Permutation-based multiple testing when fitting many generalized linear models

Riccardo De Santis*[1], Jelle J. Goeman[2], Samuel Davenport[3], Jesse Hemerik[4], and Livio Finos[5]

[1]University of Siena, Italy
[2]Leiden University Medical Center, The Netherlands
[3]University of California San Diego, United States
[4]Erasmus University Rotterdam, The Netherlands
[5]University of Padova, Italy

Abstract

The multiple testing problem appears when fitting multivariate generalized linear models for high dimensional data. We show that the sign-flip test can be combined with permutation-based procedures for assessing the multiple testing problem

1 Introduction

In high-dimensional data, such as neuroimaging and transcriptomics, it is common to fit many generalized linear regression models in parallel, each with a small sample size [Schaarschmidt et al., 2022, Love et al., 2014, Winkler et al., 2014]. Usually, the goal of the analysis is to perform hypothesis testing to find relevant associations, usually after adjusting the p-values for multiple testing.

*To contact: *riccardo.desantis2@unisi.it*

This approach comes with several challenges. In the first place, small sample sizes can make classical tests unreliable, especially in nonlinear models where exact methods are not available and tests rely on asymptotic arguments. In such cases, the usual normal approximation of the test statistic can be quite unreliable [Schaarschmidt et al., 2022]. Secondly, the generalized linear model makes some crucial assumptions which are difficult to check, especially for nonlinear models with small sample size. In particular, the detection of overdispersion over the assumed models can be quite problematic, especially since the variance is not constant among the observations even without overdispersion. Proper model checking is further hampered by the sheer number of models that need to be checked in high-dimensional data. Finally, the test statistics of the parallel models are often correlated, due to correlations in the underlying biological measurements. Classical multiple testing corrections (such as Bonferroni-Holm) are designed in order to protect against any correlation structure, but can be very conservative in the presence of such correlations [Goeman and Solari, 2014, Gao et al., 2010, Saffari et al., 2018], while taking correlations into account could increase the power of the test procedure.

In order to address all these issues we propose a permutation-based multiple testing procedure in combination with the sign-flip score test [Hemerik et al., 2020, De Santis et al., 2022]. The test shows reliable behavior for small sample sizes, and has been shown to be robust against general variance misspecification under minimal assumptions. We start by defining a test for a global null hypothesis about a multivariate regression parameter, which guarantees weak control of the familywise error rate (FWER). The global test statistics can be used to make additional inference through the Closed Testing approach [Marcus et al., 1976]. Consequently we compute adjusted p-values for each individual hypothesis through a multiple testing procedure based on the max-T method of Westfall and Young [1993], which guarantees strong control of the FWER, that is, it produces valid adjusted p-values for all possible subsetting hypotheses.

The key point of using a permutation-based solution is that the proposed method adapts to the unknown correlation structure. It is especially useful when strong correlation between the individual test statistics is present; in that situation it guarantees a relevant gain in power over alternative methods, as we will show in a simulation. Further, the sign-flip test can be applied to estimate a lower bound for the true discovery proportion (TDP), that is, the proportion of non-null coefficients over a pre-specified group of hypotheses,

using a permutation-based framework [Goeman and Solari, 2011, Andreella et al., 2023, Vesely et al., 2023, Blain et al., 2022].

The paper is organized as follows: Section 2 revisits the sign-flip score test for univariate testing; Section 3 contains the novel contribution of the paper, that is, the multivariate testing extension. Section 4 contains a simulation study.

Code to implement our methods is available in the flipscores R package [R Core Team, 2023], available from CRAN and in the pyperm python package [Davenport, 2023]. Code to reproduce the results of this paper is available at `github.com/rds/multiflip`.

2 One dependent variable

In this Section we review the derivation of the univariate sign-flip score test as detailed in Hemerik et al. [2020] and De Santis et al. [2022].

Let Y_i be the target variable which belongs to the exponential dispersion family, i.e., with a density of the form [Agresti, 2015]

$$f_{\beta,\boldsymbol{\gamma},\boldsymbol{X}_i}(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i,\phi_i)\right\},$$

where θ_i and ϕ_i are respectively the canonical and the dispersion parameter while $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. Consequently, the mean and variance of the observed outcome are defined as

$$\mu_i = E(Y_i) = b'(\theta_i); \qquad Var(Y_i) = b''(\theta_i)a(\phi_i).$$

The expected value of the vector $Y = (Y_1, \ldots, Y_n)^T$ is assumed to depend on some covariates through the relation

$$\mathbb{E}(Y) = g^{-1}(X\beta + Z\gamma)$$

where $g(\cdot)$ is the link function, the covariate X is an $n \times 1$ matrix, i.e., a column vector, such that $\beta \in \mathbb{R}$. Further, the nuisance covariates Z are a $n \times (k-1)$ matrix.

Throughout the section we are interested in the null hypothesis $H_0 : \beta = \beta_0$ against a one or two-sided alternatives. Consequently, the other parameters γ (and ϕ_i) are considered as nuisance parameters.

Let

$$D = \text{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\}; \qquad V = \text{diag}\{\text{Var}(y_i)\}.$$

The diagonal matrix of the GLM weights is defined as $W = DV^{-1}D$ with entries $w_i = (\partial \mu_i / \partial \eta_i)^2 / var(Y_i)$ [Agresti, 2015]. Finally,

$$H = W^{1/2} Z (Z'WZ)^{-1} Z' W^{1/2}$$

represents the projection matrix for GLMs.

The test is based on the *effective score*. The interpretation of the effective score is that if the true nuisance parameter is plugged in, the effective score is the residual from the projection of the marginal score for β on the space spanned by the nuisance scores [Marohn, 2002]. If the estimated nuisance parameter $\hat{\gamma}$ is plugged in, the effective score is

$$n^{1/2} S^*_{\hat{\gamma}} = X' W^{1/2} (I - H) V^{-1/2} (Y - \hat{\mu}) = \sum_{i=1}^{n} \nu^*_{\hat{\gamma},i}.$$

Note that indeed the effective score can be written as a sum of n elements $\nu^*_{\hat{\gamma},i}$, which we call the score contributions.

