Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI

Lourens Waldorp

University of Amsterdam, Roetersstraat 15, 1018 WB, the Netherlands

Abstract

The statistical analysis of fMRI time series requires accurate estimates of the variance. In practice many assumptions required for accurate variance estimation are violated. For instance, it is generally acknowledged that the model used to account for temporal autocorrelations or the model for the hemodynamic response are approximations. As a consequence tests on general linear model coefficients are not valid. Robust estimation of the variance of the general linear model (GLM) coefficients in fMRI time series is therefore essential. In this paper an alternative method to estimate the variance of the GLM coefficients accurately is suggested and compared to other methods. The alternative, referred to as the sandwich, is based primarily on the fact that the time series are obtained from multiple exchangeable stimulus presentations. The analytic results show that the sandwich is unbiased. Using this result, it is possible to obtain an exact statistic which keeps the 5% false positive rate. Extensive Monte Carlo simulations show that the sandwich is robust against misspecification of the autocorrelations and of the hemodynamic response model. The sandwich is seen to be in many circumstances robust, computationally effecient, and flexible with respect to correlation structures across the brain. In contrast, the smoothing approach can be robust to a certain extent but only with specific knowledge of the circumstances for the smoothing parameter.

Key words: robust estimation, false positive rate, neuroimaging statistics, standard errors, sandwich

1 Introduction

Brain activity maps from functional magnetic resonance imaging (fMRI) time series are becoming increasingly important in the cognitive sciences [1]. An

Email address: waldorp@uva.nl (Lourens Waldorp).

fMRI brain activity map contains thousands of volume elements (voxels) that make up the entire brain. For each of these voxels a blood-oxygenation level dependent (BOLD) time series is available. In order to increase the signal to noise ratio, exchangeable stimuli are repeated several times in experiments [2]. Since there are many voxels, analyses are often performed voxel-wise to decrease computational load (mass univariate approach). In the general linear model (GLM), the time series of each voxel is represented by a linear combination of modeled time series corresponding to a condition or effect [3]. Amplitude coefficients and their variances are then computed such that hypothesis testing can be performed on (a function of) these coefficients to, for example, test between conditions. This paper is about estimating the variance of the amplitude coefficients as accurately as possible such that hypothesis testing is valid.

Hypothesis tests on functions of parameters are greatly influenced by the estimate of the variance of the model parameters, which in turn is greatly influenced by the autocorrelations of the time series [1,4,5]. Generally, two approaches to estimating the variance of the coefficients can be distinguished: (i) transforming the data such that the time series becomes uncorrelated or "white", and (ii) transforming the data such that the data are smoothed or "colored", and then using the known, smooth structure for variance estimation [6,7]. In prewhitening, on the one hand, a model for the autocorrelations of the time series is used which should render the data uncorrelated [8]. Often an autoregressive (AR) process is used [9], but many other strategies exist [10–13]. The advantage of prewhitening is that the obtained variance estimate is smallest compared to all other unbiased estimates [14]. However, this advantage holds only if the model for the correlation structure is correct [7], which is, of course, difficult to maintain. It has been suggested that accounting for bias due to autocorrelations is not required because the estimates did not improve enough [7]. However, Marchini and Smith [7] did not consider an incorrect correlation structure, only bias due to limited length of the time series. Precoloring, on the other hand, has the advantage that the assumed correlation structure need not be correct [4]. A disadvantage is that a smoothing parameter of, for example, a Gaussian kernel needs to be chosen (see e.g., [15]). Such a decision can influence the quality of the variance estimate [7,13]. Another disadvantage of the smoothing approach is that high frequency components in the data can be attenuated [11].

In addition to misspecification of the autocorrelations, the model for the hemodynamic response is also likely to be incorrect [16]. This means that the residuals contain misspecification which is carried into the estimator of the variance of the coefficients. It is therefore important to take such misspecification into account in any statistical analysis of fMRI time series.

I agree with Friston et al. [4] that robust variance takes priority over efficient variance, regardless of whether the model for the correlations is correct or not. However, optimally a robust variance estimate should also be able to adapt to local variations of correlation structure. Variation of correlation

structure exists across different locations of the brain [9]. A variance estimate like the smoothing approach that works well on average of brain locations can therefore be improved. I suggest a robust variance estimate based on the residuals but taking into account the individual replications or events. This variance estimate adapts to correlational changes, is computationally efficient, and is robust. I show that this robust variance estimate is unbiased and as a result can be used for hypothesis testing even with few replications.

The paper is organized as follows. Section 2 introduces the differences between the true underlying process and the GLM, the working model. This section also discusses existing methods of estimating the variance of the coefficients and introduces the new, robust variance estimate. Subsequently, hypothesis testing is discussed for the different estimators. In Section 3 extensive Monte Carlo simulations are discussed to show how the different estimators perform in different circumstances for blocked and event-related designs.

2 Model specification and misspecification

In model specification a data generating process (DGP) is assumed to exist. This DGP is in general unknown and is therefore approximated by a working model. Such an approximation can be misspecified in at least two ways: (i) the model for the mean can be incorrect, and (ii) the model for the autocorrelations noise can be incorrect. An example of a misspecified model for the mean is using a gamma function as a model for the hemodynamic response when the BOLD response is in fact generated by the balloon model, see e.g., [16]. An example of misspecification of the autocorrelations is using an autoregressive model for temporal correlations, when the correlations are actually 1/frequency [1]. First, statistical assumptions of the DGP are described followed by misspecification of the GLM for fMRI data as a working model.

Data of $i = 1, \ldots, p$ time points or scans are available measured on $j = 1, \ldots, n$ independent trials or replications. The data are collected in the p-vector Y_j. The DGP for Y_j is $Y_j = g_\theta(Z) + e_j$, where $g_\theta(Z) < \infty$ is an unkown (non)linear, nonrandom function with fixed regressors $Z = (z_1, \ldots, z_m)$ and unknown parameters θ. The noise e_j has joint distribution function $F(e)$ with mean zero and unknown variance $E\{e_j e_k'\} = \Sigma$ for $j = k$ and zero otherwise. So, there is autocorrelation, but no correlations among replications.

