ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 19 (2003) 1014–1032

www.elsevier.com/locate/ynimg

Diagnosis and exploration of massively univariate neuroimaging models

Wen-Lin Luo and Thomas E. Nichols*

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 USA

Received 7 May 2002; revised 13 December 2002; accepted 3 March 2003

Abstract

The goal of this work is to establish the validity of neuroimaging models and inferences through diagnosis and exploratory data analysis. While model diagnosis and exploration are integral parts of any statistical modeling enterprise, these aspects have been mostly neglected in functional neuroimaging. We present methods that make diagnosis and exploration of neuroimaging data feasible. We use three- and one-dimensional summaries that characterize the model fit and the four-dimensional residuals. The statistical tools are diagnostic summary statistics with tractable null distributions and the dynamic graphical tools which allow the exploration of multiple summaries in both spatial and temporal/interscan aspects, with the ability to quickly jump to spatiotemporal detail. We apply our methods to a fMRI data set, demonstrating their ability to localize subtle artifacts and to discover systematic experimental variation not captured by the model.

Keywords: Diagnosis; Exploratory data analysis; Artifact; Autocorrelation; Global scaling; Interactive visualization

Introduction and motivation

Neuroimaging analyses proceed by localizing brain regions exhibiting experimental variation. A PET or fMRI experiment yields a sequence of large three-dimensional images of the subject's brain, each containing as many as 100,000 volume elements or voxels. The typical analysis strategy is marginal or "massively univariate" (Holmes, 1994), where data for each voxel are independently fit with the same model (Friston et al., 1995). Images of test statistics are used to make inference on the presence of an effect at each voxel.

The main purpose of this work is to establish the validity of inferences in neuroimaging through diagnosis of model assumptions. Hypothesis tests and P values depend on assumptions on the data, and inferences should not be trusted unless assumptions are checked. Diagnosis is usually done by the graphical analysis of residuals (Neter et al., 1996; Draper and Smith, 1998). For example, one standard tool is a scatter plot of residuals versus fitted values, useful for diagnosing nonconstant variance, curvature, and outliers. This sort of graphical analysis is not practical since it is not possible to evaluate 100,000 plots.

The other purpose of this work is to characterize signal and artifacts through exploratory data analysis (EDA; Tukey, 1977). EDA is an important step in any statistical analysis, as it familiarizes the analyst with form of the expected experimental variation, the presence of unexpected systematic variation, and the character of random variation. As with model diagnosis, traditional EDA tools are graphical and cannot be applied voxel-by-voxel exhaustively. Fortunately EDA can also be accomplished by exploring the fit and the residuals (Hoaglin et al., 1983). A model partitions data as the sum "Data = Fit + Residuals," and in neuroimaging data the fit and residuals are individually more amenable to exploration than the full data. The fit is parameterized by the user and is readily interpretable, while the residuals are homogeneous and unstructured if the model fits. Interesting features in the residuals can be found by use of statistics sensitive to structure or inhomogeneity; for example, something as simple as outlier counts per scan can quickly identify interesting scans. Diagnosis and EDA are enmeshed: Diagnosis takes the form of exploration of

* Corresponding author. Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. Fax: +1-734-763-2215.

E-mail address: nichols@umich.edu (T.E. Nichols).

1053-8119/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1053-8119(03)00149-6

diagnostic statistics, and exploration of residuals serves to understand problems identified by diagnosis.

In this work we propose a collection of tools and explicit procedures to check model assumptions and to explore fit and residuals. The two key aspects of our work are (1) images and one-dimensional summaries that characterize fit and residuals and (2) dynamic visualization tools to explore these summaries and to efficiently identify spatiotemporal regions (or voxels and scans) of interest.

We use the term "model summaries" to refer to images that assess fit or residuals at each voxel, and "scan summaries" to refer to time series (fMRI) or one-dimensional (1-D) vectors (PET, etc.) that assess fit or residuals over space. For model summaries, we use both images of linear model parameters and images of diagnostic statistics. For example, we assess linear model assumptions like normality, homoscedasticity (homogeneous variance), and independence of errors with scalar diagnostic statistics; to view these diverse measures on a common scale, we create images of $-\log_{10} P$ values. For scan summaries, we use measures which describe model fit and residuals over an image, as well as preprocessing parameters. For example, global intensity and outlier count per image both can capture transient acquisition problems, and in fMRI, head motion estimates are useful for finding scans with motion artifacts.

