NeuroImage 201 (2019) 116030

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Robust nonparametric tests of general linear model coefficients: A comparison of permutation methods and test statistics

Nathaniel E. Helwig [a,b,*]

[a] Department of Psychology, University of Minnesota, Minneapolis, MN, 55455, USA
[b] School of Statistics, University of Minnesota, Minneapolis, MN, 55455, USA

ARTICLE INFO

Keywords:
General linear model
Neuroimaging
Permutation
Randomization
Robust statistics

ABSTRACT

Statistical inference in neuroimaging research often involves testing the significance of regression coefficients in a general linear model. In many applications, the researcher assumes a model of the form $Y = \alpha + X\beta + Z\gamma + \varepsilon$, where Y is the observed brain signal, and X and Z contain explanatory variables that are thought to be related to the brain signal. The goal is to test the null hypothesis $H_0 : \beta = 0$ with the nuisance parameters γ included in the model. Several nonparametric (permutation) methods have been proposed for this problem, and each method uses some variant of the F ratio as the test statistic. However, recent research suggests that the F ratio can produce invalid permutation tests of $H_0 : \beta = 0$ when the ε terms are heteroscedastic (i.e., have non-constant variance), which can occur for a variety of reasons. This study compares the classic F test statistic to the robust W (Wald) test statistic using eight different permutation methods. The results reveal that permutation tests using the F ratio can produce accurate results when the errors are homoscedastic, but high false positive rates when the errors are heteroscedastic. In contrast, permutation tests using the W test statistic produced valid results when the errors were homoscedastic, and asymptotically valid results when the errors were heteroscedastic. In the situation with homoscedastic errors, permutation tests using the W statistic showed slightly reduced power compared to the F statistic, but the difference disappeared as the sample size n increased. Consequently, the W test statistic is recommended for robust nonparametric hypothesis tests of regression coefficients in neuroimaging research.

1. Introduction

1.1. Statistical mapping

Despite differences in imaging modalities and experimental protocols, most neuroimaging studies share a common purpose, i.e., to understand the structure and function of the brain (see Friston, 2009). This is often accomplished by exploring associations between brain images and other explanatory variables. Depending on the application, the explanatory variables could be properties of the experiment (e.g., different task conditions), properties of the subject (e.g., the individual's age), or some combination of the two. In typical studies, the goal is to test the statistical significance of certain (primary) associations while controlling for other (nuisance) associations that are thought to influence the brain images. For example, a researcher may want to explore associations between brain images collected under two different experimental conditions X and Y, while controlling for some baseline measure(s) Z obtained during a control condition.

To test the statistical significance of associations between brain signals (the response variable) and other explanatory (or predictor) variables, neuroimaging researchers often rely on some form of the general linear model (GLM) (see Chartier and Faulkner, 2008; Christensen, 2002). Using the GLM perspective, testing the significance of the primary associations in the presence of the nuisance associations involves testing the null hypothesis that some of the regression coefficients are equal to zero. Such a test is conducted by comparing the observed test statistic to the sampling distribution of the test statistic under the null hypothesis. If the observed test statistic is extreme compared to what would be expected under the null hypothesis (e.g., probability of occurrence less than $\alpha = 0.05$), the researcher concludes that the null hypothesis is untenable.

Assuming that no primary associations exist (i.e., the null hypothesis is true), researchers would arrive at the correct conclusion about 95% of the time (using an $\alpha = 0.05$ significance level) if the sampling distribution of the test statistic is known *exactly*. However, when analyzing real brain images, the sampling distribution of the test statistic is never truly known. This is because statistical models are convenient approximations

* Corresponding author. Department of Psychology and School of Statistics, University of Minnesota, Minneapolis, MN, 55455, USA.
E-mail address: helwig@umn.edu.

https://doi.org/10.1016/j.neuroimage.2019.116030
Received 13 March 2019; Received in revised form 10 July 2019; Accepted 16 July 2019
Available online 19 July 2019

used to understand observed data—not the true models which generated the observed data (see Box and Draper, 1987). As a result, some method must be used to approximate the sampling distribution of the test statistic under the null hypothesis. Clearly, the quality of the approximation will affect the quality of the researcher's conclusions about the associations in the data. In this paper, I investigate how different aspects of the approximation (i.e., inference framework, permutation strategy, and test statistic) affect conclusions about such associations.

1.2. Parametric or nonparametric?

There are two general frameworks for obtaining (an approximation of) a test statistic's sampling distribution: parametric and nonparametric methods, see Nichols (2012) for an overview. *Parametric tests* assume that the observed data follow some particular distribution (e.g., Gaussian), which makes it possible to derive the sampling distribution of the test statistic under the null hypothesis (e.g., F distribution). If the true statistical properties of the data are reasonably approximated by the assumed statistical properties of the parametric model, then the parametric approach should produce a reasonable approximation to the test statistic's sampling distribution—and, thus, reasonable conclusions. However, if the parametric assumptions are misspecified, the parametrically derived sampling distribution may be a poor approximation for the true sampling distribution of the test statistic under the null hypothesis. In such cases, the parametric approach can produce misleading conclusions about associations in the data.

To avoid such mistakes, several researchers have turned to permutation (or randomization) tests, which are a form of nonparametric statistical inference (Blair and Karniski, 1994; Holmes et al., 1996; Nichols and Hayasaka, 2003). Note that, unlike their parametric counterparts, *nonparametric tests* rely on computationally intensive methods to approximate a test statistic's sampling distribution using minimal assumptions about the data. As a result, nonparametric tests have the potential to provide valid conclusions under a wider collection of data generating conditions. This concept has been demonstrated in several recent studies, which have highlighted the robustness of nonparametric tests in brain imaging research (e.g., see Eklund et al., 2012, 2016; Winkler et al., 2016a, b). In particular, nonparametric tests have been shown to provide valid answers in situations where parametric tests fail due to violations of assumptions.

Despite their nonparametric nature, permutation tests do require some basic assumptions for valid statistical inference. Typically, the required assumptions are stated in terms of the *exchangeability* of the error terms in the GLM (e.g., Winkler et al., 2014). Note that exchangeability implies that permuting the errors would not change their joint distribution. This implicitly requires that the error covariance matrix satisfies *compound symmetry*, i.e., (i) the correlation between any two error terms is equal, and (ii) the errors are homoscedastic, i.e., have equal variances. Note that these assumptions are satisfied by the classic "iid" (*independent and identically distributed*) errors assumption, which is frequently used for parametric inference with the GLM. But these assumptions may not be satisfied for real data, which poses a challenge for both parametric and nonparametric inference.

1.3. Need for robust tests

Violations of the exchangeability assumption can occur for a variety of reasons. Focusing first on the correlation constraints, the exchangeability of the errors is likely to be violated whenever the data have some sort of nesting structure. For example, if the subjects are nested within family units, data from individuals within and between families are likely to have a different correlation structure, which violates the exchangeability assumption. In such cases, applying the standard (unrestricted) permutation test would yield invalid conclusions, e.g., inflated false positives (Winkler et al., 2015). However, assuming that the data are exchangeable within and/or between "blocks" (i.e., groups of data points), a restricted permutation procedure can provide valid nonparametric inference when nesting structures are present in the data (see Winkler et al., 2015).

Turning to the variance constraints, the exchangeability of the errors will be violated whenever the error terms have heterogeneous variance. If the error variances differ because of a known grouping structure in the data, the group-wise heteroscedasticity can be accommodated via a modified test statistic G (Winkler et al., 2014).[1] But heteroscedasticity could arise for a variety of reasons aside from a grouping structure in the data. Note that the GLM specifies the form of the conditional distribution of the brain signal Y conditioned on the explanatory effects X and Z. The conditional variance of Y given (X, Z) will depend on the joint distribution of (Y, X, Z), so whether or not the errors satisfy the homoscedasticity assumption will depend on which explanatory variables are included in the model.

