Journal of Data Science **4**(2006), 275-289

Testing for Activation in Data from FMRI Experiments

Martina Pavlicová[1], Noel Cressie[2], and Thomas J. Santner[2]
[1]Columbia University and [2]Ohio State University

Abstract: The traditional method for processing functional magnetic resonance imaging (FMRI) data is based on a voxel-wise, general linear model. For experiments conducted using a block design, where periods of activation are interspersed with periods of rest, a haemodynamic response function (HRF) is convolved with the design function and, for each voxel, the convolution is regressed on prewhitened data. An initial analysis of the data often involves computing voxel-wise two-sample t-tests, which avoids a direct specification of the HRF. Assuming only the length of the haemodynamic delay is known, scans acquired in transition periods between activation and rest are omitted, and the two-sample t-test is used to compare mean levels during activation versus mean levels during rest. However, the validity of the two-sample t-test is based on the assumption that the data are Gaussian with equal variances. In this article, we consider the Wilcoxon rank test as well as modified versions of the classical t-test that correct for departures from these assumptions. The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

Key words: Excess kurtosis, haemodynamic response function, Shapiro-Wilk test, skewness, two-sample t-test, Welch test, Wilcoxon Rank test.

1. Introduction

Functional Magnetic Resonance Imaging (FMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to FMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully *and* the data follow the Gaussian distribution. In this article, we consider the analysis of FMRI data collected in one of two states, called "activation" and "rest," based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the

activation data and the rest data are Gaussian with equal variances. In this article, we propose use of a modified two-sample test for FMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937), which is a modification of the two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for FMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in FMRI and the most common methods used to process FMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for FMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of FMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

2. FMRI Experiments

2.1 Some physiology

All neuronal activation is linked to an increase in oxygen consumption, causing a local increase in the blood flow. The body's response is to supply more oxygen than is required for the neuronal activity. Due to the different magnetic properties of oxygenated and de-oxygenated blood, the excess oxygenated blood that circulates during neuronal activation alters the magnetic properties of the venous blood, resulting in the so-called *blood oxygenation level dependent* (BOLD) signal. FMRI produces a sequence of brain images that is sensitive to changes in the BOLD signal.

In a classical FMRI experiment, the subject is scanned every few seconds to obtain an image of the brain; the subject is exposed to an experimental stimulus in some time periods, and is in a rest state during the remaining time periods. The stimulus can either be applied for brief periods in rapid, possibly random succession (an "event-related" experimental design, Josephs *et al.*, 1997), or for longer periods with interspersed rest periods (a "block" experimental design, Frackowiak *et al.*, 1997). In this paper, we focus on FMRI experiments conducted using a block experimental design.

Even though neuronal activation occurs immediately after exposure to the experimental stimulus, the vascular response evolves more slowly, resulting in

the BOLD signal. The temporal relationship between neuronal activation and the observed BOLD signal is called the haemodynamic response. To model the haemodynamic response, it is common to convolve the experimental design with a so-called haemodynamic response function (HRF). Poisson, gamma, and Gaussian distributions are used widely as HRFs (Friston *et al.*, 1994).

The region of the brain where there is neural activation is found by regressing the observed FMRI data on the expected BOLD signal, obtained as a convolution of the experimental design with the HRF. Of course, this depends on a well-specified HRF.

2.2 FMRI data

Observed FMRI data are four-dimensional, in space and time. At each time point, a three-dimensional image of the brain is acquired, called a volume. Each volume consists of voxels, and each voxel has an associated one-dimensional time series of observed signal intensities.

The most common approach to the analysis of FMRI data is to consider the voxels independently. A widely-used approach assumes a general linear model (GLM) for the voxel-wise time series (Friston *et al.*, 1995). For example, after various preprocessing steps, including prewhitening to achieve approximately independent errors, a two-sample test statistic is computed for each voxel where the two samples correspond to activation data and rest data. A voxel is declared to be significant if the test statistic exceeds some threshold. The distribution theory associated with this approach is based on the assumption of Gaussianity of the observed data and the proper specification of the HRF leading to the expected BOLD signal.