The dynamic visualization tools are used for simultaneously exploring multiple model and scan summaries and for quickly jumping from these summary margins to the full raw or residual data. We use linked orthogonal viewers to explore the images of model summaries, and parallel plots with linked cursors to study of plots of scan summaries. From a model summary image the model detail for a specific voxel can be brought up, including plots of the raw data, fitted model, residuals, and traditional diagnostic plots. From a plot of scan summaries the scan detail for a specific image can be displayed, consisting of images of studentized residuals. These tools have been implemented as statistical parametric mapping diagnosis (SPMd, http://www.sph.umich.edu/~nichols/SPMd), a toolbox for SPM (http://www.fil.ion.ucl.ac.uk/spm).

In this article we assume independent errors at each voxel. This assumption is suitable for data from PET, SPECT, VBM (Ashburner and Friston, 2000), or simple second-level fMRI models (Holmes and Friston, 1999) and for single-subject fMRI models after decorrelation or whitening. Our methods are also appropriate for fMRI covariance model building: Since the appropriate model for the signal must be found before fMRI noise can be modeled, our methods can be applied with independence regarded as a "working" autocorrelation model. Also, our autocorrelation diagnostics will capture the form and spatial heterogeneity of autocorrelation, allowing the exploration of temporal dependency before it is modeled.

There has been little previous work in neuroimaging model diagnosis (Razavi et al., 2001; Nichols and Luo, 2001). In EDA there are many data-driven tools that have found use in fMRI, including clustering (Goutte et al., 1999; Moser et al., 1999), independent components analysis (ICA software can be found, for example, at http://www.fmrib.ox.uk/fsl/melodic) (McKeown et al., 1998), and principal components analysis (PCA software can be found, for example, at http://www.madic.org/download) (Kherif et al., 2002). Our work differs from these EDA tools in that we individually explore fit and residuals, instead of raw data, and that we support our EDA with statistical summaries and P values to make inferences on the magnitude of discovered patterns (relative to a putative model).

In the next section we introduce these summaries and give specific strategies for model diagnosis. In the subsequent section we report on simulation studies that investigate the performance of the diagnostic summaries with respect to different correlation conditions. Finally, we demonstrate our tools on a fMRI data set.

Methods

Consider a general linear model fit at each voxel. For a given voxel we have

$$Y = X\beta + \varepsilon,$$

where Y is a N-vector of responses, X is a $N \times p$ matrix of p predictor variables, β is a p-vector of unknown parameters, and ε is a N-vector of unknown, random errors. This general form captures almost all used models, including ANOVA, t test, and complicated fMRI models; the predictors may also include variables accounting the global signal or, in fMRI, drift and the phase of the hemodynamic response.

To make inferences at each voxel we must assume that ε is a vector of normal random variables with expectation 0 and variance–covariance matrix $\sigma^2 V$. In this work we assume that $V = I$, which is appropriate for PET, SPECT, MEG, VBM, or long-TR fMRI data. It is also appropriate for any fMRI data when Y and X have been whitened based on an estimated autocorrelation structure (Burock and Dale, 2000; Woolrich et al., 2001).

To make inferences corrected for multiple testing problem, additional assumptions may be needed. Gaussian random field theory methods have several assumptions (Petersson et al., 1999) and a thorough assessment of them is beyond the scope of this article. Briefly, the essential requirements are univariate normality and sufficient smoothness. Normality is easy to check with the tools below, and the smoothness assumption requires the estimated FWHM smoothness to be at least three times the voxel size. For cluster size tests the other key assumption is stationary noise covariance, which is addressed elsewhere (Hayasaka and Nichols, manuscript in preparation).

