Effect of Temporal Autocorrelation Due to Physiological Noise and Stimulus Paradigm on Voxel-Level False-Positive Rates in fMRI

Patrick L. Purdon* and Robert M. Weisskoff

MGH-NMR Center, Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts, and Harvard-MIT Division of Health-Sciences and Technology, Cambridge, Massachusetts

Abstract: Statistical mapping within a binary hypothesis testing framework is the most widely used analytical method in functional MRI of the brain. A common assumption in this kind of analysis is that the fMRI time series are independent and identically distributed in time, yet we know that fMRI data can have significant temporal correlation due to low-frequency physiological fluctuation (Weisskoff et al. [1993]; *Proc Soc Magn Reson Med* 9:7; Biswal et al. [1995]; *Mag Reson Med* 34:537–541). Furthermore, since the signal-to-noise ratio will vary with imaging rate, we should expect that the degree of correlation will vary with imaging rate. In this paper, we investigate the effect of temporal correlation and experimental paradigm on false-positive rates (type I error rates), using data synthesized through a simple autoregressive plus white-noise model whose parameters were estimated from real data over a range of imaging rates. We demonstrate that actual false-positive rates can be biased far above or below the assumed significance level α when temporal autocorrelation is ignored in a way that depends on both the degree of correlation as well as the paradigm frequency. Furthermore, we present a simple method, based on the noise model described above, for correcting such distortions, and relate this method to the extended general linear model of Worsley and Friston ([1995]; *Neuroimage* 2:173–181). *Hum. Brain Mapping* 6:239–249, 1998. © 1998 Wiley-Liss, Inc.

Key words: fMRI; statistical parametric mapping; noise modeling; colored noise; brain noise

INTRODUCTION

Statistical mapping within a binary hypothesis testing framework is the most widely used analytical method in functional MRI (fMRI) of the brain. In combination with experimental paradigms organized

around a purported functional principle, this method seeks to identify regions of functional specialization within the brain [Friston et al., 1995b]. Typically, a hypothesis test is applied on a voxel-by-voxel basis to fMRI time series, creating a spatially distributed map of significance levels (P values) based on some test statistic (e.g., Student's t -test, Fourier F -test, Kolmogorov-Smirnov test). Inherent in these univariate hypothesis tests against fMRI time series is the assumption that the data are independent and identically distributed in time. In fact, the noise in fMRI data can have significant temporal correlations (or noise “coloration”) due to underlying physiological fluctuations

Contract grant sponsor: National Institute on Drug Abuse; Contract grant number: DA09467; Contract grant sponsor: Whitaker Foundation.

*Correspondence to: Patrick L. Purdon, MGH-NMR Center, 2nd Floor, Building 149, 13th Street, Charlestown, MA 02129. E-mail: ppurdon@mit.edu

Received for publication 12 August 1997; accepted 6 April 1998

[Weisskoff et al., 1993; Biswal et al., 1995]. Furthermore, since the signal-to-noise ratio will vary with imaging rate, we should expect that the degree of correlation due to physiological fluctuation will vary with imaging rate as well. The effect of this temporal correlation on the accuracy of hypothesis testing methods remains an unresolved topic of debate. Some authors argue that correlation due to temporal smoothing outweighs any intrinsic physiological correlation, essentially ignoring any intrinsic physiological correlation in the fMRI time series [Friston et al., 1995a; Worsley and Friston, 1995]. Other authors have attempted to quantify how such intrinsic correlations might bias false-positive rates (type I error) above the assumed P value, but have been limited to studies focused on a single imaging rate and experimental paradigm [Xiong et al., 1996; Zarahn et al., 1997]. However, because the correlation structure may vary with imaging rate, and because a given correlation structure may impose a characteristic time scale upon the data set, we should expect that the accuracy of P values used for inference will depend on both imaging rate and paradigm choice. In this study, using empirically obtained noise data, we numerically estimated the actual false-positive rates brought about by these conditions and determined the accuracy of P values for the Student's t -test, Kolmogorov-Smirnov test, and Fourier-based F -test.

METHODS

In order to investigate the effect of temporal correlation on false-positive rates in a systematic way, we approached the problem as follows: 1) We used a simple noise model to quantify the nature of the temporal correlations in fMRI data sets over a range of imaging rates. 2) We then synthesized activation-free noise data using the parameters determined above and performed statistical analysis on these data to estimate the effect of the noise coloration on false-positive rates.

