# Statistical Methods of Estimation and Inference for Functional MR Image Analysis

Edward Bullmore, Michael Brammer, Steve C. R. Williams, Sophia Rabe-Hesketh, Nicolas Janot, Anthony David, John Mellers, Robert Howard, Pak Sham

**Two questions arising in the analysis of functional magnetic resonance imaging (fMRI) data acquired during periodic sensory stimulation are: i) how to measure the experimentally determined effect in fMRI time series; and ii) how to decide whether an apparent effect is significant. Our approach is first to fit a time series regression model, including sine and cosine terms at the (fundamental) frequency of experimental stimulation, by pseudogeneralized least squares (PGLS) at each pixel of an image. Sinusoidal modeling takes account of locally variable hemodynamic delay and dispersion, and PGLS fitting corrects for residual or endogenous autocorrelation in fMRI time series, to yield best unbiased estimates of the amplitudes of the sine and cosine terms at fundamental frequency; from these parameters the authors derive estimates of experimentally determined power and its standard error. Randomization testing is then used to create inferential brain activation maps (BAMs) of pixels significantly activated by the experimental stimulus. The methods are illustrated by application to data acquired from normal human subjects during periodic visual and auditory stimulation.**

**Key words: time series; functional MRI; statistical mapping; regression.**

## INTRODUCTION

Functional magnetic resonance imaging (fMRI) is to structural MRI as movies are to still photography: functional MR images, like movies, show events unfolding longitudinally in time; whereas still or structural images are instantaneous. Each element in the 2-dimensional matrix of a typical functional MR image is a point in a time series. If the size of the matrix is, e.g., 128 × 64, the image is effectively comprised of 8192 series of digitized time points, $t = 1,2,3,\ldots, N$, all of an identical length, $N$, dictated by the experimental design and scanning parameters used to acquire the image.

An experimental design that has been widely used since fMRI first became available to neurobiologists and clinicians in the early 1990s is periodic sensory stimulation. In this, the experimenter exposes the subject to a regularly periodic sensory stimulus and hopes to detect a related periodic pattern in the fMRI time series recorded from those cortical regions involved in processing sensory input of the experimental modality. Two general questions of data analysis that arise from such studies are: i) what is the best way to measure temporal change in the fMRI signal apparently related to perception of the experimentally determined stimulus? and ii) how should we decide whether any such measured change is significant or not?

One simple way to estimate the experimental effect is to average images acquired during the ON and OFF periods of sensory stimulation, then subtract the average ON image from the average OFF image; thus estimating the mean ON-OFF difference in signal intensity at each pixel. An equivalent approach is to estimate the cross correlation between the time series observed at each pixel and the square or "box-car" waveform representing experimentally determined ON-OFF change in conditions (also known as the input or contrast function). Several groups have used one or other of these fundamentally identical methods to demonstrate (as anticipated) a relatively large experimental effect on $T_2$*-weighted signals recorded from temporal cortex during periodic auditory stimulation (1), and occipital cortex during periodic visual stimulation (2). However, it is clear that this general approach to estimation entails loss of power to detect activated pixels, chiefly because it ignores the inevitable delay (in the order of 5–8 s) between neuronal activation and 90% maximal hemodynamic response (3); in other words, stimulus-related changes in signal intensity will not begin and end at the same times as stimulus presentation.

More sophisticated methods of estimation have been proposed to address this problem of a hemodynamically modulated response to stimulation. For example, Bandettini *et al.* (3) adjusted the phase, $\phi$, of the contrast function by an estimate of hemodynamic delay; then estimated the cross correlation between fMRI time series and the phase-adjusted contrast function. A comparable but more formal approach has been introduced by Friston *et al.* (4). These authors modeled the hemodynamic response to neuronal activation as a Poisson function parameterized by a global estimate of temporal smoothness, $\lambda$, in the fMRI time series; and proposed that cross correlation between fMRI time series and the box-car input function should properly be estimated only after the latter has been convolved with this hemodynamic response function. In short, to persevere with cross correlation as a measure of experimental effect, one must somehow "undo" hemodynamic modulation of a presumably instantaneous neuronal response to the experimentally determined input function—either by adjusting

MRM 35:261-277 (1996)

From the Departments of Biostatistics & Computing (E.B., M.B., S.R-H., N.J., P.S.), Psychological Medicine (E.B., A.D., J.M., P.S.), Neuroscience (E.B., M.B.), and Neurology (S.C.R.W.), Institute of Psychiatry; and the Department of Old Age Psychiatry (R.H.), the Maudsley Hospital, London, United Kingdom.

Address correspondence to: Edward Bullmore, Department Neuroscience, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, U.K.

Received March 30, 1995; revised August 14, 1995; accepted August 18, 1995.

E.B. is supported by a Wellcome Trust Clinical Research Training Fellowship. P.S. is supported by a Wellcome Trust University Award (Senior Lecturer).

0740-3194/96 $3.00

the phase of the input function by an (*a priori*) estimate of hemodynamic delay (3); or by adjusting both phase and shape of the input function by convolution with a point spread function (4).

Of course, whatever method is used to estimate the experimental effect on fMRI time series, it remains to decide whether or not an observed effect is significant; in other words, whether or not the pixel at which that effect was observed should be regarded as activated by the experimental stimulus. It is possible to make such a decision simply by comparing the size of the experimental effect observed at a given pixel to some arbitrarily large value. For example, one can decide that a pixel represents activated brain tissue if the cross correlation coefficient estimated at that pixel, $r_i$, is greater than, say, 0.25, 0.5, or 0.75 (3); but this sort of decision has the obvious disadvantage that it imparts no sense of how likely we are to be mistaken if we believe it. Perhaps for this reason, a probabilistic approach has been preferred by other groups. Probabilistic decision making (e.g., as in ref. 4, on the basis of estimated $r$) typically involves three steps: i) estimating the test quotient (e.g., $rQ_i = r_i$ divided by its standard error $\mathrm{SE}(r_i)$) at each pixel in the image; ii) referring the observed values of the test quotient (e.g., $rQ_i$) to its sampling distribution under a null hypothesis; and iii) deciding that a pixel is not activated unless the probability of its test quotient under a null hypothesis (e.g., $p\ (rQ \geqslant rQ_i\,/H_o)$) is less than an arbitrarily small level, $\alpha$.

These principles may be generalized to probabilistic decision making on the basis of any estimated measure of experimental effect; but, in any case, accurate specification of the null distribution will be of crucial importance. If we can be sure that the observed values of the general test quotient, $\Pi Q_i$, are identically distributed under a null distribution of theoretically known form, e.g., normal, then it is not difficult to obtain from standard tables a critical value, CV, for a test of size $\alpha$, and accept that any observed value of the test quotient greater than CV has a probability under the null hypothesis less than $\alpha$. However, theoretical asymptotic distributions will not always be available or sufficiently accurate for significance testing in fMRI analysis (5). Groups working on positron emission tomography (PET) data analysis have recently used nonparametric or distribution free methods, such as randomization or Monte Carlo simulation, to ascertain critical values for testing the significance of activated pixel clusters (6, 7). Especially in the context of testing large image data sets, the theoretical approach has a clear advantage in terms of speed; but the computationally more intensive Monte Carlo or randomization methods have generic advantages of directness, robustness, and versatility. To paraphrase a remark by R. A. Fisher (cited in refs. 8 and 9), randomization is tedious but the results obtained by theory are valid only insofar as they are corroborated by this elementary method.

In this paper, we investigate: i) time series regression modeling to estimate the experimental effect at each pixel of functional MR images acquired during periodic sensory stimulation; and ii) randomization testing to decide which pixels are significantly activated by the experimental stimulus. We also include some thoughts on future developments.

## IMAGES

### Functional MR Image Acquisition

The data reported in this study were acquired by echoplanar imaging (EPI) using a GE Signa system (General Electric, Wisconsin) controlled by an Advanced NMR operating console (Advanced NMR, Massachusetts). One hundred $T_2$*-weighted MR images (*TE* 40 ms; *TR* 3 s) depicting BOLD contrast (10) were acquired at a field strength of 1.5 Tesla from 10, 5-mm thick, contiguous slices, with an in-plane resolution of 3 mm. Each 2D image matrix was comprised of 128 × 64 pixels, each of which had a 16-bit integer value for signal intensity.

Pilot scans in three orthogonal planes were used to define the plane of image acquisition. For experiments involving visual stimulation and the null experiment (see below), the plane of acquisition was near axial, parallel to the line of the calcarine fissure. For the experiment involving auditory stimulation, the plane of acquisition was again near axial, parallel to the line of the Sylvian fissure. All images are presented so that the left side of the image corresponds to the right side of the brain.

### Experimental Designs

Images were acquired from normal volunteer subjects under the following experimental designs:

*Photic Stimulation.* The ON condition was 30 s of 8 Hz pattern-flash photic stimulation *via* light proof stimulating goggles (model GRASS SV100); the OFF condition was 30 s of darkness. In all, five ON-OFF cycles were presented in the course of image acquisition over 5 min.

*Visual Hemifield Stimulation.* During photic stimulation (as above), the subject was wearing contact lenses, which restricted his field of vision to a unilateral hemifield (11). Two images were acquired: one with the subject exposed to visual stimulation only in his right hemifield; the other with the subject exposed to visual stimulation only in his left hemifield.

*Visual Perception of Motion.* The ON condition was 30 s of visual exposure to an animated (cartoon) film; the OFF condition was 30 s of exposure to a single frozen frame of the film. Five ON-OFF cycles were presented over the course of 5 min.

*Bimodal Stimulation.* Two ON conditions were presented at different frequencies in the course of the same experiment. One ON condition was 21 s of auditory exposure to the sound of the experimenter reading aloud from a novel; the corresponding OFF condition was 21 s of silence. The other ON condition was 27 s of photic stimulation (as above); the corresponding OFF condition was 27 s of darkness.

*Null.* In this experiment, there was no alternation between ON and OFF conditions. The subject was simply asked to lie quietly in the scanner while images were acquired in the usual way over the course of 5 min.

### Image Registration

A fully automated and objective method of image registration was used to estimate the extent of rigid motion in two spatial dimensions $\{x,y\}$ during image acquisition (12). The method finds (by a nonlinear search algorithm (13)) estimates for translation and rotation in $\{x,y\}$ that minimize the total absolute difference in gray scale values between each 2D (match) image acquired at a given point in time, $\text{fMRI}_t$, and the mean 2D (base) image created by averaging all (100) $\text{fMRI}_t$'s over time. The maximum extent of $\{x,y\}$ translation identified in any of our images was less than 0.5 mm; and the maximum angle of $\{x,y\}$ rotation was less than 0.5 degrees. Nevertheless, the match images were realigned relative to the base image, by bicubic spline interpolation, prior to any further analysis.

For the images presented in this paper, all of which were acquired from healthy volunteers who were able to remain still in the scanner, realignment alone seemed sufficient to address the problem of (minor) movement artifact. However, it may be advantageous to take further steps to remove movement-correlated components from fMRI time series prior to analysis of images acquired from more mobile subjects (14).

## ESTIMATION

### Exploratory Regression Modeling of an fMRI Time Series

A time series for exploratory analysis was obtained by averaging the individual time series observed at 156 pixels representing occipital cortex in a 2D image acquired during photic stimulation; it was expected *a priori* (2, 15) that this region of the brain should be highly activated under these experimental conditions. At each pixel, the number of points in the series was identical: $N = 100$. The averaged time series is plotted, together with the concomitant box-car input function (Fig. 1a); Figs. 1b and 1c show the corresponding periodogram and correlogram. Inspection of these plots suggested a slight negative linear trend over the course of the experiment, as well as a marked periodic or sinusoidal trend with the same (fundamental) frequency as the input function. In addition, relatively modest peaks were evident in the periodogram at frequencies corresponding to the first and second harmonics of the fundamental frequency.

