A Critique of the Use of the Kolmogorov-Smirnov (KS) Statistic for the Analysis of BOLD fMRI Data

Geoffrey K. Aguirre, Eric Zarahn, Mark D'Esposito

The use of the Kolmogorov-Smirnov (KS) statistic for testing hypotheses regarding activation in blood oxygenation level-dependent functional MRI data is critiqued on both theoretical and empirical grounds. Theoretically, it is argued that the KS test is formally unable to support inferences of interest to most neuro-imaging studies and has reduced sensitivity compared with parametric alternatives. Empirically, false-positive rates yielded by the KS test in human data collected under the null hypothesis were significantly in excess of tabular values. These excessive false-positive rates could be explained by the presence of temporal autocorrelation. We also present evidence that the distribution of blood oxygenation level-dependent functional MRI data is only slightly nonnormal, questioning the initial impetus for the use of the KS test in this context. Finally, it is noted that parametric alternatives exist that do provide adequate control of the false-positive rate and can support inferences of interest.

Key words: Kolmogorov-Smirnov; functional MRI; normality; null-hypothesis.

INTRODUCTION

A common implementation of functional MRI (fMRI) detects changes in a blood oxygenation level dependent (BOLD) signal, which reflects underlying neural activity (1, 2). A feature of experiments which employ this technique is the need to test hypotheses regarding the effect of an experimental manipulation upon the observed signal. Because the distribution of BOLD fMRI signal values has been reported to be nonnormal (3, 4), some researchers (3) have proposed the use of univariate, nonparametric tests in the analysis of fMRI data. In particular, the Kolmogorov-Smirnov (KS) statistic, which tests for differences in the cumulative probability distribution of two samples, has been widely applied (5-15) and continues so up to the writing of this report (16, 17).

It is the intention of this report to demonstrate that the KS is an invalid statistical test for use with BOLD fMRI data based on empirical and theoretical grounds. (i) Empirically, we demonstrate that the KS test does not control the false-positive rate in BOLD fMRI data. This failure of the KS test is most likely due to violation of the independence assumption. (ii) We also empirically demonstrate that BOLD fMRI data are only slightly nonnormal, perhaps obviating the need for consideration of nonparametric statistics, like the KS test. (iii) It is explained in the Discussion section that the KS statistic is theoretically incapable of supporting inferences of interest to many experimenters, namely that differences in signal magnitude have occurred as the result of experimental treatment.

METHODS

MRI Technique

Imaging was carried out on a 1.5 T SIGNA scanner (GE Medical Systems) equipped with a prototype fast gradient system for echoplanar imaging (EPI). A standard RF head coil was used with foam padding to comfortably restrict head motion. High-resolution sagittal T1-weighted images were obtained in every subject. A gradient echo, echoplanar sequence was used to acquire data sensitive to the BOLD signal at a TR = 2000 ms, TE = 50 ms. Resolution was 3.75 mm × 3.75 mm in plane, and 5 mm between planes (16 axial slices acquired). A total of 160 gradient-echo echoplanar images in time were obtained per slice in each 320-s run. Twenty seconds of “dummy” gradient and RF pulses preceded the actual data acquisition. For all analyses presented below, only voxels with a minimum signal value over time greater than a threshold (600) were considered. This signal value is consistently in a range that is well separated from nontissue voxels in our data.

fMRI Datasets

Null-hypothesis (“noise”) datasets were obtained by scanning 17 healthy, young subjects. Subjects were instructed to relax with eyes open (to otherwise mimic experimental conditions and discourage sleep). Results based on the same datasets obtained from these subjects have been previously reported (18, 19). The room was dim, but lights from the control room were visible. Equipment used in our laboratory to present stimuli during behavioral imaging experiments (including an overhead projector and LCD panel (InFocus systems panelbook 550) positioned and powered within the Faraday cage, which were connected to a Macintosh 280c computer located outside the cage) were kept on, and in their standard setups to include all sources of noise typical to our experiments. These electronic devices have been demonstrated to add only white noise to EPI data collected at our site (18).

Motion Correction

Two iterations of a six-parameter, rigid-body, least squares realignment routine were used (part of SPM96b...
package (20)). The effect of this realignment procedure has been demonstrated (20) to be very similar to that of another frequently used registration technique, the Automated Image Registration (AIR) routine (21, 22). The "spin-history" correction advocated by Friston and colleagues (23) was not applied.