The least squares estimator of β is $\hat{\beta} = (X^\top X)^{-1} X^\top Y$. A contrast vector c is a length-p row vector defining an effect

Table 1
Model summaries

Statistic	Assesses	Null Dist^n	Reference
Contrast estimates	Signal	t[a]	Appendix A
Standard deviation/PCT[b]	Artifacts		See text
Durbin–Watson	$\mathrm{Cor}(\varepsilon_i, \varepsilon_i + 1) = 0$	Beta	Durbin and Watson (1950, 1951)
Cumulative periodogram with BLUS residuals	$\mathrm{Var}(\varepsilon) = \sigma^2 I$	Uniform	Diggle (1990), Schlittgen (1989), Smirnov (1948), Stephens (1970), Theil (1971)
Cook–Weisberg score test	$\mathrm{Var}(\varepsilon_i) = \sigma^2$	χ^2	Casella (2001), Cook and Weisberg (1983)
Shapiro–Wilk	Normality	Normal[c]	Royston (1982), Shapiro et al. (1968), Stephens (1974)
Outlier count	Artifacts	Binomial	Neter et al. (1996), Ryan (1997)

[a] After standardization.
[b] Percent change threshold.
[c] After transformation.

of interest, or contrast, $c\hat{\beta}$. Central to exploration and diagnosis are the residuals

$$e = Y - \hat{Y} = Y - X\hat{\beta},$$

where $\hat{Y} = X\hat{\beta}$ are the fitted values. Note that even when the errors are homogeneous and independent, the residuals are heteroscedastic and dependent, as per $\mathrm{Cov}(e) = (I - H)\sigma^2$, where $H = X(X^\top X)^{-1}X^\top$. To visualize residuals with homogeneous variance we use studentized residuals, $r_i = e_i/\sqrt{\mathrm{diag}(I - H)\hat{\sigma}^2}$. When the dependence of the residuals is problematic we use Best Linear Unbiased residuals with a Scalar (diagonal) covariance matrix or BLUS residuals (Theil, 1971). BLUS residuals are unbiased in the sense that their expectation is zero and they are best in that their distance from the true errors is minimized in expectation.

After model fitting and calculating residuals we can compute the diagnostic statistics and summary measures. The two main components of our work are the model and scan summaries of the data and the interactive visualization tools to explore those summaries.

Model summaries

Our model summaries are images of model parameters, to represent fit, and residual summaries, to assess lack-of-fit and model assumptions (see Table 1). To provide a consistent metric for visualizing the diagnostic measures we create images of $-\log_{10} P$ values.

Exploratory summaries

Contrasts and statistic images. We prefer images of signal that are maximally interpretable, and hence use percent change contrast images in addition to t images. Note that while t images are scale-invariant and unitless, a contrast estimate $c\hat{\beta}$ has units determined by the predictors and the contrast vector, and hence a linear model with interpretable units is required for $c\hat{\beta}/\mu \times 100\%$ to be percent change (see Appendix A).

We use F images to summarize nonscalar effects, such as a subject/block effect or a hemodynamic response parameterized with a finite impulse response model. For an anatomical reference we create a grand mean image; while other high resolution images may be available, there are often misalignment problems owing to head motion or susceptibility artifacts.

Standard deviation and percent change threshold. While residual standard deviation is a key summary measure, it lacks concrete units that, say, a contrast image has (response magnitude, all other effects hold constant). To increase interpretability, we characterize residual uncertainty with the $(1 - \alpha)\%$ confidence margin of error for an effect of interest. The margin of error is the half-width of a $(1 - \alpha)\%$ confidence interval (either corrected or uncorrected). If an effect is expressed in units of percent change, we call this quantity the percent change threshold (PCT), as it is the minimum percent change needed to reach level α significance.

Another reason to use PCT is that it intuitively expresses the impact of standard deviation on power. For example, in a fMRI data set, say that a region with a PCT of 10% is found; the immediate interpretation is that no fMRI signal will be detected in that region, since BOLD signal change rarely exceeds 5% (for 1.5 T, Moonen and Bandettini, 2000) (see http://www.sph.umich.edu/~nichols/PCT for more detail). Of course, collecting more data or more subjects into a fixed effects analysis will reduce PCT.

Diagnostic summaries

From a detailed review the linear model diagnostics literature we selected the most appropriate measures for neuroimaging data (Table 1). The key diagnostic statistics are the Cook–Weisberg score test for homoscedasticity, Shapiro–Wilk test for normality, and Durbin–Watson statistic and cumulative periodogram test with BLUS residuals for independence. (Raw fMRI residuals have no energy in low frequencies due to drift modeling, causing the cumulative periodogram tests to falsely reject; the use of BLUS residuals corrects this.) We also use an image of outlier counts per voxel. For detailed definition of the statistics,

Table 2
Scan summaries

Summary	Definition	Function
Experimental predictors	Predictors from the design matrix	Provide reference to other plots.
Global signal	Average of intracerebral voxels	Assess possible global confound, bad scans
Scan outlier count	Sum of outliers over each scan	Detect transient acquisition problems, bad subjects
Image preprocessing parameters	Registration shift and rotation movement parameters	Capture aspects of artifacts or anomalies

their distribution under a null hypothesis of model fit, and how to assess them, please refer to Luo and Nichols (2002) and respective references.

