



Universiteit
Leiden
The Netherlands

Permutation-based inference for high-dimensional data.

Hemerik, J.

Citation

Hemerik, J. (2018, May 1). *Permutation-based inference for high-dimensional data*. Retrieved from <https://hdl.handle.net/1887/61825>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/61825>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/61825> holds various files of this Leiden University dissertation.

Author: Hemerik, J.

Title: Permutation-based inference for high-dimensional data

Issue Date: 2018-05-01

5

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

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. This 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.

5.1 Introduction

We consider the problem of hypothesis testing in a generalized linear model (GLM) with a correct link function, but potentially misspecified distribu-

tion. When the model is misspecified, the traditional parametric tests tend to lose their properties, since they estimate the Fisher information under incorrect assumptions.

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. 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 (Westfall and Young, 1993; Pesarin, 2001; Chung et al., 2013; Hemerik and Goeman, 2018). These methods are useful since they require few parametric assumptions. Especially when multiple hypotheses are tested, permutation methods are powerful since they can often take into account the dependence structure in the data. 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 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 scores are perfectly symmetric around zero, then our test is exact for small n , even if the score contributions have misspecified variances (Pesarin and Salmaso, 2010). A parametric test, on the other hand, is then usually not exact.

In case nuisance effects are estimated, the individual score contributions become dependent and our 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 (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 5.2 we consider the scenario that no nuisance effects need to be estimated. In Section 5.3 we show how the estimation of nuisance effects can be taken into account. Section 5.4 contains simulations and Section 5.5 an analysis of real data.

5.2 Models with known nuisance parameters

Consider random variables $\nu_1, \dots, \nu_n$, which satisfy Assumption 5.2.1 below. These will often be individual score contributions (see Section 5.3, Rao, 1948, or Hall and Mathiason, 1990, p. 86), but the results in Sections 5.2.1-5.2.3 are satisfied for any random variables satisfying this assumption.

Assumption 5.2.1. Assume that ν_i , $i \in \mathbb{N}$, are random variables which are independent of each other and have finite third moment. Suppose that for some $c < 1$, for every $i \in \mathbb{N}$, $\mathbb{P}(\nu_i = 0) \leq c$. Let $\sigma_i^2 = \text{Var}(\nu_i)$. Assume that as $n \rightarrow \infty$, $s_n^2 := \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \rightarrow c'$ for some constant $c' \geq 0$ and $\lim_{n \rightarrow \infty} (ns_n^2)^{-1} = 0$.

Throughout Section 5.2, we consider any hypothesis H_0 which implies that $\mathbb{E}\nu_i = 0$ for all $1 \leq i \leq n$. This is in particular the case if H_0 is a point hypothesis and $\nu_1, \dots, \nu_n$ are the corresponding score contributions. Generalizations to hypotheses of the form $\nu_i \geq C$, $i = 1, \dots, n$, are straightforward, as well as extensions to two-sided tests.

A key assumption throughout Section 5.2 is that $\nu_1, \dots, \nu_n$ are independent. As soon as nuisance parameters need to be estimated, however, score contributions become dependent. This issue is the topic of Section 5.3.

5.2.1 Basic sign-flipping test

Let $w \in \mathbb{N}_{>1}$ be the number of random sign-flippings to be used and $\alpha \in [0, 1)$. Define $g_1 = (1, \dots, 1) \in \mathbb{R}^n$ and for every $2 \leq j \leq w$ let $g_j = (g_{j1}, \dots, g_{jn})$ be independent and uniformly distributed on $\{-1, 1\}^n$.

Given values $T^1, \dots, T^w \in \mathbb{R}$, we let $T^{(1)} \leq \dots \leq T^{(w)}$ be the sorted values and write $T^{[1-\alpha]} = T^{(\lceil (1-\alpha)w \rceil)}$. We will need the following lemma.

Lemma 5.2.1. *Suppose that for $n \rightarrow \infty$, a vector $\mathbf{T}^n = (T^1, \dots, T^w)$ converges in distribution to a vector $\mathbf{T}$ of i.i.d. continuous variables. Then $\mathbb{P}(T^1 > T^{[1-\alpha]}) \rightarrow \lfloor \alpha w \rfloor / w$.*

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

$$A = \{(t_1, \dots, t_w) \in \mathbb{R}^w : \#\{2 \leq j \leq w : t_j < t_1\} \geq \lceil (1-\alpha)w \rceil\}.$$

Note that if $t \in \partial A$, then $t_i = t_j$ for some $1 \leq i < j \leq w$. It follows that $\mathbb{P}(\mathbf{T} \in \partial A) = 0$. 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 $\mathbb{1}_A(\mathbf{T}^n) \xrightarrow{d} \mathbb{1}_A(\mathbf{T})$.

The elements of $\mathbf{T}$ 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 , $\mathbb{P}(\mathbf{T} \in A) = \lfloor \alpha w \rfloor / w$. Thus $\mathbb{P}(\mathbf{T}^n \in A) \rightarrow \lfloor \alpha w \rfloor / w$. $\square$

Throughout the rest of Section 5.2, for every $1 \leq j \leq w$, we let

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

We now show that the basic sign-flipping test is asymptotically exact.