These observations suggested the following time series regression model to account for linear and sinusoidal trends in the data:

$$Y_t = \alpha + \beta t + \gamma\sin(\omega t) + \delta\cos(\omega t) + \gamma'\sin(2\omega t) + \delta'\cos(2\omega t) + \gamma''\sin(3\omega t) + \delta''\cos(3\omega t) + \rho t. \quad [1]$$

Here $Y_t$ is the $T_2$*-weighted signal intensity value observed at time point $t = 1,2,3,\ldots,N$; $\omega$ is the (fundamental) frequency in radians per time point of the box-car function (in this case, $\omega = 2\pi/20$); $2\omega$ and $3\omega$ are the first and second harmonic frequencies, respectively; and $\rho_t$ is a residual term (see ref. 16 for a general introduction to time series regression). This model can be written more succinctly using matrix notation:

$$Y = XB + R, \quad [2]$$

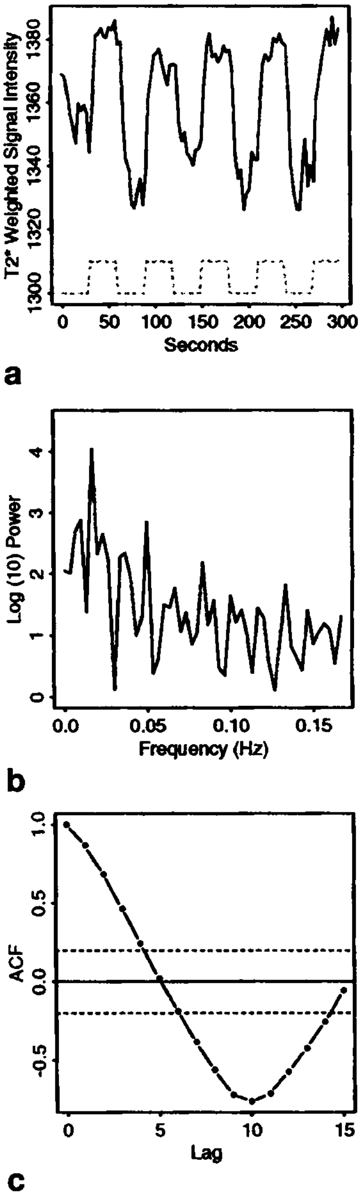

FIG. 1. (a) Time series plot of a spatially averaged fMRI signal (solid line) observed during periodic visual stimulation (dotted line); (b) periodogram of fMRI time series; (c) correlogram of fMRI time series; horizontal dotted lines are $\pm 2/\sqrt{N}$, Bartlett's approximate 95% confidence interval. Estimated AR coefficients more negative or positive than these limits are significantly different from zero.

where $\mathbf{Y}$ is an $N$-dimensional column vector of $T_2$*-weighted signal intensity values ($N = 100$); $\mathbf{B}$ is a $p$-dimensional column vector of model parameters ($P = 8$); $\mathbf{X}$ is an $N \times p$ design matrix; and $\mathbf{R}$ is an $N$-dimensional column vector of residuals. Assuming that the elements of $\mathbf{R}$ are serially independent, the $p \times p$ dimensional matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ then has diagonal elements proportional to the standard errors of the parameter estimates, and off-diagonal elements proportional to the covariance between parameter estimates.

The model was first fitted by ordinary least squares (OLS), and the OLS parameter estimates $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}, \hat{\gamma}', \hat{\delta}', \hat{\gamma}'', \hat{\delta}''\}$ are given in Table 1 with their standard errors. However, before placing too much emphasis on these

Table 1
Estimated Regression Coefficients, and Standard Errors (SE), Obtained by Various Fitting Procedures

| | OLS | | PGLS | | ARIMA | |
|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE |
| $\alpha$ | 1365.4 | 1.5 | 1366 | 2.8 | 1365.8 | 2.7 |
| $\beta$ | −0.1 | 0.026 | −0.11 | 0.05 | −0.1 | 0.05 |
| $\gamma$ | −18.14 | 1.03 | −17.93 | 1.69 | −17.83 | 1.57 |
| $\delta$ | 15.12 | 1.03 | 15.24 | 1.70 | 15.43 | 1.71 |
| $\gamma'$ | 1.9 | 1.02 | 2.14 | 1.34 | 2.23 | 1.25 |
| $\delta'$ | 1.43 | 1.02 | 1.31 | 1.34 | 1.45 | 1.23 |
| $\gamma''$ | 3.6 | 1.02 | 3.75 | 1.07 | 3.83 | 1.09 |
| $\delta''$ | 4.4 | 1.02 | 4.12 | 1.06 | 4.23 | 1.01 |
| $\zeta$ | 0.54 | 0.085 | 0.53 | 0.1 | 0.54 | 0.007 |

estimates, we must bear in mind that the method of ordinary least squares generally provides minimum variance unbiased estimates (MVUEs) of regression model parameters only if the error terms of the fit are independent and normally distributed (17). To check the validity of this crucial assumption, we examined the residual terms $\{\rho_t\}$ by time series, periodogram, and correlogram plots (Fig. 2). The correlogram suggested that the residual process was autocorrelated.

A plot of the partial autocorrelation function (PACF) was used to indicate which order of autoregressive (AR) model should be fitted to the $\{\rho_t\}$ series. The partial correlation between $\rho_t$ and $\rho_{t+k}$ is the correlation at lag $k$ after regression of $\rho_t$ on all intermediate terms $(\rho_{t+1}, \ldots, \rho_{t+k-1})$, and is zero for lags greater than the order of the AR process in the series (18). As shown in Fig. 3, the PACF was only significantly different from zero at $k = 1$; suggesting that the first order AR process,

$$\rho_t = \zeta \cdot \rho_{t-1} + \epsilon_t, \qquad [3]$$

would be an appropriate model to fit. Here $\zeta$ is the AR coefficient (estimated by OLS in Table 1), and $\epsilon_t$ is an error term. The adequacy of this AR(1) model to account for structure in the $\{\rho_t\}$ series was assessed by testing for persistent autocorrelation in its error terms, $\{\epsilon_t\}$. Figure 4 shows time series, periodogram, and correlogram plots for the $\{\epsilon_t\}$ series. There is no graphical evidence for significant autocorrelation; this visual impression was corroborated by a Box-Pierce test for white noise (19). The Box-Pierce test statistic, $Q_K$, is given by the following expression:

$$Q_K = N \cdot \sum_{1}^{K} ac_k^2, \qquad [4]$$

where $ac_k$ is the autocorrelation coefficient at lag $k = 1,2,3, \ldots, K$. Under the null hypothesis that the time series in question is serially independent, or white noise, $Q_K$ is distributed as $\chi^2$ with $K - q$ degrees of freedom, where $q$ is the order of AR process fitted to the series. Improbably large values for $Q_K$ may therefore be taken as evidence of significant serial dependency. For the $\{\epsilon_t\}$ series, $Q_{15}$ was 13.2 with 14 degrees of freedom, which was compatible with the null hypothesis ($P = 0.51$).

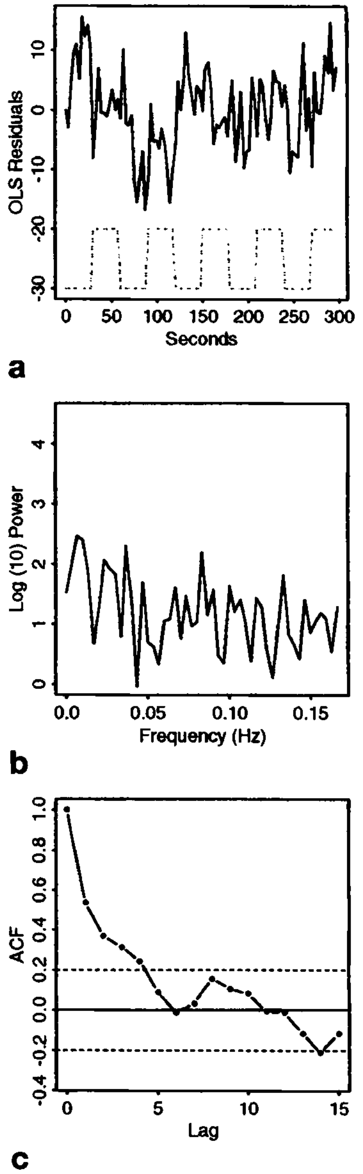

FIG. 2. (a) Time series plot of residual terms $\{\rho_t\}$ generated by OLS fit of a sinusoidal regression model (Eq. [1]) to the fMRI time series in Fig. 1a; (b) periodogram of $\{\rho_t\}$; (c) correlogram of $\{\rho_t\}$.

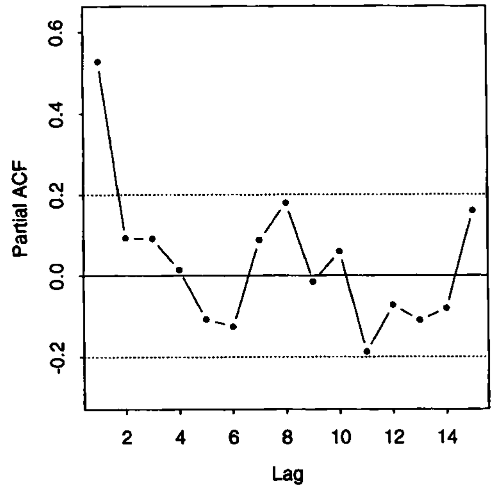

FIG. 3. Partial autocorrelation function for the OLS residual series, $\{\rho_t\}$. Horizontal dotted lines are $\pm 2/\sqrt{N}$, Bartlett's approximate 95% confidence interval; estimated partial autocorrelation coefficients more negative or positive than these limits are significantly different from zero.

## Least Squares Fitting with Autocorrelated Errors

Exploratory analysis thus far strongly suggests that we cannot assume that the error terms generated by an ordinary least squares fit of the regression model are independent; the $\{\rho_t\}$ time series is in fact (first order) autocorrelated. The chief implication of this discrepancy is that the standard errors of the regression coefficients estimated by straightforward OLS will be biased; typically, the errors will be underestimated. This bias in error estimation could in turn lead to spuriously elevated test quotients (e.g., $\hat{\gamma}/\mathrm{SE}(\hat{\gamma})$), a false sense of confidence in the coefficient estimates, and mistaken attribution of significance to linear and/or sinusoidal trends in the observed time series. For these reasons, we adopted one of a variety of alternative model fitting techniques, known as pseudogeneralized least squares (PGLS) (20–22), to acknowledge and correct for the autocorrelated structure of the $\{\rho_t\}$ series in estimating the time series regression coefficients and the standard errors of these estimates.

The series of autocorrelated residuals, $\{\rho_t\}$, was generated by a preliminary OLS fit, as described above. From these autocorrelated residuals, the first order AR coefficient, $\zeta$, was estimated; then used to transform the original terms of the regression model, as shown below in matrix notation. (Transformed terms are asterisked, *; the subscript $t$ denotes the row of the **Y** vector or **X** matrix at time point $t = 1,2,3,\ldots, N$.)

$$Y_t^* = Y_t - \zeta Y_{t-1} \qquad [5]$$

$$X_t^* = X_t - \zeta X_{t-1} \qquad [6]$$

Ordinary least squares was then used a second time to estimate the model parameters **B** in the transformed model, $\mathbf{Y}^* = \mathbf{X}^*\mathbf{B} + \mathbf{R}^*$, yielding a new series of residuals $\mathbf{R}^*$, or $\{\rho_t^*\}$. Both graphically (Fig. 5), and by the Box-Pierce test ($Q_{15} = 13.7$, $df\,14$, $P = 0.47$), the $\{\rho_t^*\}$ series was uncorrelated; and a normal quantile plot of $\{\rho_t^*\}$ was linear. These results suggest that, at least in the analysis of this spatially averaged time series, the residuals of the second OLS fit satisfy the crucial assumptions of independent and normally distributed errors.

