to \infty$, $T^{[[w]]}_{[1-\alpha]} - T^{[w]}_{[1-\alpha]}$ converges in distribution to 0. Thus there is a $W \in \mathbb{N}$ such that for all $w > W$,

$$\mathbb{P}(|T^{[w]}_{[1-\alpha]} - \sigma_0\Phi(1-\alpha)| < \epsilon') > 1 - \epsilon. \tag{3}$$

Since the distribution of $(T^n_1, ..., T^n_w)$ converges to the distribution of $(T'_1, ..., T'_w)$ as $n \to \infty$,

$$T^n_{[1-\alpha]} \xrightarrow{d} T^{[w]}_{[1-\alpha]} \tag{4}$$

as $n \to \infty$. Since in the present proof w is not fixed, we will write $T^n_{[1-\alpha]} = T^{n,w}_{[1-\alpha]}$. By results (3) and (4), for $w > W$, $\liminf_{n\to\infty} \mathbb{P}(|T^{n,w}_{[1-\alpha]} - \sigma_0\Phi(1-\alpha)| < \epsilon') > 1 - \epsilon$. Thus $\lim_{w\to\infty} \liminf_{n\to\infty} \mathbb{P}(|T^{n,w}_{[1-\alpha]} - \sigma_0\Phi(1-\alpha)| < \epsilon') = 1$.

The distribution of T^n_1, which does not depend on w, converges to a continuous distribution as $n \to \infty$. It follows that for every $\epsilon'' > 0$, there is an W' such that there

is a N such that for all $w > W'$ and $n > N$, $\mathbb{E}\big(|\mathbb{1}_{\{T_1^n > T_{[1-\alpha]}^{n,w}\}} - \mathbb{1}_{\{T_1^n > \sigma_0 \Phi(1-\alpha)\}}|\big) < \epsilon''$.
This means that $\lim_{w \to \infty} \liminf_{n \to \infty} \mathbb{E}\big(|\mathbb{1}_{\{T_1^n > T_{[1-\alpha]}^{n,w}\}} - \mathbb{1}_{\{T_1^n > \sigma_0 \Phi(1-\alpha)\}}|\big) = 0$, as was to be shown. □

Proof of Proposition 7. For every $1 \leq j \leq w$ we have
$$\tilde{S}_{\hat{\gamma}}^{*j} = c_1 S_{\hat{\gamma}}^j - c_2 \hat{\boldsymbol{\mathcal{I}}}'_{12} c_2^{-1} \hat{\boldsymbol{\mathcal{I}}}_{22}^{-1} c_1 \mathbf{S}_{\hat{\gamma}}^{(k-1),j} = c_1 S_{\hat{\gamma}}^{*j}.$$
Hence the test is identical to that of Theorem 5, since that test is unchanged if all T_j^n, $1 \leq j \leq w$, are multiplied by the same constant. □

Proof of Theorem 8. Suppose H_0 holds. Consider the $d \times j$-matrix
$$\left(n^{1/2} \sum_{i=1}^n g_{ji} \boldsymbol{\nu}_{\gamma_0,i}\right)_{1 \leq j \leq w}. \tag{5}$$
It follows from the multivariate central limit theorem (Van der Vaart, 1998) that, as $n \to \infty$, this matrix converges in distribution to a matrix with identically distributed columns which are independent of each other. Note that for every $1 \leq j \leq w$, T_j^n is a function of the j-th column of the matrix (5). Thus, with the continuous mapping theorem (Van der Vaart, 1998, Theorem 2.3) it follows that $(T_1^n, ..., T_j^n)$ also converges in distribution to a vector with continuous i.i.d. elements. The result now follows from Lemma 11. □

Proof of Theorem 9. Consider the case $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$. As in the proof of Theorem 8, under H_0, $(T_1^n, ..., T_w^n)$ converges in distribution to a vector of w i.i.d. variables. As in the proof of Theorem 5, the same is true if we take $\hat{\boldsymbol{\gamma}}$ to be a different $\sqrt{n}$-consistent estimator of $\boldsymbol{\gamma}_0$. (Again, the reason is that the effective score based on $\hat{\boldsymbol{\gamma}}$ is asymptotically equivalent to the effective score based on $\boldsymbol{\gamma}_0$.) The result now follows from Lemma 11 again. □