Theorem 5.2.1. *Suppose that Assumption 5.2.1 holds. Consider the test that rejects H_0 if and only if $T^1 > T^{[1-\alpha]}$. Then, as $n \rightarrow \infty$, the level of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$. Moreover, the statistics $T^1, \dots, T^w$ are asymptotically normal and independent with mean 0 and common variance $\lim_{n \rightarrow \infty} s_n^2$.*

Proof. Suppose H_0 holds. We will show that $\mathbf{T}^n$ converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and variance $\lim_{n \rightarrow \infty} s_n^2 \mathbf{I}$, where $\mathbf{I}$ is the $w \times w$ identity matrix. It then follows from Lemma 5.2.1 that $\mathbb{P}(T^1 > T^{[1-\alpha]}) \rightarrow \lfloor \alpha w \rfloor / w$.

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

$$\begin{aligned} \text{cov}(T^j, T^k) &= \\ \text{cov}(n^{-1/2} \sum_{i=1}^n g_{ji} \nu_i, n^{-1/2} \sum_{i=1}^n g_{ki} \nu_i) &= 0, \end{aligned}$$

since the g_{ki} , $2 \leq k \leq w$, are independent with mean 0. Hence $\mathbf{Q}_n$ converges to $\lim_{n \rightarrow \infty} s_n^2 \mathbf{I}$. Note that $\mathbf{T}^n$ is a sum of n vectors. By the multivariate Lindeberg-Feller central limit theorem (Greene (2012), p. 1123) $\mathbf{T}^n$ converges in distribution to a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\lim_{n \rightarrow \infty} s_n^2 \mathbf{I}$.

We have shown that $\mathbf{T}^n$ converges in distribution to a vector $\mathbf{T}$, say, of i.i.d. normal random variables. It now follows from Lemma 5.2.1 that $\mathbb{P}(T^1 > T^{[1-\alpha]}) \rightarrow \lfloor \alpha w \rfloor / w$. □

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 our test is exact. However, if the ν_i are symmetric, then even for small n the level is always at most α , as noted in the following proposition.

Proposition 5.2.1. *Suppose that $\nu_1, \dots, \nu_n$ are independent and continuous and that under H_0 , for each $i \in \mathbb{N}$, $\nu_i \stackrel{d}{=} -\nu_i$. Then the level of the test of Theorem 5.2.1 is at most α for any $n \in \mathbb{N}$. Moreover, if $g_2, \dots, g_w$ are uniformly drawn from $\{1, -1\}^n \setminus (1, \dots, 1)$ without replacement (so that only g_1 takes the value $(1, \dots, 1)$), then the level is exactly $\lfloor \alpha w \rfloor / w$.*

Proof. Note that $(\nu_1, \dots, \nu_n) \stackrel{d}{=} (g_{j1}\nu_1, \dots, g_{jn}\nu_n)$ for every $1 \leq j \leq w$. This means that the test becomes a basic random transformation test and all results follow directly from Theorem 1 in Hemerik and Goeman (2018). $\square$

If the g_j are drawn with replacement or the ν_i are discrete, then the level of the test of Proposition 5.2.1 is (slightly) smaller than $\lfloor \alpha w \rfloor / w$ for finite n , due to the possibility of ties among the test statistics T^j , $1 \leq j \leq w$. Otherwise the level is $\lfloor \alpha w \rfloor / w$.

Note that when the level 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.

5.2.2 Fixed transformations

In the test of Theorem 5.2.1, random transformations are used, rather than the whole set of $|\{-1, 1\}^n| = 2^n$ transformations. The first reason is that this will often be done in practice, especially when using all 2^n transformations would be computationally intensive (which can be the case in a multiple testing context or when n is large). Second, the fact that the number of random transformations w is fixed and $g_1, \dots, g_w$ are independent (due to drawing with replacement), in fact simplified the proof.

It may be of interest to use every element of $G = \{-1, 1\}^n$ exactly once, for example to be able to obtain p -values as small as 2^{-n} . The following proposition states that the test that uses all 2^n transformations once, is asymptotically equivalent to the test of Theorem 5.2.1.

Proposition 5.2.2. *Suppose that Assumption 5.2.1 holds. Let T' be the $\lceil (1 - \alpha)|G| \rceil$ -th smallest value among T^g , $g \in G$, where $T^g = n^{-1/2} \sum_{i=1}^n g_i \nu_i$. Let $\phi_{n,w}$ and ϕ'_n be the rejection functions $\mathbf{1}_{\{T^1 > T^{[1-\alpha]}\}}$ and $\mathbf{1}_{\{T^1 > T'\}}$, respectively. Assume that $\mathbb{P}(E_1) \rightarrow 0$ as $\epsilon^{-1}, n, w \rightarrow \infty$, where*

$$E_1 = \{T^{(\lceil (1-\alpha-\epsilon)w \rceil)} \leq T^1 \leq T^{(\lceil (1-\alpha+\epsilon)w \rceil)}\}.$$

Then, for $w, n \rightarrow \infty$, $\phi'_n - \phi_{n,w} \xrightarrow{d} 0$.