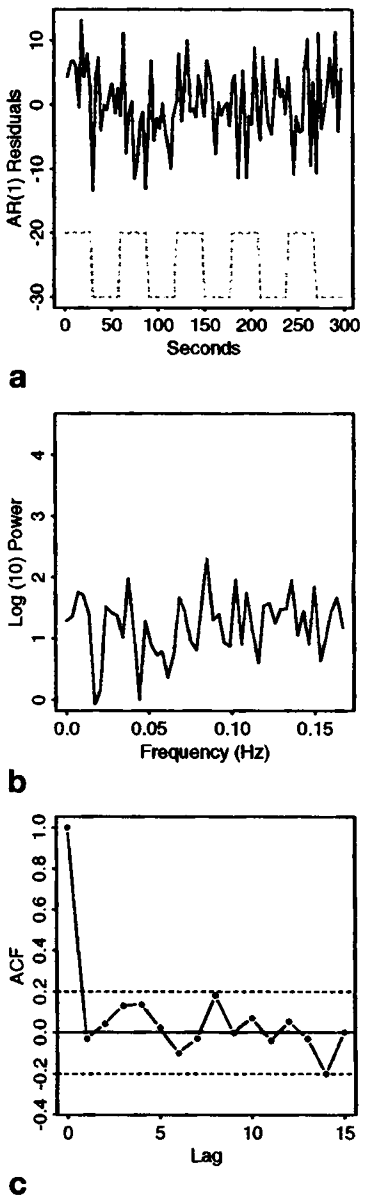

FIG. 4. (a) Time series plot of residual terms $\{\epsilon_t\}$ generated by fitting a first order autoregressive model (Eq. [3]) to the $\{\rho_t\}$ series in Fig. 2; (b) periodogram of $\{\epsilon_t\}$; (c) correlogram of $\{\epsilon_t\}$.

To check that these assumptions were more generally satisfied, we fitted the regression model by OLS and PGLS to 156 individual time series sampled from a 2D image acquired during photic stimulation. This generated 156 OLS and PGLS residual series, $\{\rho_t\}$ and $\{\rho_t^*\}$, respectively. For each residual series, we computed the Box-Pierce statistic, $Q_{15}$; and, for each $\{\rho_t^*\}$ series, the ranked, standardized residual values. Figure 6 shows a quantile-quantile (qq) plot of $Q_{15}$ estimated in 156 $\{\rho_t\}$ and $\{\rho_t^*\}$ series versus 156 random samples from the theoretical null distribution ($\chi^2$, df 14). Departures from linearity in this plot indicate discrepancy between the observed distribution of $Q_{15}$ and its theoretical null dis-

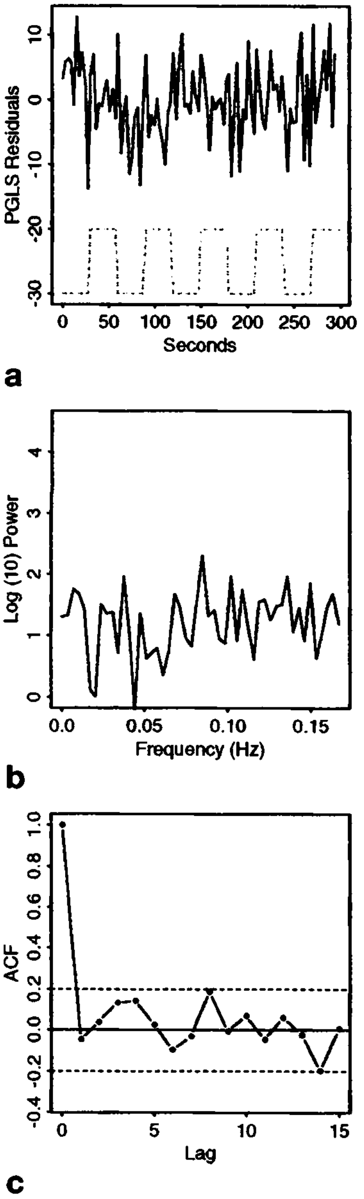

FIG. 5. (a) Time series plot of residuals $\{\rho^*_t\}$ generated by PGLS fit of a sinusoidal regression model (Eq. [1]) to the fMRI time series in Fig. 1a; (b) periodogram of $\{\rho^*_t\}$; (c) correlogram of $\{\rho^*_t\}$.

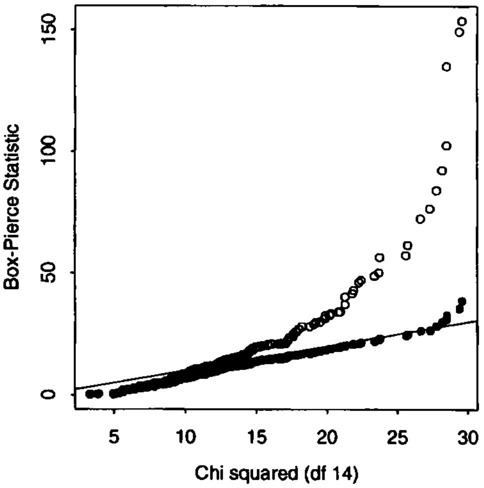

FIG. 6. Quantile-quantile plot of the Box-Pierce test statistic, $Q_{15}$, estimated in 156 OLS residual series $\{\rho_t\}$, and 156 PGLS residual series $\{\rho^*_t\}$ versus 156 random samples from the null distribution of $Q_{15}$, $\chi^2$ with 14 df. Open circles indicate OLS residual estimates; filled circles indicate PGLS residual estimates. Departure from the straight line is incompatible with a null hypothesis of serial independence.

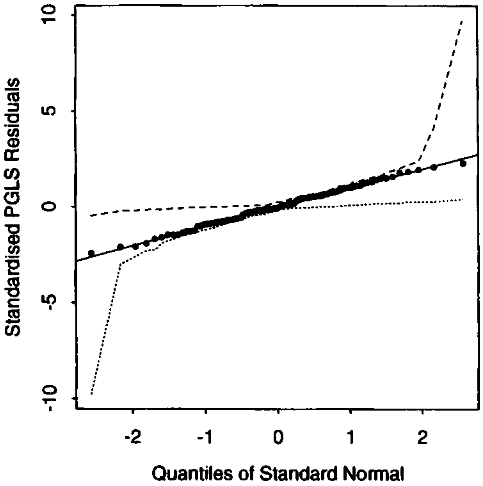

FIG. 7. Quantile-quantile plot of mean standardized PGLS residuals versus quantiles of the standard normal distribution. The dashed line indicates the maximum standardized residual value observed at each quantile; the dotted line indicates the minimum standardized residual value at each quantile. Linearity is consistent with normality.

tribution. The qq plot of $Q_{15}$ in the $\{\rho_t\}$ series is clearly nonlinear, suggesting that these residuals are significantly autocorrelated; whereas, the qq plot of $Q_{15}$ in the $\{\rho_t^*\}$ series is almost exactly linear. This plot confirms the need for treatment of OLS residual autocorrelation (already indicated by the correlogram in Fig. 2c derived from the spatially averaged time series), and provides good graphical evidence that the distribution of $Q_{15}$ in the PGLS residual series is compatible with the assumption of serial independence. Figure 7 shows another quantile plot, in which the mean standardized PGLS residuals, averaged over all 156 $\{\rho_t^*\}$ series, are plotted against quantiles of the standard normal distribution. Also shown is the range (maximum and minimum) of the standardized PGLS residuals observed at each quantile, which gives an indication of the extent of scatter of the residuals about their mean. This plot supports the hypothesis that the sampling distribution of PGLS error terms is normal.

On the basis of these graphical results, it seems justifiable to conclude that the assumption of independent and normally distributed error terms is generally satisfied after iterated OLS, or PGLS, fitting of the model. It therefore seems reasonable to regard PGLS estimates of the regression coefficients $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}, \hat{\gamma}', \hat{\delta}', \hat{\gamma}'', \hat{\delta}''\}$, and their standard errors, as best unbiased estimates of these parameters.

As shown in Table 1, there was relatively little difference between the PGLS estimates of these coefficients and the estimates obtained by straightforward OLS; however, as theoretically predicted, the standard errors of the coefficients estimated by PGLS were up to 70% greater than the biased (under)estimates of standard errors obtained by OLS fitting of the model with untransformed terms and autocorrelated residuals, $\{\rho_t\}$.

### Other Fitting Procedures

Adopting the terminology of Box and Jenkins (23), one can describe the OLS residual series $\{\rho_t\}$ as an autocorrelated integrated moving average (ARIMA) process of order (1,0,0). It is possible to fit such models, with additional regression variables, by nonlinear optimization; for example, the function arima.mle() in S-PLUS will converge on maximum (conditional) likelihood estimates of both AR and time series regression coefficients by iteration of a quasi-Newton algorithm (18, 24). We compared the estimated parameters, and their standard errors, obtained by this nonlinear optimizing function to the estimates obtained by iterated least squares. It can be seen from Table 1 that estimates of the regression coefficients and their standard errors obtained by these two fitting procedures are very similar. This is theoretically not surprising because both procedures maximize the Gaussian likelihood of $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}, \hat{\gamma}', \hat{\delta}', \hat{\gamma}'', \hat{\delta}''\}$ conditioned on $\hat{\zeta}$. In this light, it seems natural to prefer the fitting procedure that is computationally less demanding, and iterated least squares is approximately four times faster than nonlinear optimization (comparing routines for both procedures written in S-PLUS). Parenthetically, the maximum conditional likelihood estimates provided by both PGLS and arima.mle() may be slightly less efficient than estimates obtained by maximizing the full likelihood, but the difference in efficiency will be negligible when the number of data points (100) is large relative to the order of the AR process (1) in the residuals (as it is in these data); and the computational cost of maximizing the full likelihood would probably be greater than that of maximizing the conditional likelihood by nonlinear optimization (22).

### Derivation of Power and Phase

From the estimated sinusoidal regression coefficients $\{\hat{\gamma}, \hat{\delta}, \hat{\gamma}', \hat{\delta}', \hat{\gamma}'', \hat{\delta}''\}$, it is possible to derive the power and phase of each of the three periodic components in the model. For example, power at the ON-OFF frequency of stimulation, or fundamental power (FP), is

$$FP = \hat{\gamma}^2 + \hat{\delta}^2. \qquad [7]$$

It can be seen that fundamental power is equivalent to the squared amplitude of a phase shifted sine wave at the frequency of experimental stimulation, i.e.,

$$\hat{\gamma}\sin\omega t + \hat{\delta}\cos\omega t = \sqrt{\hat{\gamma}^2 + \hat{\delta}^2}\,\sin(\omega t - \phi), \qquad [8]$$

$$\phi = -atan\,\frac{\hat{\delta}}{\hat{\gamma}}.$$

So, if the OFF condition is presented first, the delay (in seconds) between stimulus and response is half the length of the ON-OFF cycle (i.e., 30 s) multiplied by

$$\frac{\pi + \phi}{\pi}; \qquad [9]$$

thus the estimated hemodynamic delay in these data is approximately 6.8 s. (In experiments where the ON condition is presented first, delay is given by half the ON-OFF cycle (in seconds) multiplied by $\phi/\pi$.)