The homoscedasticity constraint poses a practical challenge for both parametric and nonparametric tests. This is because, when working with real data, it may be difficult (or impossible) to know whether the homoscedasticity assumption is appropriate for a particular choice of covariates. Unlike violations of the correlation assumption—which typically arise due to nesting structure in the data—violations of the homoscedasticity assumption can occur for obscure reasons, e.g., subtle properties of the joint distribution of (Y, X, Z). This implies that classic tests of GLM coefficients should be conducted with caution, unless one is confident that the error terms reasonably satisfy the exchangeability assumption. When heteroscedasticity may exist, a more robust hypothesis testing approach is needed.

1.4. Robust nonparametric tests

DiCiccio and Romano (2017) recently proposed a robust permutation test for regression coefficients, which can provide asymptotically valid results in the presence of heteroscedastic errors. The approach replaces the usual F test statistic with a robust W (Wald) test statistic, which is calculated using the heteroscedasticity-consistent covariance matrix estimator proposed by White (1980). DiCiccio and Romano demonstrated that their approach provides (asymptotically) valid tests of primary effects with nuisance effects present in the model. However, these methods have only been explored (in simulation studies) using two different permutation strategies (see DiCiccio and Romano, 2017; Helwig, 2019b). Note that with nuisance effects present in the model, there are eight different permutation strategies that could be considered (see Winkler et al., 2014).

In this paper, I explore which combinations of test statistics and permutation strategies should be preferred in different (realistic) data generating scenarios. Using both simulated and real data examples, I compare all 16 possible combinations (2 test statistics by 8 permutation strategies) that could be used to conduct a permutation test of primary regression coefficients with nuisance effects in the model. The simulation study compares the methods under 60 different combinations of conditions (including homo- versus heteroscedastic errors) to determine how the various test statistics and permutation strategies perform with respect to error rates and power. The real data example compares the methods to determine if the chosen test statistic and/or permutation strategy can affect the conclusions drawn about associations between real brain images.

2. Theory

2.1. Notation and terminology

Let Y_{iv} denote the i-th subject's observed data at the v-th spatial

[1] Note that the G statistic reduces to the classic F statistic when there is no known grouping structure.

location (e.g., voxel or electrode), and let $X_i = [X_{ij}]_{j=1}^{p}$ and $Z_i = [Z_{ik}]_{k=1}^{q}$ denote the i-th subject's vectors of primary and nuisance effects. At each spatial location $v = 1, \ldots, V$, consider a GLM

$$\begin{aligned} Y_v &= \alpha_v 1_n + X\beta_v + Z\gamma_v + \varepsilon_v \\ &= \alpha_v 1_n + M\psi_v + \varepsilon_v \end{aligned} \tag{1}$$

where $Y_v = [Y_{iv}]_{i=1}^{n}$ is the $n \times 1$ *response vector*, 1_n is an $n\times$ 1 vector of ones, $M = (X, Z)$ is the $n \times r$ *design matrix* ($r = p + q$) with $M_i^\top = (X_i^\top, Z_i^\top)$ denoting the i-th row, α_v is an unknown *intercept coefficient*, $\psi_v = (\beta_v^\top, \gamma_v^\top)^\top$ is an $r \times 1$ vector that contains the combined (primary and nuisance) *slope coefficients*, and $\varepsilon_v = [\varepsilon_{iv}]_{i=1}^{n}$ is the $n \times 1$ *error vector*.

The *least squares estimates* of the slope coefficients have the form

$$\widehat{\psi}_v = \left(M_c^\top M_c\right)^{-1} M_c^\top Y_v \tag{2}$$

where M_c is the columnwise mean centered design matrix, i.e., the matrix with i-th row defined as $M_i^\top - \overline{M}^\top$ where $\overline{M} = n^{-1}\sum_{i=1}^{n} M_i$ is the mean predictor vector. The least squares estimate of the intercept has the form $\widehat{\alpha}_v = \overline{Y}_v - \overline{M}^\top \widehat{\psi}_v$ where $\overline{Y}_v = n^{-1}\sum_{i=1}^{n} Y_{iv}$ is the mean response value. The *fitted values*, or model predictions, are given by $\widehat{Y}_{iv} = \widehat{\alpha}_v + M_i^\top \widehat{\psi}_v$. In matrix notation, the fitted values can be written as $\widehat{Y}_v = HY_v$ where

$$H = n^{-1} 1_n 1_n^\top + M_c \left(M_c^\top M_c\right)^{-1} M_c^\top$$

is the *hat matrix*. The model *residuals* are defined as $\widehat{\varepsilon}_{iv} = Y_{iv} - \widehat{Y}_{iv}$, which is the difference between the response and the fitted values. In matrix notation, the residuals have the form $\widehat{\varepsilon}_v = RY_v$ where $R = (I - H)$ is the *residual forming matrix*.

2.2. Hypotheses and test statistics

The *local null hypothesis* (at the v-th spatial location) is that the primary associations are equal to zero, i.e., $H_0^v : \beta_{jv} = 0$ for all $j = 1, \ldots, p$. Note that the local null hypothesis can also be written as $H_0^v : S^\top \psi_v = 0_p$, where S is a $r \times p$ selection matrix such that $\beta_v = S^\top \psi_v$ and 0_p is a $p\times$ 1 vector of zeros. The corresponding alternative hypothesis is that at least one of the primary associations is non-zero, i.e., $H_1^v : \beta_{jv} \neq 0$ for some $j = 1, \ldots, p$. The *global null hypothesis* (across all spatial locations) is defined as $H_0 : \beta_{jv} = 0$ for all $j = 1, \ldots, p$ and all $v = 1, \ldots, V$, or otherwise $H_0 : S^\top \psi_v = 0_p$ for all $v = 1, \ldots, V$. The alternative hypothesis is $H_1 : \beta_{jv} \neq 0$ for some j, v. Note that the V local hypotheses test the significance of the primary effects separately at each spatial location, whereas the global hypothesis tests the primary associations simultaneously across all spatial locations.

In parametric and nonparametric tests, evidence against the local null hypothesis H_0^v is typically quantified using the F test statistic (see Winkler et al., 2014), which has the form

$$F_v = \frac{1}{p} \widehat{\beta}_v^\top \widehat{\Sigma}_{\beta_v}^{-1} \widehat{\beta}_v \tag{3}$$

where $\widehat{\Sigma}_{\beta_v} = S^\top \widehat{\Sigma}_{\psi_v} S$, the matrix in the middle has the form $\widehat{\Sigma}_{\psi_v} = \widehat{\sigma}_v^2 (M_c^\top M_c)^{-1}$, and $\widehat{\sigma}_v^2 = \frac{1}{n-r-1}\sum_{i=1}^{n} \widehat{\varepsilon}_{iv}^2$ is the estimated error variance. Under the assumption that the errors are iid, the matrices $\widehat{\Sigma}_{\beta_v}$ and $\widehat{\Sigma}_{\psi_v}$ are the estimated covariance matrices of $\widehat{\beta}_v$ and $\widehat{\gamma}_v$, respectively. Under the additional assumption that the errors are Gaussian—as is typically assumed in parametric applications of the GLM—the F_v statistic follows an F distribution with degrees of freedom parameters p and $n - r - 1$.