Proof. Let $0 < \epsilon < \min\{\alpha, 1 - \alpha\}$. Consider the event

$$E_2 = \{T^{(\lceil (1-\alpha-\epsilon)w \rceil)} \leq T' \leq T^{(\lceil (1-\alpha+\epsilon)w \rceil)}\}.$$

It follows from the law of large numbers that $N_\epsilon, W_\epsilon \in \mathbb{N}$ exist such that for all $n > N_\epsilon, w > W_\epsilon$,

$$\mathbb{P}(E_2) > 1 - \epsilon.$$

Hence, for all $n > N_\epsilon, w > W_\epsilon$,

$$\mathbb{E}\{(\phi'_n - \phi_{n,w}) \cdot \mathbb{1}_{E_1^c}\} < \epsilon.$$

For $\epsilon^{-1}, w, n \rightarrow \infty$, $\mathbb{P}(E_1) \rightarrow 0$. It follows that for $w, n \rightarrow \infty$,

$$\mathbb{E}(\phi'_n - \phi_{n,w}) \rightarrow 0.$$

□

Note that in Theorem 5.2.1 and Proposition 5.2.2, we did not assume continuity of the observations ν_i . For finite n , continuity is needed to avoid ties among the test statistics to guarantee exactness. However, for $n \rightarrow \infty$, $\mathbb{P}(T^j = T^k) \rightarrow 0$ for any $1 \leq j < k \leq w$ regardless of the distribution of the ν_i (unless most ν_i are constant 0, which we rule out in Assumption 5.2.1). This allows using Theorem 5.2.1 for discrete GLMs.

5.2.3 Asymptotic equivalence with parametric test

Nonparametric tests such as permutation tests are often asymptotically equivalent to their corresponding parametric test (Wald and Wolfowitz, 1944; Noether, 1949). This is also the case for our sign-flipping test. This has not yet been proven under our specific assumptions, so we provide the proof here. This proof is related to the proof of Theorem 5.2.1, but here we let w increase to infinity.

Proposition 5.2.3. *Suppose that ν'_i , $i \in \mathbb{N}$, satisfy Assumption 5.2.1 and $\mathbb{E}(\nu'_i) = 0$, $i \in \mathbb{N}$. For every i , consider $\nu_i = \nu'_i + n^{-1/2}\kappa$, with $\kappa \in \mathbb{R}$. As usual, for $1 \leq j \leq w$ let*

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

Let $\phi_{n,w} = \mathbb{1}_{\{T^1 > T^{[1-\alpha]}\}}$ and $\phi'_n = \mathbb{1}_{\{T^1 > s_n \Phi(1-\alpha)\}}$, where $s_n^2 = n^{-1} \sum_{i=1}^n \text{var}(\nu_i) = \text{var}\{T^1\}$ and Φ is the cdf of the standard normal distribution. Then $\phi_{n,w} - \phi'_n \xrightarrow{d} 0$ for $w, n \rightarrow \infty$.

Proof. For $2 \leq j \leq w$, $\lim_{n \rightarrow \infty} \text{var}(T^j) =$

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{var}\{g_{ji}(\nu'_i + \kappa n^{-1/2})\} &= \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbb{E}(\nu'_i + \kappa n^{-1/2})^2 = \\ \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbb{E}(\nu_i'^2 + 2\nu_i' \kappa n^{-1/2} + \kappa^2 n^{-1}) &= \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{var}(\nu_i) = \lim_{n \rightarrow \infty} \text{var}(T^1). \end{aligned}$$

Hence, as in the proof in Theorem 5.2.1, $\{T^2, \dots, T^w\}$ asymptotically has a multivariate normal distribution with mean $\mathbf{0}$ and variance $\lim_{n \rightarrow \infty} s_n^2 \mathbf{I}$, where $\mathbf{I}$ is the $(w-1) \times (w-1)$ identity matrix.

It is now straightforward to show that $T^{[1-\alpha]} - s_n \Phi(1-\alpha) \xrightarrow{d} 0$ for $n, w \rightarrow \infty$. Since T^1 converges to a continuous distribution, it follows that also $\phi_{n,w} - \phi'_n \xrightarrow{d} 0$ for $n, w \rightarrow \infty$. $\square$

5.2.4 Robustness

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

$$f(y_i) = \exp \left\{ \frac{y_i \eta_i - b(\eta_i)}{a_i} + c(y_i) \right\}, \quad (5.1)$$

where $\eta_i = x_i \beta + z_i \gamma$, $x_i, \beta \in \mathbb{R}$, $z_i, \gamma \in \mathbb{R}^{k-1}$. Here β is the coefficient of interest and γ (which usually includes the intercept) may need to be estimated, as discussed in Section 5.3. 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 $\sum_{i=1}^n \nu_i = \frac{\partial}{\partial \beta} l(\beta | \mathbf{x}, \mathbf{z}, \mathbf{y})|_{\beta=\beta_0}$ is

$$\sum_{i=1}^n \frac{x_i (y_i - b'(\eta_i))}{a_i} \Big|_{\beta=\beta_0} = \sum_{i=1}^n \frac{x_i (y_i - \mathbb{E}(y_i))}{a_i} \Big|_{\beta=\beta_0}. \quad (5.2)$$

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

$$\sum_{i=1}^n x_i (y_i - \mu_i) \Big|_{\beta=\beta_0} = \sum_{i=1}^n x_i \{y_i - \exp(z_i \gamma)\}.$$

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

$$\sum_{i=1}^n \frac{x_i(y_i - \eta_i)}{\sigma^2} \Big|_{\beta=\beta_0}. \quad (5.3)$$

Apart from some mild assumptions, the main assumption made in Theorem 5.2.1 is that $\mathbb{E}(\nu_i) = 0$, $i = 1, \dots, n$. 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 5.2.1 holds.