The working model specifies an approximation to the DGP for the mean and the variance of the data. In the GLM a linear function $X\beta$ is used as an approximaion to the mean $E\{Y_j\} = g_\theta(Z)$, where X is a $p \times k$ matrix and β a k-vector of coefficients. The noise is assumed to have temporal correlations but remains unspecified for the moment. Then the working model on replication j is $Y_j = X\beta + r_j$, where the residual $r_j = g_\theta(Z) - X\beta + e_j$ contains both the modeling error $g_\theta(Z) - X\beta$ and noise e_j. The variance of the residual r_j is again Σ since the modeling error is fixed (but see below for the estimated residual). The model $X\beta$ could correspond to the DGP, that is $g_\theta(Z) = X\beta$,

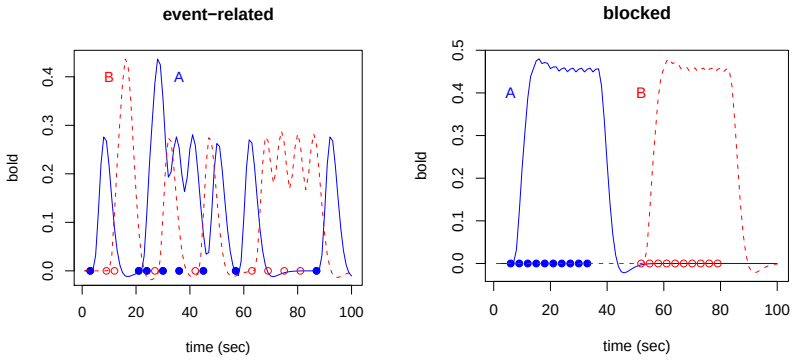

Fig. 1. Convolution of the HRF and the stimulus function for an event-related (left) and a blocked design (right). Stimulus presentation latencies for condition A (solid blue) are indicated with filled circles, and open circles for condition B (dashed red). Parameters are: $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$, $c = 0.35$ [18].

but in general they are different. It is assumed that the matrix X has full column rank, $r(X) = k$, such that $X'X$ is nonsingular.

The main parameters of interest in fMRI are the amplitude parameters β of the BOLD response time series. To model the delayed response, a hemodynamic response function (HRF) is used, convolved with the stimulus presentation timing of the experiment. A possible HRF used in analyses is a double gamma function [17,18]. The stimulus ("on-off") function is given by $s(t) = 1$ for all time points t that the stimulus is present and zero otherwise. An example of the convolution of the time series is given in Figure 1. The experiment can either be event-related or blocked [1,19]. In an event-related design each presentation in a sequence can belong to any of the conditions, whereas in a blocked design a sequence of presentations for a particular condition is given in blocks (see e.g. [1,20]). An example of each is given in Figure 1. The convolutions form the columns of the design matrix X. The design matrix X can also include temporal derivatives to account for latencies in the BOLD signal [21,22].

When the coefficients are estimated, a function of the estimate $\hat{\beta}$ is usually tested, which is called a contrast. The variance of a contrast $c'\hat{\beta}$ is then $c'\text{var}\{\hat{\beta}\}c$. A possible test of the contrast is the F-test

$$F = k_n \frac{(c'\hat{\beta} - a)^2}{c'\text{var}\{\hat{\beta}\}c}, \tag{1}$$

where k_n is a factor to obtain the correct null distribution for the hypothesis $c'\hat{\beta} = a$ [18]. This statistic is approximately F distributed with degrees of freedom dependent on the estimate of the contrast variance. It is clear from the definition that the statistic, and therefore the false positive rate, is directly influenced by the contrast variance. This paper is about finding

a robust estimate of this contrast variance such that inference concerning β through hypothesis testing is valid.

2.1 Estimation

A general way of estimating the coefficients and their variance is explained, after which the four different methods of defining an estimator are discussed. This follows mostly the presentations of [7,12]. The four methods are also summarized in Table 1.

Let S be a nonsingular $p \times p$ matrix and premultiply the data, model, and residual with S such that $SY_j = SX\beta + Sr_j$. Then the variance of the residual r_j is $S\Sigma S'$. The least squares estimate is $\hat{\beta} = (X'S'SX)^{-1}X'S'S\bar{Y}$, where $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$. Because the model is misspecified, $\hat{\beta}$ is biased, that is

$$E\{\hat{\beta}\} = (X'S'SX)^{-1}X'S'Sg_\theta(Z) = \beta^* \tag{2}$$

The mean β^* can be described as a least squares approximation to the unknown function $g_\theta(Z)$, which is very different from linearization of $g_\theta(Z)$ in terms of a first order Taylor expansion. The main difference between the least squares and Taylor approximation, is that the first describes the nonlinear function on the whole range of Z, whereas the latter is accurate only in a neighborhood of a specific Z (see [23] for more details on this). The variance of $\hat{\beta}$ is

$$\text{var}\{\hat{\beta}\} = \frac{1}{n}(X'S'SX)^{-1}X'S'S\Sigma S'SX(X'S'SX)^{-1}. \tag{3}$$

An estimate of the residual is given by

$$\hat{r}_j = (I_p - H_{SX})Sg_\theta(Z) - H_{SX}S\bar{e} + Se_j, \tag{4}$$

where $H_{SX} = SX(X'S'SX)^{-1}X'S'$ and $\bar{e} = \frac{1}{n}\sum_{j=1}^{n} e_j$. The mean and variance of the estimated residual are

$$E\{\hat{r}_j\} = Q_{SX}Sg_\theta(Z) \qquad \text{var}\{\hat{r}_j\} = \frac{1}{n}Q_{SX}S\Sigma S'Q_{SX} + \frac{n-1}{n}S\Sigma S', \tag{5}$$

where $Q_{SX} = I_p - H_{SX}$. These results are different from other derivations in three ways (see e.g., [6,7]): (i) the estimator $\hat{\beta}$ is biased because the incorrect model is used for the mean, (ii) the expectation of the estimated residual is not zero because $\hat{\beta}$ is biased, and (iii) the variance of the estimated residual $\hat{r}_j$ contains two terms, one with the design matrix X and one without X, because the number of replications is taken into account. Especially this last point will be used to our advantage, as described below.

The easiest least squares estimate is ordinary least squares (OLS). This is obtained by assuming that the noise variance is $\Sigma = \sigma^2 I_p$ and setting $S = I_p$. Then the variance of the OLS estimate $\hat{\beta}_O$ is obtained by estimating the scalar

noise variance σ^2, which is estimated by the sum of the squared residuals [1]. The OLS estimator of the variance of $\hat{\beta}_O$ is then $\hat{V}_O = \hat{\sigma}_O^2 (X'X)^{-1}$. This estimator is biased because the estimator $\hat{\beta}_O$ is biased. From (4) it is easily seen that the bias term in the nominator is $g_\theta(Z)' Q_X g_\theta(Z)$. It is well known that if there are autocorrelations then OLS will lead to variance estimates that are too small (see also simulation section below), see e.g., [4,24,25].