Regarding the estimator $\hat{\boldsymbol{\gamma}}$, we make the same assumptions of De Santis et al. [2022], in particular that $\hat{\boldsymbol{\gamma}}$ is a $\sqrt{n}$-consistent estimate of the true parameter $\boldsymbol{\gamma}_0$.

The test is performed by means of random sign flipping of score contributions. Define $g_1 = (1, \ldots, 1) \in \mathbb{R}^n$ and for every $2 \leq j \leq w$ let $g_j = (g_{j1}, \ldots, g_{jn})$ be independent and uniformly distributed on $\{-1, 1\}^n$. For $1 \leq j \leq w$, let the superscript j denote that g_j has been applied. The transformed test statistics are defined as

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

The derivation of the flipped effective score statistic in this matrix notation is straightforward. Note that the effective score is the product of an $1 \times n$ and an $n \times 1$ vector. This means that we can write it multiplying the $n \times 1$ matrix by the $n \times n$ sign-flipping matrix G_j, with diagonal entries $g_{j1}, \ldots, g_{jn}$. The transformed test statistic is therefore

$$S^{*,j}_{\hat{\gamma}} = n^{-1/2} X' W^{1/2} (I - H) V^{-1/2} G_j (Y - \hat{\mu}).$$

The test based on sign-flipping effective scores is asymptotically exact, as the following theorem states. This result coincides with Theorem 2 in Hemerik et al. [2020].

Theorem 1 (Hemerik et al., 2020)**.** *For every* $1 \leq j \leq w$*, consider the statistic* $T_j^n = S_{\hat{\gamma}}^{*j}$ *and let* $T_{(1)}^n \leq ... \leq T_{(w)}^n$ *be the sorted statistics. Consider the test that rejects if* $T_1^n > T_{\lceil (1-\alpha)w \rceil}^n$*. As* $n \to \infty$*, under* H_0 *the rejection probability converges to* $\lfloor \alpha w \rfloor / w \leq \alpha$*.*

We will now recall the definition of the test proposed in De Santis et al. [2022], which is a recent upgrade of the test above. This recent adaptation often improves the small-sample performance of the test by standardization of the effective score. We call the resulting test statistic the standardized score statistic, defined as

$$S_{\hat{\gamma}}^S = S_{\hat{\gamma}}^* / Var\{S_{\hat{\gamma}}^*\}^{1/2}, \tag{1}$$

where

$$Var\{S_{\hat{\gamma}}^*\} = n^{-1} X^T W^{1/2} (I - H) W^{1/2} X.$$

The transformed test statistic, for a generic flip G_j, is defined as

$$S_{\hat{\gamma}}^{S,j} = S_{\hat{\gamma}}^{*,j} / Var\{S_{\hat{\gamma}}^{*,j}\}^{1/2}, \tag{2}$$

where

$$Var\{S_{\hat{\gamma}}^{*,j}\} = n^{-1} X^T W^{1/2} (I - H) G_j (I - H) G_j (I - H) W^{1/2} X.$$

The test based on sign-flipping standardized scores is asymptotically exact, as the following Theorem states (for a discussion about the faster convergence, with respect to the effective scores, see De Santis et al. [2022]). This result coincides with Proposition 1 in De Santis et al. [2022].

Theorem 2 (De Santis et al., 2022)**.** *For every* $1 \leq j \leq w$*, consider the statistic* $T_j^n = S_{\hat{\gamma}}^{S,j}$ *and let* $T_{(1)}^n \leq ... \leq T_{(w)}^n$ *be the sorted statistics. Consider the test that rejects if* $T_1^n > T_{\lceil (1-\alpha)w \rceil}^n$*. As* $n \to \infty$*, under* H_0 *the rejection probability converges to* $\lfloor \alpha w \rfloor / w \leq \alpha$*.*

Some robustness properties of this test are discussed in Hemerik et al. [2020] and in De Santis et al. [2022]. In particular, the proposed test is proven to be asymptotically exact in case of general variance misspecification under minimal assumptions.

3 Multiple responses

Suppose there are $m \geq 2$ dependent variables $Y^1, \ldots, Y^m$. For each response Y^l we have n independent observations $Y_1^l, \ldots, Y_n^l$, which follow some model in the exponential dispersion family. We consider m null hypotheses $H_1, \ldots, H_m$, where H_l is the hypothesis that $\beta^l = \beta_0^l$. Like before, we assume $\hat{\boldsymbol{\gamma}}^l$ is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\gamma}_0^l$.

This section introduces the multiple testing procedure. It involves computing the effective score for each of the dependent variable $Y^1, \ldots, Y^m$. The effective score for the l-th response is then

$$S_{\hat{\boldsymbol{\gamma}}}^{*,l} = n^{-1/2} X'(W^l)^{1/2}(I - H^l)(V^l)^{-1/2}(Y^l - \hat{\mu}^l) = n^{-1/2} \sum_{i=1}^{n} \nu_{\hat{\boldsymbol{\gamma}},i}^{*l}.$$

Analogously to Section 2, we define the sign-flipped effective score test statistic for a generic flip matrix G_j as

$$S_{\hat{\boldsymbol{\gamma}}}^{*,j,l} = n^{-1/2} X'(W^l)^{1/2}(I - H^l)(V^l)^{-1/2} G_j (Y^l - \hat{\mu}^l) = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_{\hat{\boldsymbol{\gamma}},i}^{*l}.$$

Further, we have

$$Var\left\{S_{\hat{\boldsymbol{\gamma}}}^{*,j,l}\right\} = X'(W^l)^{1/2}(I - H^l)G_j(I - H^l)G_j(I - H^l)(W^l)^{1/2}X.$$

while the standardized score is

$$S_{\hat{\boldsymbol{\gamma}}}^{S,j,l} = S_{\hat{\boldsymbol{\gamma}}}^{*,j,l} Var\left\{S_{\hat{\boldsymbol{\gamma}}}^{*,j,l}\right\}^{-1/2}.$$

We can build a multivariate test statistic which takes into account the standardization of the joint variance-covariance matrix as follows. Let

$$H_0 = \bigcap_{l \in L} H_L : \beta^l = \beta_0^l \tag{3}$$

where L is a set containing any combination of the m hypotheses. Note that H_0 can be either a global or a partial null hypothesis for a subset of parameters. For simplicity, we will derive the procedure for the global null hypothesis, but it is exactly analogous for partial null hypotheses of any kind.