As a concrete example, consider the normal model with identity link function, which assumes that $\text{var}(Y_1) = \dots = \text{var}(Y_n)$. 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, since the estimated variance does not have the assumed chi-squared distribution. In Section 5.4 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. $\text{var}(Y_i) = \mu_i$ is assumed, but in reality $\text{var}(Y_i) > \mu_i$, 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 covariates, then again the parametric test loses its properties. Theorem 5.2.1, 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 5.4.2). Such latent nuisance variables will increase the variance of Y_i , however, which poses a problem for the 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 $n \rightarrow \infty$), since $\mathbb{E}(Y_i - \mu_i)$ becomes dependent on X_i under H_0 .

5.3 Taking into account nuisance estimation

Consider independent and identically distributed pairs (Y_i, X_i) , $i = 1, \dots, n$ where $X_i \in \mathbb{R}^d$, $d \in \mathbb{N}$, and $Y_i \in \mathbb{R}$ has distribution $\mathbb{P}_{\beta, \gamma_0, X_i}$, which depends on parameter of interest $\beta \in \mathbb{B} \subseteq \mathbb{R}$ and unknown nuisance parameter $\gamma_0 \in \mathbb{G} \subseteq \mathbb{R}^{k-1}$. We will discuss the issues arising from estimating γ_0 and propose a solution, which allows us to obtain an asymptotically exact test.

As above, we consider the null hypothesis $H_0 : \beta = \beta_0$. Generalizations to hypotheses of the form $\beta_0 \geq C$ are straightforward, as well as generalizations to two-sided tests.

Suppose that $\mathbb{P}_{\beta, \gamma, X_i}$ has a density f_{β, γ, X_i} with respect to some dominating measure. Let γ denote an arbitrary element of $\mathbb{G}$. For $1 \leq i \leq n$ write

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

where we assume the derivative exists. The value ν_i is the score for the i -th observation. Under H_0 , $\mathbb{E}(\nu_{\gamma_0, i}) = 0$, $i = 1, \dots, n$. The score for all n observations simultaneously is $n^{1/2}S_\gamma$, where

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

Assume that $\hat{\gamma}$ is a $\sqrt{n}$ -consistent estimate of γ_0 . For every $1 \leq i \leq n$, let

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

denote the $(k-1)$ -vector of score contributions for the nuisance parameters. Let

$$\boldsymbol{S}_\gamma^{(k-1)} = n^{-1/2} \sum_{i=1}^n \boldsymbol{\nu}_{\gamma, i}^{(k-1)} \in \mathbb{R}^{k-1}$$

be the vector of nuisance scores.

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

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

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

5.3.1 Asymptotically exact test

When the nuisance parameter γ_0 is unknown, it needs to be estimated, which is typically done by maximizing the likelihood of the data under the null hypothesis. Consequently, the distribution of $S_{\hat{\gamma}}$ can be substantially different than that of S_{γ_0} , the score based on the true nuisance parameters. Indeed, 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 the asymptotic variance of the *effective score*, defined below. The effective information is usually 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 γ_0 makes the summands $\nu_{\hat{\gamma},1}, \dots, \nu_{\hat{\gamma},n}$ underlying $S_{\hat{\gamma}}$ correlated, in such a way that $\text{var}(S_{\hat{\gamma}}) < \text{var}(S_{\gamma_0})$ (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 5.3.1). Hence, using $\nu_{\hat{\gamma},1}, \dots, \nu_{\hat{\gamma},n}$ in the test of Theorem 5.2.1 can lead to a conservative test, even as $n \rightarrow \infty$.

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 $\hat{\gamma}$. 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 $\nu_{\hat{\gamma},i}^*$, $i = 1, \dots, n$ (which we assume have nonzero variance for $\hat{\gamma} = \gamma_0$) are defined as

$$\begin{aligned} S_{\hat{\gamma}}^* &= S_{\hat{\gamma}} - \hat{\mathbf{I}}_{12}' \hat{\mathbf{I}}_{22}^{-1} S_{\hat{\gamma}}^{(k-1)}, \\ \nu_{\hat{\gamma},i}^* &= \nu_{\hat{\gamma},i} - \hat{\mathbf{I}}_{12}' \hat{\mathbf{I}}_{22}^{-1} \nu_{\hat{\gamma},i}^{(k-1)}, \end{aligned}$$

so that

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

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

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

As discussed, $S_{\hat{\gamma}}$ is not generally asymptotically equivalent to S_{γ_0} . The effective score however is the residual from the projection of the score on the space spanned by the nuisance scores. Hence $S_{\gamma_0}^*$ is uncorrelated with the nuisance scores $\mathbf{S}_{\gamma_0}^{(k-1)}$ (Marohn, 2002). Consequently, as noted in the proof below, under mild regularity assumptions $S_{\hat{\gamma}}^* = S_{\gamma_0}^* + o(1)$, i.e. asymptotically the effective score it is not affected by $\hat{\gamma}$.