Creation of Statistical Maps

In the creation of statistical maps and assessment of false-positive rates, an "assumed" behavioral task was used. The temporal structure of the assumed paradigm was designated to be a boxcar with an 80-s period (i.e., two 40-s epochs; 0.0125 Hz). This low paradigm frequency is comparable to that used in many "blocked" fMRI experiments that have been analyzed with the KS statistic (20-s epochs, refs. 5, 15; 30-s epochs, refs. 7, 8, 10, 12, 16; 40-s epochs, refs. 11, 14; 60-s epochs, ref. 13).

A KS statistical map was created for each dataset. The maximum distance \( D \) between the cumulative probability distributions \( S_{\text{on}}(x) \) of the "on" and "off" data points was computed at each voxel (24):

\[
D = \max_{-\infty < a < \infty} |S_{\text{on}}(x) - S_{\text{off}}(x)|
\]

Manipulations of the time series data are often undertaken to reduce low-frequency noise. The effect of two of these manipulations was examined (i) removal of a linear slope from the time series data to remove signal "drift," and (ii) removal of frequencies below that of the paradigm frequency (i.e., below 0.0125 Hz). KS maps were also generated for the datasets with each of these treatments.

Calculation of \( \alpha \) Thresholds

Critical, voxelwise \( D \) thresholds were calculated using (24):

\[
\text{Probability}(D > \text{observed}) = Q_{KS}(\lambda) = 2\sum_{i=1}^{\infty} (-1)^{i-1}e^{-2i^2\lambda^2}
\]

\[
\lambda = \left[ \sqrt{N_e} + 0.12 + 0.11 / \sqrt{N_e} \right] D
\]

\[
N_e = \frac{N_{\text{on}}N_{\text{off}}}{N_{\text{on}} + N_{\text{off}}}
\]

where \( N_{\text{on}} \) and \( N_{\text{off}} \) are the number of observations in the on and off periods, respectively. Mapwise \( D \) thresholds were also calculated by using Bonferroni correction for the multiple comparisons conducted across voxels. For the 17 maps analyzed under the different treatments, the mapwise critical \( D \) for a mapwise \( \alpha = 0.05 \) corresponded to a mean, voxelwise \( \alpha \) value of 3.9e-6 (range, 3.2e-6 to 4.8e-6).

False-Positive Rate Comparisons

Two different measures were used to characterize false-positive rates in the null-hypothesis datasets. One (hereafter referred to as FP1) was the percentage of voxels in each dataset for which the \( D \) value exceeded a voxelwise \( \alpha \) threshold. The second false-positive measure (FP2) was the proportion of datasets that contained at least one voxel with a \( D \) value that was greater than a Bonferroni corrected, mapwise threshold. These two measures will differ if a small subset of the voxels within a dataset exceed the mapwise threshold, although not significantly affecting the overall voxelwise proportion. False-positive rates were measured for a range of \( \alpha \) values (0.1-0.0001) for both of these measures. The ratio of observed to expected false-positive rates was calculated for each \( \alpha \) examined for each measure (FP1 and FP2). The hypothesis that the central tendency of FP1 was different from 5\% for an \( \alpha = 0.05 \) was tested with the signed rank test (two-tailed). The hypothesis that FP2 was greater than 5\% for an \( \alpha = 0.05 \) was tested with the binomial distribution (one-tailed, because there was no power to test that FP2 was less than 5\%). Both hypotheses were tested for each treatment of the "noise" data.

Computer-generated noise, with both Gaussian and uniform distributions, was used to ensure that our implementation of the KS statistic yielded tabular false-positive rates when the assumption of independence was met.

Test of Normal Distribution

Satisfaction of the normality assumption for parametric tests requires that the residuals of an analysis be normally distributed. Because we have collected BOLD fMRI data under the null hypothesis, the distribution of these raw data can be expected to conform to the distribution of the residuals of any parametric test that validly models introduced experimental variance. Thus, if the distribution of our raw fMRI data is not grossly nonnormal, then parametric tests as a class are be expected to be valid with sufficient degrees of freedom (vis a vis, the normality assumption). Additionally, recent applications of parametric statistics to fMRI data have included some degree of high-pass filtering. We also characterized the normality of the fMRI data after two treatments designed to reduce low-frequency noise: (i) removal of linear "drift" and (ii) removal of frequencies below that of the assumed paradigm. These are the same treatments for which the false-positive rate using the KS test was examined (see above). The distribution of the data after these manipulations is expected to conform to that of the residuals of any parametric test (validly modeled). The experimentally introduced variance is well modeled.