For initial data analysis, it is enough for us to know the length of the haemodynamic *delay* between neural activation and changes in the BOLD signal (Banddettini *et al.*, 1993). This knowledge is used to omit scans acquired in transition periods between possibly "activated" BOLD signals and "resting" BOLD signals. The delay between the neural activation and changes in the BOLD signal depends on many different factors; the type of stimuli, the duration of each stimulus, and the brain activation regions can all effect the length of the delay. Empirical studies have proposed methods for estimating HRFs that can adapt to different experimental designs. By using the block designs described in Section 2.1 and deleting transition data in our preliminary analysis, we have a sample of data acquired under activation and a second sample of data acquired under rest. In the next section, we describe four possible two-sample tests that might be used to test for the presence of activation at each voxel.

2.3 Two-sample tests

The null hypothesis of no difference between the means of two populations can be investigated with appropriate two-sample tests. In what follows, we summarize the four tests to be compared where, under activation the voxel data are $\underline{Y}_a = \{Y_i\}_{i\in \boldsymbol{A}}$ and, under rest the voxel data are $\boldsymbol{Y}_r = \{Y_j\}_{j\in \boldsymbol{R}}$; here $\boldsymbol{A}$ and $\boldsymbol{R}$ denote the activation and rest acquisition times, respectively.

The classical two-sample t-test

The classical two-sample t-test assumes:

(A1) Observations $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$ are uncorrelated.

(A2) The observations within each of $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$ have identical Gaussian distributions; that is,

$$\boldsymbol{Y}_f \;\sim\; Gau(\mu_f * \mathbf{1},\ \sigma_f^2 * \boldsymbol{I}); \qquad f \in \{a, r\}.$$

(A3) $\sigma_a^2 = \sigma_r^2$.

To test the hypothesis:

$$H_0 :\ \mu_a \le \mu_r \ \text{ versus } \ H_1 :\ \mu_a > \mu_r, \tag{2.1}$$

the classical two-sample t-test uses test statistic,

$$T \equiv \frac{\overline{Y}_a - \overline{Y}_r}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_r}\right)\left(\frac{(n_a-1)s_a^2+(n_r-1)s_r^2}{n_a+n_r-2}\right)}}, \tag{2.2}$$

with

$$\overline{Y}_f = \frac{1}{n_f}\sum_{i\in\mathbf{F}} Y_i \quad \text{and} \quad s_f^2 = \frac{\sum_{i\in\mathbf{F}}\left(Y_i - \overline{Y}_f\right)^2}{n_f - 1}; \qquad f \in \{a, r\},$$

where $\mathbf{F}$ is the set of activation times $\boldsymbol{A}$ (rest times $\boldsymbol{R}$) if $f = a$ ($f = r$), and n_a (n_r) is the number of the observations in the sample $\boldsymbol{Y}_a$ ($\boldsymbol{Y}_r$).

If Assumptions (A1), (A2), and (A3) are satisfied, the classical two-sample t-test with significance level α is:

$$\begin{aligned} &\text{Accept } H_0 \ \text{ if } \ T < t_d(1-\alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where $t_d(1-\alpha)$ is the $100(1-\alpha)$ percentile of the t distribution on $d = n_a + n_r - 2$ degrees of freedom.

The Welch test

The Welch test (Welch, 1937) is used to test the same hypotheses (2.1), but it assumes only (A1) and (A2); that is, it is possible that $\sigma_a^2 \neq \sigma_r^2$. Welch (1937) has shown that under the null hypothesis H_0, the test statistic

$$T^* \equiv \frac{\overline{Y}_a - \overline{Y}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}} \tag{2.3}$$

has approximately a t distribution with

$$e \equiv \frac{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)}{\left(\frac{\sigma_a^4}{n_a^2(n_a-1)} + \frac{\sigma_r^4}{n_r^2(n_r-1)}\right)} \tag{2.4}$$

degrees of freedom. In practice, the population variances σ_a^2, σ_r^2 in (2.4) are estimated from data using sample variances s_a^2, s_r^2. The Welch test with significance level α is:

$$\begin{aligned} &\text{Accept } H_0 \text{ if } \ T^* < t_e(1-\alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where the cut-off value $t_e(1-\alpha)$ is based on fractional degrees of freedom and is obtained by interpolation of the $t_d(1-\alpha)$ cut-off levels based on the nearest integers d to e.