Scan summaries

Our scan summaries are vectors where each element assesses a single image (see Table 2). Since we do not have an explicit spatial model to evaluate, our scan summaries use more ad hoc measures. We use the experimental predictors, outlier count per scan, global signal, and some image preprocessing parameters such as registration shift and rotation movement parameters. These measures capture motion, physiological, scanner artifacts or possible confounding variables; for multisubject studies they may identify anomalous subjects. While inference is not feasible on scan summaries, we do compute reference values under the null hypothesis when possible. For example, the expected outlier count is easy to compute as the number of voxels times the Gaussian tail area of the outlier cutoffs.

Diagnosis strategies

A dozen summary images and plots are an improvement over looking at every diagnostic plot and residual image, but these summary measures are of limited use if just examined one-by-one and not linked to the full data. Hence we use a dynamic graphical tool to simultaneously view several summaries, linked to residual plots and images. Specifically we use four viewers (Fig. 1) to efficiently explore the summaries and, as guided by the summaries, the fit and residuals.

It is not immediately clear, however, in what order these summaries should be examined, how the tools should be used with the summaries, and how the results of investigations should be applied to the final analysis of the data. In this section we give an outline of strategies to simultaneously (1) check assumptions, (2) explore expected and unexpected variability, and (3) address problems found (see Table 3). In short, we move from summaries to detail and perform exploration and diagnosis of noise before exploration of signal.

Step 1: explore plots of scan summaries

We use parallel plots of scan summaries to find scans affected by artifacts and acquisition problems (Fig. 1a). We first check for systemic problems. For example, whether the global signal is related to experimental effects or if there is excessive movement. Second, we check for transient problems, like the jumps or spikes in the global signal, the outlier count per scan, and the motion parameters. Usually, these jumps correspond to head movements and acquisition artifacts. Third, we check the relationships between different scan summaries, for instance, whether movement jumps and outlier spikes coincide or if global spikes coincide with outliers or movements. From this information, we note which scans are possibly corrupted or may be influential to the data analysis. The origin of the spikes can be investigated in detail in Step 4.

Step 2: explore images of model summaries

Model summaries are displayed next with linked orthogonal slice viewers (Fig. 1b) to do both diagnosis and exploration. In the diagnostic model summaries we pay special attention to regions with both significant diagnostic statistics and anticipated experimental effects.

For exploratory purposes we search images of signal and noise, focusing on the noise first, principally with a PCT image. We window the PCT image such that the modal PCT value is half the maximal intensity (see Appendix B); this gives middle gray the interpretation of typical sensitivity and white that of less than one-half typical sensitivity. Regions with large PCT are noted as possible sites of Type II errors. We next check t or F images of nuisance effects, such as drift; with use of the model detail, interesting features in the image of drift magnitude can often lead to discovery of artifacts. Finally we explore the expected signal, as measured with percentage change, t or F images. We localize interesting activations and note any broad patterns. In particular, extensive positive or negative regions indicate a subtle signal (or artifact) that would not be evident in a thresholded statistic image. For any notable region discovered, by diagnosis or exploration, we check its model detail.

Step 3: explore model detail

For a given voxel we examine the model detail plots (Fig. 1c), doing so interactively with images of model summaries to characterize the sources of the significant experimental or diagnostic statistics. From the plots of data with fit and residuals, we can not only assess the goodness-of-fit