We characterized the normality of the distribution of the fMRI time series data at each voxel within each of the 17 datasets using the D'Agostino-Pearson \( K^2 \) omnibus test of normality (25):

\[
K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)
\]

where \( Z(b_1) \) and \( Z(b_2) \) are the normal approximations to the sample estimates of skewness and kurtosis (see ref. 25.
for details). $K^2$ has a $\chi^2$ distribution with 2 $df$. This statistic is more powerful than the traditional $\chi^2$ and Kolmogorov tests and is sensitive to deviations in skewness and kurtosis (25). The application of the test to normally distributed, computer-generated noise (160 data points) yielded tabular false-positive rates. Uniformly distributed, computer-generated noise yielded a true-positive rate of 99.99% when $\alpha$ was held at 5%, demonstrating the high sensitivity of the test.

The normality measure was calculated as the percentage of voxels in each dataset for which the $K^2$ value exceeded the voxelwise $\alpha$ threshold. Under the null-hypothesis of normal distribution at every voxel within the brain, the theoretically expected value of this measure is 5%. The hypothesis that the central tendency of the normality measure was different from 5% was tested with the signed rank test (two-tailed). Note that the evaluation of normality on a voxelwise basis is preferable to simply pooling all time-series data from all voxels in a given subject, as voxelwise variance may not be stationary over space.

RESULTS

The circles in Fig. 1 represent the median ratio of observed to expected voxelwise false-positive rates (FP1) for a range of $\alpha$ values. The median FP1 measure obtained for an $\alpha = 0.05$ was 9.6% for the raw time-series, 8.2% for the data with linear “drift” removed, and 13.1% for the data with frequencies removed below that of the assumed task. The set of 17 FP1 measures was judged significantly different from 5% for all three treatments ($P < 0.0001$, signed rank test for all three treatments). Thus, in datasets collected while subjects performed no experimental task and were exposed to no time-locked stimuli, the KS statistic yielded voxelwise false-positive rates significantly higher than tabular values.

The squares in Fig. 1 represent the ratio of observed to expected false-positive rates (FP2) for a range of $\alpha$ values. The FP2 measure obtained for an $\alpha = 0.05$ was significantly greater than the nominal 5% level ($P < 1e-7$, binomial distribution) for all three treatments. For the data in which low frequency confounds were removed, the mapwise false-positive rate for a tabular $\alpha = 0.05$ was 94%. That is, of 17 subjects studied, 16 were judged by the KS test to have brain areas with significant responses to a nonexistent task.

Finally, the datasets were examined to determine whether BOLD fMRI data are normally distributed on a voxelwise basis. The percentage of voxels in each dataset with a significantly nonnormal distribution of signal values was determined. Under the null-hypothesis (i.e., that every voxel has a normal distribution), this measure should be 5%. For the raw time-series data, the median deviation rate ($\pm \frac{1}{2}\text{IQR}$ interquartile range) was 13.5% ± 2.5%. Removal of linear drift from the time series improved the normality of the data for 14 of 17 datasets and resulted in a median deviation from normality of 11.1% ± 3.0%. Removal of frequencies below that of the assumed paradigm improved the normality of 16 of 17 datasets and produced a median measure of 11.3% ± 2.3%. In all cases, the rejection of normality measure was found to be significantly greater than 5% ($P < 0.0001$, signed rank test). Given the considerable sensitivity of the test (99.99% for uniformly distributed data), a rejection rate of approximately 11% (of the voxels) suggests only a slight, albeit extant, deviation from normality for the dataset as a whole.

DISCUSSION

Specificity

When the KS statistic was used to analyze human, BOLD fMRI data collected under the null-hypothesis, both voxelwise and mapwise false-positive rates were found to be significantly in excess of tabular values. Given that statistical maps generated for neuroimaging studies are frequently thresholded and searched for excursions, the observed mapwise false-positive rates are particularly relevant and worrisome. The treatment most similar to that used in many studies, namely six-parameter realignment motion correction and removal of linear drift, produced a mapwise false-positive rate of 65%. It is worth noting that the mapwise threshold (voxelwise $\alpha = 3.9e-6$, significantly greater than 5% ($P < 0.0001$, signed rank test). Given the considerable sensitivity of the test (99.99% for uniformly distributed data), a rejection rate of approximately 11% (of the voxels) suggests only a slight, albeit extant, deviation from normality for the dataset as a whole.
corresponding to mapwise $\alpha = 0.05$) used here used a Bonferroni correction to account for the multiple statistical comparisons conducted over the entire brain volume. Mapwise thresholds that do not adequately control for multiple-comparisons in the face of hypotheses tested over extended volumes (9, 11–13, 15) will, of course, result in commensurately higher false-positive rates.