The CW test

The CW test (Cressie and Whitford, 1986) also tests hypotheses (2.1), but makes only Assumption (A1); that is, it is possible that the data are non-Gaussian with unequal variances. To account for this, we use the same statistic T^* given by (2.3) as Welch, but modify its null distribution according to the skewnesses α_{3a}, α_{3r} and the excess kurtoses α_{4a}, α_{4r} of the non-Gaussian activation and rest distributions, respectively.

By calculating the Cornish-Fisher expansion of T^*, Cressie and Whitford (1986) show that under Assumption (A1) and H_0, the distribution of T^* is approximately that of the random variable,

$$V = U + \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{6\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}(U^2-1) - \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{2\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}U^2 - \frac{1}{2}gUZ, \tag{2.5}$$

where U, Z are i.i.d. $N(0,1)$ and

$$g \equiv \left\{ \frac{\frac{\sigma_a^4}{n_a^3}(\alpha_{4a}+2)+\frac{\sigma_r^4}{n_r^3}(\alpha_{4r}+2)}{\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^2} - \frac{\left(\frac{\alpha_{3a}\sigma_a^3}{n_a^2}-\frac{\alpha_{3r}\sigma_r^3}{n_r^2}\right)^2}{\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^3} \right\}^{1/2}. \tag{2.6}$$

The CW test with significance level α is

$$\text{Accept } H_0 \text{ if } \; T^* < v(1-\alpha)$$
$$\text{Accept } H_1 \text{ otherwise,}$$

where $v(1-\alpha)$ is the $100(1-\alpha)$ percentile of the distribution of V, obtained by simulation. As for the Welch test, the population moments in (2.5) and (2.6) are estimated from data using sample versions; see Section 3.3.

The Wilcoxon rank (WR) test

The WR test (Wilcoxon, 1945) makes only assumption (A1), as does the CW test. In addition, it assumes that the distribution function $F(y)$ of the observations $\boldsymbol{Y}_r$ is continuous and the distribution function of the observations $\boldsymbol{Y}_a$ is $F(y-\delta)$, for $\delta \in \mathbb{R}$. Then the WR statistic tests the hypotheses,

$$H_0 : \; \delta \leq 0 \;\; \text{versus} \;\; H_1 : \; \delta > 0. \tag{2.7}$$

In order to test (2.7), the WR test sums the ranks of each of the $\boldsymbol{Y}_a$ values in the combined sample of $N = n_a + n_r$ data consisting of the $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$ values ordered from smallest to largest. Let R_i denote the rank of Y_i; $i \in \boldsymbol{A}$. The test statistic for the WR test is

$$W = \sum_{i \in \boldsymbol{A}} R_i.$$

An exact p-value is then computed based on the null distribution ($\delta = 0$) of W, which is obtained by considering all possible $N!$ permutations of ranks of the $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$. However, this is computationally demanding for large n_a and n_r. For large n_a and n_r, we approximate the distribution of the centered and scaled version of W,

$$W^* = \frac{W - .5 - n_a(n_a+n_r+1)/2}{\sqrt{n_a n_r (n_a+n_r+1)/12}},$$

with a standard normal (Hollander and Wolfe, 1999). Hence the WR test with significance level α is:

$$\text{Accept } H_0 \text{ if } \; W^* < z(1-\alpha)$$
$$\text{Accept } H_1 \text{ otherwise,}$$

where $z(1-\alpha)$ is the $100(1-\alpha)$ percentile of the Gaussian distribution with zero mean and unit standard deviation.

3. Methods of Analysis and Comparisons

In this section, we continue to consider inference based on a single generic voxel. Simultaneous inference involving all voxels is considered in Section 4.