The standard error of fundamental power, SE(FP), is a function of the standard errors of its two constituent parameter estimates, SE($\hat{\gamma}$) and SE($\hat{\delta}$), and the covariance between them, cov$\{\hat{\gamma},\hat{\delta}\}$. From the matrix $(\mathbf{X}^{*T}\mathbf{X}^{*})^{-1}$, for $\zeta$ in the range [0, 1], cov$\{\hat{\gamma},\hat{\delta}\}$ is never greater than 0.2% of the variance in $\hat{\gamma}$ and $\hat{\delta}$; the contribution of cov$\{\hat{\gamma},\hat{\delta}\}$ to the standard error of FP is therefore practically negligible. Assuming $\hat{\gamma}$ and $\hat{\delta}$ are independently normal, SE(FP) is:

$$\mathrm{SE(FP)} = \sqrt{4(\hat{\gamma}^2\mathrm{SE}(\hat{\gamma}) + \hat{\delta}^2\mathrm{SE}(\hat{\delta})) + 2(\mathrm{SE}(\hat{\gamma})^4 + \mathrm{SE}(\hat{\delta})^4)}. \qquad [10]$$

Under the null hypothesis that there is no experimentally determined periodicity in observed fMRI time series, the expected values of $\hat{\gamma}$ and $\hat{\delta}$ are zero and the standard error of FP is given by the simpler expression:

$$\mathrm{SE(FP)} = \sqrt{2(\mathrm{SE}(\hat{\gamma})^4 + \mathrm{SE}(\hat{\delta})^4)}. \qquad [11]$$

These formulae can immediately be generalized to estimate power at the first and second harmonic frequencies, P1 and P2, and their standard errors (SE(P1) and SE(P2)). The estimates and standard errors of these parameters derived from a PGLS fit of the model to the spatially averaged time series are given in Table 2.

### Comparison to Other Estimators of Experimental Effect

To compare this method (fitting a sinusoidal regression model by PGLS) to alternative estimators of the experimental effect in fMRI time series, we must first introduce some more general notation. Let the pattern of ON-OFF experimental stimulation be denoted by a square or box-car function, $\{\mathrm{BOX}_t\}$, which has value 1 during the ON condition and value $-1$ during the OFF condition. We can then say that the observed time series, after removal of linear trend, $\{Y_t\}$, is proportional to $\{\mathrm{BOX}_t\}$ plus noise:

$$Y_t = A \cdot \mathrm{BOX}_t + \mathrm{Err}_t. \qquad [12]$$

Here $A$ is the amplitude of the box-car function and Err is an error term, with expected value zero. Then three possible measures of experimental effect—(i) the mean difference in signal intensity during ON and OFF conditions, (ii) the cross-covariance between the observed fMRI time series and the box-car input function, and (iii)

Table 2
Fundamental and Harmonic Power Estimates Derived from Sinusoidal Regression Coefficients

| | Estimate | Standard error | Coefficient of variation |
|---|---|---|---|
| FP | 553.7 | 5.75 | 0.01 |
| P1 | 6.29 | 3.59 | 0.57 |
| P2 | 31.1 | 2.26 | 0.07 |

an ordinary least squares fit to the linear model in Eq. [12]—will all have the same expected value of $A$.

However, as already discussed, this model is inadequate to cope with hemodynamically mediated delay between the input function and the observed fMRI time series. To include hemodynamic delay in the model (e.g., as in Bandettini *et al.* (3)), we can rewrite Eq. [12] as follows:

$$Y_t = A \cdot \mathrm{BOX}_{t-d} + \mathrm{Err}_t, \qquad [13]$$

where $d$ is the number of time points between stimulus and observable response. As shown in Fig. 8, we have fitted this model by least squares to the spatially averaged fMRI time series using various integer values of $d$ (points marked by $x$). We assume that the first point in the fMRI series is observed 3 s after the start of the experiment, so delay between stimulus and response in seconds is $3^*(d+1)$. It can be seen that the goodness-of-fit (GOF = residual sum of squares divided by total sum of squares) of this model is a function of delay. As expected, the worst fit to the data is given by $d = 0$ (corresponding to a delay of 3 s), which is effectively fitting Eq. [12]. The best fit by this method is with $d = 2$, corresponding to a delay of 9 s. It should be noted that the smaller the residual sum of squares, the greater the estimated cross correlation, so worse fitting models will tend to underestimate the size of the experimental effect in terms of $\hat{r}$.

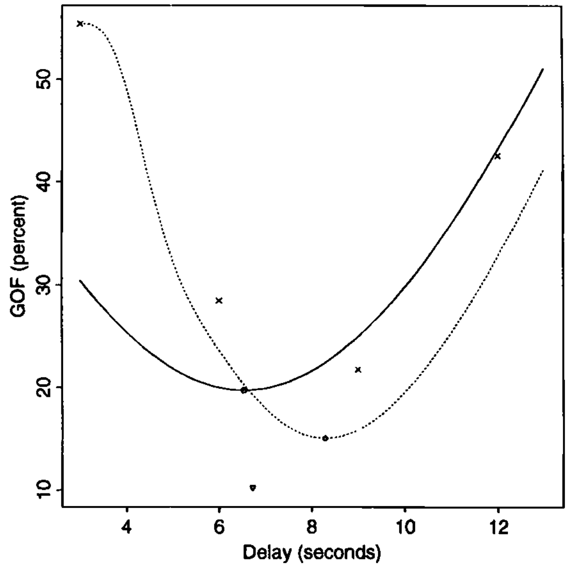

FIG. 8. Plots of GOF = residual sum of squares divided by total sum of squares versus delay, in seconds, for alternative estimators of the experimental effect in the spatially averaged fMRI time series. Diagonal crosses, X, mark the GOF obtained by fitting the model of Eq. [13] with variable integer values of $d$. The best fit obtainable by this method has GOF = 21.7%; the worst fit (equivalent to fitting the model of Eq. [12]) has GOF = 55.3%. The dotted line shows the GOF obtained by fitting the model of Eq. [15] with variable values of $\lambda$ used to parameterize the Poisson (hemodynamic response) function; the diamond indicates the best fit by this method (GOF = 15.0%). The solid line shows the GOF obtained by fitting the model of Eq. [14], with a pure sine wave as the smooth periodic function $f$, and variable phase relative to the box-car function; the circle indicates the best fit by this method (GOF = 19.7%). The inverted triangle indicates the GOF, and estimated hemodynamic delay, obtained by fitting a sinusoidal regression model (Eq. [1]) using the method of PGLS (GOF = 10.2%).

To take account of differences in shape (as well as phase or delay) between the observed time series and the box-car function, it is possible to modify Eq. [13] thus:

$$Y_t = A \cdot f(t-d) + \mathrm{Err}_t, \qquad [14]$$

where $f$ is a general smooth periodic function. We fitted Eq. [14] by OLS to the spatially averaged time series using two different functions: (i) a phase shifted sine function, $f(t-\mathrm{d}) = \sin(\omega(t-d))$, and (ii) the box-car function convolved with a Poisson function parameterized by $\lambda$ ( $= 3d$),

$$Y_t = A \cdot \sum_j \mathrm{BOX}_{t-j} \frac{\lambda^{3j} e^{-\lambda}}{(3j)!} + \mathrm{Err}_t, \qquad [15]$$

used by Friston *et al.* (4). As shown in Fig. 8, the GOF obtained by these two methods is a function of $\alpha$ and $\lambda$, respectively. The best possible fit to these data obtained by the method of Friston *et al.* (4) (point marked by a diamond) is better than the best possible fit obtained by the phase shifted sine wave (point marked by a circle); and this may reflect the relative success of these functions in approximating the "squareness" of this particular fMRI waveform (see Fig. 1a).

Overall, our method provides the best fit to these data (point marked by an inverted triangle), and to other time series we have comparatively analyzed, probably for two reasons: (i) adequate treatment of residual autocorrelation, as discussed above; (ii) use of six sine and cosine terms in the model, which allow a more flexible approximation to the phase and shape of the main periodic trend in the series than is possible by the relatively under-parameterized alternative methods.

### Descriptive Image Analysis and Mapping

We fitted the sinuoidal regression model (Eq. [1]) by PGLS to the multiple fMRI time series comprising a single 2D slice of the image acquired during photic stimulation. The size of the image matrix was 128 × 64 and the total number of time series was therefore 8192, each of length $N = 100$. Pixels representing only nonbiological background noise typically had much lower signal intensity value than pixels representing skull or brain. To reduce computational overheads, pixels with initial signal intensity value less than an empirically determined threshold of 200 were excluded from analysis; this reduced the total number of time series to 1811. This set of suprathreshold pixels will be referred to subsequently as the search volume, and the number of pixels in the search volume will be designated SV.

Linear, sinusoidal, and AR coefficients were individually estimated for each time series in the search volume, and the sinusoidal coefficients used to derive estimates of power at the fundamental and harmonic frequencies. The observed distribution of the linear and AR coefficient estimates are summarized by box plots in Fig. 9a.

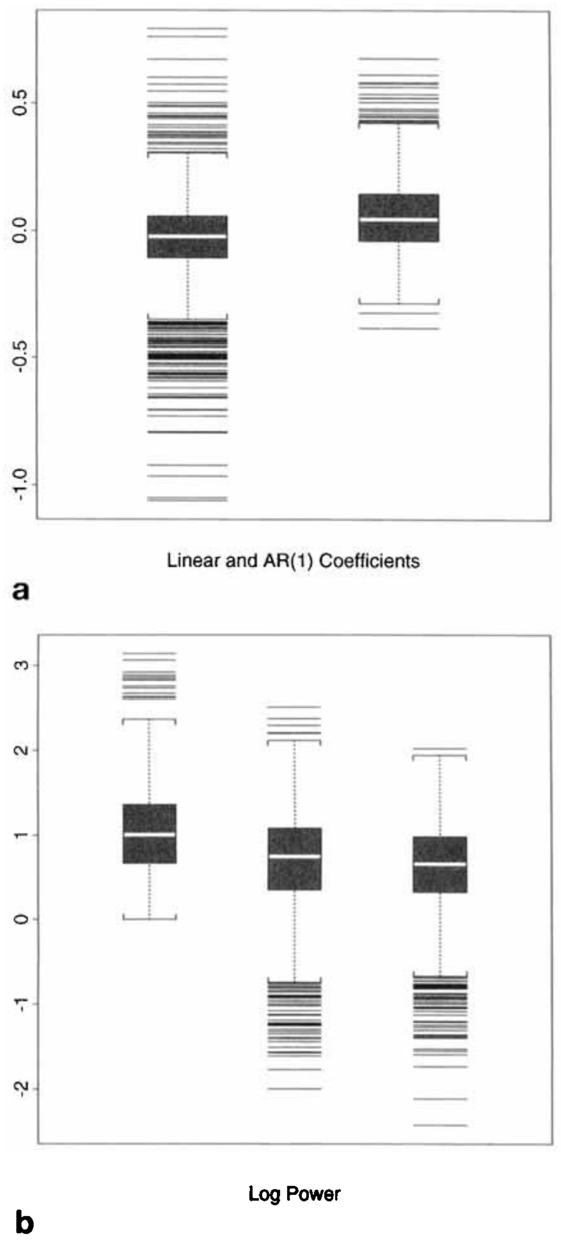

FIG. 9. (a) Box plots of the linear trend coefficient ($\hat{\beta}$) and first order AR coefficient ($\hat{\zeta}$) estimated at 1811 pixels in a 2D image acquired during photic stimulation. (b) Box plots of log power at various frequencies estimated as above. From left to right: log power at the fundamental frequency, the first and second harmonic frequencies.

The distribution of $\hat{\beta}$ is approximately symmetrical; the distribution of $\hat{\zeta}$ is slightly positively skewed. The non-negative values of power at the fundamental and harmonic frequencies were log transformed before box plotting (Fig. 9b). The spatial distributions of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\zeta}$, FP, P1, and P2 over all 1811 pixels in the search volume are shown by descriptive parametric maps in Fig. 10. It can be seen that large positive values of log FP and log P2 are concentrated at pixels representing occipital cortex; and positive values of $\hat{\zeta}$ tend to be found at pixels at least partly representative of cerebrospinal fluid (CSF).