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

Like Marohn (2002), we assume that

$$S_{\hat{\gamma}} = S_{\gamma_0} - \mathcal{I}'_{12} \sqrt{n}(\hat{\gamma} - \gamma_0) + o(1),$$

$$\mathbf{S}_{\hat{\gamma}}^{(k-1)} = \mathbf{S}_{\gamma_0}^{(k-1)} - \mathcal{I}_{22} \sqrt{n}(\hat{\gamma} - \gamma_0) + o(1),$$

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

Theorem 5.3.1. *Consider the test of Theorem 5.2.1 with $T^j = S_{\hat{\gamma}}^{*j}$, $1 \leq j \leq w$. As $n \rightarrow \infty$, the level of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$.*

Proof. Suppose that H_0 holds. Note that

$$\begin{aligned} S_{\hat{\gamma}}^* &= S_{\hat{\gamma}} - \hat{\mathcal{I}}'_{12} \hat{\mathcal{I}}_{22}^{-1} \mathbf{S}_{\hat{\gamma}}^{(k-1)} = S_{\hat{\gamma}} - \mathcal{I}'_{12} \mathcal{I}_{22}^{-1} \mathbf{S}_{\hat{\gamma}}^{(k-1)} + o(1) = \\ &= S_{\gamma_0} - \mathcal{I}'_{12} \sqrt{n}(\hat{\gamma} - \gamma_0) - \mathcal{I}'_{12} \mathcal{I}_{22}^{-1} \left\{ \mathbf{S}_{\gamma_0}^{(k-1)} - \mathcal{I}_{22} \sqrt{n}(\hat{\gamma} - \gamma_0) \right\} + o(1) = \\ &= S_{\gamma_0}^* + o(1). \end{aligned}$$

Let $2 \leq j \leq w$ and

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

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

Note that

$$S_{\hat{\gamma}}^j = S_{\hat{\gamma}}^{j+} - S_{\hat{\gamma}}^{j-} = \left\{ S_{\gamma_0}^{j+} - \frac{1}{2} \sqrt{n} \mathcal{I}'_{12}(\hat{\gamma} - \gamma_0) \right\} - \left\{ S_{\gamma_0}^{j-} - \frac{1}{2} \sqrt{n} \mathcal{I}'_{12}(\hat{\gamma} - \gamma_0) \right\} + o(1) =$$

$$S_{\gamma_0}^{j+} - S_{\gamma_0}^{j-} + o(1) = S_{\gamma_0}^j + o(1).$$

The intuitive reason why $S_{\hat{\gamma}}^j = S_{\gamma_0}^j + o(1)$, is that the estimation of $\hat{\gamma}$ does not cause the summands underlying $S_{\hat{\gamma}}^j$ to be correlated (as is the case for the original score $S_{\hat{\gamma}}$). Similarly we find that $S_{\hat{\gamma}}^{(k-1),j} = S_{\gamma_0}^{(k-1),j} + o(1)$ and conclude that $S_{\hat{\gamma}}^{*j} = S_{\gamma_0}^{*j} + o(1)$.

Let $\mathbf{T}^n$ be as in the proof of Theorem 5.2.1, with ν_i replaced by $\nu_{\gamma_0,i}^*$. Suppose H_0 holds and $\hat{\mathcal{I}} = \mathcal{I}$, so that the summands underlying $\mathbf{T}_j^n$ are independent. For every $1 \leq i \leq n$, $\mathbb{E}(\nu_{\gamma_0,i}^*) = 0$. The elements of $\mathbf{T}^n$ are uncorrelated and have common variance $\text{var}(\nu_{\gamma_0,1}^*)$. Hence, by the Lindeberg-Feller central limit theorem (Greene, 2012, p. 1123), $\mathbf{T}^n$ converges in distribution to $N(\mathbf{0}, \text{var}(\nu_{\gamma_0,1}^*) \mathbf{I})$. We supposed that $\hat{\mathcal{I}} = \mathcal{I}$ to use the central limite theorem, but the asymptotic distribution of $\mathbf{T}^n$ is the same if $\hat{\mathcal{I}}$ is any consistent estimator of $\mathcal{I}$.

Let $\hat{\mathbf{T}}^n$ be as in the proof of Theorem 5.2.1, with ν_i replaced by $\nu_{\hat{\gamma},i}^*$. For every $1 \leq j \leq w$, $S_{\hat{\gamma}}^{*j} = S_{\gamma_0}^{*j} + o(1)$. Thus $\hat{\mathbf{T}}^n$ and $\mathbf{T}^n$ are asymptotically equivalent. The result now follows from Lemma 5.2.1. $\square$

The test of Theorem 5.3.1 has a parametric counterpart, which uses that under H_0 , S_{γ}^* is asymptotically normal with known variance, the effective information (Marohn, 2002; Hall and Mathiason, 1990). As $n, w \rightarrow \infty$, this test becomes equivalent to the test of Theorem 5.3.1, as the following proposition says.

Proposition 5.3.1. *Proposition 5.2.3 (asymptotic equivalence with the parametric test) still holds if for $1 \leq j \leq w$ we take*

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

Proof. In case $\hat{\gamma} = \gamma_0$, the proof is analogous to that of Proposition 5.2.3. It is left to show that as $n, w \rightarrow \infty$, the test of Theorem 5.3.1 with $\hat{\gamma} = \gamma_0$ becomes equivalent to the test of Theorem 5.3.1 with $\hat{\gamma}$ any other $\sqrt{n}$ -consistent estimator. This follows in a straightforward way since for w fixed, $\mathbf{T}^n$ and $\hat{\mathbf{T}}^n$ (as in the proof of Theorem 5.3.1) are asymptotically equivalent. $\square$

5.3.2 Robustness

In Section 5.2.4 it was explained that the test of Theorem 5.2.1 is usually robust against misspecification of the variance of the score. The test of Theorem 5.3.1 is also robust against certain forms of variance misspecification. This is in particular the case if $S_{\hat{\gamma}}$ and $\mathbf{S}_{\hat{\gamma}}^{(k-1)}$ are misspecified by the same factor, see Proposition 5.3.2. This is for example the case if the variance misspecification factor is independent of the covariates.