3.1 Application of two-sample tests to FMRI data

Let **T** be the set of acquisition times of the observed intensities associated with the given voxel. Assuming the subject was exposed to only one type of neural activation, **T** can be divided into three groups: the time points $\boldsymbol{A}$ where activation of the BOLD signal is expected, the time points $\boldsymbol{R}$ during which the BOLD signal is expected to be in a rest state, and the time points **B** corresponding to the transition periods between the activation and the rest times. An example of such a division of time points is illustrated in Figure 1. In the two-sample tests considered in this article, one sample corresponds to $\boldsymbol{A}$ and other sample corresponds to $\boldsymbol{R}$; intensities corresponding to **B** are omitted from further analysis.

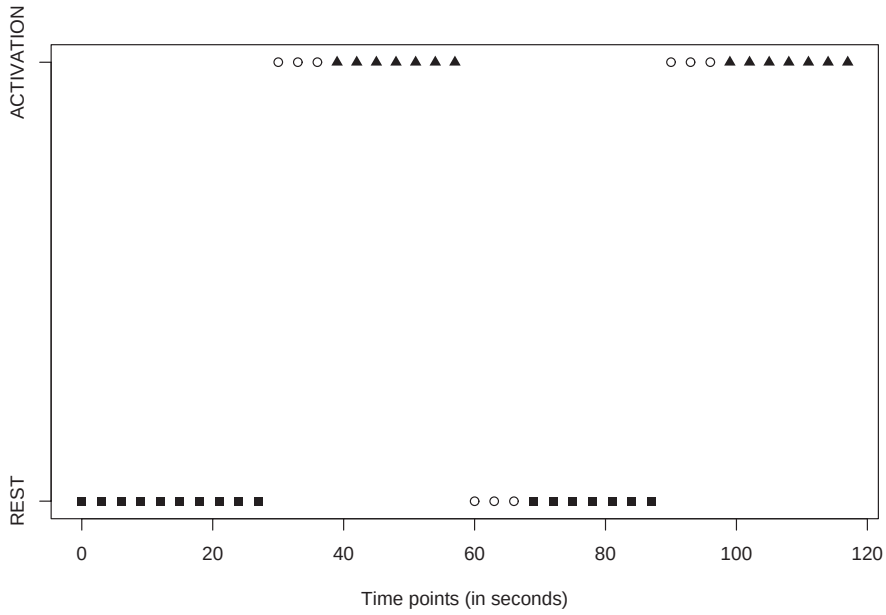

Figure 1: Hypothetical example of 40 observed intensities. Each rest and activation time period is 30 seconds long, and the haemodynamic delay is assumed to be 9 seconds. The time points **A** when the BOLD signal is expected to be in the activated state are denoted by ▲, the time points **R** when the BOLD signal is in the rest state are denoted by ■, and the time points **B** of transition periods between the rest and activation state of the BOLD signal are denoted by ○.

Consider the two-sample tests of H_0 versus H_1 given in Section 2. For a given voxel and a given test, accepting the alternative hypothesis H_1 means that the associated voxel is declared to be activated by the experimental stimulus.

3.2 Simulated FMRI data

Six datasets were obtained from 3 healthy volunteers (1 female, 2 males) using a 1.5T Signa scanner. The data were collected under rest conditions; that is, the subjects were not exposed to any stimulus during the experiment and they were instructed to relax in the scanner with their eyes closed. One such rest dataset was obtained from the first male subject (30 years old), two rest datasets were obtained from the second male subject (27 years old), and three rest datasets were obtained from the female (30 years old). Each dataset consisted of 200 volumes, every observed volume contained 28 slices, and each slice had 64x64 voxels. These datasets were preprocessed for motion correction and prewhitened to make the time series uncorrelated (using the software FEAT, which is part of the FSL package; see Smith *et al.*, 2001).

Figure 2: An example of activation clusters superimposed on one volume of the artificial-activation dataset. The three images depict samples of 3 axial views (the center image is positioned in the middle of the brain, the left image is positioned inferior to the middle, and the right image is positioned superior to the middle).