A first idea to perform the test might be given by a full standardization. We can indeed build a global test statistic based on the form of a Mahalanobis

distance as follows. Let $S^{*,j,L}_{\hat{\gamma}}$ be the vector with element $S^{*,j,l}_{\hat{\gamma}}$, $1 \leq l \leq m$. Note that by definition it is an m-dimensional vector. Call $Var\{S^{*,j,L}_{\hat{\gamma}}\}$ the corresponding variance-covariance matrix, which has dimension $m \times m$ and is assumed (for the moment) known. Let the joint test statistic be

$$T^n_{j,L} = \left(S^{*,j,L}_{\hat{\gamma}}\right)' Var\left\{S^{*,j,L}_{\hat{\gamma}}\right\}^{-1} \left(S^{*,j,L}_{\hat{\gamma}}\right) \tag{4}$$

If we assume the normality of the response we have the following theorem;

Theorem 3. *Assume to fit a normal linear regression model. Let $T^n_{j,L}, 1 \leq j \leq w$ as defined in (4). Let $T^n_{(1)} \leq ... \leq T^n_{(w)}$ be the sorted statistics. Consider the test that rejects if $T^n_1 > T^n_{\lceil (1-\alpha)w \rceil}$. Under H_0, the test is an exact α level test.*

Proof. Trivially, note that $S^{*,j,L}_{\hat{\gamma}}$ is normally distributed with zero mean and that the test statistic is a quadratic form. The expected value of each test statistic is equal to m for every flip, while the variance is equal to $2m$ for every flip. This follows from standard properties of random vectors. It implies that all the test statistics share the first two moments. Thus we can directly apply Theorem 1 of Hemerik and Goeman [2018] to show that the test derived is an α-level test for finite sample size. □

The procedure outlined has some issues. First of all, the test is exact only for normal responses with known variance. Otherwise, we have to follow asymptotic arguments related to the asymptotic normal distribution of the effective score statistic [Marohn, 2002]. Another weak point is the fact that for each flip it is required to invert an $m \times m$ matrix for each flip, which might be unfeasible for large values of m. Further, we have to estimate the correlation between responses, which must be assumed to be known except for a limited number of parameters; for instance, we might choose to assume the correlation of the responses to be equal between different units. Moreover, in order to perform an overall analysis with post-hoc validity, we might choose a closed testing approach [Marcus et al., 1976], which requires to perform all the 2^m possible tests, which can be very demanding for growing m.

A fast alternative, which is more appealing for large values of m, consists of doing marginal standardization of the test statistic. We will start from the effective scores, showing that we are able to derive a valid procedure for both the flipscores approaches. Finally, we will see that we still have a remarkable result for the normal linear model.

Let $\boldsymbol{M}^n$ be the w-by-m matrix with (j,l)-th entry equal to $S^{*j,l}_{\hat{\boldsymbol{\gamma}}}$. The following lemma will be fundamental in proving that the proposed multiple testing methods are asymptotically exact.

Lemma 4. *Let $\boldsymbol{M}^n$ as defined above. Then, for $n \to \infty$, $\boldsymbol{M}^n$ converges in distribution to $\boldsymbol{M}$, where all rows of $\boldsymbol{M}$ have the same multivariate normal distribution.*

Proof. Let $\boldsymbol{M}^n_0$ be the w-by-m matrix with (j,l)-th entry equal to $S^{*,j,l}_{\boldsymbol{\gamma}_0}$. Note that $\boldsymbol{M}^n_0$ is based on knowledge of the true nuisance parameters $\boldsymbol{\gamma}_0$. The consequence is that each entry of the matrix $\boldsymbol{M}^n_0$ is the sum of n independent (flipped) score contributions. Further, note that each row of $\boldsymbol{M}^n_0$ is uncorrelated with the other rows, due to the independence of the flips. Finally, the correlation structure within each row coincides with the correlation structure of the contributions $\nu^{*1}_{\boldsymbol{\gamma}_0}, \ldots, \nu^{*m}_{\boldsymbol{\gamma}_0}$. Consequently, the multivariate central limit theorem [Van der Vaart, 1998] implies that $\boldsymbol{M}^n_0$ converges in distribution to some matrix $\boldsymbol{M}_0$, which has identically distributed multivariate normal rows.

Now it is left to show that $\boldsymbol{M}$, i.e., the matrix based on *estimated* nuisance parameters, also has identically distributed multivariate normal rows. As shown in the proof of Theorem 2 in Hemerik et al. [2020], we have $S^{*,j,l}_{\hat{\boldsymbol{\gamma}}} = S^{*,j,l}_{\boldsymbol{\gamma}_0} + o_p(1)$, $1 \leq j \leq w$, $1 \leq l \leq m$. This means that $\boldsymbol{M}$ is asymptotically equivalent to $\boldsymbol{M}_0$. Hence the result holds. □

If instead of effective scores we used standardized effective scores to fill the matrix $\boldsymbol{M}^n$, then Lemma 4 will still hold, since the standardized effective scores are asymptotically equivalent to the unstandardized effective scores. This is detailed in De Santis et al. [2022].

A special case of interest relates to homoscedastic linear regression models. The matrix of weights becomes an identity matrix, which causes some simplifications in the formulas; in particular

$$S^{*,j,l}_{\hat{\boldsymbol{\gamma}}} = X'(I-H)G_j(Y^l - \hat{\mu}^l)$$

and

$$Var\left\{S^{*,j,l}_{\hat{\boldsymbol{\gamma}}}\right\} = X'(I-H)G_j(I-H)G_j(I-H)X.$$

while the standardized score is still

$$S^{S,j,l}_{\hat{\boldsymbol{\gamma}}} = S^{*,j,l}_{\hat{\boldsymbol{\gamma}}} Var\left\{S^{*,j,l}_{\hat{\boldsymbol{\gamma}}}\right\}^{-1/2}.$$

For the linear regression model, using the standardized scores, we are able to get finite sample results, as stated in the following Lemma. This reflects that we still get the second-moment null-invariance property for linear regression models with normal response, while it is not true in general.