To test the local null hypothesis H_0^v, DiCiccio and Romano (2017) proposed using a Wald test statistic, which has the form

$$W_v = \widehat{\beta}_v^\top \widehat{\Omega}_{\beta_v}^{-1} \widehat{\beta}_v \tag{4}$$

where $\widehat{\Omega}_{\beta_v} = S^\top \widehat{\Omega}_{\psi_v} S$, the selection matrix S satisfies $\widehat{\beta}_v = S^\top \widehat{\psi}_v$, and the other matrix is defined as $\widehat{\Omega}_{\psi_v} = \Theta^{-1} \widehat{\Theta}_v \Theta^{-1}$ with $\Theta = M_c^\top M_c$ denoting the design crossproduct matrix, and $\widehat{\Theta}_v = M_c^\top \widehat{D}_{\varepsilon_v} M_c$ denoting a weighted crossproduct matrix where $\widehat{D}_{\varepsilon_v}$ is a diagonal matrix that contains the squared residuals $(\widehat{\varepsilon}_{1v}^2, \ldots, \widehat{\varepsilon}_{nv}^2)$ on the main diagonal.[2] The matrices $\widehat{\Omega}_{\beta_v}$ and $\widehat{\Omega}_{\psi_v}$ are the asymptotic covariance matrices of $\widehat{\beta}_v$ and $\widehat{\gamma}_v$, respectively, under a broad collection of data generating assumptions (see DiCiccio and Romano, 2017; White, 1980). Specifically, assuming some basic regularity conditions,[3] the W_v statistic approximately follows a χ^2 distribution with p degrees of freedom for large n. This is assuming that the data (Y_{iv}, X_i, Z_i) are iid samples from some joint distribution, such that the conditional distribution of Y_{iv} given (X_i, Z_i) follows the model in Equation (1) with $E(\varepsilon_{iv}|X_i, Z_i) = 0$.

2.3. Permutation tests in practice

Given a chosen test statistic and permutation strategy (see next subsections), conducting a permutation test is rather simple. Let T_v denote the observed test statistic (F_v or W_v) for the v-th spatial location, and let $T_{\max} = \max_v T_v$ denote the maximum test statistic across all V spatial locations. To test the global null hypothesis $H_0 : \beta_v = 0_p$ for all $v = 1, \ldots, V$, a permutation test uses the permutation distribution of the test statistic $T_{\max}$ as a surrogate for the true (unknown) sampling distribution. The *permutation distribution* is formed by calculating the test statistic $T_{\max}$ for a large number of permutations of the data (e.g., $R = 9999$). The *permutation p-value* for testing H_0 is obtained by calculating the proportion of the $R + 1$ test statistics (R permutations plus 1 observed) that are as or more extreme than the observed test statistic $T_{\max}$. To test the local null hypothesis $H_0^v : \beta_v = 0_p$, the observed test statistic T_v can be compared to the permutation distribution of $T_{\max}$, which controls the familywise error rate across the multiple tests (see Holmes et al., 1996).

2.4. Choosing a test statistic

As a preliminary step, a researcher must choose which test statistic to use for the permutation test, i.e., F or W. The choice between the two statistics will depend on what one is willing to assume about the statistical properties of the brain images—conditioned on the particular choice of covariates. The F test statistic is appropriate for situations when the error terms satisfy the iid assumption, but may fail to produce valid results when the errors contain heteroscedasticity. This is because the F test statistic is a *pivotal quantity* when the errors are iid, but will fail to be pivotal (even asymptotically) if the error variances depend on the predictors. Note that a pivotal quantity has a sampling distribution that does not depend on the unknown parameters. Consequently, the F test statistic should be preferred for iid errors, but used with caution whenever the errors may be heteroscedastic.

The W test statistic is more appropriate for situations when then errors are independent but may be heteroscedastic. Unlike the F statistic, the W statistic can produce asymptotically valid results when the error variances depend on the predictors in the model. If the X_i's are independent of the (Y_{iv}, Z_i) vectors, then the permutation test using the W statistic will be exact (for a particular permutation scheme). When dependence exists, the permutation test using W will be asymptotically level α given that W is an asymptotically pivotal quantity under the stated assumptions. Compared to the F statistic, the drawbacks of the W statistic

[2] When the model does not contain nuisance parameters, the diagonals of $\widehat{D}_{\varepsilon_v}$ can be defined as $(Y_{iv} - \overline{Y}_v)^2$, which are the squared residuals under the null hypothesis.

[3] The result requires some basic regularity conditions, i.e., that the Φ and $\widehat{\Phi}_v$ matrices are nonsingular, and that the observed variables have finite fourth moments.

Table 1
Eight permutation methods for testing $H_0 : \beta = 0$ with nuisance parameters γ in the model. All models are assumed to have an intercept, which is excluded from the depictions for notational simplicity.

Code	Method	Permutation Scheme
DS	Draper-Stoneman (1966)	$Y = PX\beta + Z\gamma + \varepsilon$
OS	O'Gorman-Smith (2005/8)	$Y = PR_zX\beta + Z\gamma + \varepsilon$
MA	Manly (1986)	$PY = X\beta + Z\gamma + \varepsilon$
FL	Freedman-Lane (1983)	$(PR_z + H_z)Y = X\beta + Z\gamma + \varepsilon$
TB	ter Braak (1992)	$(PR_m + H_m)Y = X\beta + Z\gamma + \varepsilon$
SW	Still-White (1981)	$PR_zY = X\beta + \varepsilon$
KC	Kennedy-Cade (1996)	$PR_zY = R_zX\beta + \varepsilon$
HJ	Huh-Jhun (2001)	$PQ'R_zY = Q'R_zX\beta + \varepsilon$

Notes. P is a permutation matrix, $R_z = I - H_z$ is the residual-forming matrix with only Z in the model. $R_m = I - H_m$ is the residual-forming matrix with $M = (X, Z)$ in the model. For the Huh-Jhun method, $R_z = QQ'$ with $Q'Q = I$. *Permute X methods*: Draper and Stoneman (1966) and the reviewer in O'Gorman (2005) and Smith (see Nichols et al., 2008). The OS method could also be called the "Dekker" method, given that Dekker et al. (2007) proposed a similar permutation strategy (Winkler et al., 2016a). *Permute Y and include Z methods*: Manly (1986), Freedman and Lane (1983), and ter Braak (1992). *Permute Y and partial out Z methods*: Still and White (1981), Kennedy and Cade (1996), and Huh and Jhun (2001).

are (i) the increased computational burden, and (ii) the loss of statistical power in the situation when the iid errors assumption is reasonable. Note that point (i) is because $\widehat{\Omega}_{\beta_v}^{-1}$ is more costly to compute than $\widehat{\Sigma}_{\beta_v}^{-1}$, and point (ii) is because the common error variance σ_v^2 is not estimated with the W statistic.

2.5. Choosing a permutation strategy

As a next step, a researcher must choose which permutation strategy to use for the permutation test. With no nuisance predictors in the model, the permutation distribution is typically computed by permuting[4] the response variable (DiCiccio and Romano, 2017; Winkler et al., 2014). When nuisance parameters are present in the model, several different permutation approaches are possible: see Table 1 for a summary of eight different approaches considered by Winkler et al. (2014). While several authors have compared subsets of these methods under limited circumstances (e.g., Anderson and Legendre, 1999; Anderson and Robinson, 2001; O'Gorman, 2005; Huh and Jhun, 2001), there has been only one thorough comparison of all eight methods. Based on a series of simulation studies, Winkler et al. (2014) recommended the FL (Freedman-Lane) and OS (O'Gorman-Smith) permutation strategies, which "produce the best results in terms of control over error rates and power" (p. 385).

It should be noted that the simulation studies of Winkler et al. (2014) explored heteroscedastic scenarios where the error variance depends on a known grouping structure, but did not explore more general scenarios where the error variance depends on a (possibly unknown) function of continuous covariates. Recent research suggests that classic permutation tests of GLM coefficients (using the F statistic) can produce high false positive rates in this more general scenario, whereas using the W statistic can produce (asymptotically) valid results (DiCiccio and Romano, 2017; Helwig, 2019b). However, these recent studies only explored a limited number of data generating conditions (uncorrelated X and Z) and permutation strategies (DS and FL). As a result, when the errors are heteroscedastic, the optimal combination of test statistic and permutation strategy remains to be explored. In the following sections, I compare all 16 possible combinations (2 test statistics by 8 permutation strategies) for permutation tests of regression coefficients.