## INFERENCE

### Ascertaining the Null Distribution

Assume, for the moment, that the experimental effect in fMRI time series may be well estimated by sinusoidal power, FP, at the (fundamental) frequency of stimulation. The next question is how to decide, on the basis of fundamental power, whether or not a given pixel is activated. In keeping with general principles of probabilistic decision making, we can derive a fundamental power quotient, $FPQ_i = FP_i/\mathrm{SE}(FP_i)$, from the estimate of fundamental power, and its standard error, at each pixel; and refer each pixel's observed power quotient to the distribution of *FPQ* under a null hypothesis. A good choice of form for the null distribution of *FPQ* is clearly essential to success, and there are arguably three ways we can ascertain this null distribution: (i) by theory, (ii) by experiment, (iii) by randomization. In this section, we briefly discuss the relative merits of these three alternatives.

The quickest and cheapest method is to derive a parametric form for the null distribution from normal theory. From Eqs. [7] and [11] above, we have

$$FPQ = \frac{\hat{\gamma}^2 + \hat{\delta}^2}{\sqrt{2(\mathrm{SE}(\hat{\gamma})^4 + \mathrm{SE}(\hat{\delta})^4)}}. \qquad [16]$$

From the matrix $(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}$, for $\zeta$ in the range [0, 1], the difference between $\mathrm{SE}(\hat{\gamma})$ and $\mathrm{SE}(\hat{\delta})$ is never greater than 0.5%. We neglect this small difference and assume that the standard errors of $\hat{\gamma}$ and $\hat{\delta}$ are equal, i.e., $\mathrm{SE}(\hat{\gamma}) = \mathrm{SE}(\hat{\delta}) \equiv \mathrm{SE}$; we can then rewrite Eq. [16]:

$$FPQ = \frac{1}{2}\left[\left(\frac{\hat{\gamma}}{\mathrm{SE}}\right)^2 + \left(\frac{\hat{\delta}}{\mathrm{SE}}\right)^2\right]. \qquad [17]$$

If we further assume that, under the null hypothesis, the quotients $\hat{\gamma}/\mathrm{SE}$ and $\hat{\delta}/\mathrm{SE}$ are sampled from a *t* distribution, which is approximately standard normal when the number of points in each fMRI time series is in the order of 100, then the squared quotients will each be distributed under the null hypothesis as $\chi^2$ with 1 df, and the theoretically derived null distribution of FPQ is therefore:

$$FPQ \sim \frac{\chi^2_2}{2}; \qquad [18]$$

that is, a scaled $\chi^2$ distribution with 2 df. (For $N \ll 100$, a closer approximation would be the *F* distribution with 2 and $N-2$ df).

The experimental way of ascertaining the null distribution is the most expensive, but arguably the least approximate, of the three alternatives. Before or after acquisition of an image during periodic sensory stimulation, an image can be identically acquired under conditions that would not be expected to determine any periodic response. Power in the time series at each pixel of this null image can then be estimated, and the distribution of FPQ under a less formal null hypothesis, that observed values of $FPQ_i$ are not determined by periodic stimulation, can be directly sampled. One attraction of this approach is that values of $FPQ_i$ in the null image will be

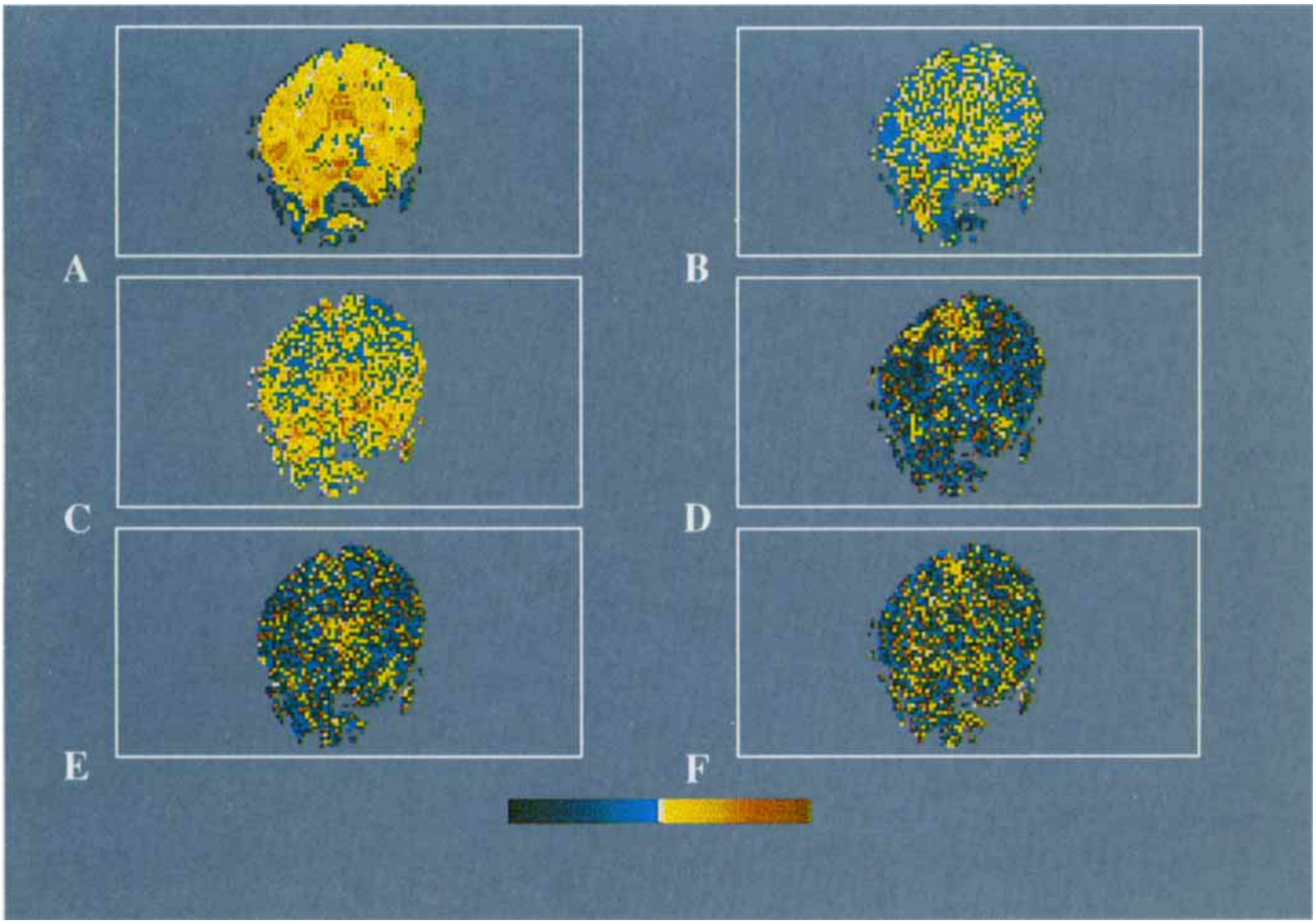

FIG. 10. Descriptive parametric maps. (a) intercept, $\hat{\alpha}$; (b) linear trend coefficient, $\hat{\beta}$; (c) first order AR coefficient, $\hat{\zeta}$; (d) log fundamental power, FP; (e) log first harmonic power, P1; (f) log second harmonic power, P2. The limits of the color range, from dark blue to dark orange, correspond to minimum and maximum observed values of each parameter. All maps are orientated with the left posterior quadrant of the brain at the upper right quadrant of the map.

spatially autocorrelated, and experimental ascertainment should therefore yield a null distribution that reflects the degree of nonindependence in the observed $FPQ_i$s.

Finally, it is possible to ascertain the null distribution by randomization; that is, by randomly reorganizing the order of signal intensity values in each observed time series and estimating FPQ for each randomized time series. Values of $FPQ_i$ in the randomized series will not be determined by periodic stimulation, so the randomized null distribution of *FPQ* can be used to test the same (relatively informal) null hypothesis as the experimental null distribution.

We experimentally ascertained the *FPQ* distribution from 1961 estimates of $FPQ_i$ in a single 2D slice of the null image (see *Images* for details of acquisition parameters and experimental design), and compared this to the null distributions obtained (i) by independently, randomly permuting each time series in the null image; and (ii) by randomly sampling 1961 values from the theoretical null distribution. The location and variance of the theoretical null distribution were significantly different from the location and variance of both the experimental and randomized null distributions (Wilcoxon test statistics = 2.43 and 3.38, respectively, $P < 0.05$ in both cases; $F = 0.62$ and 0.62, respectively (df 1960, 1960), $P < 0.05$ in both cases); whereas, there was no significant difference, by the same tests of location and variance, between the experimental and randomized null distributions (Wilcoxon test statistic = −0.94, $P = 0.34$; $F = 0.99$ (df 1960, 1960), $P = 0.9$).

We conclude that, although it is the least costly to ascertain, the theoretical null distribution is an unacceptably imperfect approximation to the "gold standard" of the experimental null distribution. Randomization, on the other hand, yields a null distribution virtually indistinguishable from that obtained by estimating $FPQ_i$ at each pixel of the (spatially correlated) null image; and, although computationally more time demanding than experimental ascertainment, it is more economical in terms of scanner time. The randomized null distribution thus seems to represent the best balance between approximation and cost.

### Randomization Testing

Test quotients, $\Pi Q_i$, were derived for each parameter estimate $\Pi = \{\hat{\beta}, \hat{\zeta}, \text{FP}, \text{P1}, \text{P2}\}$, at each pixel, by the general formula:

$$\prod Q_i = \frac{\prod_i}{\mathrm{SE}(\prod_i)}, \qquad [19]$$

where $\mathrm{SE}(\Pi_i)$ is the standard error of parameter estimate $\Pi$ at pixel $i = 1,2,3,\ldots, \mathrm{SV}$.

Our null hypothesis was that the observed values of $\Pi Q_i$ were not determined by periodic sensory stimulation; or, to put it another way, could equally have arisen by chance. This hypothesis was judged untenable, and a given pixel was consequently said to be activated if $\Pi Q_i$ exceeded a threshold value for that quotient, designated CV. Randomization testing was used to set the critical value such that, under the null hypothesis, the resulting inferential brain activation map (BAM) would include an arbitrary number of false positive pixels, expected to be "activated" by chance.

In the case of the 2D image slice already described, each time series in the search volume (SV = 1 811) was independently permuted once to yield a set of 1811 randomly reorganized time series. The randomized time series were then subject to analysis in exactly the same way as the observed time series, producing a randomized distribution, $\mathbb{R}$, for each estimated parameter's test quotient. Permutation of the observed time series and analysis of the resulting randomized series could be repeated NPERM times, so that the ultimate size of the randomized distributions was NPERM * SV = RAN.

To derive appropriate critical values, CV, from the corresponding randomized distribution, $\mathbb{R}$, one must first set the level of significance, or pixel-wise probability of a false positive, $\alpha$. For a one-tailed test (of the condition that $\Pi Q_i > \mathrm{CV}^{\mathrm{upper}}$), $\mathrm{CV}^{\mathrm{uppor}}$ is then defined as the $(1-\alpha)$th quantile in the randomized null distribution. For a two-tailed test (of the condition that $\Pi Q_i < \mathrm{CV}^{\mathrm{lower}}$ or $\Pi Q_i > \mathrm{CV}^{\mathrm{upper}}$), $\mathrm{CV}^{\mathrm{upper}}$ and $\mathrm{CV}^{\mathrm{lower}}$ are defined as the $(1-\alpha/2)$th and $(\alpha/2)$th quantiles, respectively, in the randomized null distribution.

We applied these principles (see refs. 8, 9 for general introductions to randomization testing) to inferential mapping of test quotients derived from the parameter estimates $\hat{\beta}$, $\hat{\zeta}$, FP, P1 and P2. A randomized distribution for each parameter's test quotient was generated by 10 permutations of the observed data, so that RAN for each $\mathbb{R}$ was 10*1,811 = 18,110. In other words, the 100 points in each observed time series were permuted 10 times to yield 10 randomized time series. This operation was repeated independently at each pixel in the image, yielding 10 randomized images that had temporal activity under the null hypothesis but the same expected spatial structure as the observed image.