We created activation datasets by essentially adding a signal having *known magnitude and location* of the activation to each preprocessed rest dataset. The signal component was calibrated against an image acquired from a previous unrelated visual-activation FMRI experiment; see Figure 2 for an example. By applying the signal in the locations acquired from a previous visual experiment, we avoided the possibility of applying the signal near so-called default regions (regions which show decreased neuronal activity during the activation of the stimulus) and their confounding effects on the simulated signal. The activation datasets alternated blocks of 10 time points of rest with 10 time points of activation. The average peak-signal change, defined as a ratio between the average of

the intensities under the activation and the average of the intensities measured during the rest periods for the most activated voxel, was set to be 3%. Each dataset contains 200 time points; the three sets of time points $\boldsymbol{A}$, $\boldsymbol{R}$, and $\mathbf{B}$ were obtained assuming a haemodynamic delay of 3 time periods, resulting in $n_a = 70$ and $n_r = 73$.

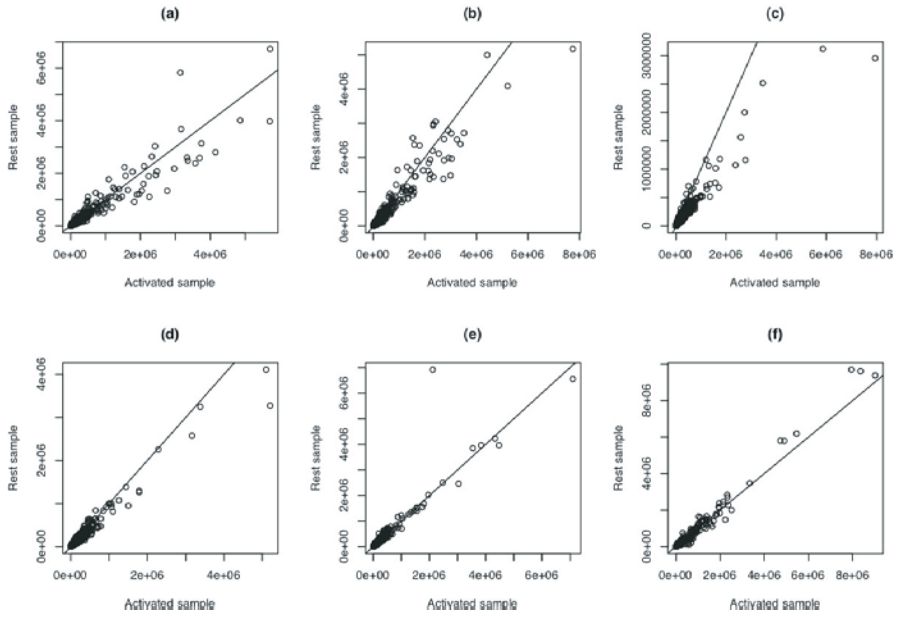

Figure 3: Estimated sample variances of activated samples versus rest samples for each voxel which is located in subject's brain area (only 22,340 voxels from the total of $64 \times 64 \times 28 = 114,688$ voxels were located in the brain area). The six panels (a)-(f) correspond to the six datasets.

3.3 Violations of equal variances and Gaussianity assumptions

Several methods were used to assess the degree of departure of the activation datasets from (A2) and (A3). Consider a generic voxel and recall from Section 1 that $\boldsymbol{Y}_a = \{Y_i\}_{i \in \boldsymbol{A}}$ make up the so-called "activated" sample and $\boldsymbol{Y}_r = \{Y_j\}_{j \in \boldsymbol{R}}$ make up the "rest" sample.

To investigate the violation of Assumption (A3) given in Section 2, thereby allowing $\sigma_r^2 \neq \sigma_a^2$, we computed the sample variances for $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$ for each voxel in each activation dataset. The pairs of sample variances of active and rest samples for all voxels that are located in subject's brain (out of all $64 \times 64 \times 28 = 114,688$ voxels, only $22,340$ of them were located in subject's brain) are plotted in Figure 3; the 45-degree line corresponding to equal variances is superimposed.

In all panels, and especially in 3(c), we see some points far from the diagonal, which suggests that the assumption of homogeneity is violated for three voxels. A formal F-test ($\alpha = 0.05$) of equal variances detected $1,225$ out of $22,340$ (5.5%) brain voxels to have significantly different sample variances, and visual inspection of these voxels indicated no spatial pattern. This indicates that, overall, unequal variances may not be a serious problem for these FMRI data.