Lemma 5. *Assume a multivariate linear regression model with normal response, where the covariance between responses is equal among different units, that is,*

$$E[(Y^l - \mu^l)(Y^p - \mu^p)'] = \sigma_{pl}\boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Let $\boldsymbol{M}^{S,n}$ be the w-by-m matrix with (j,l)-entry equal to $S_{\hat{\gamma}}^{S,j,l}$. Then $\boldsymbol{M}^{S,n}$ has independent rows with the same multivariate normal distribution.

Proof. The rows of the matrix $\boldsymbol{M}^{S,n}$ are uncorrelated, due to the independence of the flips. Within the row the correlation structure is

$$Cov\left\{S_{\hat{\gamma}}^{S,j,l}, S_{\hat{\gamma}}^{S,j,p}\right\} = Var\left\{S_{\hat{\gamma}}^{*,j,l}\right\}^{-1/2} E\left[S_{\hat{\gamma}}^{*j,l}\left(S_{\hat{\gamma}}^{*,j,p}\right)'\right] Var\left\{S_{\hat{\gamma}}^{*,j,p}\right\}^{-1/2}$$

Using the result

$$E[(Y^l - \mu^l)(Y^p - \mu^p)'] = \sigma_{pl}\boldsymbol{I}_n$$

we get, after simple computations,

$$Cov\left\{S_{\hat{\gamma}}^{S,j,l}, S_{\hat{\gamma}}^{S,j,p}\right\} = \sigma_{pl}/(\sigma_p\sigma_l),$$

which is independent of the flip, where σ_l is the variance of the l-th response. It follows that the matrix $M^{S,n}$ has independent and identically distributed multivariate normal rows for finite sample size. □

From Lemma 4 (or 5) we can build an asymptotically exact local α-level test for any composite hypothesis as follows. Here we derive the results for a two-sided alternative. Let H_0 as defined in (3). Define $\psi : \mathbb{R}^k \to \mathbb{R}$ as a function non decreasing in its argument. We can derive a global flipped test statistic as

$$T_j^n = \psi\left(|S_{\hat{\gamma}}^{S,j,1}|, \ldots, |S_{\hat{\gamma}}^{S,j,L}|\right). \tag{5}$$

The following theorem shows that from this test statistic we can get an asymptotic α-level test.

Theorem 6. *For every $1 \leq j \leq w$, consider the statistic T_j^n as defined in (5) and let $T_{(1)}^n \leq ... \leq T_{(w)}^n$ be the sorted statistics. Consider the test that rejects if $T_1^n > T_{\lceil (1-\alpha)w \rceil}^n$. As $n \to \infty$, under H_0 the rejection probability converges to $\lfloor \alpha w \rfloor / w \leq \alpha$.*

Proof. Lemma 4 implies that the test statistics $T_1^n, \ldots, T_w^n$ are asymptotically independent and identically distributed. Note that, by the definition of ψ, high values of T_1 shows evidence against the null hypothesis H. Hence, Lemma 1 of Hemerik et al. [2020] directly applies and we derive that we get an asymptotically α-level test. □

Theorem 6 implies that we can build an asymptotic valid local test for the hypothesis (3), which therefore guarantees weak control of the FWER. Note that for the linear model we can obtain finite sample results, because we can directly apply Lemma 5 and it follows that the test has finite sample size properties.

Theorem 7. *Assume to fit a multivariate linear regression model with normal response. For every $1 \leq j \leq w$, consider the statistic T_j^n as defined in (5) and let $T_{(1)}^n \leq ... \leq T_{(w)}^n$ be the sorted statistics. Consider the test that rejects if $T_1^n > T_{\lceil (1-\alpha)w \rceil}^n$. Under H_0 the rejection probability is $\leq \alpha$.*

Proof. Lemma 5 implies that the test statistics $T_1^n, \ldots, T_w^n$ are independent and identically distributed. Note that, by the definition of ψ, high values of T_1 shows evidence against the null hypothesis H. Hence, Theorem 1 of Hemerik and Goeman [2018] directly applies and the test derived is an α-level test for finite sample size. □

Multiple choices of the function ψ are available [Pesarin, 2001] and the choice will influence the power properties in different settings. We can subsequently apply the closed testing approach [Marcus et al., 1976] to build a procedure which guarantees strong control of the FWER by computing the 2^n intersection tests, and this procedure is optimal in the sense that every multiple testing procedure is equal or can be improved by applying the closed testing principle [Goeman et al., 2021] .

However, the number of tests might be unfeasible for large values of m. A dramatic shortcut is given by selecting the maximum of the test statistics as combining function in the following way. This is called the max-T approach. For sake of completeness, we will give a direct proof of its validity.

There are roughly two versions of the max-T method by Westfall and Young [Westfall and Young, 1993, Westfall and Troendle, 2008, Meinshausen et al., 2011]: the single-step method and the sequential method. The single-step method is simpler and faster, while the sequential method is more powerful. The single-step max-T method, based on the matrix $\boldsymbol{M}^n$ of test statistics, is defined as follows. Here we formulate a version that employs two-sided tests. For every $1 \leq j \leq w$, let m_j be the maximum of the test statistics $|\boldsymbol{M}^n_{j,l}|$, $1 \leq l \leq m$. Let $m_{(1)}, \ldots, m_{(w)}$ be the sorted values $m_1, \ldots, m_w$. Then the multiple testing method rejects all hypotheses with index l for which $|\boldsymbol{M}^n|_{1,l} > m_{(\lceil (1-\alpha)w \rceil)}$. The sequential max-$T$ method is defined as follows; after the first step defined above is completed, the procedure is continued in a step-down way. We remove from the matrix $\boldsymbol{M}^n$ all rows corresponding to the hypotheses rejected in the first step, then we apply again the same procedure described above. The procedure can be continued until we have no more rejections. Note that the test with standardized test statistic is defined in the same way.