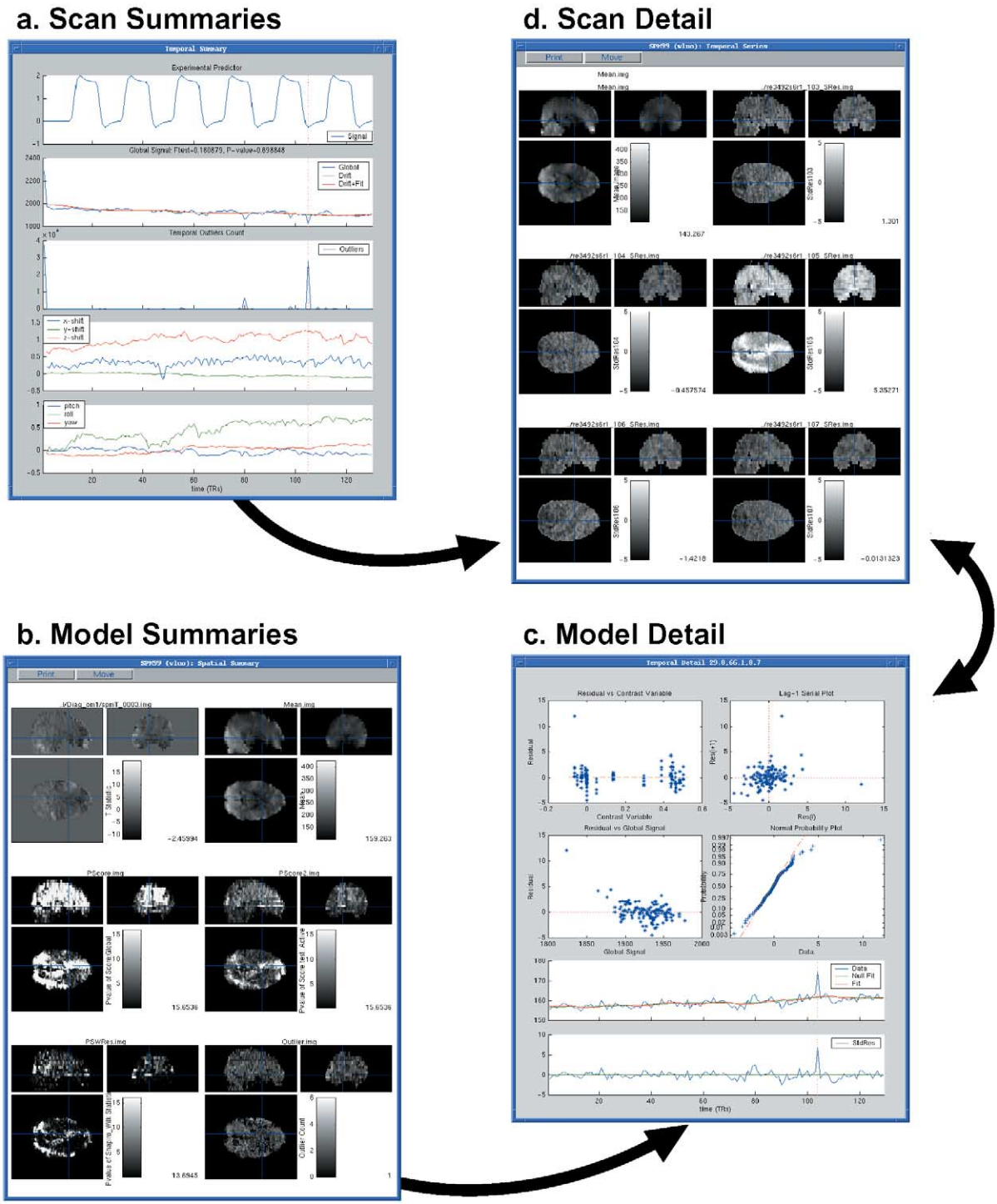


Fig. 1. We use a dynamic graphical tool to efficiently navigate through the various diagnostic measures. We use plots of scan summaries (a) and images of model summaries (b) to find individual voxels and scans of interest; a single voxel is studied with plots of model detail (c) and a series of residual images are viewed in scan detail (d).

of the model to the intrinsic signal (important for fMRI BOLD data), but also identify unmodeled signals, that is, any systematic variation not captured by the model. Also, from the plot of residuals, we note possible outlier scans. We reference these with the outlier count per scan and characterize their spatial extent with the scan detail. Furthermore, we use the diagnostic residual plots to check the specificity of the significant diagnostic statistics. For example, if a voxel is large in the image of Cook–Weisberg homogeneous variance statistic, we use a residual plot versus predictor variable to verify that system heteroscedasticity and not a single outlier is responsible.