Another estimator is obtained by assuming that there are autocorrelations and these are estimated. Then set S such that the estimate of the noise variance is $\hat{\Sigma} = SS'$ [8]. This is known as (feasible) generalized least squares (GLS), also sometimes called prewhitening. The variance of the GLS coefficient $\hat{\beta}_G$ is often written as a product of a scalar variance and a correlation matrix, $\Sigma = \sigma^2 R$. Then the estimate of σ^2 using $\hat{\beta}_G$ in the residuals is obtained similarly to OLS and is referred to as $\hat{\sigma}_G^2$. The correlation matrix R can be estimated by any number of suggested algorithms. Often an AR(p) process is assumed for R with $p = 1, 2$ [9,18], or sometimes higher [26]. Other GLS methods include transforming the time series to the frequency domain [10–12], and transforming the time series to the wavelet domain, retaining the correlation structure to obtain an estimator for R [13]. The variance of the coefficient $\hat{\beta}_G$ estimated by GLS is $\hat{V}_G = \hat{\sigma}_G^2 (X' \hat{R}^{-1} X)^{-1}$. It is known that if the model for the variance is correct, then GLS is most efficient, i.e., the estimator attains the Cramér-Rao lower bound of the variance of all unbiased estimates [14]. The problem is that it is very difficult to find an unbiased estimate of R, even for large time series (large p), not in the least because the model used for the temporal correlations is incorrect [4,27,28]. If no correct model is known, then GLS could lead to very inaccurate variance estimates for the coefficients β. Friston et al. [4] show clearly that assuming an incorrect model for the noise correlations can lead to variance estimates that are too high or too low (see also the section Monte Carlo simulations).

The third estimator is obtained by assuming that $\Sigma = \sigma^2 R$, with R a correlation matrix, and setting S such that $SRS' \approx S\hat{R}S'$ [29]. So, the temporal correlations in the time series are dominated by a smoothing matrix S such that the true temporal correlations become irrelevant to estimating the variance of the coefficient $\hat{\beta}_S$. This is sometimes called the smoothing approach or precoloring. Then σ^2 is estimated by $\hat{\sigma}_S^2$, which is the average squared residuals divided by the degrees of freedom [29]. The estimator $\hat{\sigma}_S^2$ is biased if $\hat{\beta}_S$ is biased. The correlation matrix R needs to be estimated, which can be done in the same manner as described above for GLS, e.g. with an AR(p) model [18]. The variance estimator for the coefficient $\hat{\beta}_S$ using a smoothing matrix S is equation (3) with $\hat{\Sigma} = \hat{\sigma}_S^2 \hat{R}$, which is referred to as $\hat{V}_S$. The smoothing matrix is often generated by the Gaussian function $\exp[-(i-j)^2/2\tau^2]$, where i is the row and j the column of SS' and τ^2 is the variance [30]. Suggested values for τ^2 are 4 to 8 s^2. An advantage of $\hat{V}_S$ is that it is robust against using an incorrect model for R, which is likely to be the case. However, it is in general difficult to set S such that $SRS' \approx S\hat{R}S'$ for each correlation structure [7]. Friston et al. [4] suggest a bandpass filter for S which minimizes the the squared differ-

ence for a contrast between the true and estimated variance over all possible (autoregressive) correlations in the time series. This will result on average in a reasonable estimate for all voxels with different correlation strengths which is computationally efficient. Optimally, however, one would like to use the same estimator for each voxel that somehow adapts to the particular correlation strengths of that voxel. Such a robust estimator is described next.

The fourth and final estimator considered here is obtained by assuming there are autocorrelations and setting $S = I_p$. The basic idea is to use the fact that the variance of the estimated residual in equation (5) has two components, one is orthogonal to the design matrix and the other contains only the true variance if $S = I_p$. The estimate $\hat{\beta}_O$ is used and from (4) it can be deduced that with $S = I_p$ for the estimated residuals $\hat{r}_j$

$$E\{\hat{r}_j\hat{r}_j'\} = Q_X\left(g_\theta(Z)g_\theta(Z)' + \frac{1}{n}\Sigma\right)Q_X + \frac{n-1}{n}\Sigma. \tag{6}$$

Hence, any estimator of the GLM coefficients using these squared residuals and containing X will make the bias part vanish and leave only the true variance Σ in the expectation. So, we now use these estimated residuals from the OLS estimate $\hat{\beta}_O$ to estimate the correlation structure of the noise based on n replications

$$W = \frac{1}{n-1}\sum_{j=1}^{n}\hat{r}_j\hat{r}_j' = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - X\hat{\beta}_O)(Y_j - X\hat{\beta}_O)', \tag{7}$$

The variance estimate of $\hat{\beta}_O$ using $\hat{\Sigma} = W$ in (3) with $S = I_p$ is referred to as $\hat{V}_W$. Because of the expectation of the squared residuals, the estimate $\hat{V}_W$ is unbiased, that is

$$E\{\hat{V}_W\} = \frac{1}{n}(X'X)^{-1}X'E\{W\}X(X'X)^{-1} = \frac{1}{n}(X'X)^{-1}X'\Sigma X(X'X)^{-1}. \tag{8}$$

It works because of the two-part variance in (5) where the second part contains only the true variance. And there are two parts in the variance because we took into account the number of replications obtained in the experiment. This estimator is in other contexts sometimes referred to as the sandwich estimator [31]. In general the sandwich can be shown to be consistent, i.e. the estimator will be correct for large n (note the difference of the asymptotics with GLS) [23]. In this particular case where the design matrix is fixed, the sandwich estimator is even unbiased, which is usually not the case. As a consequence, the sandwich is accurate for few number of replications n. The fact that the sandwich is unbiased without any specification of smoothing or a model for the noise correlation structure is especially appealing. Another advantage is that because the residuals are used, the sandwich estimator adapts itself according to the correlation structure of each voxel. So, it is flexible, computationally efficient, and robust. These facts of the sandwich can be used to create an exact test, shown in the next section.