The following theorem states that the single-step and sequential max-T methods provide strong asymptotic FWER control. Write $FWER_n$ to indicate potential dependence of the FWER on n.

Theorem 8. *For both the single-step and sequential max-T method,* $\limsup_{n\to\infty}(FWER_n) \leq \alpha$.

Proof. For both the single-step and the sequential max-T method, the argument is as follows. Recall that $\boldsymbol{M}^n$ converges in distribution to $\boldsymbol{M}$. Let $\boldsymbol{M}_{\mathcal{N}}$ be the submatrix of $\boldsymbol{M}$ that only contains the rows corresponding to the true hypotheses. If the matrix $\boldsymbol{M}$ is used as input for the multiple testing procedure, then strong FWER control follows directly from the fact that the rows of $\boldsymbol{M}_{\mathcal{N}}$ are exchangeable, i.e. swapping rows does not change the distribution of the matrix.

If instead $\boldsymbol{M}^n$ is used as input, then FWER control follows from the continuous mapping theorem, since $\boldsymbol{M}^n$ converges to $\boldsymbol{M}$ in distribution. This finishes the proof.

Finally, note that for the procedure that uses standardized scores for linear regression model the control of the familywise error rate is obtained for finite sample size, as the matrix $\boldsymbol{M}^{S,n}$ has an exact multivariate normal distribution. □

4 Simulation Study

4.1 Univariate

We first show a simulation study for univariate tests in Figure 1. We test $H_0 : \beta = 0$ setting the regression parameters $(\beta, \gamma) = (0, 1)$. The correlation between covariates is equal to 0.5 while a total of 100 000 simulations have been run. We compare the Flipscores approach with the three standard parametric competitors, namely the Wald, Score and Likelihood Ratio (LRT) tests. The Flipscores test is based on 2 000 random flips (we refer to Hemerik and Goeman [2018] for a discussion about the use of a limited number of random flips). The x axis represents the number of hypotheses tested, while the y axis represents the ratio between the Empirical type I error and the nominal level α. A Bonferroni correction is applied to the number of hypotheses tested.

We can observe that the Flipscores is always inside the 95% simulation confidence bands. The Score test is slightly conservative while the two other parametric methods are less satisfactory for opposite reasons. More simulations can be find in the appendix for different sample size, value of nuisance coefficient and correlation.

4.2 Multivariate

Figures 2 and 3 represent a multivariate simulation. We set a total of 1 000 dependent variables. For 20% they have $\beta = 1$ and remaining 80 % have $\beta = 0$. We set $\gamma = -1$, and the correlation between covariates equal to 0.5. Again, 2 000 random flips are used.

For controlling the FWER the proposed method is used for the Flipscores approach, while the other parametric competitors are corrected with the Bonferroni-Holm procedure.

The x axis of both figures represents the average observed correlation between the dependent variables. Figure 2 represents the empirical FWER of the true null hypothesis. The flipscores is closed to the nominal level, the LRT is anticonservative for low correlation, while the other two methods are highly conservative. Figure 3 represents the empirical power for the true alternatives. The flipscores has greater power, especially for higher correlation, while the other methods do not take advantage of the correlation structure. We observe similar results for different settings in the appendix.

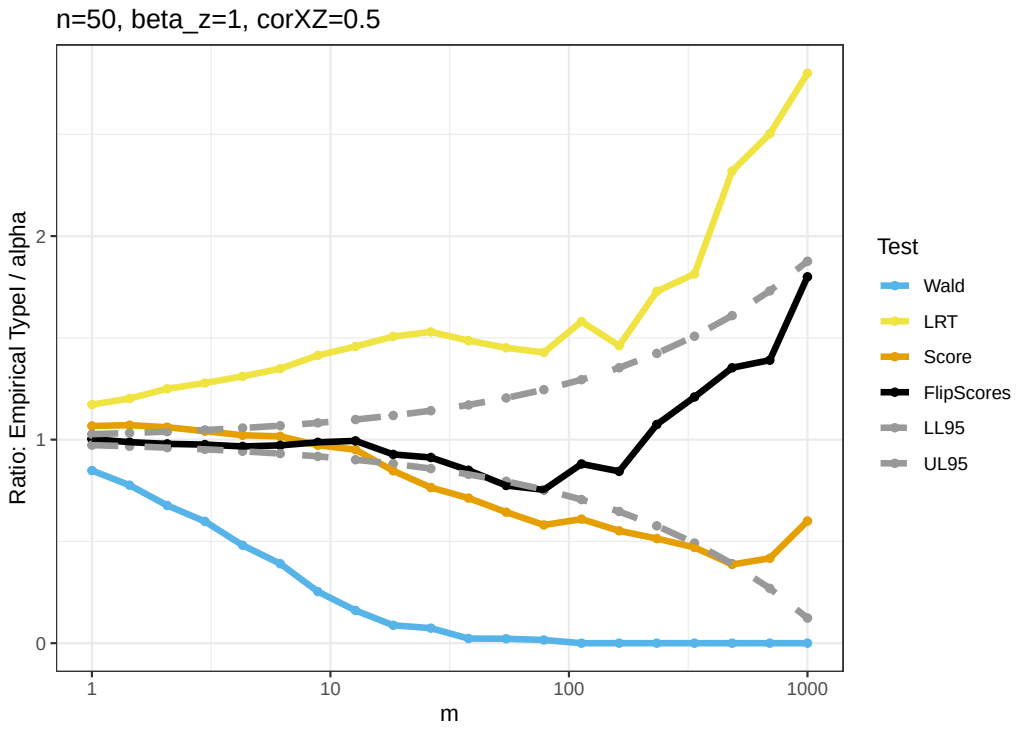

Figure 1: Logit model, univariate. On abscissa the Bonferroni-adjusted alpha for the control of FWER at level 5%, i.e. .05/m