Because the linear ($\hat{\beta}$) and AR ($\hat{\zeta}$) coefficient estimates could have both negative and positive values, two-tailed tests were applied pixel-by-pixel to $\hat{\beta}Q_i$ and $\hat{\zeta}Q_i$. Because power at any frequency can have only positive values, significance of $FPQ_i$, $\mathrm{P1}Q_i$, and $\mathrm{P2}Q_i$ was assessed by a one-tailed pixel-by-pixel test. For each of the maps in Fig. 11, $\alpha = 5.5 \cdot 10^{-4}$; this is equivalent to one false positive pixel per image under the null hypothesis. Table 3 gives the corresponding critical values, ascertained by randomization and theory, for the tests of each quotient.

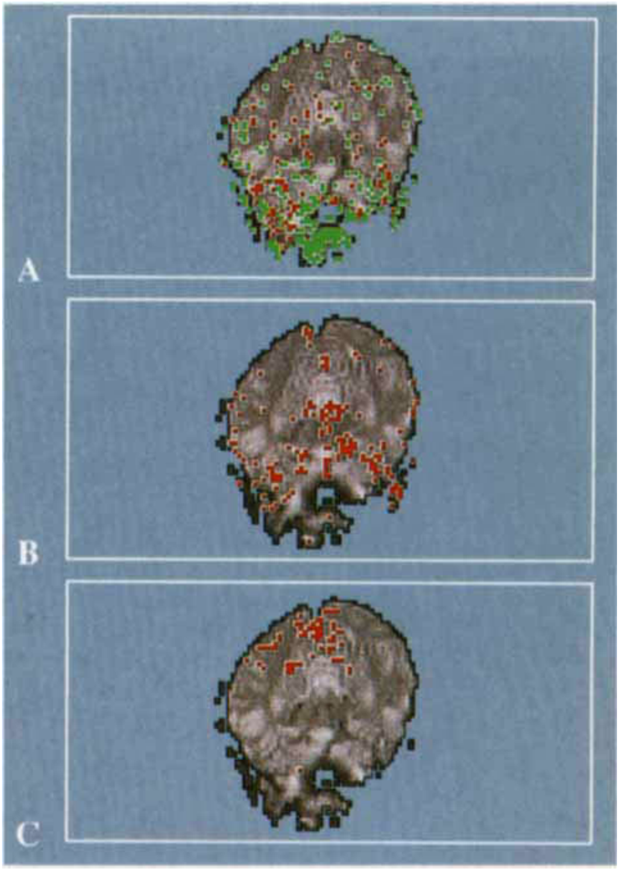

FIG. 11. Inferential brain activation maps (BAMs). The probable number of error pixels per image (eppi) is 1; $\alpha = 5.5 \cdot 10^{-4}$. (a) Linear trend quotient, $\hat{\beta}Q$; (b) first order AR quotient, $\hat{\zeta}Q$; (c) fundamental power quotient, *FPQ*. Activated pixels are colored and superimposed on a gray scale map of the intercept terms ($\hat{\alpha}^*$) estimated by PGLS fitting of the multiple regression model at each pixel. Red indicates the pixel locations of quotient values that exceeded the upper critical value in a one- or two-tailed test; green indicates the locations of quotient values that were less than the lower critical value in a two-tailed test. BAMs for power at the first and second harmonic frequencies are not shown because in neither case was more than 1 pixel activated, as expected by chance. All maps are orientated with the left posterior quadrant of the brain at the upper right quadrant of the map.

## Sensitivity and Specificity of Brain Activation Maps

If the expected number of error pixels per image (eppi) is initially set greater than 1, less specific but more sensitive maps of regional brain activation are generated. Figure 12 illustrates six versions of the brain activation map (BAM), derived from the fundamental power quotient, with the expected number of error (false positive) pixels per image, eppi, ranging from 1 to 100. Table 4 gives the pixel-wise probability of Type I error, $\alpha$; *threshold value*, $\mathrm{CV}^{\mathrm{uppor}}$; and number of activated pixels, NPIX, for each map.

The specificity of each map is $1-\alpha$; and its sensitivity is $1-\beta$, where $\alpha$ is the probability of type I error (or a false positive pixel), and $\beta$ is the probability of Type II error (or a false negative pixel). To estimate $\beta$ for each $\mathrm{BAM}^j$ in Fig. 12 ($j$ = 1, 2, 3, 4, 5, 6), we need to know the number of pixels that represent regions of the brain physiologi-

Table 3

Critical Values (CV) of the Randomized[a] and Theoretical Null Distributions of Linear ($\beta Q$), AR ($\zeta Q$), and Fundamental and Harmonic Power (FPQ, P1Q, P2Q) Quotients

| | Test | $CV^{upper}$ [b] | $CV^{lower}$ [b] |
|---|---|---|---|
| $\beta Q$ | 2 tailed | 4.09<br>(3.46) | −4.07<br>(−3.46) |
| $\zeta Q$ | 2 tailed | 2.96<br>(3.46) | −4.42<br>(−3.46) |
| FPQ | 1 tailed | 11.33<br>(7.37) | |
| P1Q | 1 tailed | 9.34<br>(7.37) | |
| P2Q | 1 tailed | 8.52<br>(7.37) | |

[a] Size of the randomization distributions = 18,110; eppi = 1; SV = 1,811; $\alpha = 5.5 \cdot 10^{-4}$.

[b] Critical values derived from the theoretical null distribution are given in parentheses. To obtain these values, $\beta Q$ and $\zeta Q$ are assumed to be normally distributed; FPQ, P1Q, and P2Q are assumed to be distributed approximately as $\chi^2_2$ divided by 2 (see text for details).

cally activated by the experimental stimulus; this number, NTRUE, is the size of the target population of pixels we wish to identify as activated. NTRUE is the sum of the number of pixels correctly classified as activated, CA, plus the number of pixels incorrectly classified as unactivated, IU, i.e., NTRUE = CA + IU. The number of pixels actually classified as activated in a given map, NPIX, is the sum of the number of pixels correctly classified as activated, CA, plus the number of pixels incorrectly classified as activated, IA, i.e., NPIX = CA + IA. Combining these two equations, we have

$$\text{NTRUE} = \text{NPIX} + \text{IU} - \text{IA}. \qquad [20]$$

The expected value of IA is approximately $\alpha$ (SV-NTRUE); and the expected value of IU is $\beta$ NTRUE. Substituting into Eq. [20], we have

$$\text{NTRUE} = \text{NPIX} - \alpha(\text{SV} - \text{NTRUE}) + \beta\text{NTRUE}. \qquad [21]$$

If we assume that in BAM$^6$, with NPIX = 306, eppi = 100, and $\alpha = 0.055$, all the pixels in the target population have been identified as activated (i.e., $\beta = 0$), we can estimate the number of physiologically activated pixels in the image, NTRUE, from Eq. [21] as 218. We can then rearrange the same equation to solve for sensitivity, $(1-\beta)$, with NTRUE given:

$$1 - \beta^j = \frac{\text{NPIX}^j}{\text{NTRUE}} + \alpha^j\left(1 - \frac{\text{SV}}{\text{NTRUE}}\right). \qquad [22]$$

Substituting the estimated value of NTRUE into Eq. [22], we can estimate the sensitivity $(1-\beta)^j$ for each of the other five brain activation maps in the series (see Table 4). Figure 13 shows receiver operating characteristic (ROC) curves of sensitivity $(1-\beta)$ versus false positive rate ($\alpha$) for each map. Over the first three maps in the series, it can be seen that as the false positive rate increases incrementally (from $5.5 \cdot 10^{-4}$ to $5.5 \cdot 10^{-3}$), sensitivity increases substantially (from 0.28 to 0.52). If it is of paramount importance that regional brain activation should be depicted specifically, but not with undue exclusivity, then clearly the third map (with eppi = 10, $\alpha = 5.5 \cdot 10^{-3}$) is the optimal representation among those in this series.

If, instead of optimistically assuming that $\beta$ is zero in the BAM with $\alpha = 0.055$, we allow that $\beta$ at that level of $\alpha$ may be, say, 0.1, or 0.25, we can obtain comparable curves of sensitivity versus false positive rate. As shown in Fig. 13, this family of ROC curves suggests that, regardless of the precise value of $\beta$ in the range [0, 0.25] the gain in sensitivity as eppi is increased from 1 to 10 is substantially greater than the concomitant loss of specificity; and, for a given curve, the rate of increase in sensitivity decreases as a function of eppi. In other words, assuming only that $\beta < 0.25$, $\alpha \sim 10^{-3}$ will generally be a good choice for the pixel-wise probability of false positive activation.

## Computational Aspects

A program was written in the C language to run under UNIX on a network of Sun SPARCstations (Sun Microsystems Inc., California). The program estimates experimentally determined power by PGLS fitting a sinusoidal regression model, and infers significance of activation by randomization testing. Standard C functions for OLS fitting by the Gauss-Jordan algorithm, for random number generation, and for sorting of the randomization distributions (13), were incorporated. The higher level S-PLUS language (18, 24, 25) was used for exploratory analysis and graphics. C shell scripts (26) were used to generate a user interface, to coordinate calls to image read and write functions, and to control batch processing.

The time taken to analyze a single 128 × 64 × 100 image in this way is largely dependent on the number of times the observed time series are randomly permuted to generate the randomized distributions, $\mathbb{R}$. It is generally recommended that the size of the sampled randomization distribution (RAN) should be related to the probability of Type I error in an individual test (27). For $\alpha \sim 0.05$, RAN should be at least 5,000; and for $\alpha \sim 0.001$, RAN should be at least 10,000. We wished to create maps over a range of Type I error probabilities, with the most conservative map having $\alpha \sim 10^{-4}$. We therefore permuted the observed time series 10 times to generate randomized distributions of size 18,110; and empirically confirmed that this size of distribution was sufficient to give stable or convergent critical values for a one-tailed test of the condition: $FPQ_i > CV^{upper}$.

Central processing unit time for estimation, inference after 10 permutations, and mapping, is approximately 9 min on a Sun SPARC 10 workstation for a single 2D image; processing time for a 10-slice volume is therefore approximately 1.5 h. If inference at the most conservative level of significance is not required, then the number of permutations can be correspondingly reduced, and processing time abbreviated.