Table 1
The four methods of estimation and their corresponding variance.

type	mean	variance
OLS	$\hat{\beta}_O = (X'X)^{-1}X'\bar{Y}$	$\hat{V}_O = \hat{\sigma}^2_O(X'X)^{-1}$
GLS	$\hat{\beta}_G = (X'\hat{R}X)^{-1}X'\hat{R}\bar{Y}$	$\hat{V}_G = \hat{\sigma}^2_G(X'\hat{R}X)^{-1}$
S	$\hat{\beta}_S = (X'S'SX)^{-1}X'S'S\bar{Y}$	$\hat{V}_S = \hat{\sigma}^2_S(X'S'SX)^{-1}X'S'S\hat{R}S'SX(X'S'SX)^{-1}$
W	$\hat{\beta}_O = (X'X)^{-1}X'\bar{Y}$	$\hat{V}_W = \frac{1}{n}(X'X)^{-1}X'WX(X'X)^{-1}$

2.2 Hypothesis testing

Contrasts are used to create a function of the coefficient that will allow to test for differences between conditions. For example, a single contrast could be $c' = (1, -1)$, to test between the amplitudes of different conditions. An F-test can be used to test the null hypothesis $H_0 : c'\hat{\beta} = a$ against the alternative $H_A : c'\hat{\beta} \neq a$. Depending on which estimator for β and which variance estimate is used, a specific F-test will result. For the simple contrast like $c' = (1, -1)$ and $a = 0$ the F-test is the square of the t-test. In general, for a set of q independent contrasts, collected in the $q \times k$ matrix C, the F-test is [32]

$$F = \frac{n-q}{nq}(C\hat{\beta} - a)'(C\hat{V}C')^{-1}(C\hat{\beta} - a), \tag{9}$$

which under H_0 is distributed approximately as F with degrees of freedom dependent on the statistic for the variance $\hat{V}$ (see Table 1). If OLS or GLS is used, then the statistics F_O and F_G are approximately $F(q, p-k)$ distributed. If the smoothing approach is used then usually the so-called Satterthwaite approximation f_S to the degrees of freedom is used, which depends on both the autocorrelation and the design [29,7]. So, for the smoothing approach, the statistic F_S is approximately $F(q, f_S)$ distributed. Finally, if the sandwich estimator is used, an exact test F_W exists which is $F(q, n-q)$ distributed, provided the data are multivariate normal, that is if $F(e) = N_p(0, \Sigma)$ (see appendix for details on this). The degrees of freedom do not contain the length of the time series (p) because the correlation structure of the time series is entirely estimated from the information of the replications. The fact that it is an exact test means that even for very small number of replications n the F statistic is very accurate, i.e. has a false positive rate of 5%, say. The assumption of multivariate normal noise in fMRI is important, of course, and has been investigated. It appears that the assumption of Gaussian noise is valid in general for low and high signal to noise ratios and is very accurate when considering difference images, as is often the case in fMRI analyses [33].

3 Monte Carlo simulations

In this section Monte Carlo simulations are used to show in which circumstances each of the four variance estimates works best. This is done by considering four variables: (i) the autocorrelation of the time series, (ii) misspecification of the correlation structure, (iii) misspecification of the mean model, and (iv) the type of design. The focus these simulations is on model misspecification instead of specific models for the HRF and autocorrelations. In so doing the results of these simulations apply to many different situations with different models but similar misspecification.

3.1 Data generation

A time series is created of fMRI data of length $p = 100$ seconds. The data generating process is linear in the parameters, $g_\theta(Z) = Z\theta$. The columns of the design matrix $Z = (z_1, z_2)$ are generated according to the double gamma function and represent time series corresponding to two different experimental conditions A and B of either an event-related or a blocked design [3]. The event-related design was generated using random stimulus presentations with 8 presentations per condition in the 100 second interval with the constraint that the interstimulus interval was at least 2 seconds. In the blocked design there was one block for each of the two conditions with 10 stimulus presentations in each block. The exact designs used are shown in Figure 1. The parameter θ represents the amplitude of the BOLD response corresponding to a condition. Noise e_j is added to the signal $Z\theta$ which is $N_p(0, \Sigma)$ for $j = 1, \ldots, n$ with $\Sigma = \sigma^2 R$. The correlation matrix $R = (\rho)_{ij}$ is induced by either an AR(1) or AR(2) process, which are repectively $U(t) = \phi U(t-1) + \varepsilon(t)$ and $U(t) = \gamma_1 U(t-1) + \gamma_2 U(t-2) + \varepsilon(t)$, where $\varepsilon(t)$ is white noise [34]. The coefficients of the AR(2) process have been sampled from the upper right quadrant of the stationary area: $0 < \gamma_1 + \gamma_2 < 1$ [34]. A single parameter is created to indicate strength of dependence in the time series $\phi = \gamma_1 + \gamma_2$, which is varied from 0.2 to 0.9, with γ_1 at most 0.1 larger than γ_2. This also reflects the possible differences in correlation structure as found between voxels. The variance of the time series at $t = 0$ is taken as $\sigma_0^2 = 1$. Then the data are $Y_j = Z\theta + e_j$ for $j = 1, \ldots, n$. The variance of the noise is set such that the signal to noise ratio (SNR) for the time series is approximately one for the average over replications. This is achieved by multiplying the variance of the noise by the number of replications. As a consequence the number of replications is irrelevant, only the SNR is important which is set at an appropriate low level (see [35]).

3.2 Estimation

Estimation with the working model $Y_j = X\beta + r_j$ is performed using a different HRF, $h(t)^*$, which is a single gamma function [1]. The resulting time series

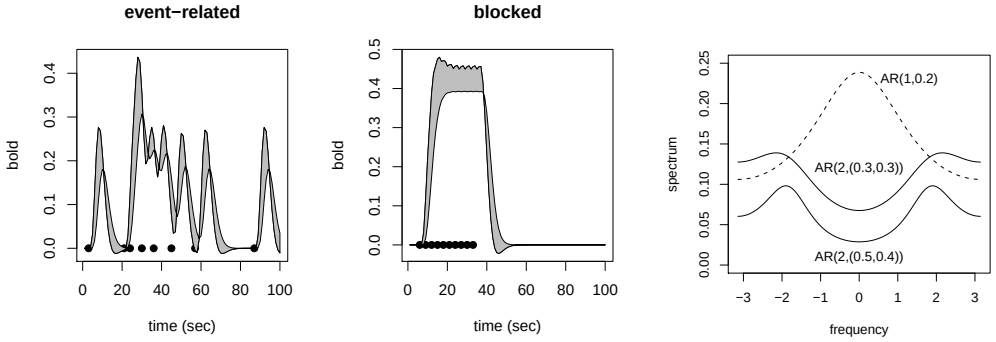

Fig. 2. Left and middle: Misspecification of the HRF for condition A with the largest relative difference of 0.278 for the event-related design and 0.149 for the blocked design. Right: Three spectra of AR processes are displayed as a function of frequency for $[-\pi, \pi]$ [36]. The AR(1) process was generated with parameter $\phi = 0.2$, and the two AR(2) process are generated with $\gamma_1 = \gamma_2 = 0.3$ and $\gamma_1 = 0.5$ and $\gamma_2 = 0.4$.