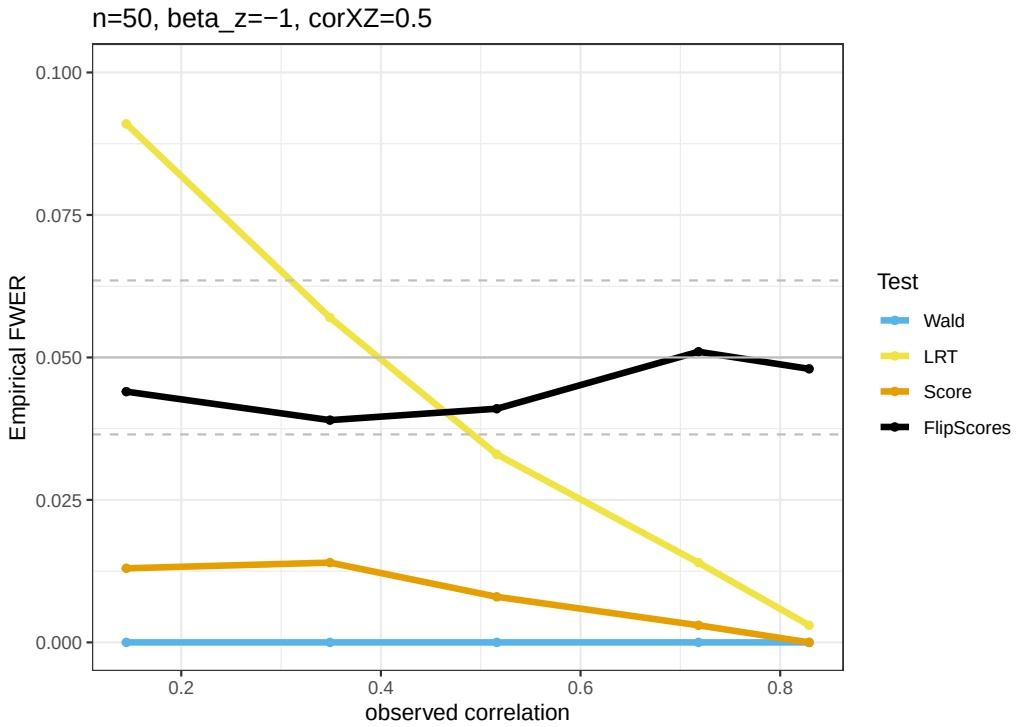

Figure 2: Logit model, multivariate

Indeed, the proposed method seems to be more satisfactory especially when high correlation is present.

5 Conclusion

This paper builds on recent state-of-the-art developments in high-dimensional inference and semi-parametric statistics. We provide the first permutation-type approach for powerful, robust multiple testing in GLMs with many responses. This represents an important step in the development of permutation methods for complex data.

In future work, we aim to zoom in on applications of our approach to neuroimaging data and RNA-Seq data. Since GLMs are so widely applicable, we expect there will be many more applications where our approach proves to be useful.

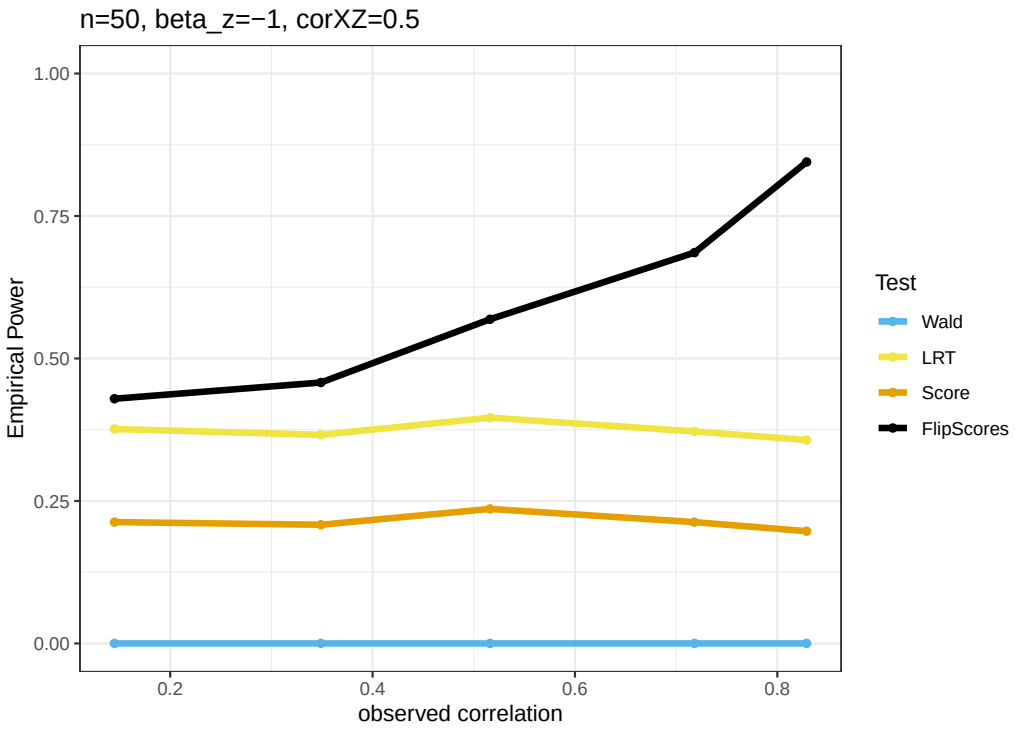

Figure 3: Logit model, multivariate

References

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

Angela Andreella, Jesse Hemerik, Livio Finos, Wouter Weeda, and Jelle Goeman. Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*, 2023.

Alexandre Blain, Bertrand Thirion, and Pierre Neuvial. Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492, 2022.

Samuel Davenport. pyperm. 2023. URL https://github.com/sjdavenport/pyperm.

Riccardo De Santis, Jelle J Goeman, Jesse Hemerik, and Livio Finos. Inference in generalized linear models with robustness to misspecified variances. *arXiv preprint arXiv:2209.13918*, 2022.