Quantifying the temporal autocorrelations in fMRI time series: Noise modeling

Noise in fMRI time series has been modeled in a variety of ways. Some authors have taken the view that noise is the result of underlying white noise shaped by the brain's hemodynamic response [Friston et al., 1994], while others have taken a more empirical view in which noise is thought to come from a mechanism separate from the hemodynamic response of activation. To this end, some authors have used traditional time-series methods [Bullmore et al., 1996;

Locascio et al., 1997], while others have used more ad hoc power spectral modeling methods [Zarahn et al., 1997; Lange and Zeger, 1997]. In this paper, we take the empirical view and employ a first-order autoregressive (AR) plus white-noise model. The AR component represents the excess low-frequency noise observed in fMRI time series, while the white-noise component represents scanner noise. Mathematically, a noise time series of this type can be represented as the output of a linear, time-invariant (LTI) system with additive white noise

$$x[n] = w[n] * h[n] + v[n] \quad (1)$$

where $x[n]$ is the noise time series, $*$ denotes the convolution operation, $h[n]$ is the impulse response of the LTI system with discrete-time Fourier transform (DTFT) $H(e^{j\omega}) = (1 - q)/(1 - qe^{-j\omega})$, and $w[n]$ and $v[n]$ are white noises of variance A_C and A_W , respectively. The power spectrum of this AR plus white noise model is given by

$$\begin{aligned} S_{xx}(e^{j\omega}) &= A_W + \frac{A_C(1 - q)^2}{(1 - qe^{-j\omega})(1 - qe^{j\omega})} \\ &= A_W + \frac{A_C}{(1 + 4q(1 - q)^{-2} \sin^2(\omega/2))}. \end{aligned} \quad (2)$$

The A_W and A_C parameters represent the amounts of white and colored noise in the spectrum, respectively, while q represents the degree of correlation between adjacent samples of the AR process. Figure 1 provides a plot of what this spectrum looks like for arbitrary parameter values.

Noise estimation and false-positive rate analysis

We estimated the power spectrum on a voxelwise basis in a visual stimulus data set using a rectangular window periodogram. Subjects were presented a full-field flickering checkerboard stimulus in a "boxcar" paradigm, alternating between 10 sec of stimulus and 15 sec without stimulus for 5 cycles. Data were collected using a General Electric Signa 1.5 T scanner modified by ANMR for Echo-Planar Imaging (EPI). Gradient Echo images, $TE = 50$ msec, were acquired using a quadrature head coil in an oblique plane passing through the visual cortex at TRs of 200, 500, 1000, 2500, and 5000 msec. An ROI was drawn over cortical gray matter, avoiding regions which contained an obvious activation signal. The average power spectrum over this ROI was then fitted to Equation (2), using an iterative nonlinear least-squares method.

Figure 1.

Example of AR plus white-noise power spectrum (parameter values $q = 0.65$, $A_W = 1.8$, $A_C = 3.6$).

Using the estimates of q , A_C , and A_W from the empirical studies, 64×64 images were synthesized with noise parameters analogous to TRs of 5000, 2500, 1250, 625, and 312.5 msec, without spatial correlation. Parameters for TRs of 1250, 625, and 312.5 msec were obtained by linear interpolation between the actual measured values for A_C and A_W , and by direct evaluation of $q = \exp(-TR/\tau)$ with $\tau = 15$ sec as obtained from the empirical data (see Results). The TR times were chosen so that the overall simulated experiment time would be the same (160 sec) for each TR value at power-of-two data lengths (facilitating use of the Fourier F-test based on the commonly used power-of-two FFT algorithm [Press et al., 1992]). These “null” data sets were then analyzed with the t-test, Kolmogorov-Smirnov (KS) test, and Fourier F-test with assumed off-on stimulus paradigms of 0.025 Hz and 0.05 Hz, corresponding to assumed off-on periods of 40 sec and 20 sec, respectively. Specifically, for the t-test and KS test, the “off” and “on” samples were considered as two separate groups and then compared using the commonly available “C” computer code for the t- and KS tests [Press et al., 1992]. This calculation was done on a voxel-by-voxel basis to generate a P value for each voxel.