form the columns of X in the working model, such that $Z \neq X$, and as a result $\theta \neq \beta$. The main difference between the functions is that there is no undershoot using the single gamma function. Additionally, a parameter is varied in the single gamma function to vary the degree of misspecification. At the largest misspecification the induces a reduction of amplitude to about 30% and a delay of about 2 seconds, shown in Figure 2. To quantify the difference between the DGP and working model, the relative difference between the functions is computed, defined as the sum of the absolute difference between the functions divided by their sum over the whole range. This relative difference was for the event-related design between 0.072 and 0.278, and for the blocked design between 0.075 and 0.149. The lowest relative difference is solely due to selecting the incorrect single gamma function. The largest effect of misspecification is in the event-related design. This is to be expected since the shape of the HRF is more important in event-related designs [1].

The misspecification in the correlation structure for GLS and the smoothing approach is created by using as a working model an AR(1) instead of an AR(2). The amount of misspecification depends on the correlation strength of the generated structure with AR(2), see Figure 2. It is clear that estimating the correlation structure using an AR(1) process will capture mostly frequencies around zero, whereas it will represent poorly frequencies further away from zero.

The smoothing approach requires setting the smoothing matrix S by the parameter τ^2. The value of this parameter depends on both the correlation strength and the design. Therefore, we first looked at the effect on the variance estimate for different values of correlation strength ϕ and τ^2. As can be seen in Figure 3, there is no absolute correct value of τ^2 for both event-related and blocked designs and all correlation strengths when only the correlation structure is misspecified. The value of $\tau^2 = 8$ seems to be most optimal in the sense

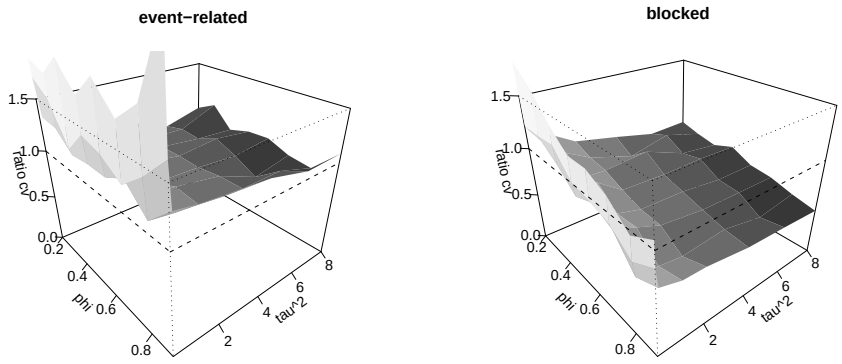

Fig. 3. Ratios of estimated and true contrast variance for event-related and blocked designs as a function of correlation strength ϕ and smoothing parameter τ^2 for the smoothing approach.

that it is robust against correlation strength, especially in the event-related design. This value is used in the simulations for the smoothing approach unless specified otherwise.

To compare the four approaches three measures are discussed: the ratio of estimated to true contrast variances, the false positive rate, and power. The contrast tested is $c' = (1, -1)$. The true contrast variance is obtained by computing the variance from the $N = 500$ simulations of the estimate $\hat{\beta}$ for each of the methods. Note that the true variance is defined differently from that defined in [4], where a second order approximation to the mean squared error was used. The bias formulation ignored stochasticity of the estimated correlation matrix $\hat{R}$ which was approximated to the second order. Let D denote the true variance obtained from the N simulations. The ratio of contrast variance is then $c'\hat{V}c/c'Dc$. If the estimated variance is good then the ratio will be 1, it is overstimated if the ratio is larger than 1, and it is underestimated if the ratio is smaller than 1.

The false positive rate or size of a test is the probability of a test to reject the null hypothesis when it is true. The false positive rate (FPR) is set at 5%. It is expected that when the contrast variance is underestimated then the FPR will be too high, that is, higher than 5%; and when the contrast variance is overestimated, the FPR will be too low. In relation to FPR, power is compared between methods as a function of effect size. Power refers to the probability of rejecting the null hypothesis when it is incorrect. Power should be close to 1 given a sufficient effect size. Effect size η is here defined as the difference between amplitude parameters divided by the true contrast variance. If the FPR is too low then the power will also be low, and when the FPR is too high, the power will be high.

3.3 Results

We first look at the contrast variance when the assumptions about the correlation structure and HRF are correct. Then we determine the effect of misspecification of the autocorrelations on the contrast variance, FPR, and power. And finally we look at possible interactions of misspecification of the autocorrelations and the HRF.

When both the HRF and autocorrelations are correctly specified all methods should work well, except OLS when there are autocorrelations. In Figure 4

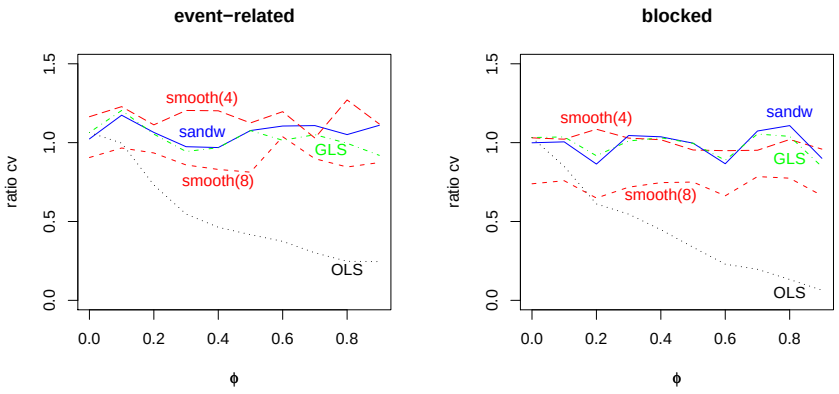

Fig. 4. Ratios of estimated and true contrast variance when the correlation structure is correctly specified as an AR(1) process as a function of the AR(1) parameter ϕ. The methods displayed are: OLS (black, dotted line), GLS (green, dashed-dotted line), smooth with $\tau^2 = 8$ (red, dashed line), smooth with $\tau^2 = 4$ (red, long-dashed line), and sandwich (blue, solid line).