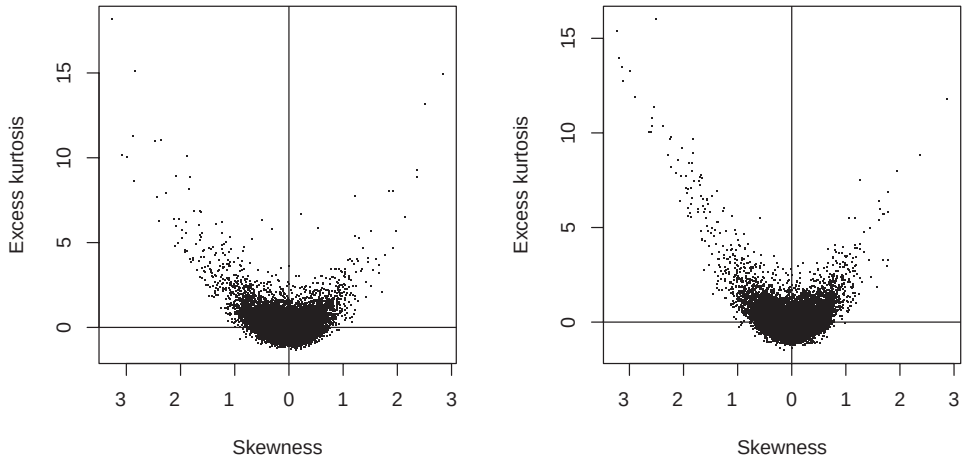

Figure 4: Sample skewness versus sample excess kurtosis for all voxels from one of the six datasets. Panel (a) Activation data, and Panel (b) Rest data.

To investigate departures from Gaussianity, Assumption (A2), we computed the sample skewness and sample excess kurtosis for $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_r$, for all six activation datasets. For the activation sample these are:

$$\begin{aligned} \widehat{\alpha}_{3a} &= \frac{\sqrt{n_a}\sum_{i\in\boldsymbol{A}}(Y_i - \overline{Y}_a)^3}{\{\sum_{i\in\boldsymbol{A}}(Y_i - \overline{Y}_a)^2\}^{3/2}}, \\ \widehat{\alpha}_{4a} &= \frac{n_a\sum_{i\in\boldsymbol{A}}(Y_i - \overline{Y}_a)^4}{\{\sum_{i\in\boldsymbol{A}}(Y_i - \overline{Y}_a)^2\}^2} - 3, \end{aligned}$$

and likewise we computed $\widehat{\alpha}_{3r}$ and $\widehat{\alpha}_{4r}$ for the rest sample.

To illustrate graphically the relationship between skewness and kurtosis, we chose one activation dataset. The pairs $(\widehat{\alpha}_{3a}, \widehat{\alpha}_{4a})$ for the 22,340 brain voxels from one activation dataset are plotted on the left panel of Figure 4, and the pairs $(\widehat{\alpha}_{3r}, \widehat{\alpha}_{4r})$ are plotted on the right panel. For Gaussian data, the plotted pairs should be very close to the origin. In Figure 4, we observe strong departures from zero skewness and zero excess kurtosis in both panels. Thus, we might expect an improvement in hypotheses testing for activation using the CW test or the WR test over the classical two-sample t-test or the Welch test.

More formally, we calculated the Shapiro-Wilk test (e.g., Royston, 1982) for normality ($\alpha = .05$) for each voxel and rest/activation combination. For the dataset used in Figure 4, Table 1 summarizes the number (out of 22,430) of brain voxels that were significantly non-Gaussian. About 12% of activated samples and about 11% of rest samples were declared significant by the Shapiro-Wilk test; if the samples were Gaussian, we would expect only 5% to be declared significant. More than 20% of voxels were declared significant in at least one of the activated or rest samples.

Table 1: Brain-voxels declared significant using Shapiro-Wilk test ($\alpha = .05$), based on one of the six datasets.