While the Fourier F-test is not as common as the t- or KS tests, we describe it in this paper both because of its potential usefulness in fMRI analysis in general and because it has an intuitive interpretation that will be useful in understanding the results to follow. The Fourier F-test is used to detect periodic signals in a

background of white noise. In a repeating block designed fMRI experiment, most of the signal power in the BOLD activation signal will be contained in the fundamental paradigm frequency, so in this context we use the Fourier F-test to detect a sinusoidal signal at the paradigm frequency. The intuition behind this test is that we are comparing the power in the frequency of interest to the average power in the other frequencies, rejecting the null hypothesis when the signal at the frequency of interest has much more power than the average power of the other frequencies. The Fourier F-test is performed by computing the periodogram $I(\omega_k)$ (i.e., a simple FFT-based estimate of the power spectrum) of a single-voxel time series and constructing the following F-statistic:

$$F = \frac{I(\omega_p)}{\left[\sum_{k=0}^{N-1} I(\omega_k) - I(0) - 2I(\omega_p) \right] / (N-3)} \sim F(2, N-3)$$

$$I(\omega_k) = N^{-1} \left| \sum_{n=0}^{N-1} x[n] \exp(-j\omega_k n) \right|^2, \quad \omega_k = 2\pi k/N \quad (3)$$

where N is the total number of data points, $F(2, N-3)$ is an F distribution with 2 and $N-3$ degrees, and ω_p is the frequency of the paradigm, which in this expression must be a multiple of $2\pi/N$ (although a more general expression does exist for cases where this is not true) [Brockwell and Davis, 1991].

The result of the individual tests is an expected false-positive probability, i.e., the probability that the test statistic could have been that large or larger by chance just from noisy data. Following the brain-mapping tradition, we will refer to this value as the “ P value”. If we simulate data for N pixels, for an arbitrary significance level, α , one would expect αN of those pixels to have P values less than or equal to α . Thus for an activation-free data set, once we choose α , all pixels with $P < \alpha$ are false-positive activations. So, to test the accuracy of these statistics when the noise is not white, we inspected the resulting P -value maps for each test for false positives over a continuum of assumed significance levels. At each assumed significance level α , the number of pixels with $P < \alpha$ were counted and divided by the total number of pixels in the image to estimate the false-positive rate. Ten separate 64×64 data sets for each TR were synthesized and analyzed in this fashion with each of the above statistical tests. The false-positive rates for each TR and statistical test were averaged and plotted against the assumed significance level α to create the false-positive characteristic (FPC). Data that meet the

TABLE I. Parameter estimates for first-order autoregressive plus white noise model

TR	A_w	A_c	q	ARpow	ARpow/ A_w
200 msec	0.1	0.3	0.98	0.003	0.030
500 msec	0.2	1.0	0.96	0.020	0.102
1,000 msec	0.4	2.3	0.93	0.083	0.209
2,500 msec	1.0	3.5	0.85	0.284	0.284
5,000 msec	1.8	3.6	0.65	0.764	0.424

assumptions of the given statistical test would have false-positives rates equal to α , resulting in a linear FPC with unity slope, while those which violate the assumptions of the test would have a nonlinear relationship between false-positive rate and α (i.e., the false-positive rate is biased).

RESULTS

In this section we present the results of the noise analysis, illustrating how the noise characteristics change with TR. We then present the results of the statistical analysis, demonstrating that the false-positive rates deviate from the assumed significance levels in a way which depends on both the TR and the paradigm frequency.

Noise estimation

Parameter estimates for q , A_c , and A_w are shown in Table I for the original TR values. The q parameter (and hence the degree of correlation) increased with the imaging rate, behaving as if $q = \exp(-TR/\tau)$ with $\tau \approx 15$ sec for all TR's, suggesting that there may be an underlying continuous-time decay process that creates the noise correlation, consistent across imaging rates. In addition to the above noise parameters, Table I also shows the calculated AR power, the variance in the noise attributable to the AR process (i.e., $1/(2\pi) \times$ the integral over one cycle of the AR term in Equation (2)), and a comparison of this to the white noise variance A_w . The ratio of AR power to white noise power increased with TR, consistent with the notion that the AR noise estimated is related to an actual physiological process whose signal-to-noise ratio increases with the TR.