it is clearly seen that for the event-related and blocked design both the sandwich and GLS perform equally well for any value of ϕ. As expected, OLS is close to one only when $\phi = 0$. In the event-related design the contrast variance of the smoothing approach with $\tau^2 = 8$ is quite close to one, but the contrast variance for this τ^2 is underestimated in the blocked design. In the blocked design the contrast variance is very accurate for all values of ϕ when $\tau^2 = 4$. So, when the model for the noise variance is correct the sandwich is almost exactly the same as the minimum variance GLS regardless of design. The smoothing approach, on the other hand, depends strongly on the design, different smoothing parameters are required for accurate contrast variance estimates.

If there is misspecification in the correlation structure, then the contrast variance of a robust estimator should still be accurate for all levels of correlation strength. It is clear from Figure 5 that now OLS and GLS perform poorly. OLS always underestimates the true contrast variance and GLS either underestimates or overestimates contrast variance. Both the smoothing approach and the sandwich are robust for misspecification of the correlation structure

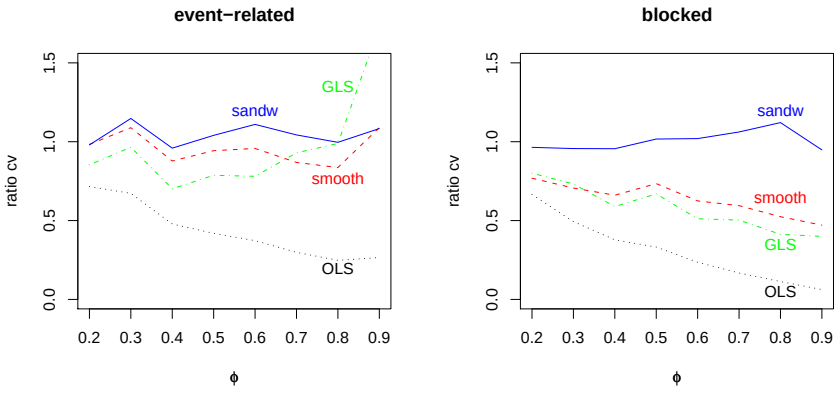

Fig. 5. Ratios of estimated and true contrast variance when the correlation structure is misspecified for the four methods for both the event-related and blocked design as a function of correlation strength ϕ.

in the event-related design. However, in the blocked design only the sandwich is robust at all levels of correlation strength. As a consequence the smoothing approach has a slightly higher FPR than the nominal 5% in the event-related design but a dramatically higher FPR in the blocked design, shown in Figure 6. This was expected because from Figure 5 the contrast variance was underestimated and so the FPR is expected to be too high. In contrast, the

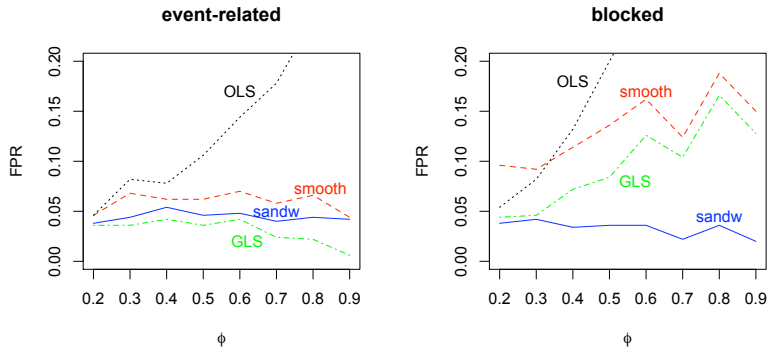

Fig. 6. False positive rate as a function of correlation strength ϕ for the event-related and blocked design when the correlation structure is incorrect.

sandwich has FPR slightly below the nominal 5% in both designs because it overestimated the contrast variance slightly. In accordance with the contrast variance and FPR results, the power of the smoothing approach is slightly higher than that of the sandwich, as can be seen in Figure 7. The power for the blocked design is comparable.

In addition to misspecification of the correlation structure the HRF model can be misspecified. To look at possible interactions with correlation strength, we varied both HRF misspecification and correlation strength. As can be seen

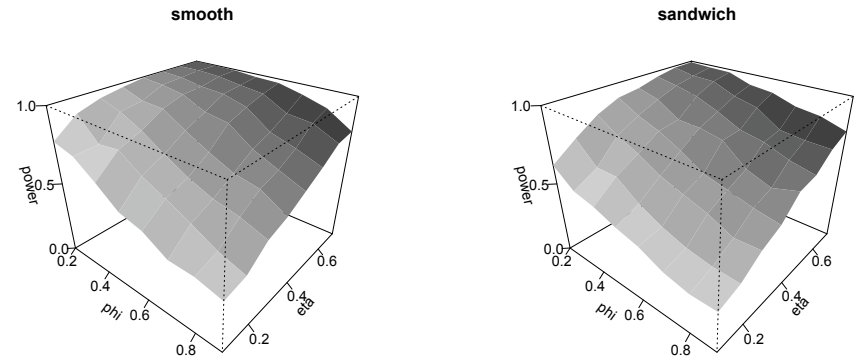

Fig. 7. Power for the event-related design as a function of correlation strength ϕ and effect size η.

in Figure 8, for the event-related design the sandwich is more accurate than the smoothing approach, which underestimates the contrast variance. But there is only a small effect of HRF misspecification for both the sandwich and smoothing approach. For the blocked design, on the other hand, the smoothing approach underestimates contrast variance greatly. Accordingly, the FPR of