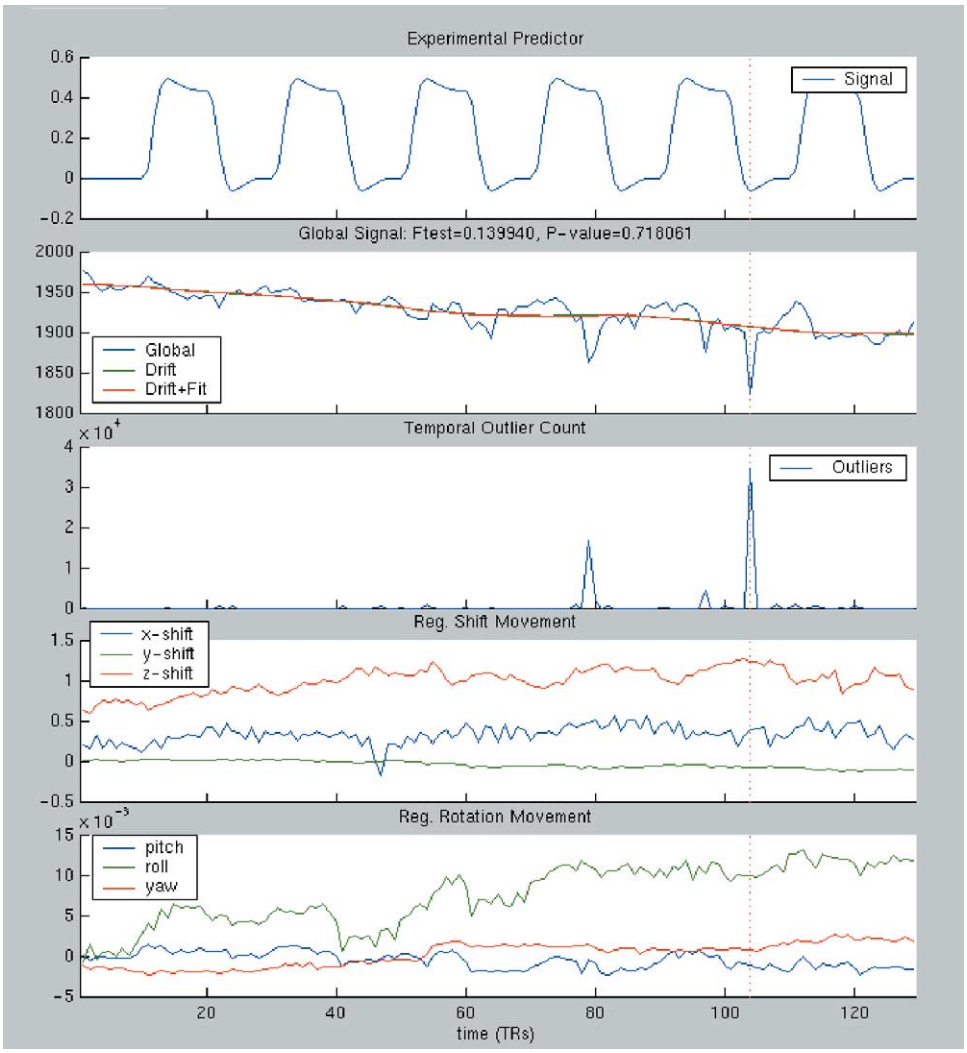


Fig. 2. Plots of scan summaries for the full 129-scan data set. Plots show, from top to bottom, experimental predictor, global signal, outlier count, translational head movements, and rotational head movements.

Step 4: explore scan detail

As guided by the model detail or plots of scan summaries, we use a sequence of studentized residual images (Fig. 1d) to spatially localize problems identified in previous steps. Fixed and spinning maximum intensity projection (MIP) images are helpful to give an overall indication of the problems. When examining the series of residual images, we note the spatial and temporal/interscan extent of the problem. For example, in fMRI, an artifact confined to a single slice in a single volume suggests shot noise, while an extended artifact may be due to physiological sources or model deficiencies.