False-positive analysis

The FPCs for each TR and statistical test are shown below in Figures 2–7, constructed by averaging the FPCs for each of the 10 trials conducted under each TR

and paradigm period. The sample variance for each of the FPC curves was less than 8×10^{-3} over all values of α in all cases. Note that since the Kolmogorov-Smirnov statistic D has a discrete-valued probability distribution, its FPC is piecewise constant. In general, at the low-frequency (40-sec) paradigm, the FPC tends to bow upwards, indicating that there are more false positives than expected from the assumed significance level α . For instance, for the Fourier F-test, at an assumed significance level of $\alpha = 0.05$ and at a TR of 625 msec, the actual false-positive rate is 0.16, three times greater than the expected value given by α . For the t-test and KS test under the same situation, the actual false positives are 0.12 and 0.1, respectively, roughly twice as great as the assumed α . As we move to smaller α , the bias in the false-positive rate becomes much worse relative to the assumed α . For instance, at an α of 0.02 for the Fourier F-test at a TR of 625 msec, the false-positive rate is approximately 0.095, nearly five times the expected value, with similar results for the KS test and t-test. At the high-frequency (20-sec) paradigm, there are uniformly fewer false positives than in the lower-frequency paradigm, resulting in fewer false positives than the assumed α in some cases. For instance, for the Fourier F-test at a TR of 5000 msec and at $\alpha = 0.05$, the low-frequency paradigm gives a false-positive rate of 0.06, whereas for the high-frequency case the false-positive rate is 0.03, a bit more than half the assumed α . At a given paradigm frequency, false positives tend to decrease with increasing TR, with the exception of TR = 312 msec at the low-frequency paradigm.

Across both TR and paradigm frequencies, all three tests show both similar trends and quite similar actual biases. That is, the FPC curves for the three different tests were very similar at each paradigm/TR combination.

DISCUSSION

Our simulation demonstrates that the disparity between the actual false-positive rate and the assumed significance level depends on both the imaging rate and the paradigm frequency. In this section, we 1) provide some interpretations for the behavior of the FPC curves, 2) comment on aspects of the noise modeling, 3) describe a simple way of correcting the false-positive bias, and 4) relate this method to an existing method based on the general linear model of Worsley and Friston [1995].

Figure 2.

False-positive characteristic (FPC) for Fourier F-test under low-frequency paradigm (40-sec period).

Figure 4.

FPC for Student's t-test under low-frequency paradigm (40-sec period).

Interpreting the behavior of the FPC

The varying behavior of the FPC at a given paradigm frequency due to different TRs depends upon the specific noise present at a given TR. Conceptually, however, we can describe the dependence of the FPC on the imaging rate in terms of the effective number of degrees of freedom [Worsley and Friston, 1995]. As the imaging rate and the degree of correlation increase (since the AR correlation parameter q increases with

imaging rate), the effective number of degrees of freedom decreases relative to that of the assumed sampling distribution, resulting in an increase in the actual false-positive rate with imaging rate by this account.

The influence of the paradigm frequency can be understood in an intuitive way by considering the two limiting cases at a given TR: 1) a paradigm where the first half of the experiment has no stimulus, while the stimulus is applied continuously in the second half

Figure 3.

FPC for Fourier F-test under high-frequency paradigm (20-sec period).

Figure 5.

FPC for Student's t-test under high-frequency paradigm (20-sec period).

Figure 6.

FPC for Kolmogorov-Smirnov test under low-frequency paradigm (40-sec period).

(i.e., one period of a boxcar paradigm), and 2) a paradigm where the stimulus is “off” for one time sample and “on” for the next (i.e., the highest frequency paradigm that we are able to sample). Suppose we attempt to analyze activation-free data for differences in the mean (i.e., a t-test), with known noise parameters. In the first scenario, we will see more false positives than the assumed significance value, due to a reduction in the effective degrees of freedom, as described earlier (i.e., an FPC bowing upwards). How-

Figure 7.

FPC for Kolmogorov-Smirnov test under high-frequency paradigm (20-sec period).

ever, for the second scenario, the high degree of correlation means that neighboring samples differ by only a small fraction plus a (relatively) small white-noise component, and hence the resulting “off” and “on” data sets will be very similar. The difference in the sample means of the “off” and “on” populations will thus be very small, leading us to detect fewer false positives than expected (i.e., an FPC bowing down).

For paradigms of arbitrary frequency, we can develop an approximate frequency domain relationship with a similar interpretation (See Appendix A for derivation). The end result is that the variance of the difference in sample means will be determined principally by the noise power at the fundamental paradigm frequency:

$$\text{var} [|\mu_{\text{on}} - \mu_{\text{off}}|^2] \propto S_{xx}(e^{j\omega_p}). \quad (4)$$

An assumption of white noise corresponds to assuming that the power spectrum of the noise is flat, with a value equal to the average value of $S_{xx}(e^{j\omega})$ over any interval of 2π . The power spectrum of the first-order AR noise plus white noise described in Equation (2) has an approximate $1/\omega^2$ dependence, seen by taking a small-angle approximation on $\sin(\omega/2)$, so that at low frequencies it is larger than average and at high frequencies it is smaller than average. Hence, for low frequencies we will tend to underestimate the variance of the difference in means, resulting in more false positives than expected. As we increase the paradigm frequency, the actual variance approaches and slides below the average value, so we will tend to detect fewer and fewer false positives, at some point detecting fewer false positives than expected. Figure 8 provides an illustration of this.