[4] Note that if the errors are assumed to be symmetric, the permutation distribution can be computed by permuting and/or resigning the response variable.

3. Methods

The simulation and real data analyses were implemented in R (R Core Team, 2019) using the **nptest** R package (Helwig, 2019a). The code and data needed to replicate the analyses are included with the Supplementary Online Materials that accompany this article.

3.1. Simulation study

I designed a simulation study to compare the performance of the different test statistics (see Equations (3) and (4)) and permutation strategies (see Table 1) under data generating conditions with and without heteroscedastic errors. The simulation study manipulates four factors: (i) the data generating distribution (2 levels: multivariate normal and multivariate t_ν with $\nu = 5$ degrees of freedom), (ii) the correlation between X and Z (3 levels: $\varphi_{XZ} \in \{0, 1/3, 2/3\}$), (iii) the sample size (5 levels: $n \in \{10, 25, 50, 100, 200\}$), and (iv) the true β coefficient (2 levels: $\beta \in \{0, 0.25\}$). Throughout the simulation, the nuisance effect was fixed at $\gamma = 1/2$. Note that the $\beta = 0$ condition is used to examine the type I error rates, whereas the $\beta = 0.25$ condition is included to examine each method's power. The multivariate normal (MVN) condition is used to explore the methods when the error terms are homoscedastic, whereas the multivariate t_ν (MVT) condition is included to compare the methods with heteroscedastic error terms.[5]

For each of the 60 combinations of the simulation parameters (2 distribution $\times$ 3 φ_{XZ} $\times$ 5 n $\times$ 2 β), I generated 10,000 independent samples from the corresponding model. For each generated sample, the null hypothesis $H_0 : \beta = 0$ was tested using the eight different permutation methods in Table 1 as well as a parametric approach. For each method, the null hypothesis was tested using both the F and W test statistics. For the permutation tests, the number of resamples was set at $R = 9999$, and the p-values were computed as described in Section 2.3, i.e., by comparing the observed test statistic to the permutation distribution. For the parametric tests, the F statistic was compared to an F distribution with $\nu_1 = 1$ and $\nu_2 = n - 3$ degrees of freedom, and the W statistic was compared to a χ^2 distribution with 1 degree of freedom. An $\alpha = 0.05$ significance level was used for all hypothesis tests.

3.2. Real EEG data

To demonstrate the importance of ignoring heteroscedasticity in multivariate permutation tests of neuroimaging data, I use the open-source electroencephalography (EEG) dataset from the UCI Machine Learning Repository (Dua and Graff, 2019). The data were originally collected by Henri Begleiter and colleagues (Zhang et al., 1995) and were posted to the UCI repository by Lester Ingber (Ingber, 1997, 1998). The dataset consists of event-related potentials (ERPs) collected from control and alcoholic subjects participating in a visual-stimulus study. The ERPs were recorded for 1 s after the presentation of a stimulus from the Snodgrass and Vanderwart (1980) set of stimuli. The subjects were shown a single stimulus (S1) that was directly followed by either a matching stimulus (S2m) or a non-matching stimulus (S2n). Subjects participated in multiple trials, and for each trial the data were collected using a 61-channel EEG cap at a rate of 256 Hz.

The P300 component of the ERP is often of interest in visual-stimulus ERP studies, given its relation to cognitive processing. In this example, I explore whether the P300 amplitude[6] during the S1 condition is (linearly) related to the P300 amplitude during the S2m and S2n conditions. Specifically, for each electrode, consider the model in Equation (1) where Y_{iv} is the i-th subject's P300 at electrode v during condition S2m or S2n, X_i is the i-th subject's P300 at electrode PZ during the S1 condition, and Z_i

[5] See Appendix A for details on the conditional error variance for MVN and MVT data.

[6] See Appendix B for details on the calculation of the P300 amplitude.

is the i-th subject's P300 at electrode OZ during the S1 condition. Note that the OZ activity is included as a covariate to control for the initial visual system activity originating in the occipital lobe. The goal is to test the global null hypothesis $H_0 : \beta_v = 0$ for all $v = 1, \ldots, 61$, as well as the local null hypothesis $H_0^v : \beta_v = 0$ at each electrode. As in the simulation study, the results were compared using all combinations of test statistic (see Equations (3) and (4)) and permutation strategies (see Table 1) using $\alpha = 0.05$ significance level.

4. Results

4.1. Type I error rate

Consider the 30 cells of the simulation design where $\beta = 0$, i.e., where the null hypothesis $H_0 : \beta = 0$ is true. For each combination of test statistic and inference method, the type I error rate was defined as the proportion of the replications (out of 10,000) were the p-value for testing $H_0 : \beta = 0$ was less than $\alpha = 0.05$. Due to space limitations, the type I error rates plotted in Fig. 1 show the results for a selected subset (i.e., 12 cells) of the simulation. See the Supplementary Materials for the full set of results in table format.

First focus on the results for the multivariate normal data in Fig. 1a. For both the F and W test statistics, we see that the permutation tests generally produced accurate results for all n and φ_{XZ}. The only noteworthy exceptions are (i) with $n = 10$ subjects the Kennedy-Cade (KC) method produced type I error rates slightly larger than the nominal $\alpha = 0.05$ level, and (ii) when $\varphi_{XZ} \neq 0$ the Still-White (SW) method produced inaccurate error rates that decreased as φ_{XZ} increased. The parametric test using the F test statistic produced accurate results for all n and φ_{XZ}, whereas the parametric test using the W test statistic only produced accurate results for large n (as expected).

Now consider the results for the multivariate t_5 data depicted in Fig. 1b. For all methods (permutation and parametric), using the F test statistic resulted in inflated type I error rates that increased as n increased. In contrast, using the W test statistic produced results that became more accurate as n increased. The parametric (Wald) test using the W test statistic only performed well for large n, which was expected. In contrast, most of the permutation tests using the W test statistic performed well. The noteworthy exception was the Still-White (SW) method: as before, when $\varphi_{XZ} \neq 0$ the SW method produced inaccurate error rates that decreased as φ_{XZ} increased. The Huh-Jhun (HJ) and Kennedy-Cade (KC) methods produced more accurate results than the other methods, particularly for small n. Note that several of the permutation strategies showed slightly inflated error rates for small n, which is not surprising given that permutation tests using the W statistic are only valid asymptotically when the error variances depend on the predictors.

4.2. Statistical power

Now consider the 30 cells of the simulation design where $\beta = 0.25$, i.e., where the null hypothesis $H_0 : \beta = 0$ is false. For each combination of test statistic and inference method, the power was defined as the proportion of the replications (out of 10,000) were the p-value for testing $H_0 : \beta = 0$ was less than $\alpha = 0.05$. Fig. 2 plots the power for a selected subset (i.e., 12 cells) of the simulation design. Tables containing the complete set of power results (for all 30 simulation cells) are provided in