		Activated samples		
		Significant	Not significant	Total
Rest samples	Significant	647	2095	2742 (12.3%)
	Not significant	1774	17824	19598
Total		2421 (10.8%)	19919	22340

The spatial distribution of the voxels declared significant is shown in Figure 5; while they are distributed fairly homogeneously between regions of the brain, there is some indication that, within a region, they can clump together.

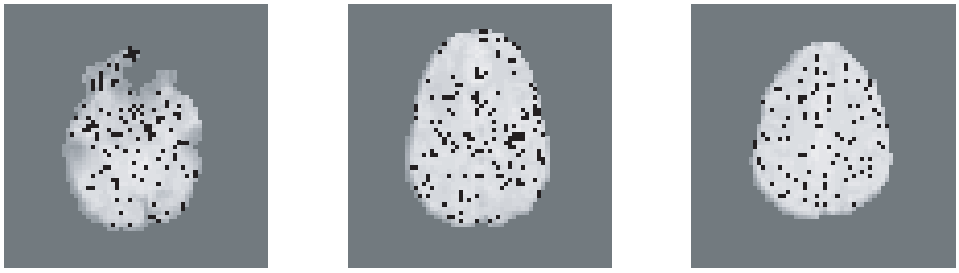

Figure 5: Spatial distibution of voxels for which activated samples violate the Gaussianity assumption (Shapiro-Wilk test; $\alpha = .05$). The three images shown correspond to the three views of the brain given in Figure 2.

4. Results

All four two-sample tests were used to test for activation in each voxel. We obtained p-values as in Section 2 where the p-value for the CW test was obtained from simulation of the random variable given by (2.5) and that for the WR test was obtained from the standard normal approximation to W^*.

Because of the multiple hypotheses being tested (one for each brain voxel), the voxels declared as active were obtained by comparing the p-values with $\alpha^* \equiv \alpha/\{\#\text{ of brain voxels}\}$ with $\alpha = .05$. This is the voxelwise Bonferroni-adjusted level of significance based on an overall level of significance of $\alpha = .05$. Voxels with p-values less than or equal to α^* were pronounced active. Because the activation pattern of each dataset was known, we can estimate and compare the sizes and powers of the two-sample tests.

Let $\mathcal{A}$ denote the set of voxels to which an activation signal has been added and $\mathcal{R}$ the set of voxels with no added activation. Let $\mathcal{A}_{\text{right}}$ denote the voxels in $\mathcal{A}$ declared to be active, and let $\mathcal{A}_{\text{wrong}}$ denote the voxels in $\mathcal{A}$ not declared active. All voxels from category $\mathcal{R}$ can be similarly divided into $\mathcal{R}_{\text{right}}$, those non-activated voxels not declared active, and $\mathcal{R}_{\text{wrong}}$, those non-activated voxels which were declared active.

The achieved size of each test was estimated by

$$\widehat{\alpha} \equiv (|\mathcal{R}_{\text{wrong}}| \,/\, |\mathcal{R}|),$$

where $|\mathcal{C}| \equiv \#$ voxels in the region $\mathcal{C}$ of the brain. The quantity $\widehat{\alpha}$ is also called the false-positive rate and should be comparable to the desired familywise level of significance α $(= .05)$. If $\widehat{\alpha} < \alpha$, the test is conservative. The power of each test was estimated by

$$\widehat{\pi} \equiv (|\mathcal{A}_{\text{right}}| \,/\, |\mathcal{A}|)\,,$$

which is the true-positive rate.

Table 2 lists the estimated sizes and powers of each test for all six simulated FMRI datasets. All four tests were consistently very conservative, with the Wilcoxon test being the most conservative. The classical t-test and Welch test had equivalent power, which was consistently greater than that of the Wilcoxon test. The CW test was the most powerful test, uniformly over the six datasets.

Table 2: Estimated size and power of the four two-sample tests for the six datatsets.