The paradigm dependence of the Fourier-based F-test can be seen in a similar way by directly examining the formula for the F-statistic from Equation (3). Its denominator can be interpreted as the average of the power spectral components minus those at DC (zero frequency) and $\pm\omega_p$, while the numerator can be thought of as the power-spectral density at the paradigm frequency. Thus, the Fourier-based F-statistic compares the power at the paradigm frequency to the average power in all frequencies (except DC). As with the t-test, given the $1/\omega^2$ power spectrum of first-order AR noise, low-frequency paradigms will result in more false positives than expected, since power at low frequency will be greater than average, while high-frequency paradigms will result in fewer false positives than expected, since power at high frequency will be lower than average.

Comments on noise modeling

Modeling the power spectrum as a rational function as in Equation (2), while arbitrary, has a number of advantages over other (also arbitrary) methods proposed previously for fMRI time series [Friston et al., 1994; Zarahn et al., 1997]. First, we can easily synthesize simulated data of the appropriate shape by using a linear, constant-coefficient difference equation implementation of Equation (2) [Oppenheim and Schaffer, 1989], facilitating a Monte Carlo study like this one aimed at understanding the basic relationships between temporal correlation, experimental paradigm, and P values in statistical maps. Second, this model, written in the discrete time domain, properly accounts for the fact that the fMRI time series are sampled data subject to aliasing. In contrast, other authors [Friston et al., 1994; Zarahn et al., 1997] have essentially chosen to model the power spectrum in the continuous-time domain (i.e., ignoring the inherent periodicity present in the frequency spectra of any sampled data), an approximation which would require additional steps or assumptions before such a model could be applied properly to the power spectrum of sampled data (from an FFT, for instance).

As described in the previous section, the noise model used in this study has an approximate $1/\omega^2$ dependence on the power spectrum. This $1/\omega^2$ dependence in the *power* spectrum is analogous to the “ $1/f$ ” dependence in the *magnitude* spectrum used by Zarahn et al. [1997] and is characteristic of noise spectra produced by filtering white noise with some form of first-order linear low-pass filter. In contrast, the term “ $1/f$,” as used in the nonlinear dynamics and complex systems literature, refers to the $1/f$ *power* spectrum observed in complex systems exhibiting self-similar dynamics. Since the brain certainly qualifies as a complex system, such self-similar dynamics may well exist when observed over very long time scales, but for now we should make a distinction between the “ $1/f$ ” noise in complex systems and the simpler $1/\omega^2$ LTI low-pass filtered noise described here and in Zarahn et al. [1997].

Correcting false-positive bias: A “whitening” filter

We can see from the previous discussion that a simple degrees-of-freedom correction is not sufficient to compensate for all the effects of this correlated noise, since the paradigm itself also affects the underlying null distribution. However, it may be possible to correct the false-positive bias by removing the temporal correlation with a “whitening filter.” The basic idea

is to filter the fMRI time series in such a way that the power spectrum of the noise becomes flat, i.e., we are forcing the noise to become “white.” For example, if the noise in fMRI time series is well-described by a rational power spectrum, such as in Equations (1) and (2), we can apply a whitening filter consisting of the inverse of the minimum-phase spectral factor of the noise power spectrum in Equation (1) [Papoulis, 1991] (see Appendix B for a detailed development):

$$S_{xx}(z) = \frac{A_c(1-q)^2 + A_w(1 - qz^{-1})(1 - qz)}{(1 - qz^{-1})(1 - qz)} = \frac{1}{H_w(z)H_w(z^{-1})} \quad (5)$$