The use of the KS statistic (as well as many other parametric and nonparametric tests) requires that observations be sampled independently from the two populations. We have recently identified and characterized temporal autocorrelation within these null-hypothesis BOLD fMRI datasets under spatially unsmoothed (18) and smoothed (19) conditions. This intrinsic autocorrelation is well modeled by a function with a reciprocal of frequency term, i.e., increasing power was present at lower frequencies. As has been demonstrated in the current report for the nonparametric KS test, parametric statistics, such as the $t$ test, were also shown to yield unacceptable false-positive rates when used to test an assumed low-frequency behavioral paradigm (18, 19). Simulations with computer-generated data in which this form of dependence was introduced (results not shown) indicate that this temporal autocorrelation is a sufficient explanation for the inflated false-positive rates observed here using the KS test. It should be noted that averaging or concatenating multiple fMRI scans will not ameliorate the violation. Additionally, because spatial smoothing further increases the relative power of low frequencies (due to greater spatial coherency of low frequencies relative to high frequencies) (19), false-positive rates can be expected to be higher still in datasets that have been spatially smoothed.

Recently, Xiong and colleagues (26) examined, among other statistics, the specificity of the KS statistic within null-hypothesis, human fMRI data. Their measure was the same as the FPI (voxelwise) measure presented here. Interestingly, Xiong et al. reported that the KS statistic yielded an empirical false-positive rate that was half (more conservative than) the nominal $\alpha$. It is important, however, to note that Xiong et al. used a measure of significance that differs from that used here. Furthermore, the authors did not validate the expected false-positive rate of their measure with computer-generated noise (Xiong, personal communication), hence, it is not possible to determine conclusively whether their implementation of the KS test yielded overly conservative measures because of properties of the fMRI data or because of properties of their test.

There are several caveats to be considered regarding the results presented here. First, the examination of the null-hypothesis rate was conducted for an assumed paradigm at 0.0125 Hz. This frequency was chosen because of its similarity to those used in many neuroimaging studies. It might be expected that experiments conducted at higher temporal frequencies (e.g., (9)), in which null-hypothesis power is low relative to other frequencies, will not yield such egregious false-positive rates. This conclusion, however, should be a matter of empirical validation. It is also important to consider the possibility that the temporal autocorrelation previously observed (18) in these datasets is a consequence of the particular scanner/pulse-sequence/reconstruction algorithm/motion correction method used here. We have presented data previously (18) that suggests that temporal autocorrelation cannot be entirely explained by subject motion. In addition, several other groups have presented null-hypothesis power spectra that show a 1/frequency curve. These other sites have used single-shot EPI (27–29) and spiral-imaging (30) to collect data at 1.5 and 4 T (29). Thus, it seems that at least the general pattern of increasing power at lower frequencies is present in fMRI data collected at other sites, at different field strengths, and with different pulse sequences. Almost certainly, however, the particular pattern of autocorrelation and, by implication, the false-positive rates that we have observed are idiosyncratic to this lab. Thus, the most prudent course of action would seem to be an empirical characterization of the false-positive rate for any site/analysis technique.

Normality and Sensitivity

The rationale for the use of nonparametric statistics, such as the KS test, is the absence of distributional assumptions regarding the data under study. Parametric statistics, on the other hand, require that the residuals of the analysis be normally distributed. Although previous reports have claimed that BOLD fMRI data are not normally distributed, there are reasons to question these assertions. Baker and colleagues (3) obtained gradient-echo EPI data at a TR of 133 ms from primary visual cortex during darkness and photic stimulation. The distributions of the signal values from the two task conditions were found to be normal. However, when Baker et al. resampled their data at 1 Hz, the data were found to no longer have a normal distribution. This nonnormality was attributed to signal pulsatility produced by respiratory and cardiac fluctuations. However, cardiac and respiratory pulsations are not present in our data (18). This absence is most likely the result of our longer TR (2 s), which renders our data less sensitive to $T_1$-dependent in-flow effects, believed to be the cause of these pulsatile signals (29). Because of these empirically supported, theoretical arguments which suggest that data collected at different TRs will have differing contributions from pulsatile signals, the test of Baker and colleagues is inappropriate, because they are examining the effects of a longer TR by resampling their data collected at a shorter TR. The only other report (4) that has been offered as evidence of a non-Gaussian distribution of fMRI data actually found that the distribution of 2DFT fast low-angle shot data was not significantly different from a Gaussian distribution as judged by a KS test ($P = 0.34$). Thus, both of the previous reports that have examined the distribution of fMRI data have found it to be normally distributed.