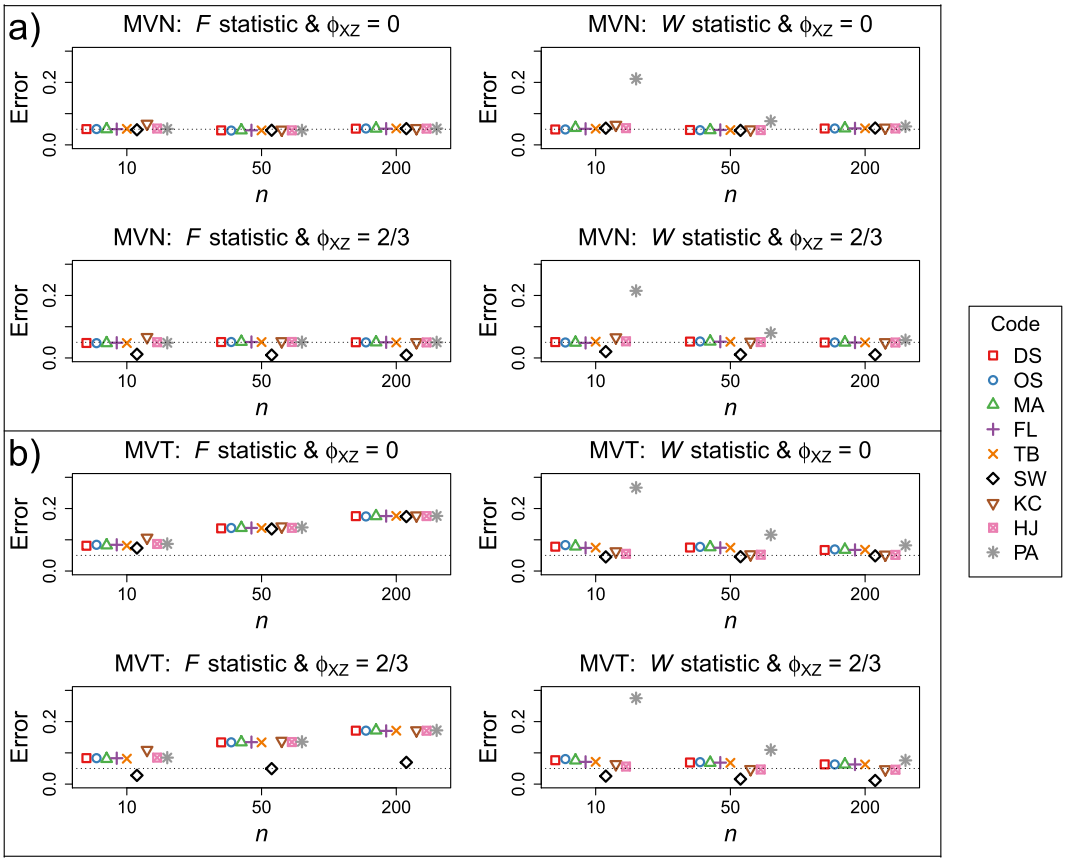

Fig. 1. Simulation Type I Error. Within each subplot, the Type I error rate is plotted for each method (PA = parametric) at three sample sizes, and the nominal rate of $\alpha = 0.05$ is denoted with a dotted line. The top subplots (labeled "a") are the results for the multivariate normal data (MVN), whereas the bottom plots (labeled "b") are the results for the multivariate t data with $\nu = 5$ degree of freedom (MVT). Within each group of subplots, the columns denote the different test statistics (F left, W right), and the rows denote two different correlations between X and Z, i.e., $\varphi_{XZ} = 0$ and $\varphi_{XZ} = 2/3$.

the Supplementary Materials. For all combinations of simulation conditions, we found that the power increased as n increased using both the F and W test statistic, which was expected.

Focusing first on the top row of Fig. 2a, it is evident that using the W test statistic (when the errors are iid) leads to a slight reduction of power ($\approx$ 2–4% less power) for small n, but the difference disappears as n increases. Comparing the first and second rows of Fig. 2a, we see that increasing the correlation between X and Z reduces the power for a given n, with the reduction being largest for the moderate sample sizes ($\approx$ 10–14% less power). Note that the reduction is more modest for large n ($\approx$ 4% less power), and we do not see this sort of reduction in the power when the correlation is $\varphi_{XZ} = 1/3$, see the tables in the Supplementary Materials.

Now turn to the power plots in Fig. 2b. Note that using the F test statistic results in more power than using the W test statistic; however, this is not the focus given that the F test statistic did not control the type I error rate for the MVT data (see Fig. 1b). The focus is on comparing the righthand portions of Fig. 2a and b. Compared to the MVN data, the power estimates for the MVT data showed noteworthy reductions that were largest for $n = 100$ ($\approx$ 19–22%) and still noticeable for $n = 200$ ($\approx$ 13–14%). For small-to-moderate sample sizes, the KC and HJ methods had slightly reduced power compared to the other permutation methods; however, it should be noted that the other methods also showed increased type I error rates at these sample sizes (see Fig. 1b).

4.3. Real EEG data

Aside from the Still and White (1981) method, all of the permutation strategies produced similar results when using the same test statistic (see the Supplementary Materials). The adjusted p-values for testing $H_0^v : \beta_v = 0$ using the Huh and Jhun (2001) method are plotted in Fig. 3. With the matching (S2m) condition as the response, the null hypothesis is only rejected for electrode PZ using both the W and F test statistic. In contrast, with the non-matching (S2n) condition as the response, the null hypothesis is rejected for 9 electrodes using the W test statistic and 17 electrodes using the F statistic. Comparing the two significant sets: the W test statistic rejects H_0^v for {CP1, CP2, CPZ, P1, P2, P3, PO1, POZ, PZ}, and the F statistic rejects H_0^v for the electrodes rejected by W as well as {AF1, AF2, CP3, FP1, FP2, FPZ, P4, PO2}. Note that using the F statistic (which assumes homoscedasticity) would result in stronger effects as well as different conclusions.

To explore whether heteroscedasticity may be the responsible for the differences in the results, I use a combination of regression diagnostic plots (Cook and Weisberg, 1983) to visualize potential violations of homoscedasticity, and the Breusch-Pagan test (Breusch and Pagan, 1979) to test the null hypothesis of homoscedasticity. The results are depicted in Fig. 4, which suggests that heteroscedasticity may be driving the differences in the results obtained from the F and W test statistics. In particular, (i) the nonparametric regression estimates (denoted by red lines in the diagnostic plots) reveal potential relationships between the location and scale, and (ii) the Breusch-Pagan test rejects the null hypothesis (of homoscedasticity) for six electrodes. Given these results, the additional relationships that were declared significant by the F test statistic should be interpreted with caution.

5. Discussion

5.1. Summary of findings

The simulation and real data results demonstrate that the chosen test statistic and permutation strategy can affect the validity of conclusions drawn from permutation tests of brain images. The results suggest that the F and W statistics perform similarly when the errors are homoscedastic, but the W statistic should be preferred when the errors are

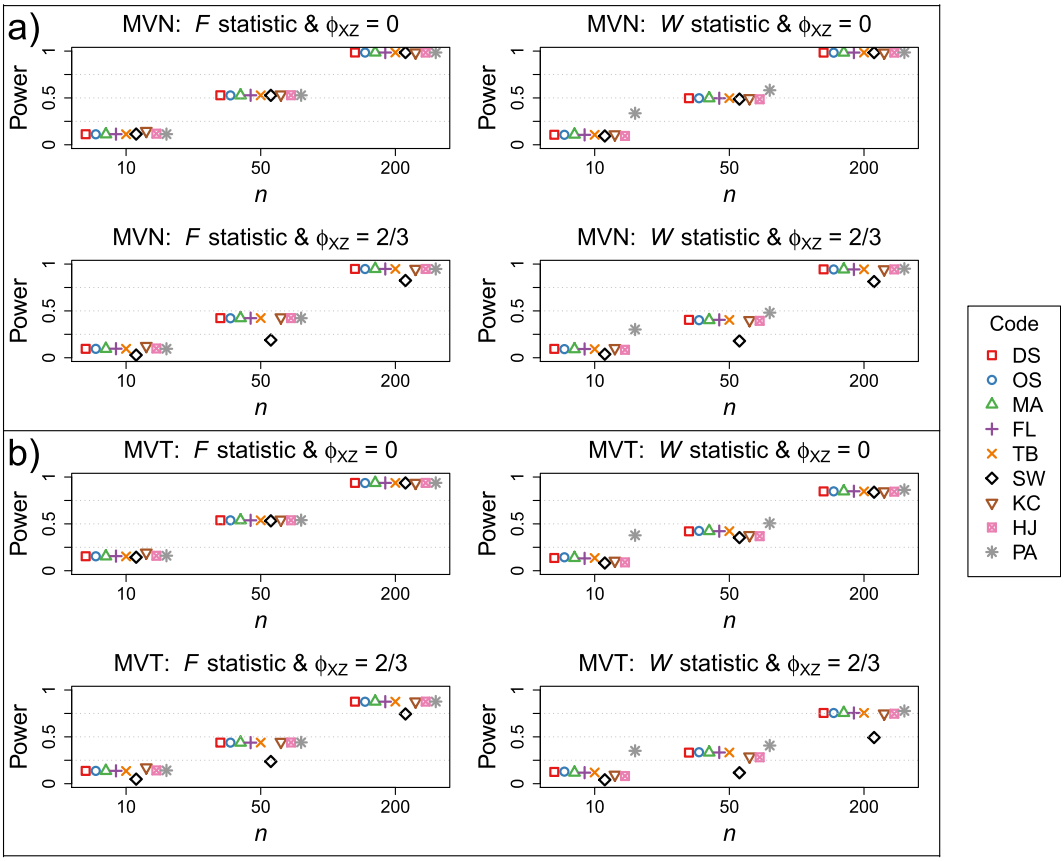

Fig. 2. Simulation Power. Same organization as Fig. 1, except the y-axis is power.