Proposition 5.3.2. *Suppose that $\hat{\mathbf{T}} = n^{-1} \mathbf{X}' \hat{\mathbf{W}} \mathbf{X}$, where $\mathbf{X}$ is an $n \times k$ design matrix with i.i.d. rows and $\hat{\mathbf{W}}$ a weight matrix. Consider a misspecification factor $c_1 > 0$ and misspecified scores*

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

Further, for $c_2 > 0$ consider the misspecified weight matrix $\tilde{\mathbf{W}} = c_2 \hat{\mathbf{W}}$. Let $\tilde{\mathbf{T}} = n^{-1} \mathbf{X}' \tilde{\mathbf{W}} \mathbf{X}$ be the misspecified average Fisher information. Let $\tilde{\nu}_{\hat{\gamma},i}^ = \tilde{\nu}_{\hat{\gamma},i} - \tilde{\mathbf{T}}_{12}' \tilde{\mathbf{T}}_{22}^{-1} \tilde{\nu}_{\hat{\gamma},i}^{(k-1)}$ be the misspecified effective scores, $i = 1, \dots, n$. Consider the test of Theorem 5.3.1, with $S_{\hat{\gamma}}^{*j}$, $j = 1, \dots, w$, replaced by the misspecified effective score*

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

As $n \rightarrow \infty$, the level of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$.

Proof. For every $1 \leq j \leq w$ we have

$$\tilde{S}_{\hat{\gamma}}^{*j} = c_1 S_{\hat{\gamma}}^j - c_2 \hat{\mathcal{I}}_{12}' c_2^{-1} \hat{\mathcal{I}}_{22}^{-1} c_1 \mathbf{S}_{\hat{\gamma}}^{(k-1),j} = c_1 S_{\hat{\gamma}}^{*j}.$$

Hence the test is identical to that of Theorem 5.3.1, since that test is unchanged if all T^j are multiplied by a constant. $\square$

Proposition 5.3.2 is useful, since it tells us that if in a GLM $\text{var}(Y_i)$ is misspecified by a constant, such that $\hat{\mathbf{W}}$ and the scores are misspecified by a constant, the resulting test is still asymptotically exact. In Proposition 5.3.2 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. We could slightly relax this assumption, allowing these factors to be random. Indeed, in practice the test can be very robust against heteroscedasticity (see Section 5.4). The variance misspecification is not generally allowed to depend on the covariates, since then $S_{\hat{\gamma}}$ and $\mathbf{S}_{\hat{\gamma}}^{(k-1)}$ can be misspecified by different factors asymptotically. There are exceptions however, see Sections 5.3.3 and 5.4.1.

The test based on the basic scores $\nu_{\hat{\gamma},i}$ is valid as soon as $\mathbb{E}(\nu_{\hat{\gamma},i}) = 0$, and hence it may sometimes be preferred over the test of Theorem 5.3.1. The test based on the basic score is asymptotically conservative if the score S_{γ_0} is correlated with the nuisance scores $\mathbf{S}_{\gamma_0}^{(k-1)}$, i.e. when $\mathcal{I}_{12} \neq 0$. Hence it can be useful to redefine the covariate of interest such that $\mathcal{I}_{12} = 0$. When $\hat{\mathbf{W}} = b\mathbf{I}$, $b > 0$, this means orthogonalizing the covariate of interest with respect to the nuisance covariates. When the model is potentially misspecified, then the weights and hence $\mathcal{I}_{12}$ are not asymptotically known, but the user could substitute a best guess for the weights.

5.3.3 An example

As discussed, the test of Theorem 5.3.1 is often not asymptotically exact if the variance misspecification depends on the covariates. An important exception is the case where the model is

$$Y_i \sim N(\gamma_0 + \beta X_i', \sigma^2) \quad i = 1, \dots, n, \quad (5.4)$$

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 $1 \leq i \leq n$,

$$\begin{aligned}\nu_{\hat{\gamma},i} &= x'_i(y_i - \hat{\mu}_i)/\sigma^2, \\ \nu_{\hat{\gamma},i}^{(k-1)} &= (y_i - \hat{\mu}_i)/\sigma^2.\end{aligned}$$

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

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

Thus, the effective score contributions are exactly the basic score contributions after centering $x'_1, \dots, x'_n$ around 0. Similarly, if $x'_1, \dots, x'_n$ are already centered, then $\nu_{\hat{\gamma},i}$ and $\nu_{\hat{\gamma},i}^*$ coincide, since then $\hat{\mathcal{T}}_{12} = 0$.