For data containing an activation signal, the resulting “whitened” signal is given by

$$\begin{aligned} x[n] &= \zeta[n] + a[n] \\ \tilde{x}[n] &= x[n] * h_w[n] = \zeta[n] * h_w[n] + a[n] * h_w[n] \\ &= \zeta[n] * h_w[n] + \tilde{a}[n] \\ \tilde{a}[n] &= a[n] * h_w[n] \end{aligned} \quad (6)$$

where $\zeta[n]$ is the colored noise, $a[n]$ is the activation signal (i.e., the assumed neuronal response smoothed by the hemodynamic response), $h_w[n]$ is the impulse response of the whitening filter, $\zeta[n] * h_w[n]$ is the whitened noise, and $\tilde{a}[n]$ is the modified activation signal. Thus, following the whitening step, signal detection (i.e., correlation analysis or regression) would be done based on the $\tilde{a}[n]$ signal. For instance, a t-test is essentially equivalent to regression against a square wave of the appropriate frequency and duty cycle, so in this whitening framework one would assume a square wave for $a[n]$ and use the appropriate $\tilde{a}[n]$ for regression. In the case of the Fourier-based F-test, this correction corresponds simply to normalizing the periodogram components by the power-spectral density of the underlying noise. Note that in a linear regression context, this whitening filter is equivalent to a weighted least-squares estimate, using the (temporal) covariance matrix implied by the autocorrelation function corresponding to Equation (2) (i.e., the inverse DTFT of $S_{xx}(e^{j\omega})$) as our weighting matrix. The advantage here is that, again because of the choice of a rational DTFT domain model of the power spectrum, we can easily implement Equation (6) with a linear, constant-coefficient difference equation which is $O(N)$ in computational complexity, compared to the general $O(N^3)$ complexity of the matrix inversion inherent in weighted

Figure 8.

Illustration of relationship between stimulus paradigm frequency and distortion of P values.

least squares. A plot comparing the FPCs of whitened and unwhitened data for the $TR = 312$ msec data set is given in Figure 9. Note how the FPC of the whitened data closely matches that of the ideal FPC.

Comparisons and connections with the extended general linear model of Worsley and Friston [1995]

The extended general linear model (E-GLM) of Worsley and Friston [1995] has been used by some authors [Zarahn et al., 1997] as a means of correcting P -value distortions. The E-GLM method, taken as explicitly stated in Worsley and Friston [1995], does not address the issue of correcting distortions due to intrinsic physiological correlation, though it does provide the correct expressions to account for temporal smoothing imposed during postprocessing. However, following the work by Zarahn et al. [1997], we can make a simple modification to the GLM method which will account for the temporal autocorrelation. In what follows we briefly review the original E-GLM, we describe the modification and demonstrate its ability to correct for P -value distortion, and, finally, we compare this modified E-GLM method to that of the whitening-filter method described earlier.

The original E-GLM postulates a linear model $X = G\beta + e$, where X is the data, G is a matrix of postulated covariate waveforms, β is the covariate coefficient vector, and e is Gaussian white noise of variance σ^2 . The data X are then smoothed with a matrix K whose

rows consist of the hemodynamic response function, yielding an equation $KX = G^*\beta + Ke$, where $G^* = KG$, which yields a nonoptimal least-squares solution for the estimator of B , $b = (G^{*T}G^*)^{-1}G^{*T}KX$ [Worsley and Friston, 1995]. The modification consists of 1) replacing the hemodynamic response function with the “shaping filter” of the noise (the inverse of the whitening filter or, equivalently, the impulse response of $1/H_w(e^{j\omega})$) in K , and then 2) replacing KX with X_C , the observed physiologically correlated data, in the expressions for the estimator and residual vectors (yielding $b = (G^T G^*)^{-1} G^T X_C$ and $r = R X_C$). Their expression for the effective degrees of freedom is then used to obtain the voxelwise P values. The rationale here is that the observed data X_C are already temporally correlated, i.e., they have already been operated on by the shaping filter K , so we need not operate on them again. The FPC generated by this modified extended general linear model (ME-GLM), applied to $TR = 312$ msec data, is shown in Figure 10. This FPC is nearly identical to the ideal FPC, with performance comparable to the whitening-filter method. In this scheme, we are regressing against covariates G filtered by the shaping filter K (i.e., $G^* = KG$), a by-product of the original E-GLM method which may or may not be desired.

While both the ME-GLM and whitening filter produce good corrections for false-positive bias, with a speed advantage in favor of the whitening filter, the ME-GLM method does not provide an optimal solution to the problem posed [Friston and Worsley, 1995]. This nonoptimal formulation was chosen by Worsley

Figure 9.

FPC for whitening-filter method from a $TR = 312$ -msec data set.

Figure 10.

FPC for ME-GLM from a TR = 312-msec data set.

and Friston [1995] to increase robustness of the solution, since the functional form for the K matrix used in their framework would have resulted in an ill-conditioned matrix inversion if a fully optimal weighted least-squares solution had been chosen. However, the AR plus white-noise model used in this study does not suffer from this problem, since the white-noise term prevents the spectrum from ever decaying to zero, and can provide the optimal weighted least-squares solution. In preliminary simulations, we have found nearly identical robustness between these two methods.