X Gao, L.C. Becker, D.M. Becker, J.D. Starmer, and M.A. Province. Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, 34(1):100–105, 2010.

Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

Jelle J Goeman and Aldo Solari. Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014.

Jelle J Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics, Ann. Statist.*, 49(2):1218–1238, 2021.

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

Jesse Hemerik, Jelle J Goeman, and Livio Finos. Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):841–864, 2020.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3): 655–660, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335748.

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

Nicolai Meinshausen, Marloes H Maathuis, Peter Bühlmann, et al. Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

A. Saffari, M.J. Silver, P. Zavattari, L. Moi, A. Columbano, E.L. Meaburn, and F. Dudbridge. Estimation of a significance threshold for epigenome-wide association studies. *Genetic Epidemiology*, 42(1):20–33, 2018.

F Schaarschmidt, C Ritz, and L.A. Hothorn. The tukey trend test: Multiplicity adjustment using multiple marginal models. *Biometrics*, 78:789–797, 2022.

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

Anna Vesely, Livio Finos, and Jelle J Goeman. Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):664–683, 2023.

Peter H Westfall and James F Troendle. Multiple testing with minimal assumptions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(5):745–755, 2008.

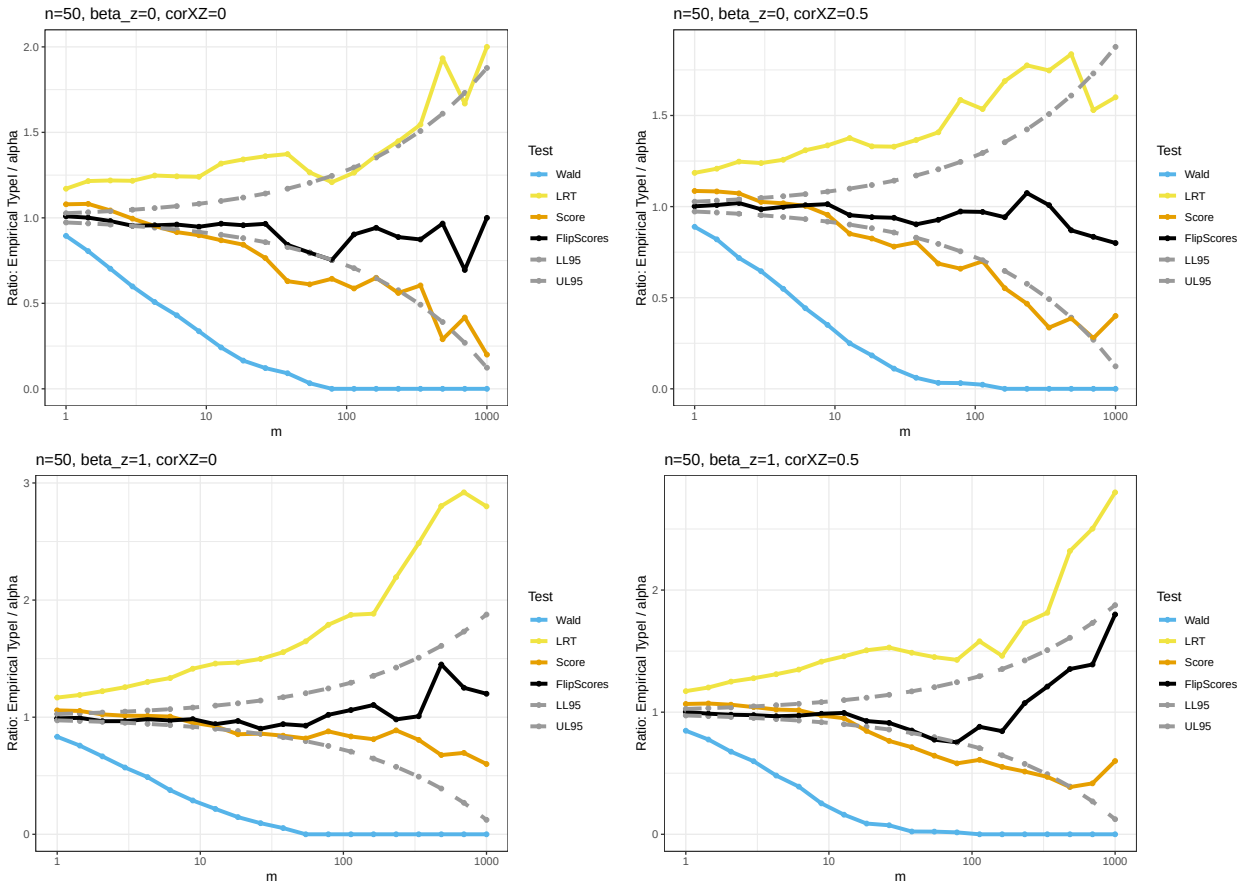

Figure 4: Logit model, univariate

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

Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

6 Appendix A

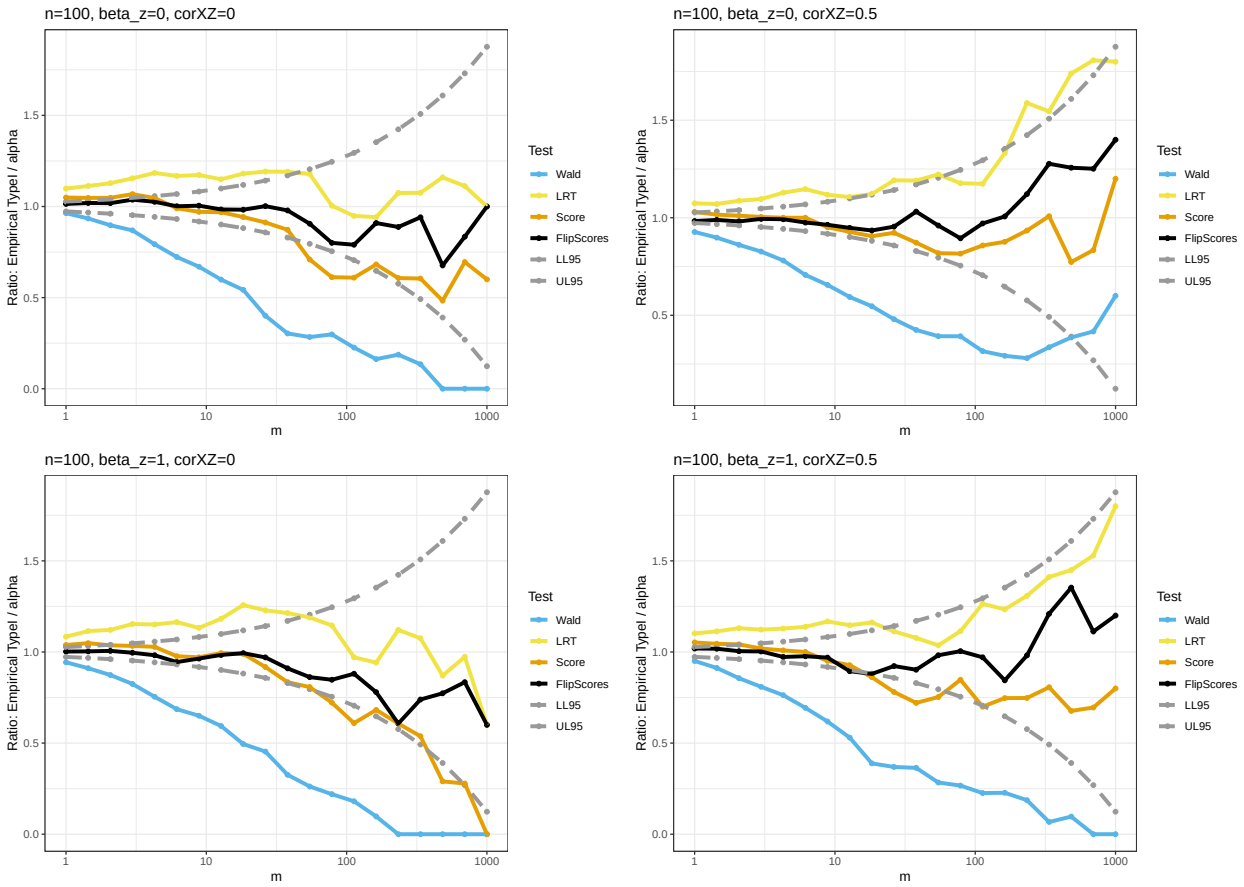

Figure 5: Logit model, univariate

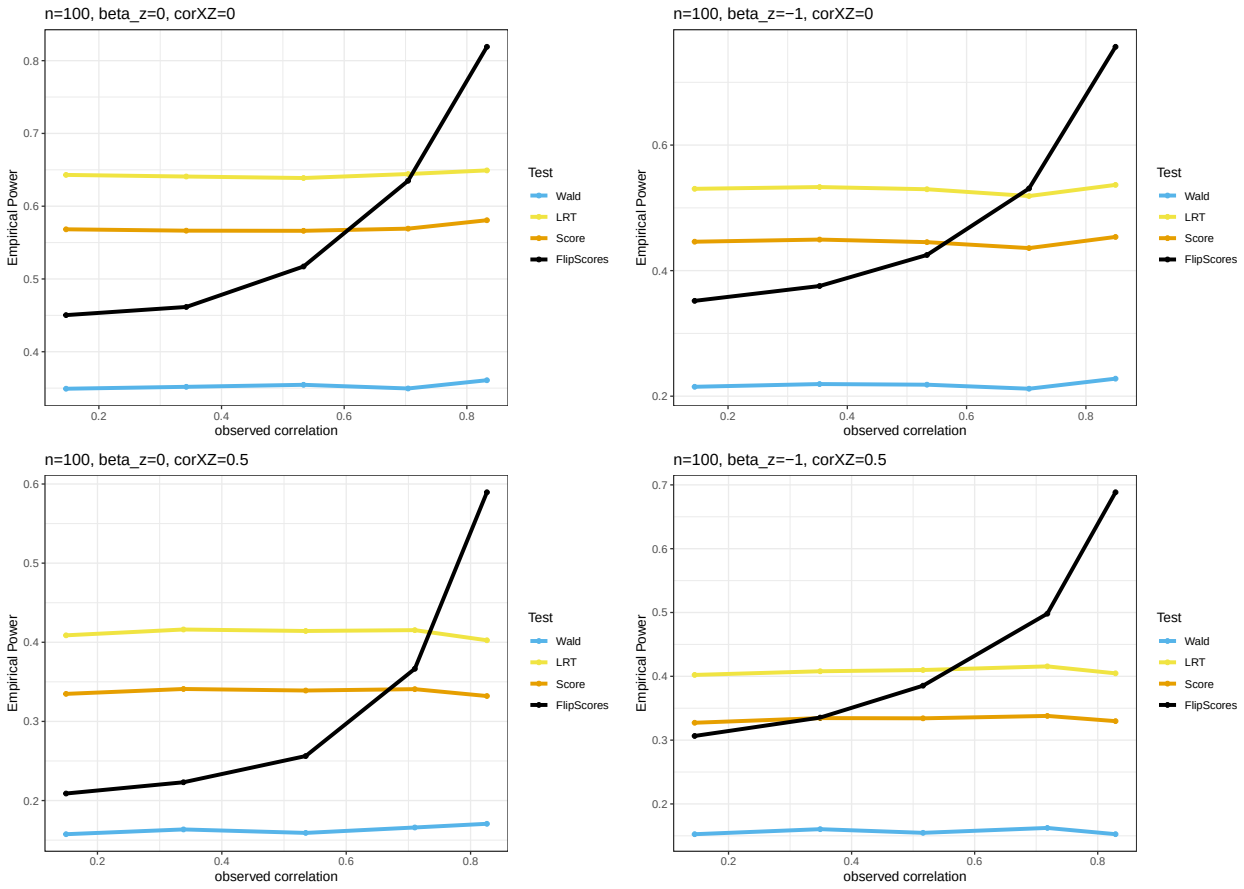

Figure 6: Logit model, multivariate