heteroscedastic. When the primary (X) and nuisance (Z) effects were uncorrelated, the different permutation methods performed similarly to one another, especially as the sample size n increased. When the primary and nuisance effects were correlated, most of the permutation methods performed similarly with one noteworthy exception: the permutation method proposed by Still and White (1981) did not provide valid results when X and Z were correlated. Overall, I found the permutation method of Huh and Jhun (2001) in combination with the W test statistic to provide the best combination of type I error control and power.

5.2. Choice of test statistic

The simulated and real data examples illustrate that the chosen test statistic has the greatest effect on the results. The simulation study reveals that permutation tests using the F and W statistics perform similarly when the errors are iid. Note that with iid errors, I found that (i) the permutation tests produced accurate type I error rates for all examined sample sizes using both the F and W statistics (see Fig. 1a), and (ii) for small to moderate samples, the power of the permutation tests with the W statistic was slightly reduced compared to using the F statistic (see Fig. 2a). However, when the errors are heteroscedastic, the simulation clearly reveals that the F statistic can lead to inflated type I error rates, which increased as the sample size n increased (see Fig. 1b). In contrast, for large sample sizes, the W statistic can produce valid results when the error variances are heteroscedastic.

These results imply that when a permutation test (with the F statistic) is applied to data with heteroscedastic errors, a false positive finding can occur with probability much higher than the desired significance level. For example, in the simulation study, a false positive finding occurred about 17.5% of the time (using an $\alpha = 0.05$ significance level) when the data were generated from a multivariate t_ν distribution with $n = 200$ subjects. Furthermore, it is worth noting that the data generating scenario where the F statistic fails is rather realistic, given that the multivariate t_ν distribution has a similar shape as the multivariate normal. So, with a slight twist of the multivariate normal scenario, the permutation testing procedure can produce misleading results when the F statistic is used.

5.3. Choice of permutation strategy

For a given test statistic, the examples reveal that the different permutation strategies produce similar results—aside from the method of Still and White (1981), which did not work well when X and Z were correlated. When the errors were homoscedastic, the permutation method of Kennedy and Cade (1996) produced slightly inflated type I error rates for $n = 10$ subjects, but the error rates became accurate as n increased. The inflation occurred for both the F and W statistic, suggesting that the KC method should be avoided for small samples. The other permutation methods (aside from SW) produced accurate type I error rates for all n when the errors were homoscedastic (see Fig. 1a).

When the errors were heteroscedastic, the different permutation strategies showed noteworthy differences in their performances (see Fig. 1b). Focusing on the results with the W statistic, many of the permutation methods produced inflated type I error rates when the errors were heteroscedastic. The exception was the permutation strategy of Huh and Jhun (2001), which produced relatively accurate type I error rates for all combinations of data generating conditions (when using the W statistic). The permutation strategy of Kennedy and Cade (1996), in combination with the W statistic, produced accurate type I error rates as long as $n \geq 50$. The other permutation strategies resulted in inflations of the type I error rate ($\approx$ 1–3%) that decreased as n increased—but still persisted for $n = 200$ subjects.

5.4. Practical implications

These findings have important implications for the use of permutation tests in brain imaging research. The F test statistic is typically used for permutation tests of regression coefficients in neuroimaging studies when the researcher has no prior knowledge about "variance groups" present in the data (see Winkler et al., 2014). Consequently, the F statistic is the default test statistic that a neuroimaging researcher would choose when testing $H_0 : \beta = 0$ given a sample of n independent subjects from the same population (note that the F and G statistics are equivalent in this case). As I have demonstrated, the F ratio may not be a good default choice for the test statistic in permutation tests of regression coefficients. This is because the error terms in the regression model may not be homoscedastic, even if the n subjects are an independent sample from some population. Whether or not the errors are homoscedastic will depend on which covariates (i.e., X and Z) are included in the model, so the homoscedasticity assumption must be considered on a case by case basis.

This study highlights the need to focus on the data generating process when conducting permutation tests of regression coefficients. Past discussions of permutation tests for neuroimaging have (correctly) noted that exchangeable errors are needed for valid permutation inference with the F statistic (e.g., see Winkler et al., 2014). However, there has been limited discussion of what data generating processes can be expected to produce conditional distributions with (i) expected values equaling the regression fitted values, and (ii) exchangeable errors (note that exchangeability implies homoscedasticity). The results in this paper reveal that these two assumptions can be attained with multivariate normal data, but the second assumption can breakdown with slight departures from multivariate normality. Consequently, the W test statistic should be the default choice for permutation tests of GLM coefficients, unless one is rather certain that the errors are homoscedastic.

5.5. Cautions and considerations

When interpreting these results and recommendations, it is important to remember that the asymptotic validity of the W test statistic requires the model to be correctly specified. In other words, the approach assumes that the conditional mean of Y_{iv} given (X_i, Z_i) follows the linear model in Equation (1) with $E(\varepsilon_{iv}|X_i, Z_i) = 0$. This implicitly presumes that the model contains the correct covariates, and that each covariate's effect is correctly included in the model (e.g., conditionally linear and measured without error). Of course, when working with real data, the linear model will never be absolutely correct, given that it is merely an approximation used to understand associations in the data. Consequently, in practice, the quality of the results will depend on how well the assumed linear model approximates the (unknown) true properties of the observed neuroimaging data.

This reveals that the choice of which covariates to include in (or exclude from) the model plays a pivotal role in the interpretation of the final result. As mentioned in the Introduction, explanatory variables in linear models could relate to properties of the experiment, characteristics of the subject, or both. However, in any given application, it can be difficult to determine which combinations of covariates should be included in the model. This is particularly true when working with large databases that contain many potential covariates, such as the Human Connectome Project (Van Essen et al., 2012), and/or when motion parameters are considered as potential covariates (e.g., see Johnstone et al., 2006). Thus, in practice, one should give careful consideration to (and, ideally, theoretical justification for) the particular choice of covariates included in the model.