Dataset	TEST							
	Classical t-test		Welch		CW		WR	
	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$
1	.496E-4	.289	.496E-4	.288	.992E-4	.305	0	.277
2	0	.208	7.466E-4	.208	19.927E-4	.224	4.479E-4	.200
3	0	.221	0	.217	0	.235	0	.202
4	.583E-4	.233	.583E-4	.233	1.750E-4	.253	.583E-4	.224
5	0	.205	0	.205	0	.223	0	.188
6	0	.239	0	.237	27.527E-4	.251	1.101E-4	.227

Table 3 gives a more detailed comparison of the classical t-test and the CW test for one of the datasets. While 626 out of 2,173 activated brain voxels were correctly detected as significant by both tests, 37 additional activation voxels were correctly detected by the CW test that were not identified by the classical t-test. Only one activation voxel was identified by the classical t-test that was missed by the CW test.

Table 3: Comparison of the performance of the CW test and the classical t-test, based on one of the six datasets

			CW test			
			Voxels from $\mathcal{A}$		Voxels from $\mathcal{R}$	
			$\mathcal{A}_{\text{right}}$	$\mathcal{A}_{\text{wrong}}$	$\mathcal{R}_{\text{right}}$	$\mathcal{R}_{\text{wrong}}$
Classical t-test	Voxels from $\mathcal{A}$	$\mathcal{A}_{\text{right}}$	626	1	·	·
		$\mathcal{A}_{\text{wrong}}$	37	1509	·	·
	Voxels from $\mathcal{R}$	$\mathcal{R}_{\text{right}}$	·	·	20165	1
		$\mathcal{R}_{\text{wrong}}$	·	·	0	1

5. Discussion and Conclusions

While the results were obtained from only one type of scanner, the 1.5T Signa GE, and with FMRI data for three subjects, they show that FMRI data can exhibit both unequal variances and non-Gaussianity. Using the Shapiro-Wilk test, more than 20% of voxels in the dataset were declared significant in one or both of the rest or activated samples. We believe that more powerful scanners will lead to data that are even more non-Gaussian, since their finer spatial resolution involves less averaging of the response.

The Welch test is valid for unequal variances but when non-Gaussianity is suspected, the CW test accounts for both. The WR test is a nonparametric analog of the classical t-test. In the six datasets studied in Section 3, non-Gaussianity was a bigger problem than unequal variances. The results in Section 4 showed that the CW test performed better than the other three tests. These results suggest that the CW test should replace any standard use of the classical parametric or nonparamentric two-sample tests based on FMRI data.

Acknowledgement

This research was supported by the Office of Naval Research under grants N00014-99-1-0214 and N00014-02-1-0052 and by the National Science Foundation Grant DMS-0406026. The authors would like to thank Antonio Algaze and

Petra Schmalbrock for providing the FMRI data and the members of FMRIB, Oxford UK for initial consultation about simulating activation FMRI datasets. Perceptive comments by the referees led to improvements in the exposition and strengthening of our conclusions.

References

Bandettini, P. A., Jesmanowicz A., Wong, E. C. and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine* **30**, 161-173.

Cressie, N. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal* **28**, 131-148.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. and Mazziotta, J. C. (1997). *Human Brain Function*. Academic Press.

Friston, K. J.,. Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189-210.

Friston, K. J., Jezzard, P. and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* **2**, 69-78.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods, 2nd edn.* John Wiley and Sons.

Josephs, O., Turner, R. and Friston, K. J. (1997). Event-related fMRI. *Human Brain Mapping* **5**, 243-248.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* **31**, 115-124.

Smith, S. M., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, J., Leibovici, D., Ripley, B., Woolrich, M. and Zhang, Y. (2001). FSL: New tools for functional and structural brain image analysis. *NeuroImage* **13**, S249.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

Received November 5, 2004; accepted May 11, 2005.

Martina Pavlicová
Department of Biostatistics
Mailman School of Public Health
Columbia University
722 West 168th Street, 6th Floor
New York, NY 10032

Noel Cressie
Department of Statistics
404 Cockins Hall
1958 Neil Ave
The Ohio State University
Columbus, OH 43210
ncressie@stat.ohio-state.edu

Thomas J. Santner
Department of Statistics
404 Cockins Hall
1958 Neil Ave
The Ohio State University
Columbus, OH 43210