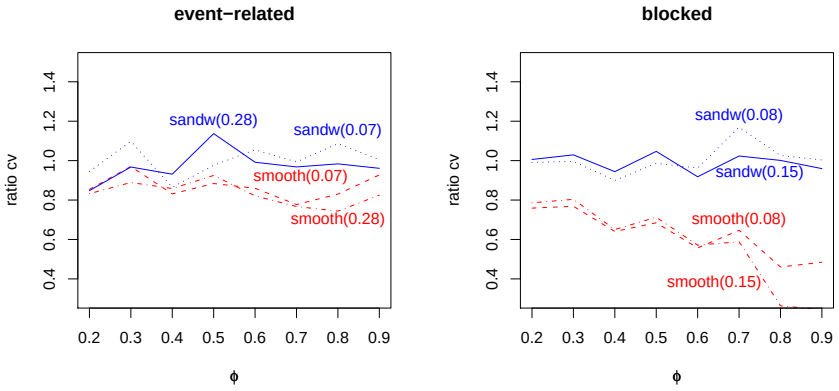

Fig. 8. Ratios of estimated and true contrast variance for the event-related and blocked design when both the correlation structure and HRF model are incorrect. Two cuts of both the sandwich (blue) and smoothing approach (red) variance estimates are shown, at $\delta = 0.07$ and 0.28 for event-related, and at $\delta = 0.08$ and 0.15 for blocked design.

the smoothing approach in the event-related design is too low, around 2.5%. This is due to overcompensation of the degrees of freedom f_S in the smoothing approach. When there are no autocorrelations f_S is high and when there are autocorrelations f_S is low. When the HRF is modeled incorrectly, f_S is too low so that the FPR is too low. In the blocked design the FPR behaves as

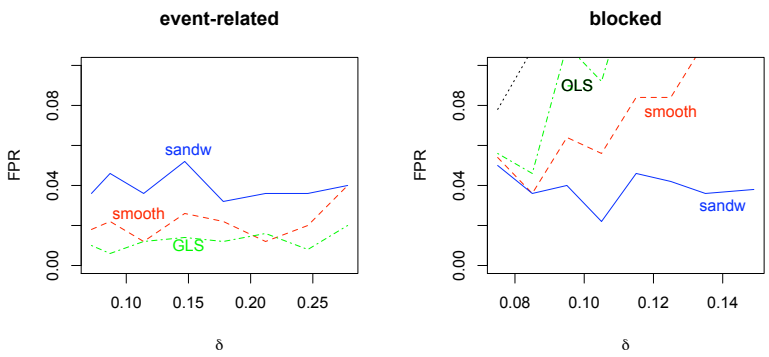

Fig. 9. False positive rate as a function of relative difference δ for the event-related and blocked design when both the correlation structure and HRF model are incorrect. The correlation strength was $\phi = 0.9$.

expected for the smoothing approach: the contrast variance is underestimated which leads to overestimated FPR. The sandwich remains in both designs relatively stable around 5%. The power behaves as expected in this case (not shown): for the smoothing approach the power is similar to that in Figure 7 for the event-related design and higher for the blocked design. The power of the sandwich is similar to that of Figure 7.

4 Discussion

It has been repeatedly shown that the false positive rate in fMRI brain activity maps can be quite high if the assumptions of the method are violated (see e.g., [4,7]). Therefore, the robustness of the variance estimator of the GLM coefficients is an important issue. It has been shown here that the sandwich is unbiased and accordingly an exact F-test with the sandwich exists. Additionally, misspecifications in both autcorrelation and HRF model is accommodated by the sandwich for both event-related and blocked designs. In contrast, the smoothing approach is affected by both autocorrelation and HRF misspecification. Additionally, the smoothing approach requires a smoothing parameter which must be specified for each correlation structure to get accurate results. In contrast, the sandwich variance has two main advantages to the smoothing approach: (i) the sandwich adapts to local changes in correlation structure, whereas the smoothing approach does not, and (ii) no model or parameter needs to be determined for accurate results with the sandwich.

The potential of the application of the sandwich to real data is large. For example, we have applied the sandwich to real fMRI data in Weeda et al. [37]. In that paper we took a multivariate approach to model the GLM coefficients using Gaussian shaped functions. Using an incorrect shape function and incorrect autocorrelation assumptions, we showed that the contrast variance is still accurate of the sandwich. Using the sandwich we were able to find a plausible

set of areas of brain activity in an auditory task.

Another area where the sandwich can be used is random effects analysis [38], which is our current work. The first level of a two-level random effects model requires only an OLS estimate of the coefficient of each subject, and its sandwich. Then at the second level, the group effects are estimated with OLS again, and another sandwich is formed which is simply the sandwich of the first level variance with the group design for all subjects.

References

[1] G. Sarty, Brain activity maps from fMRI time series data, Oxford University Press, 2006.

[2] A. Dale, R. Buckner, Selective Averaging of Rapidly Presented Individual Trials Using fMRI, Human Brain Mapping 5 (1997) 329–340.

[3] K. Friston, C. Frith, R. Turner, R. Frackowiak, Characterizing evoked hemodynamics with fMRI, NeuroImage 2 (1995) 157–165.

[4] K. Friston, O. Josephs, E. Zarahn, A. Holmes, S. Rouquette, J.-B. Poline, To smooth or not to smooth?, NeuroImage 12 (2000) 196–208.

[5] L. Waldorp, H. Huizenga, R. Grasman, The Wald test and Cramer-Rao bound for misspecified models in electromagnetic source analysis, IEEE Transactions on Signal Processing 53 (9) (2005) 3427–3435.

[6] K. Friston, A. Holmes, J.-B. Poline, P. Grasby, S. Williams, R. Frackowiak, R. Turner, Analysis of fMRI time-series revisited, NueroImage 2 (1995) 45–53.

[7] J. Marchini, S. Smith, On bias in the estimation of autocorrelations for fMRI voxel time series analysis, NeuroImage 18 (2003) 83–90.

[8] K. J. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, A. Evans, A general statistical analysis for fMRI data, NeuroImage 15 (2002) 1–15.

[9] J. Locascio, P. Jennings, C. Moore, S. Corkin, Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging, Human Brain Mapping 5 (3) (1997) 168–193.

[10] E. Zarahn, G. Aguirre, M. D'Esposito, Empirical of BOLD fMRI statistics: I spatially unsmoothed data collected under the null-hypothesis conditions, NeuroImage 5 (1997) 179–197.

[11] J. Marchini, B. Ripley, A new statistical approach to detecting significant activation in functional MRI, NeuroImage 12 (2000) 366–380.

[12] M. Woolrich, B. Ripley, M. Brady, S. Smith, Temporal autocorrelation in univariate linear modeling of FMRI data, NeuroImage 14 (2001) 1370–1386.

[13] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. Carpenter, M. Brammer, Colored noise and computational inference in neurophysiological time series analysis: Resampling methods in time and wavelet domains, Human Brain Mapping 12 (2001) 61–78.