# MAPS

So far, we have described development and application of our methods in the context of a few 2D functional MR images. To illustrate the general applicability of these methods, and to provide some informal proof of their

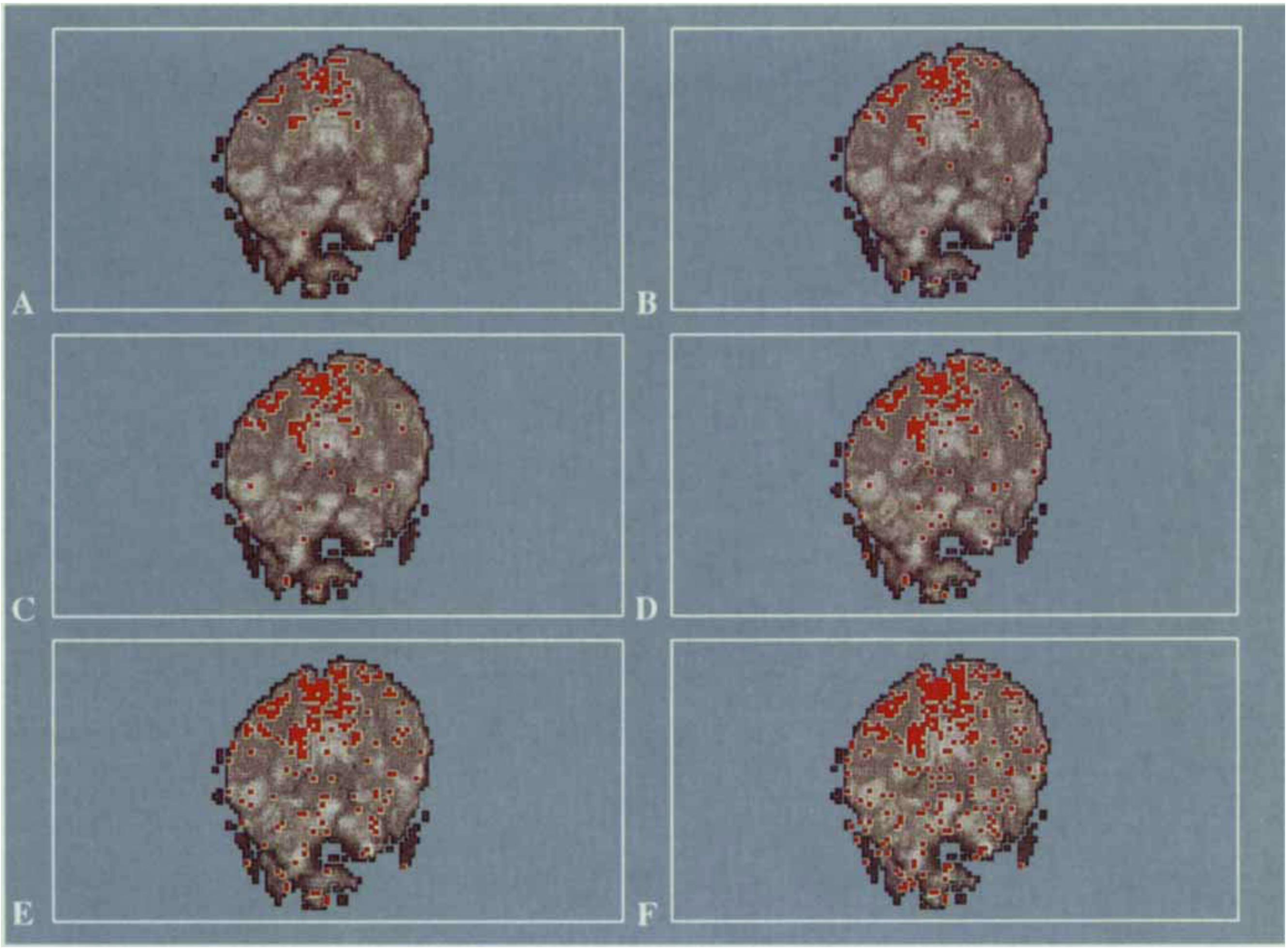

FIG. 12. Six brain activation maps of the fundamental power quotient with different numbers of error pixels per map. (a) eppi = 1; (b) eppi = 5; (c) eppi = 10; (d) eppi = 25; (e) eppi = 50; (f) eppi = 100. For other inferential statistics concerning these images, see Table 4. All maps are orientated with the left posterior quadrant of the brain at the upper right quadrant of the map.

validity, the brain activation maps obtained from a number of other studies of the visual system are shown in Fig. 14; the figure legend incorporates some commentary on these maps.

## DISCUSSION AND CONCLUSIONS

In this paper, we have presented two main methodological innovations in analysis of functional MRI time series: estimation of the experimentally determined effect by sinusoidal regression modeling; and inference of significant activation by randomization testing.

### Modeling

In an fMRI study of periodic sensory stimulation, the experimenter has control over the frequency of stimulation, and it is natural to suppose that a cardinal feature of the experimentally determined effect will be a periodic trend in fMRI time series at the frequency of stimulation. All methods of fMRI analysis previously discussed have this basic premise in common. However, the experimenter will not have control over the phase or waveform of the periodic trend he or she seeks to measure in the data. A box-car input function may elicit an instantaneous (zero phase) and identically shaped (square) output function at the level of neuronal activation; but we cannot directly observe such a pristine neural response by fMRI, even if it occurs (4). The output functions we can observe are less ideal, having been hemodynamically modulated. In other words, the phase and waveform of the putative neural response have been altered by a hemodynamic response function. This hemodynamic response function is obviously not under direct experimental control; and, although it has not yet been well characterized physiologically, it is probably locally variable. The estimation problem is therefore how to measure the size of a periodic trend that has a determined and consistent frequency, but undetermined and inconsistent phase and shape, over several thousand pixels in an image.

The approach we have developed is to measure the size of the experimental effect by fitting a sinusoidal regression model to the time series at each pixel. The main advantage of this method is that a periodic trend of given size will be equally well estimated whatever its phase and, to a lesser extent, whatever its shape. This is in contrast to some other methods of estimation, which by assuming *a priori* that the phase and shape of response

Table 4
Inferential Statistics for Pixel-by-Pixel One-Tailed Tests of the Fundamental Power Quotient, FPQ, in Each of Six Brain Activation Maps[a]

| eppi | $\alpha$ | CV$^{\text{upper}}$ [b] | NPIX | Sensitivity[c] $(1-\beta)$ | Specificity $(1-\alpha)$ |
|---|---|---|---|---|---|
| 1 | $5.5 \cdot 10^{-4}$ | 11.33<br>(7.37) | 63 | 0.28 | 0.99 |
| 5 | $2.8 \cdot 10^{-3}$ | 7.67<br>(6.03) | 105 | 0.46 | 0.99 |
| 10 | $5.5 \cdot 10^{-3}$ | 6.58<br>(5.29) | 123 | 0.52 | 0.99 |
| 25 | $1.4 \cdot 10^{-2}$ | 5.33<br>(4.30) | 166 | 0.66 | 0.98 |
| 50 | $2.8 \cdot 10^{-2}$ | 4.42<br>(3.59) | 222 | 0.80 | 0.97 |
| 100 | $5.5 \cdot 10^{-2}$ | 3.50<br>(2.89) | 306 | 1 | 0.94 |

[a] RAN = 18, 110; SV = 1, 811.
[b] Critical values derived from the theoretical null distribution are given in parentheses.
[c] Assuming $\beta = 0$ when $\alpha = 0.055$ (see text for details).

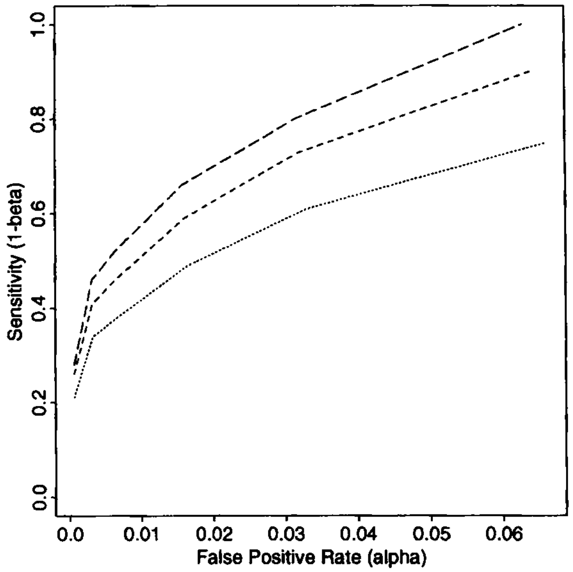

FIG. 13. Plots of sensitivity $(1-\beta)$ versus false positive rate $(\alpha)$ for each brain activation map shown in Fig. 12, assuming various values for $\beta$ when $\alpha = 0.055$. Long dashed line, $\beta$ assumed = 0; short dashed line, $\beta$ assumed = 0.1; dotted line, $\beta$ assumed = 0.25.

is the same over the whole image, will yield biased estimates of the size of the experimental effect at pixels where the phase or shape happens to differ from expectation (see Fig. 8).

Another way of framing this comparison is in terms of the amount of information about each time series in the image that results from estimating the experimental effect in various ways. If we assume *a priori* that we know the phase and shape of response over the whole image then of course we cannot detect any differences in phase and shape of response that there may in fact be between pixels; whereas, by sinusoidal regression, one can obtain estimates of the phase and (in terms of relative power at harmonic frequencies) the shape of response individually for each pixel. Such extra information might well prove to be of value in identifying structure within a set of activated pixels; for example, one could construct the correlation matrix for the power and phase descriptors estimated at each activated pixel and use this as the basis for a principal component or factor analysis to elucidate patterns of functional connectivity (28). At least until such possibilities have been explored, it seems imprudent to prefer less informative (as well as less precise) methods of estimation.

Sinusoidal regression is not the only way in which spatially varying delay and dispersion can be accomodated in estimation of the experimental effect. It is possible, for example, to refine the method of Friston *et al.* (4) so that the Poisson parameter, $\lambda$, is individually estimated at each pixel; and this has been shown to yield more focused maps of brain activation than are obtained by assuming that $\lambda$ is spatially invariant (29). But a sinusoidal regression model has the distinctive advantage of being linear, and therefore quick to fit. There are, however, technical issues to consider concerning the most appropriate choice of fitting procedure.

We have shown that the residuals of an ordinary least squares fit may often be (first order) autocorrelated. Residual autocorrelation is demonstrable in an image acquired without the subject being exposed to periodic sensory stimulation, and generally seems to be most evident in time series observed at pixels at least partly representative of cerebrospinal fluid. One explanation for these observations is that residual autocorrelation is an endogenous phenomenon, perhaps related to pulsation of CSF or blood vessels. It is theoretically predictable that signal changes at the frequency of the cardiac cycle, when aliased to the frequency range sampled with *TR* = 3 s, should be manifest as positive autocorrelation; and, if the cardiac cycle is indeed the source of residual autocorrelation, a longer repetition time might be expected to alleviate the problem (albeit at the cost of temporal resolution of experimentally determined signal change). But, whatever the cause of residual autocorrelation, the fact of its presence in these data renders OLS insufficient as a means of model fitting. The alternative

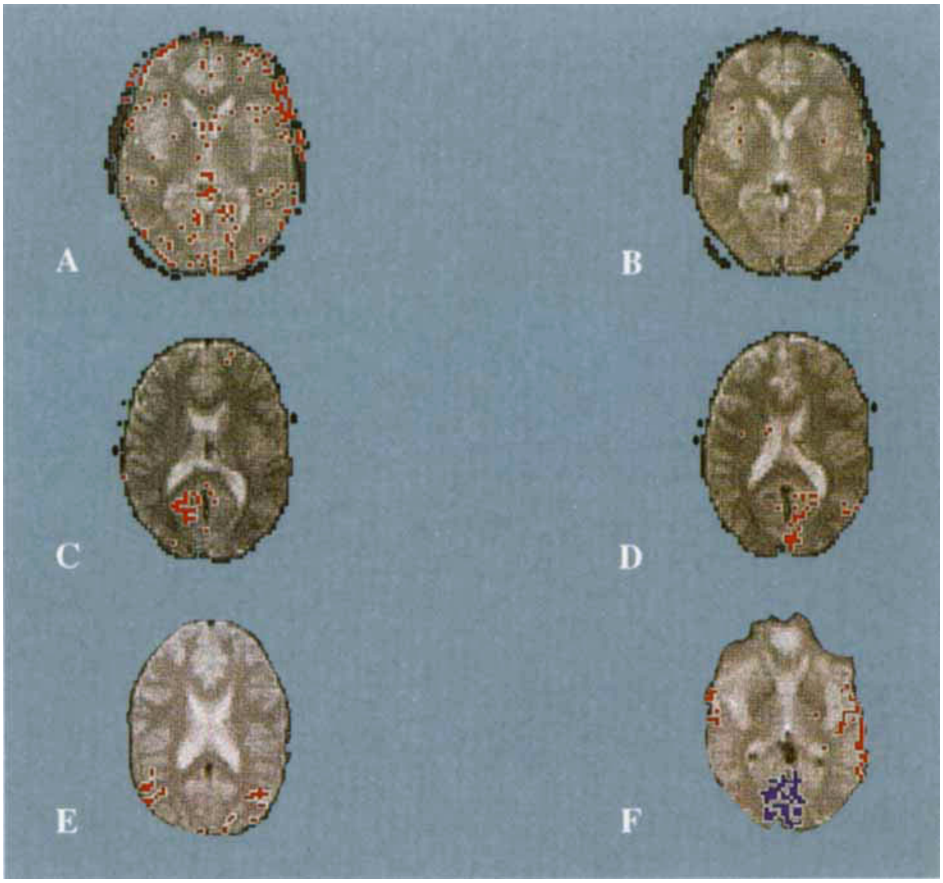

FIG. 14. Brain activation maps obtained from the null image and several pilot studies of the visual system. All maps are orientated with the left anterior quadrant of the brain at the upper right quadrant of the map. Maps of (a) $\hat{\zeta}Q$ and (b) *FPQ* (eppi = 10) obtained from a single slice of the null image. Note that even after image realignment, and in the absence of periodic sensory stimulation, there is significant positive autocorrelation in excess of chance expectation. The number of pixels in the map of *FPQ* is 8, 2 less than expected by chance. Maps of *FPQ* (eppi = 5) obtained from comparable slices of two images acquired during photic hemifield stimulation: (c) the map obtained from left hemifield stimulation, showing predominantly right sided occipital cortex activation; (d) the map obtained from right hemifield stimulation, showing predominantly left sided occipital cortex activation. (e) Map of *FPQ* (eppi = 1) obtained from an image acquired during visual perception of motion. Cortical activation is mainly localized to the region of the lateral occipital sulcus and inferior temporal sulcus. (f) Map of *FPQ* (eppi = 5) obtained from a single slice of an image acquired during bimodal (visual and auditory) stimulation. Pixels activated by auditory stimulation are colored red, and are mainly located in bilateral temporal regions; pixels activated by visual stimulation are colored blue.