The test of Theorem 5.3.1 is not always asymptotically exact if $S_{\hat{\gamma}}$ and $S_{\hat{\gamma}}^{(k-1)}$ are misspecified by different factors. However, if $\hat{\mathcal{T}}_{12} = 0$, then this does not apply anymore. The test of Theorem 5.3.1 then remains asymptotically exact and reduces to the test based on the basic score. For the model (5.4), this means that even if the misspecification of $\text{var}(Y_i)$ 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 $\mu^1 \leq \mu^2$ or $\mu^1 \geq \mu^2$). 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, 2010; Lehmann and Romano, 2005). Under mild assumptions, we obtain an asymptotically exact test for this problem. Pesarin and Salmaso (2010) already suggested sign-flipping residuals to solve this problem. This is equivalent to flipping scores in our linear model (5.4) with $|x'_1| = \dots = |x'_n|$.

5.4 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.

5.4.1 Overdispersion, heteroscedasticity and estimated nuisance

In Sections 5.4.1 and 5.4.2 the assumed model was Poisson, but in fact $Y_1, \dots, Y_n$ 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 $\text{var}(X) = \text{var}(Z) = \text{var}(Z^l) = 1$. (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 5.4.1 we took $\gamma^l = 0$. 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 $\text{var}(Y_i) = \mu_i + \mu_i^2$.

The assumed model, however, was Poisson, i.e. $\text{var}(Y_i) = \mu_i$ 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 5.4.1 the linear predictor was correct as well.

In Figure 5.4.1 the estimated rejection probabilities 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 5.4.1, this test was anti-conservative for small n . 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 $\nu_{\gamma,i}$. We took $w = 200$ (taking w larger gives similar results, see also Marriott,

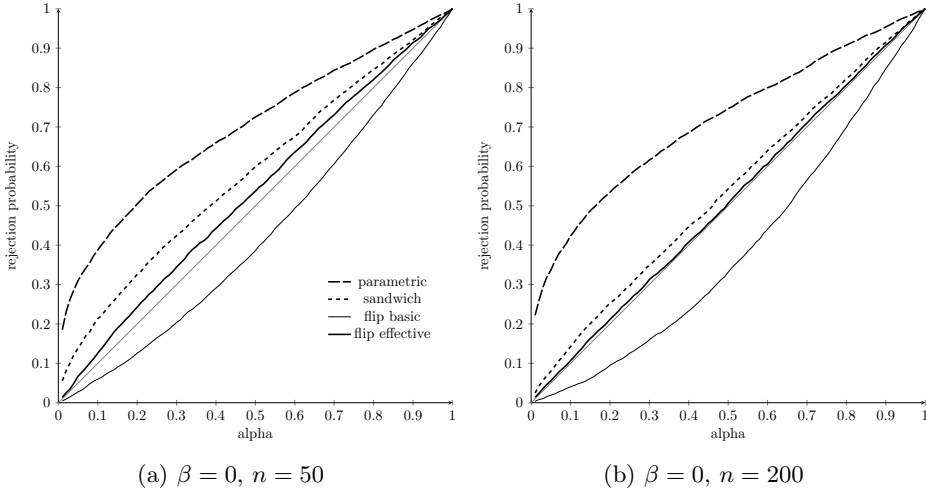


Figure 5.1: Estimated rejection probabilities for four tests under misspecified variance and estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

1979). Due to the estimation of γ_0 , the variance of the score was shrunk and the test was conservative, as explained in Section 5.3.1.

Finally, we used the sign-flipping test of Theorem 5.3.1, which is based on the effective scores $\nu_{\hat{\gamma},i}^*$. In Section 5.3.2 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 level of the test was approximately α . This illustrates that the test has some additional robustness, which we have not theoretically shown.

5.4.2 Ignored nuisance

The same simulations were performed as in Section 5.4.1, but with $\gamma_0^l = 1$. Since $\gamma_0^l = 0$ was assumed, Z_i^l represented an ignored, latent variable. Figure 5.4.2 shows similar results as Figure 5.4.1. The parametric test was even more anti-conservative than in Section 5.4.1. 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 5.4.1.

The test of Theorem 5.3.1 was still nearly exact for $n = 200$, even though μ_i was misspecified. (Even marginally over Z_i^l , μ_i was misspecified.

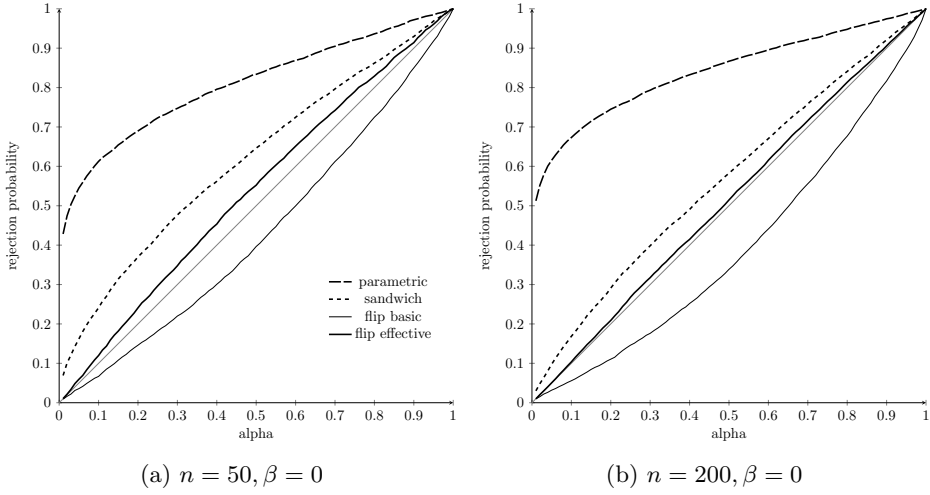


Figure 5.2: Estimated rejection probabilities for four tests under misspecified variance, estimated nuisance and ignored nuisance. The null hypothesis was $H_0 : \beta = 0$.