There is an important conceptual difference between the original E-GLM and the whitening and ME-GLM methods presented here. The former considers noise correlations to come from the hemodynamic response function, while the latter methods take a more empirical view of the underlying noise. While it would be convenient if the underlying noise in the MR imaging

process were temporally shaped by the hemodynamic response function, in practice the noise includes both white components and correlated components with time constants that seem longer than the hemodynamic response function. As shown above, the strategy of using the ME-GLM model can be effective in correcting distortions in the false-positive rate, but the appropriate K matrix must be determined from the actual noise, not the hemodynamic response. Zarahn et al. [1997] described further elaborations to the E-GLM which use empirical estimates of noise to account for temporal autocorrelation along with a separate hemodynamic smoothing kernel to shape the covariate waveforms. We omit a detailed analysis of such methods for brevity, but we suggest that it should yield results similar to the ME-GLM. Table II provides a summary comparison of how the various methods described here handle the issue of noise modeling and hemodynamic smoothing of covariate waveforms.

Finally, it is important to point out that the specific coefficients for the AR process, which characterize an underlying biological variation and not instabilities in the scanner, depend on spatial resolution, pulse sequence, field strength, etc., and may even vary strongly between individual subjects and regions of the brain. As a result, the specific distortions in false-positive rates should not be generalized beyond the examples described.

CONCLUSIONS

We have shown that when temporal autocorrelation in fMRI data sets is ignored, the voxel-level P values which result are seriously distorted, in a way that depends upon both imaging rate and paradigm choice. Furthermore, we have developed a simple method for correcting these P -value distortions and have used the intuition behind this method to suggest a modification

TABLE II. Comparison of features in various fMRI analysis methods

	Methods			
	E-GLM	ME-GLM	Zarahn et al. [1997]	Whitening filter and related methods
Source of underlying noise	Shaping from hemodynamic response function	Empirical noise model	Empirical noise model	Empirical noise model
Smoothing of covariate waveforms	Smoothing by hemodynamic response	Smoothing by noise-shaping filter	Smoothing by hemodynamic response	Smoothing by hemodynamic response
Uses optimal estimator?	No	No	No	Yes

to the modified general linear model that can also correct the P -value distortions. Because the accuracy of P -value assignment is essential to the technique of statistical mapping, it is important that the effects of this temporal autocorrelation are properly accounted for while creating such maps.

ACKNOWLEDGMENTS

The authors thank Robert Savoy and Kathy O'Craven for their help in acquiring the human data. P.L.P. is supported by the Whitaker Foundation.

REFERENCES

- Aguirre GK, Zarahn E, D'Esposito M (1997): The Kolmogorov-Smirnov (KS) statistic fails to control type I error in the analysis of BOLD fMRI data. *Magn Reson Med* (submitted).
- Biswal B, Yetkin FZ, Haughton VM, Hyde JS (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med* 34:537–541.
- Brockwell PJ, Davis RA (1991): *Time Series: Theory and Methods*, New York: Springer-Verlag, pp 330–336.
- Bullmore E, Brammer M, Williams SCR, Rabe-Hesketh S, Janot N, David A, Mellers J, Robert H, Sham P (1996): Statistical methods of estimation and inference for functional MR image analysis. *MRM* 35:261–277.
- Friston KJ, Jezzard P, Turner R (1994): Analysis of functional MRI time-series. *Hum Brain Mapp* 1:153–171.
- Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SCR, Frackowiak RSI, Turner R (1995a): Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
- Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSI (1995b): Statistical parametric maps in functional imaging: A general linear approach. *Hum Brain Mapp* 2:189–210.
- Lange N, Zeger S (1997): Non-linear Fourier time series analysis for human brain mapping by functional MRI. *J R Stat Soc [C] Appl Stat* 46:1–26.
- Locascio JJ, Jennings PJ, Moore CI, Corkin S (1997): Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Hum Brain Mapp* 5:168–193.
- Oppenheim AV, Schaffer RW (1989): *Discrete-Time Signal Processing*, NJ: Prentice-Hall.
- Papoulis A (1991): *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, pp 401–404.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992): *Numerical Recipes in C: The Art of Scientific Computing*, New York: Cambridge University Press.
- Weisskoff RM, Baker J, Belliveau J, Davis TL, Kwong KK, Cohen MS, Rosen BR (1993): Power spectrum analysis of functionally-weighted MR data: What's in the Noise? *Proc Soc Magn Reson Med* 1:7.
- Worsley KJ, Friston KJ (1995): Analysis of fMRI time-series revisited—Again. *Neuroimage* 2:173–181.
- Xiong J, Gao JH, Lancaster JL, Fox PT (1996): Assessment and optimization of functional MRI analyses. *Hum Brain Mapp* 4:153–167.
- Zarahn E, Aguirre GK, D'Esposito M (1997): Empirical analyses of BOLD fMRI statistics: I. Spatially unsmoothed data collected under null-hypothesis conditions. *Neuroimage* 5:179–197.