Proof of Proposition 10. By Hall and Mathiason (1990), $n^{-1/2} \sum_{i=1}^n \boldsymbol{\nu}_{\hat{\gamma},i}^*$ has an asymptotic $N(\mathbf{0}, \boldsymbol{\mathcal{I}}^*)$ distribution under $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Analogously to the one-dimensional case at Proposition 6, for $2 \leq j \leq w$, the vector $n^{-1/2} \sum_{i=1}^n g_{ji} \boldsymbol{\nu}_{\hat{\gamma},i}^*$ is asymptotically the difference of two mutually independent $N(\frac{1}{2}\boldsymbol{\mathcal{I}}^*\boldsymbol{\xi}, \frac{1}{2}\boldsymbol{\mathcal{I}}^*)$ vectors (Hall and Mathiason, 1990), so that it also has an asymptotic $N(\mathbf{0}, \boldsymbol{\mathcal{I}}^*)$ distribution (under $\boldsymbol{\beta} = \boldsymbol{\beta}^n$). As in the proof of Theorem 8, by the multivariate central limit theorem, the $d \times (w-1)$ matrix $\big(n^{-1/2} \sum_{i=1}^n g_{ji} \boldsymbol{\nu}_{\hat{\gamma},i}^*\big)_{2 \leq j \leq w}$ converges to a matrix with $w-1$ independent $N(\mathbf{0}, \boldsymbol{\mathcal{I}}^*)$ columns as $n \to \infty$. Hence, by the continuous mapping theorem, as $n \to \infty$, $(T_2^n, ..., T_w^n)$ converges in distribution to a vector of $w - 1$ i.i.d. variables (under $\boldsymbol{\beta} = \boldsymbol{\beta}^n$), which follow the asymptotic distribution which T_1^n has under $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

The result now follows as at the end of the proof of Proposition 6. □

References

Agresti, A. *Foundations of linear and generalized linear models.* John Wiley & Sons, 2015.

Boos, D. D. On generalized score tests. *The American Statistician*, 46:327–333, 1992.

Canay, I. A., Romano, J. P., and Shaikh, A. M. Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030, 2017.

Chung, E. Y. and Romano, J. P. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013.

Cox, D. R. and Hinkley, D. V. *Theoretical statistics*. CRC Press, 1979.

Cox, D. R. and Reid, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39, 1987.

Fisher, R. A. *The design of experiments*. Oliver and Boyd, 1935.

Freedman, D. A. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302, 2006.

Ganong, P. and Jäger, S. A permutation test for the regression kink design. *Journal of the American Statistical Association*, 113(522):494–504, 2018.

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

Goeman, J. J., Van De Geer, S. A., and Van Houwelingen, H. C. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

Goeman, J. J., Van Houwelingen, H. C., and Finos, L. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, 98 (2):381–390, 2011.

Greene, W. H. *Econometric analysis*. Harlow: Pearson Education Limited, 2012.

Hall, W. and Mathiason, D. J. On large-sample estimation and testing in parametric models. *International Statistical Review/Revue Internationale de Statistique*, 58(1):77–97, 1990.

Hemerik, J., Solari, A., and Goeman, J. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649, 2019.

Hemerik, J. and Goeman, J. J. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018a.

Hemerik, J. and Goeman, J. J. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155, 2018b.

Kauermann, G. and Carroll, R. J. The sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals. 2000.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.

Maas, C. J. and Hox, J. J. Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2):127–137, 2004.

Marohn, F. A comment on locally most powerful tests in the presence of nuisance parameters. *Communications in Statistics-Theory and Methods*, 31(3):337–349, 2002.

Marriott, F. H. C. Barnard's Monte Carlo tests: How many simulations? *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):75–77, 1979.

Pauly, M., Brunner, E., and Konietschke, F. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2): 461–473, 2015.

Pesarin, F. Some elementary theory of permutation tests. *Communications in Statistics-Theory and Methods*, 44(22):4880–4892, 2015.

Pesarin, F. *Multivariate permutation tests: with applications in biostatistics*, volume 240. Wiley Chichester, 2001.

Pesarin, F. and Salmaso, L. Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics*, 22(5): 669–684, 2010a.

Pesarin, F. and Salmaso, L. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010b.

Rao, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge Univ Press, 1948.

Rayner, J. The asymptotically optimal tests. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):337–345, 1997.

Rippon, P. and Rayner, J. C. Generalised score and Wald tests. *Advances in Decision Sciences*, 2010.

Solari, A., Finos, L., and Goeman, J. J. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014.

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9): 5116–5121, 2001.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

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

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

Winkler, A. M., Ridgway, G. R., Douaud, G., Nichols, T. E., and Smith, S. M. Faster permutation inference in brain imaging. *NeuroImage*, 141:502–516, 2016.