Possibly the estimation of the intercept corrected for the misspecification.)

A conclusion from the simulations of Sections 5.4.1 and 5.4.2, 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 5.3.1 was substantially less anti-conservative (while having similar power, see Section 5.4.3).

5.4.3 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 5.4.3. The estimated probabilities are based on 20,000 simulation loops.

Since the model was correct, asymptotically there was no better choice than the parametric test. The sign-flipping test of Theorem 5.3.1 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).

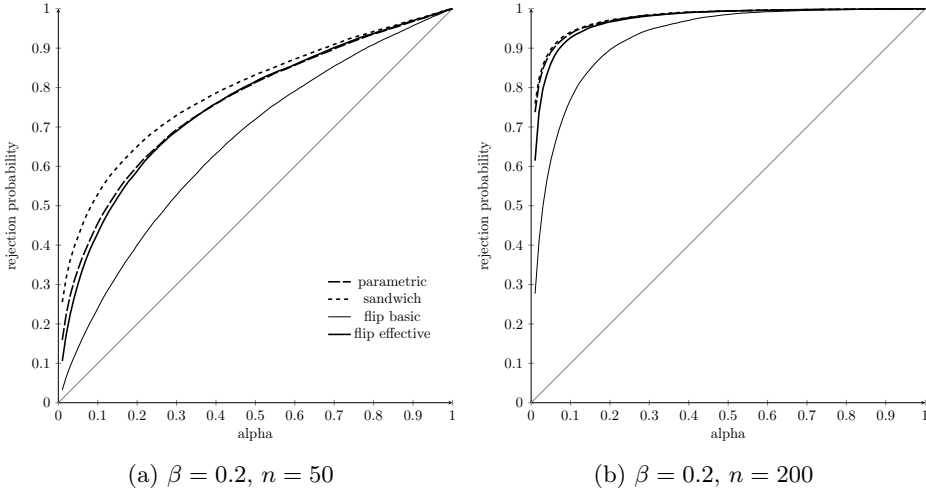


Figure 5.3: Power comparison of four two-sided tests under the correct model, with estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

5.4.4 Strong heteroscedasticity

When a Gaussian linear model is considered with $Y_i \sim N(\beta x_i, \sigma^2)$, $x_1 = \dots = x_n = 1$ 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 5.2.1 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 the level of the t-test was far from nominal for most α , see Figure 5.4a. The sign-flipping test did not need to estimate the variance. In this setting the test has level $\lfloor \alpha w \rfloor / w$ exactly if the transformations $g_1, \dots, g_w$ are drawn without replacement, since the observations are symmetric, see Proposition 5.2.1. (We drew $g_1, \dots, g_w$ 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 $\sigma_1 = \dots = \sigma_{10} = 1$. Figure 5.4b, based on 10^5 repeated simulations, shows that the tests had virtually the same power.

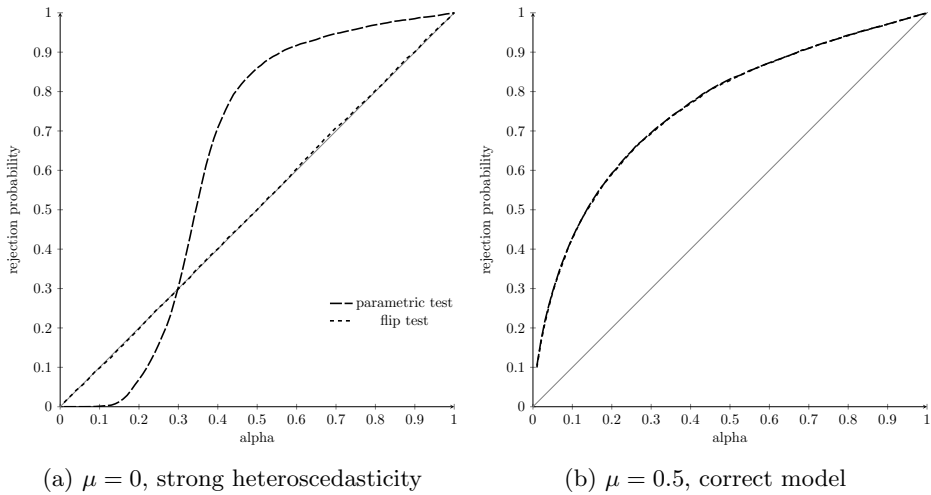


Figure 5.4: Comparison of the one-sample t-test and the sign-flipping test. The null hypothesis was $H_0 : \mu = 0$.

5.5 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

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

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 5.4, 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 $\nu_{\gamma,i}$, $i = 1, \dots, 54$ (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 5.3.1 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 5.4.1.

5.6 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 robust to these types of misspecification and is often still asymptotically exact. It often has similar power to the parametric score test, if the model is correct.

When nuisance parameters are estimated, the 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 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 under constant overdispersion, and simulations illustrate additional robustness.

Despite its potential conservativeness, the test based on the basic score may be preferred over that based on the effective score, for its robustness. Moreover, the basic test does not require computing the Fisher information. This might provide an advantage in applications where this computation is time-consuming.