6. Conclusion

Permutation tests have the potential to provide robust nonparametric inference about relationships between variables in neuroimaging studies. However, in many applications, permutation tests of regression coefficients are conducted using the F test statistic—which may not be robust to violations of the exchangeable errors assumption. For robust nonparametric hypothesis tests of regression coefficients, the W (Wald)

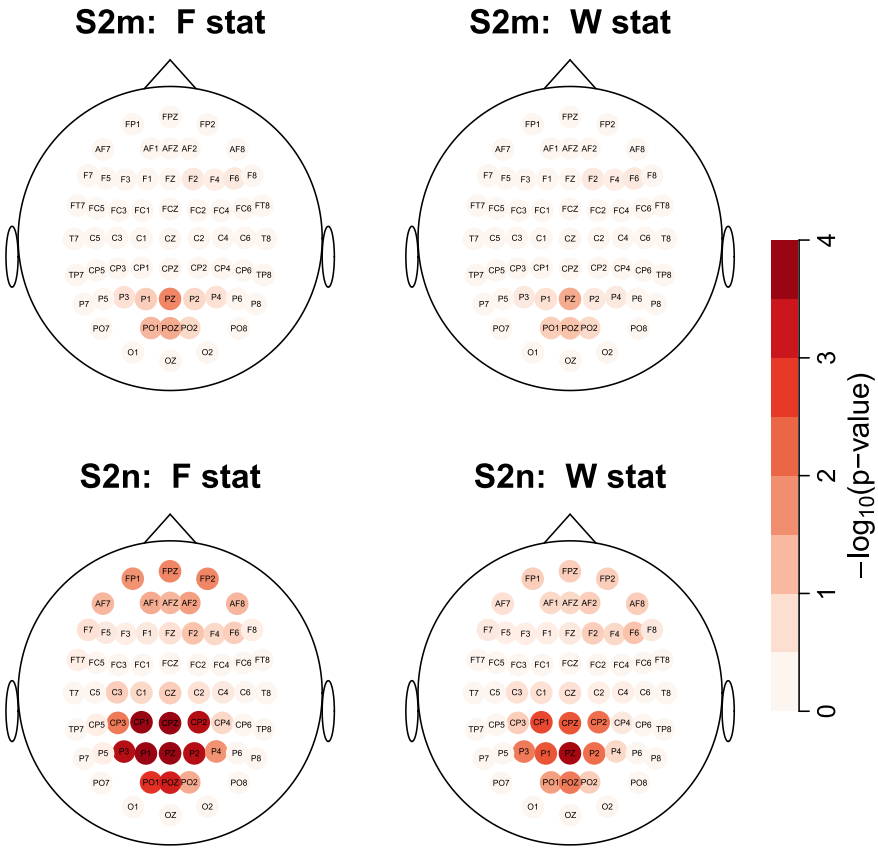

Fig. 3. Real Data Results. The (negative log 10) adjusted p-values for testing the null hypothesis at each electrode using the Huh-Jhun method. The rows show the results for the matching (top) and non-matching (bottom) conditions, whereas the columns show the results using the *F* (left) versus the *W* (right) test statistics. Plots created using the **eegkit** R package (Helwig, 2018).

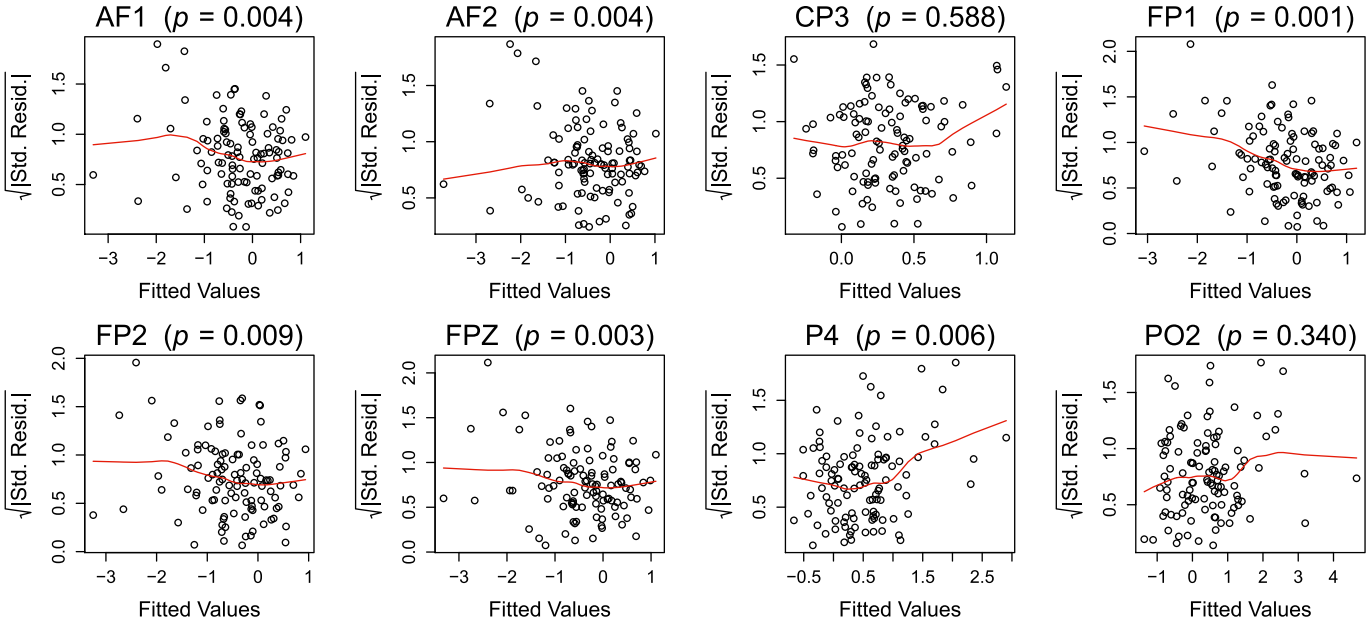

Fig. 4. Real Data Heteroscedasticity. Scale-location diagnostic plots for each of the electrodes that was rejected by the *F* statistic but not the *W* statistic. The title of each panel gives the *p*-value for the Breusch-Pagan test of the null hypothesis of homoscedasticity. Note that when the errors are homoscedastic, we should (i) observe no systematic relationship between the fitted values (*x*-axis) and the square-root of the (absolute value of the) standardized residuals (*y*-axis) in the diagnostic plots, and (ii) retain the null hypothesis that the errors are homoscedastic in the Breusch-Pagan test.

test statistic proposed by DiCiccio and Romano (2017) should be preferred. In this study, the permutation strategy of Huh and Jhun (2001) was found to provide the best control of the type I error, but all permutation methods, except that of Still and White (1981), performed reasonably well for large n. In practice, I recommend comparing permutation testing results obtained from different permutation strategies and test statistics. Note that comparing results obtained from different methods can be useful for (i) assessing the sensitivity of the solution, and (ii) identifying possible violations of assumptions.

Acknowledgements

Funded by a Single Semester Leave award from the University of Minnesota and NIH grants 1U01MH108150-01A1 and 1R01MH112583-01A1.

Appendix A

Throughout the simulation, $p = q = 1$ so that the random vector is sampled from a trivariate distribution. Define the covariance matrix of $U = (X, Z, Y)^\top$ as

$$\Phi = \begin{bmatrix} \Phi_M & \varphi_{MY} \\ \varphi_{MY}^\top & \varphi_Y \end{bmatrix}$$

where Φ_M is the covariance matrix of $M = (X, Z)^\top$, and φ_{MY} contains the covariance between the elements of M and Y. The least squares regression coefficients are defined as $\psi = \Phi_M^{-1}\varphi_{MY}$, which is the population equivalent of Equation (2). If U follows a multivariate normal distribution with mean zero and covariance matrix Φ, then the conditional distribution of Y given $M = M_i$ is univariate normal with mean $\mu_{Y|M} = M_i^\top\psi$ and homogeneous variance $\sigma^2_{Y|M} = \varphi_Y - \varphi_{MY}^\top\Phi_M^{-1}\varphi_{MY}$. In contrast, if U follows a multivariate t_ν distribution with mean zero and covariance matrix Φ, then the conditional distribution of Y given $M = M_i$ is univariate $t_{\nu+2}$ with mean $\mu_{Y|M}$ and heterogeneous variance $\xi_i\lambda\sigma^2_{Y|M}$ where $\lambda = \frac{\nu-2}{\nu}$ and $\xi_i = (\nu + \lambda^{-1}M_i^\top\Phi_M^{-1}M_i)/(\nu + 2)$ depends on the given M_i (see Ding, 2016).