APPENDIX A

We wish to derive an approximate frequency-domain relationship to illustrate the relationship between stimulus paradigm frequency and FPC distortion. Let us define a paradigm waveform $p[n]$ consisting of a square wave that has a value of -1 for samples corresponding to no stimulus (call this group “Off”) and a value of 1 for samples corresponding to a stimulus (call this group “On”). For simplicity, let $p[n]$ have zero mean (i.e., the number of samples in the Off group is equal to that in On group). The expected variance of the difference in the sample means of the Off and On groups can be expressed in the frequency domain as an integral over the noise spectrum, $S_{xx}(e^{j\omega})$, weighted by the frequency content of the paradigm:

$$\begin{aligned}\mu_{\text{on}} - \mu_{\text{off}} &= \frac{1}{N/2} \sum_0^{N-1} x[n]p[n] \\ &= \frac{1}{\pi N} \int_{-\pi}^{\pi} X(e^{j\omega})P^*(e^{j\omega})d\omega \quad (\text{Parseval's relation}) \\ |\mu_{\text{on}} - \mu_{\text{off}}|^2 &= \frac{1}{(\pi N)^2} \int_{-\pi}^{\pi} X(e^{j\omega})P^*(e^{j\omega})d\omega \int_{-\pi}^{\pi} X^*(e^{j\lambda})P(e^{j\lambda})d\lambda \\ E[X(e^{j\omega})X^*(e^{j\lambda})] &\approx S_{xx}(e^{j\lambda})\delta(\omega - \lambda) \quad \text{for large } N \\ \therefore \text{var}[|\mu_{\text{on}} - \mu_{\text{off}}|^2] &= E[|\mu_{\text{on}} - \mu_{\text{off}}|^2] \approx \frac{1}{(\pi N)^2} \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|P(e^{j\omega})|^2 d\omega \quad (7)\end{aligned}$$

where μ_{on} and μ_{off} are defined as before, $X(e^{j\omega})$ and $P(e^{j\omega})$ are discrete-time Fourier transforms of $x[n]$ and $p[n]$, respectively, $*$ denotes the complex conjugate operation, $S_{xx}(e^{j\omega})$ is the power spectrum of $x[n]$, and $\delta(\omega)$ is the Dirac delta function. Since the paradigm waveform is a square wave, for large N it can be approximated as a series of delta functions whose fundamental component corresponds to the paradigm frequency. If we use the fact that the higher-order Fourier coefficients of $p[n]$ are much smaller than the fundamental, we can approximate the result of Equation (7) as

$$\text{var}[|\mu_{\text{on}} - \mu_{\text{off}}|^2] \propto \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\delta(\omega - \omega_p)d\omega = S_{xx}(e^{j\omega_p}). \quad (8)$$

APPENDIX B

Expressing Equation (2) as a Z-transform, factoring it, and inverting it, we have

$$\begin{aligned}
 S_{xx}(z) &= \frac{A_C(1-q)^2 + A_W(1-qz^{-1})(1-qz)}{(1-qz^{-1})(1-qz)} = \frac{K(1-\gamma z^{-1})(1-\gamma z)}{(1-qz^{-1})(1-qz)} = \frac{1}{H_w(z)H_w(z^{-1})} \\
 H_w(z) &= \frac{\sqrt{K}(1-qz^{-1})}{(1-\gamma z^{-1})} \\
 \gamma &= \frac{D}{2} \pm \left(\frac{D^2}{4} - 1 \right)^{1/2}, \quad D = \left(\frac{(A_C + A_W) - 2A_C q + (A_C + A_W)q^2}{A_W q} \right) \\
 K &= A_W q / \gamma
 \end{aligned} \tag{9}$$

where $H_w(z)$ is the whitening filter and γ , D , and K are derived variables as given above. Note that γ is a pole of $H_w(z)$ and must be chosen to be stable (i.e., magnitude less than one) in this context.