We applied here a highly sensitive test for deviations from normality to our null-hypothesis data on a voxelwise basis. The relatively small proportion of voxels in which the null hypothesis of normality was rejected suggests that as a whole, voxelwise distributions do not differ greatly from a Gaussian distribution. We did not determine, however, whether this rejection rate reflects a small population of highly nonnormal voxels, or a gen-
eral tendency of the dataset as a whole to slightly non-normal distributions. The question then becomes, whether these deviations from normality are sufficient to render parametric statistics invalid. Relatively modest high-pass filtering was noted here to improve the normality of the distribution. It is possible that parametric analyses that use more stringent filtering will yield residuals that further approach normality. We also note that with sufficient degrees of freedom, the requirement for normality is moderated (due to the central limit theorem). Thus, given the considerable number of observations present in many fMRI experiments and the minor degree of deviation from normality observed here, parametric statistics may be appropriate. Of course, the only final justification of these statements can be obtained by empirical tests of the null-hypothesis distribution of parametric tests (18, 19). It should be noted that when assumptions of normality are met, the KS test will have reduced power compared with parametric alternatives (31).

Inference

The KS test is sensitive to any deviations from the null-hypothesis that two samples are drawn from the same population. Thus, despite erroneous claims to the contrary (10, 12, 13, 15, 17), the KS test cannot be used to infer that a difference in the mean level of signal exists between two experimental conditions. Significant KS scores could reflect sample differences in only variance or higher distributional moments, such as skewness. This ambiguity in inference is problematic, given the current model of the BOLD effect, as outlined below.

Local increases in neural-activity have been demonstrated to produce local increases in blood flow (32), which in turn engender a delayed decrease in local [deoxyhemoglobin] (33). Decreases in [deoxyhemoglobin] result in decreased local susceptibility gradients, thus increasing the $T_2$ weighted fMRI signal (29). Boynton and colleagues (30) demonstrated that increases in fMRI signal intensity can be induced by brief (3-s) and longer (24-s) periods of visual stimulation (which are ostensibly accompanied by roughly 3-s and 24-s periods of neural activity in V1). These fMRI signal increases were sustained when the duration of visual stimulation exceeded the time-to-peak of the putative impulse response function of the BOLD hemodynamic system. Moreover, increases in stimulus contrast, proposed to produce increased neural activity, resulted in monotonic increases in fMRI signal intensity. As a result, hypotheses that posit a period of increased neural activity in response to a given stimulus or behavioral condition can be evaluated by testing for a positive effect upon fMRI signal intensity. The interpretation of increases in variance (or higher moments) between two conditions, in the absence of any change in mean intensity, is currently less clear.

CONCLUSIONS

A rationale occasionally offered for the use of the KS statistic is that "a consensus does not yet exist regarding the analysis of fMRI data." This is certainly true and is likely to remain the case, as different analytical techniques will be used to address different hypotheses. A consensus should exist, however, as to the requirements that any univariate statistical test of fMRI data should meet. First, such a test should be empirically demonstrated to control type I error in fMRI data. Computer simulations are inadequate, given that actual data may violate assumptions that are implicit in desk-top models. The KS statistic was found here to produce voxelwise and mapwise false-positive rates significantly in excess of tabular values when used to analyze human BOLD fMRI data collected under the null-hypothesis. Second, a statistical test should maximize power. When assumptions of normality are met, parametric statistics are more powerful than nonparametric statistics such as the KS. Third, a statistical test must be capable of supporting the inferences that are of interest to the experimenter. Given the current model of the physiologic basis of the BOLD signal, changes in neural activity are expected to give rise to changes in fMRI signal intensity. The KS test is unable to differentiate these changes from changes in higher distributional moments. We therefore conclude that the KS test is an inappropriate statistical measure for use with BOLD fMRI data. We note, however, that alternative univariate methods do exist that meet these criteria. For example, an implementation of the general linear model for autocorrelated observations introduced by Worsley and Friston (34) has been demonstrated empirically to control the false-positive rate while providing acceptable sensitivity (18, 19).

REFERENCES

9. R. Buckner, P. Bandettini, K. O'Craven, R. Savoy, S. Petersen, M.