procedure we have used, known as PGLS, is essentially iterated OLS; but, between the first and second OLS fit, the original terms of the model are transformed with respect to first order autocorrelation in the residuals of the first fit. We have shown that the residuals of the second OLS fit are independent and normally distributed; this means that the maximum conditional likelihood estimates of the model parameters obtained by PGLS are best unbiased estimates of those parameters, and a valid basis for inference. There are other fitting procedures (e.g., nonlinear optimization) that might be used to obtain parameter estimates with equally desirable statistical properties, but we have preferred PGLS on the grounds of greater computational speed.

In the future, it will be particularly interesting to explore an aspect of fMRI time series analysis that has not so far been considered. Our method of estimating the experimental effect (in common with the other methods discussed) assumes that signal intensity changes due to sensory stimulation are not just periodic, but tonically periodic. In other words, we assume that the amplitude of response to the first presentation of the stimulus is the same as the amplitude of response to the last (e.g., fifth) presentation of the stimulus. Yet this may not always be the case. We have, for example, observed time series during photic stimulation that showed a phasic, habituating pattern of response; the amplitude of the first few experimentally determined cycles being much greater than the amplitude of the last few cycles. This suggests that the rate at which periodicity in fMRI time series decays over the course of image acquisition may be locally variable, and perhaps functionally relevant. One approach to estimation of this decay rate would be to multiply the regression model by an exponential term.

### Decision Making

Creating brain activation maps can be regarded as an exercise in binary decision making: for each pixel in the image we have to decide whether it is activated or not.

We have preferred to make this decision by referring the value of the fundamental power quotient observed at each pixel, $FPQ_i$, to its null distribution, ascertained by independently, randomly permuting each time series in a given 2D image slice. It is reasonable to wonder whether this technique is both necessary, given the far greater speed of significance testing by asymptotic theory, and sufficient, given that observed values of *FPQ* will probably be spatially correlated (29).

We have addressed these questions by an experiment. An image was acquired (at a financial cost) while the subject was not exposed to any periodic sensory stimulation, and $FPQ_i$ was estimated for each pixel. We assumed that the resulting distribution of *FPQ* was the best possible approximation to the true form of the null distribution, because it had been directly sampled from a spatially correlated image while the null hypothesis (that observed values of $FPQ_i$ were not determined by periodic stimulation) was patently true. In relation to this experimental null distribution, the theoretical null distribution (though relatively cheap and quick to ascertain) was found to be inadequately approximate; whereas the form of the randomized null distribution was virtually indistinguishable from that of the experimental "gold standard." On this basis, randomization seems both necessary and sufficient to cope with significance testing in fMRI analysis.

It is conceivable that, in future, we may understand why the theoretical null distribution differs to the extent that it does from both experimental and randomized null distributions; in which case, it may be possible to obtain appropriate critical values more speedily than by randomization and more cheaply than by experiment. Meanwhile, we do not consider the computational costs of randomization to be impractical. Running compiled (C) code on a Sun SPARC 10 workstation, it takes approximately 1.5 h to process entirely a volume of 10 2D functional MR image slices. Furthermore, the integrated program for sinusoidal regression modeling and randomization testing has been incorporated in a C shell script for batch processing. This makes it convenient to process several volumes overnight, using machines that would otherwise be idle. The financial cost is minimal.

The general strategy of decision making by randomization is also adaptable to more complex situations, where we wish to model the spatiotemporal process in an image. The theoretical null distribution may then be (even approximately) intractable. For example, it is very likely that truly activated pixels will tend to be spatially clustered, while falsely "activated" pixels will tend to be spatially dispersed. We might therefore be able to enhance the decision making process by incorporating information about each pixel's spatial context, as well as its own time series. An attractive way of making such higher dimensional decisions is to estimate a spatial statistic, that somehow describes clustering in the neighborhood of each pixel, as well as estimating a time series statistic, such as FPQ, to describe the experimental effect independently at each pixel. It is a daunting problem to derive from theory a good form for a bivariate null distribution, against which to test the two (spatial and temporal) statistics estimated for each pixel in the image. Yet it has already been shown (30) that such bivariate null distributions can be ascertained by Monte Carlo simulation for PET images, and the resulting brain activation maps are more sensitive than those obtained by methods that ignore the spatial dimensions of the data. We are working on an analogous development of randomization testing to enhance decision making in fMRI analysis.

## ACKNOWLEDGMENTS

The authors thank Brian Everitt, Graham Dunn, Bob Turner, Darren Hague, Chris Andrew, and Andy Simmons for technical support and valuable discussion. Thanks are also due to two anonymous referees for their constructive criticism of earlier drafts of this paper.

## REFERENCES

1. S. M. Rao, P. A. Bandettini, E. C. Wong, N. Z. Yetkin, T. A. Hammeke, W. M. Mueller, R. S. Goldman, G. L. Morris, P. G. Antuono, L. D. Estowski, V. M. Haughton, J. S. Hyde, *in* "Proc., SMRM, 11th Annual Meeting, 1992," p. 1827.
2. A. M. Blamire, S. Ogawa, K. Ugurbil, D. Rothman, G. McCarthy, J. M. Ellerman, F. Hyder, Z. Rattner, R. S. Shulman, Dynamic mapping of the human visual cortex by high speed magnetic resonance imaging. *Proc. Natl. Acad. Sci. (USA)* **89,** 11069–11073 (1992).
3. P. A. Bandettini, A. Jesmanowicz, E. C. Wong, J. S. Hyde, Processing strategies for time-course data sets in functional MRI of the brain. *Magn. Reson. Med.* **30,** 161–173 (1993).
4. K. J. Friston, P. Jezzard, R. Turner, The analysis of functional MRI time series. *Human Brain Mapping* **1,** 153–171 (1994).
5. J. R. Baker, R. M. Weisskoff, C. E. Stern, D. N. Kennedy, A. Jiang, K. K. Kwong, L. B. Kolodny, T. L. Davis, J. L. Boxerman, B. R. Buchbinder, V. J. Wedeen, J. W. Belliveau, B. R. Rosen, Statistical assessment of functional MRI signal change, *in* "Proc., SMR, 2nd Annual Meeting, 1994," p. 626.
6. J. Poline, B. M. Mazoyer, Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise pixel clusters. *J. Cereb. Blood Flow Metab.* **13,** 425–437 (1993).
7. A. P. Holmes, R. C. Blair, J. D. G. Watson, I. Ford, Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.*, in press.
8. E. S. Edgington, "Randomisation Tests," Marcel Dekker, New York, 1980.
9. P. J. Good, "Permutation Tests," Springer Verlag, New York, 1994.
10. S. Ogawa, T. M. Lee, A. R. Kay, D. W. Tank, Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. (USA)* **87,** 9868–9872 (1990).
11. A. S. David, T. Matharu, R. Rosselson, J. Cutting, Soft contact lenses with partial occlusion for prolonged lateralisation of visual input. *Neuropsychologia* **29,** 263–268 (1991).
12. N. Lange, Some computational and statistical tools for paired comparisons of digital images. *Stat. Meth. Med. Res.* **3,** 23–41 (1994).
13. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, "Numerical Recipes in C. The Art of Scientific Computing," Cambridge University Press, Cambridge, 1992.
14. K. J. Friston, S. Williams, R. Howard, R. S. J. Frackowiak, R. Turner, Movement-related effects in fMRI time series. *Human Brain Mapping*, in press.
15. R. Turner, P. Jezzard, H. Wen, K. K. Kwong, D. Le Bihan, T. Zeffiro, R. S. Balaban, Functional mapping of the human

visual cortex at 4 and 1.5 Tesla using deoxygenation contrast EPI. *Magn. Reson. Med.* **26,** 277 (1993).

16. C. W. Ostrum, "Time Series Analysis: Regression Techniques," Sage Publications, Beverly Hills, 1978.
17. S. D. Silvey, "Statistical Inference," Chapman and Hall, London, 1975.
18. W. N. Venables, B. D. Ripley, "Modern Applied Statistics with S-Plus," Springer Verlag, New York, 1994.
19. G. E. P. Box, D. A. Pierce, Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *J. Am. Stat. Assoc.* **65,** 1509–1526 (1970).
20. D. A. Hibbs, *in* "Sociological Methodology (1973–1974)" (H. L. Costner, Ed.), p. 252, Josey-Bass, San Francisco, 1974.
21. D. Cochrane, G. H. Orcutt, Application of least squares regression to relationships containing autocorrelated error terms. *J. Am. Stat. Assoc.* **44,** 32 (1949).
22. P. Bloomfield, *in* "Statistical Theory and Modelling. In honour of Sir David Cox, FRS" (D. V. Hinkley, N. Reid, E. J. Snell, Eds.), p. 152, Chapman and Hall, London, 1991.
23. G. E. P. Box, G. M. Jenkins, "Time Series Analysis: Forecasting and Control," Holden-Day, San Francisco, 1976.
24. J. M. Chambers, T. J. Hastie, "Statistical Models in S," Wadsworth & Brooks/Cole, Pacific Grove, California, 1992.
25. B. S. Everitt, "A Handbook of Statistical Analyses Using S-Plus," Chapman and Hall, London, 1994.
26. G. Anderson, P. Anderson, "The UNIX C Shell Field Guide," Prentice-Hall, Englewood Cliff, NJ, 1986.
27. B. F. J. Manly, "Randomisation and Monte Carlo Methods in Biology," Chapman and Hall, London, 1991.
28. K. J. Friston, Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping* **2,** 56–78 (1994).
29. N. Lange, Improved estimates of local fMRI signal magnitude, delay and dispersion for generic visual stimulation. *Human Brain Mapping* **S1,** 150 (1995).
30. J. B. Poline, B. M. Mazoyer, Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Trans. Med. Imaging* **13,** 702–710 (1994).

## ERRATUM

In the article "A Model for Magnitization Transfer in Tissue," by C. Morrison and R. M. Henkelman (Vol. 33, 475–482, 1995), Table 3 shows incorrect value for muscle. The corrected values are MAX MT = 0.66, $\Delta_{max}$ = 13.6 (kHz), $2\pi\omega_{1max}$ = 850 (Hz), R = 70 ± 4 ($s^{-1}$), $1/R_AT_{2A}$ = 22 ± 4, and $T_{2B}$ = 7.6 ± 0.5 ($\mu s$). All other values are as originally published.

Dr. Simon Graham is thanked for discovering this error.