Appendix B

The P300s were defined using the following approach. For each subject, the trial-level ERP data were (i) were bandpass filtered using a third-order Butterworth filter with cutoffs of 1 Hz and 50 Hz, (ii) re-referenced (from CZ) to an average reference, and (iii) averaged across the time interval 200–400 ms. Each subject's prototypical P300s during each condition (i.e., Y_i, X_i, and Z_i) was calculated by averaging the results of step (iii) across 10 trials of the experiment in each condition. Trials were excluded from the average if any electrode (a) had a maximum absolute voltage of at least 100 μV and/or (b) displayed a constant voltage across the entire trial. Of the original $N = 122$ subjects ($N_a = 77$ alcoholics and $N_c = 45$ controls), there were 11 subjects that were excluded due to insufficient data, i.e., less than 10 trials in each condition that met conditions (a) and (b). This resulted in a sample size of $n = 111$ subjects ($n_a = 72$ alcoholics and $n_c = 39$ controls).

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116030.

References

Anderson, M.J., Legendre, P., 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Stat. Comput. Simul. 62, 271–303. https://doi.org/10.1080/00949659908811936.

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43, 75–88. https://doi.org/10.1111/1467-842X.00156.

Blair, R.C., Karniski, W., 1994. Distribution-free statistical analyses of surface and volumetric maps. In: Thatcher, R.W., Hallett, M., John, E.R., Huerta, M. (Eds.), Functional Neuroimaging: Technical Foundations. Academic Press, San Diego, California, pp. 19–28.

Box, G., Draper, N., 1987. Empirical Model Building and Response Surfaces. John Wiley & Sons, New York.

Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287–1294. https://doi.org/10.2307/1911963.

Chartier, S., Faulkner, A., 2008. General Linear Models: an integrated approach to statistics. Tutor. Quant. Methods Psychol. 4, 65–78. https://doi.org/10.20982/tqmp.04.2.p065.

Christensen, R., 2002. Plane Answers to Complex Questions: the Theory of Linear Models, third ed. Springer-Verlag, New York.

Cook, D.R., Weisberg, S., 1983. Diagnostics for heteroscedasticity in regression. Biometrika 70, 1–10. https://doi.org/10.1093/biomet/70.1.1.

Dekker, D., Krackhardt, D., Snijders, T.A.B., 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72, 563–581. https://doi.org/10.1007/s11336-007-9016-1.

DiCiccio, C.J., Romano, J.P., 2017. Robust permutation tests for correlation and regression coefficients. J. Am. Stat. Assoc. 112, 1211–1220. https://doi.org/10.1080/01621459.2016.1202117.

Ding, P., 2016. On the conditional distribution of the multivariate t distribution. Am. Statistician 70, 293–295. https://doi.org/10.1080/00031305.2016.1164756.

Draper, N.R., Stoneman, D.M., 1966. Testing for the inclusion of variables in linear regression by a randomisation technique. Technometrics 8, 695–699. https://doi.org/10.2307/1266641.

Dua, D., Graff, C., 2019. UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., Knutsson, H., 2012. Does parametric fMRI analysis with SPM yield valid results? —an empirical study of 1484 rest datasets. Neuroimage 61, 565–578. https://doi.org/10.1016/j.neuroimage.2012.03.093.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 113, 7900–7905. https://doi.org/10.1073/pnas.1602413113.

Freedman, D., Lane, D., 1983. A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. 1, 292–298. https://doi.org/10.2307/1391660.

Friston, K.J., 2009. Modalities, modes, and models in functional neuroimaging. Science 326, 399–403. https://doi.org/10.1126/science.1174521.

Helwig, N.E., 2018. eegkit: toolkit for electroencephalography data. URL: http://CRAN.R-project.org/package=eegkit.Rpackageversion1.0-4.

Helwig, N.E., 2019a. nptest: Nonparametric Tests. URL: https://CRAN.R-project.org/package=nptest.Rpackageversion1.0-0.

Helwig, N.E., 2019b. Statistical nonparametric mapping: multivariate permutation tests for location, correlation, and regression problems in neuroimaging. Wire Comput. Stat. 2, e1457. https://doi.org/10.1002/wics.1457.

Holmes, A.P., Blair, R.C., Watson, J.D.G., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16, 7–22. https://doi.org/10.1097/00004647-199601000-00002.

Huh, M.H., Jhun, M., 2001. Random permutation testing in multiple linear regression. Commun. Stat. Theor. Methods 30, 2023–2032. https://doi.org/10.1081/STA-100106060.

Ingber, L., 1997. Statistical mechanics of neocortical interactions: canonical momenta indicators of electroencephalography. Phys. Rev. E 55, 4578–4593. https://doi.org/10.1103/PhysRevE.55.4578.

Ingber, L., 1998. Statistical mechanics of neocortical interactions: training and testing canonical momenta indicators of EEG. Math. Comput. Model. 27, 33–64. https://doi.org/10.1016/S0895-7177(97)00265-3.

Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox, A.S., Davidson, R.J., Oakes, T.R., 2006. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum. Brain Mapp. 27, 779–788. https://doi.org/10.1002/hbm.20219.

Kennedy, P.E., Cade, B.S., 1996. Randomization tests for multiple regression. Commun. Stat. Simulat. Comput. 25, 923–936. https://doi.org/10.1080/03610919608813350.

Manly, B., 1986. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. Res. Popul. Ecol 28, 201–218. https://doi.org/10.1007/BF02515450.

Nichols, T.E., 2012. Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62, 811–815. https://doi.org/10.1016/j.neuroimage.2012.04.014.

Nichols, T.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12, 419–446. https://doi.org/10.1191/0962280203sm341ra.

Nichols, T.E., Ridgway, G.R., Webster, M.G., Smith, S.M., 2008. GLM permutation: nonparametric inference for arbitrary general linear models. Neuroimage 41, S72.

O'Gorman, T.W., 2005. The performance of randomization tests that use permutations of independent variables. Commun. Stat. Simulat. Comput. 34, 895–908. https://doi.org/10.1080/03610910500308230.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.Rversion3.6.0.

Snodgrass, J.G., Vanderwart, M., 1980. A standardized set of 260 pictures: norms for the naming agreement, familiarity, and visual complexity. J. Exp. Psychol. Hum. Learn. Mem. 6, 174–215. https://doi.org/10.1037/0278-7393.6.2.174.

Still, A.W., White, A.P., 1981. The approximate randomization test as an alternative to the *F* test in analysis of variance. Br. J. Math. Stat. Psychol. 34, 243–252. https://doi.org/10.1111/j.2044-8317.1981.tb00634.x.

ter Braak, C.J.F., 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel, K.H., Rothe, G., Sendler, W. (Eds.), Bootstrapping and Related Techniques, Lecture Notes in Economics and Mathematical Systems, vol. 376. Springer, pp. 79–86.

Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Della Penna, S., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., Yacoub, E., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018.

White, H., 1980. A heteroscedasticity-consistent covariance matrix and a direct test for heteroscedasticity. Econometrica 48, 817–838. https://doi.org/10.2307/1912934.

Winkler, A.M., Ridgway, G.R., Douaud, G., Nichols, T.E., Smith, S.M., 2016a. Faster permutation inference in brain imaging. Neuroimage 141, 502–516. https://doi.org/10.1016/j.neuroimage.2016.05.068.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060.

Winkler, A.M., Webster, M.A., Brooks, J.C., Tracey, I., Smith, S.M., Nichols, T.E., 2016b. Non-parametric combination and related permutation tests for neuroimaging. Hum. Brain Mapp. 37, 1486–1511. https://doi.org/10.1002/hbm.23115.

Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., Smith, S.M., 2015. Multi-level block permutation. Neuroimage 123, 253–268. https://doi.org/10.1016/j.neuroimage.2015.05.092.

Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W., Litke, A., 1995. Event related potentials during object recognition tasks. Brain Res. Bull. 38, 531–538. https://doi.org/10.1016/0361-9230(95)02023-5.