[14] T. Ferguson, A course in large sample theory, Chapman and Hall, Bury st Edmunds, 1996.

[15] R. Henson, Analysis of fMRI Timeseries: Linear Time-Invariant Models, Event-related fMRI and Optimal Experimental Design, in: R. S. Frackowiak, J. T. Ashburner, W. D. Penny, S. Zeki, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price (Eds.), Human Brain Function, 2nd Edition, Academic Press, 2004, Ch. 10.

[16] K. Friston, A. Mechelli, R. Turner, C. Price, Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics, NeuroImage 12 (2000) 455–477.

[17] G. Glover, Deconvolution of Impulse Response in Event-Related BOLD fMRI, NeuroImage 9 (1999) 416–429.

[18] K. Worsley, Statistical analysis of activation images, in: P. Jezzard, P. Matthews, S. Smith (Eds.), Functional MRI: An introduction to methods, Oxford University Press, 2001, Ch. 14, pp. 251–270.

[19] K. Friston, P. Fletcher, O. Josephs, A. Holmes, M. Rugg, R. Turner, Event-related fMRI: Characterizing differential responses, NeuroImage 7 (1998) 30–40.

[20] S. Huettel, A. Song, G. Mccarthy, Functional Magnetic Resonance Imaging, Sinauer Associates, New York, 2004.

[21] C. Liao, K. Worsley, J.-B. Poline, J. Aston, G. Duncan, A. Evans, Estimating the delay of the response in fMRI data, NeuroImage 16 (2002) 593–606.

[22] R. Henson, C. Price, M. Rugg, R. Turner, K. Friston, Detecting latency differences in event-related BOLD respondes: application to words versus nonwords and initial versus repeated face presentations, NeuroImage 15 (2002) 83–97.

[23] H. White, Using least squares to approximate unknown regression functions, International Economic Review 21 (1) (1980) 149–170.

[24] J. MacKinnon, H. White, Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, Journal of Econometrics (1985) 305–325.

[25] H. Huizenga, P. Molenaar, Estimating and testing the sources of evoked potentials in the brain, Multivariate Behavioral Research 29 (1994) 237–267.

[26] T. Gautama, M. Hulle, Estimating the global order of the fMRI noise model, NeuroImage 26 (2005) 1211–1217.

[27] Y.-G. Wang, X. Lin, Effects of variance-function misspecification in analysis of longitudinal data, Biometrics 61 (2005) 413–421.

[28] M. Crowder, On the use of the working correlation matrix in using generalised linear models for rempeated measures, Biometrika 82 (2) (1995) 407–410.

[29] K. Worsley, K. Friston, Analysis of fMRI time-series revisited - again, NeuroImage 2 (1995) 173–181.

[30] R. S. Frackowiak, J. T. Ashburner, W. D. Penny, S. Zeki, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, Human Brain Function, Academic Press, 2004.

[31] K.-Y. Liang, S. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1) (1986) 13–22.

[32] G. Seber, C. Wild, Nonlinear regression, Toronto: John Wiley and Sons, 1989.

[33] A. Wink, J. Roerdink, BOLD noise assumptions in fMRI, International Journal of Biomedical Imaging (2006) 1–11.

[34] T. Amemiya, Advanced econometrics, Oxford: Basil Blackwell, 1985.

[35] S. Smith, T. Nichols, Threshold-Free Cluster Enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference, NeuroImage 44 (1) (2009) 83–98.

[36] M. Priestly, Spectral analysis and time series, Vol. 1, Academic Press, New York, 1981.

[37] W. Weeda, L. Waldorp, I. Christoffels, H. Huizenga, Activated Region Fitting: a robust high power method for fMRI analysis using parameterized regions of activation, Human Brain Mapping (2009) (in press).

[38] C. Beckmann, M. Jenkinson, S. Smith, General multilevel linear modeling for group analysis in FMRI, NeuroImage 20 (2003) 1052–1063.

[39] M. Bilodeau, D. Brenner, Theory of multivariate statistics, New York: Springer-Verlag, 1999.

Appendix

To prove the distributional result of the statistic F_W we assume three things: (i) the DGP as stated in section 2, (ii) the working model of section 2, and (iii) the noise is multivariate normal, i.e. $F(e) = N_p(0, \Sigma)$. Then, to prove that F_W is central F distributed with degrees of freedom q and $n - q$, we need to show: (i) the variance $C\hat{V}_W C'$ is Wishart distributed, (ii) $C\hat{\beta}$ and $\hat{V}_W$ are independent, and (iii) the degrees of freedom are q and $n - 1$ (see e.g., [39] chap. 7 and 8). (i) By proposition 7.4 of [39] we have that if $(n-1)\hat{V}_W \sim W_k(n-1, V)$ then $(n-1)C\hat{V}_W C' \sim W_q(n-1, CVC')$, where $V = \text{var}\{\hat{\beta}_O\}$. So, if $\hat{V}_W$ is Wishart distributed we are done. Rewrite $\hat{V}_W$, such that if $U_j = (X'X)^{-1}X'r_j$, then $n(n-1)\hat{V}_W = \sum_{j=1}^{n} U_j U_j'$. Now U_j is $N_k(0, (n-1)V)$. This is seen by noting that $E\{U_j\} = (X'X)^{-1}X'Q_X g_\theta(Z) = 0$ and $\text{var}\{U_j\} = \frac{n-1}{n}(X'X)^{-1}X'\Sigma X(X'X)^{-1}$, because of the variance of the residuals. Then by definition $(n-1)\hat{V}_W \sim W_k(n-1, V)$. For (ii), to show independence of $C\hat{\beta}_O$ and $\hat{V}_W$, it is sufficient to show independence of $\hat{\beta}_O$ and U_j. Because the data are normal by assumption, the covariance of $\hat{\beta}_O$ and U_j needs to be zero to show independence. Since the covariance of $(\bar{Y}', r_j')'$ is $\frac{1}{n}Q_X\Sigma$, it then follows that $C\hat{\beta}_O$ and $\hat{V}_W$ are independent. To show (iii), that the degrees of freedom are q for the numerator and $n-1$ for the denominator, proposition 8.2 of [39] is used. It implies that if $C\hat{\beta}_O - a \sim N_q(0, CVC')$ and $(n-1)C\hat{V}_W C' \sim W_q(n-1, CVC')$, then $F_W \sim F(q, n-q)$. The first part is true under H_0 and from the variance of the OLS estimate $\hat{\beta}_O$, and the second part was shown in (i). □