Step 5: remediation

Several approaches can be applied to address problems identified by previous steps. For the problem scans discovered in Step 1, we may possibly remove them from the analysis. We are very judicious on removing scans; we only discard an observation if we are convinced that there are deterministic measurement or execution errors (Barnett and Lewis, 1994), like gross movements or spike noise. Another approach to outliers is Windzorization (Wilcox, 1998), the shrinkage of outliers to the outlier threshold. In our experience, outliers indicate corrupted data, and we prefer to eliminate such data rather than retaining them in modified form. Corresponding scans and design matrix rows must be deleted together; the design matrix should then be checked that it is not exactly nor nearly rank deficient. In addition to omitting scans, we may find problem voxels in Step 2. These voxels can simply noted and ignored or explicitly masked from the analysis.

The other approach is to modify the model. For example, if we find incorrectly modeled experimental variation from the model detail, we may consider adding other variables to the model to improve the fit. In contrast, if the global signal is found to be significantly correlated with the experimental paradigm, it may be preferable to omit this confound as a covariate entirely (Aguirre et al., 1998).

Table 3
Diagnosis strategies

Step	Action
1. Explore scan summaries	Check for systemic problems. Check for transient problems. Check for relationships between summaries.
2. Explore model summaries	Check for violations of assumptions. Explore noise, nuisance variability. Explore experimental signal.
3. Explore model detail	Check for unmodeled, systematic variation. Note possible problem scans. Check specificity of significant diagnostic statistics.
4. Explore scan detail	Check temporal/inter-scan extent of problem. Check spatial extent of problem.
5. Remediation	Remove problem scans. Modify model. Mask out problem regions.
6. Resolution	Declare significant activation valid, or Declare significant activation as questionable. Describe unmodeled and artifactual variation.

After removing possible outliers and/or modifying the model, we refit the model and repeat the above processes again until we are satisfied that experimental inferences are valid and that gross artifactual variation has been omitted or at least characterized.

Step 6: resolution

After all the analyses and diagnoses are done, we summarize the results of diagnosis and exploration. We declare each significant region as valid, questionable or artifactual. A valid activation has assumptions clearly satisfied, while a questionable region has some significant diagnostics but exploration of the fit and residuals has shown the activation to be believable. Artifactual activation is clearly due to outliers or acquisition artifacts which could not be remedied. In brain volumes with no significant activation, it is also important to report on regions with significant diagnostic statistics. The source of these significant diagnostics may be related to the unmodeled signal or artifactual variation and may be the source of new neuroscientific hypotheses, a new type of physiological signal, or simply problems for physicists to solve.

Simulation studies

In this section we examine the performance of the model summary diagnostic statistics using simulated data sets. All of our model summary statistics have been well studied under the null hypothesis of model fit (see respective references), so extensive simulations under the null are not in order. Extensive evaluations under alternatives, on the other hand, are problematic because the space of the alternatives is very large, consisting of all combinations of possible types model lack of fit: autocorrelation, outliers, model misspecification, heteroscedasticity, etc. Some evaluations under alternatives can be found in the references (for example, in Shapiro et al. (1968), they investigate the performance of Shapiro–Wilk under alternative distributions and different sample size and show that the Shapiro–Wilk test exhibits sensitivity to nonnormality over a wide range of alternative distributions.). Hence we only investigate the alternative of greatest concern, that of autocorrelation. We do this in part to demonstrate the sensitivity of our dependency statistics (Durbin–Watson and cumulative periodogram), but primarily to characterize the specificity of the other measures under the violation of their independence assumption.

Simulation methods

Time series data were simulated from an 84-observation model, corresponding to a publicly available data set (http://www.fil.ion.ucl.ac.uk/spm/data), single-subject epoch auditory fMRI activation data; we used such a short-length time series to characterize the small-sample limitations of our diagnostic statistics. The simulated data were composed of the sum of two series: One was the fixed response effect including nine covariates corresponding to intercept, a experimental condition, and seven drift terms; the other series was the random error, which was either white noise, a first-order autoregressive processes with different degree of correlation (0.1–0.5), or an order-12 autoregressive process. The parameters of these covariates and the 12 AR parameters were obtained from a real data set. A linear regression model was fit to the simulated data and residuals were created; we computed six diagnostic statistics, Durbin–Watson (DW), cumulative periodogram (CP) with BLUS residuals (Theil, 1971), Shapiro–Wilk (SW), outliers, and two Cook–Weisberg score tests, with respect to global signal (CW-G) and predicted values (CW-P). We also calculated a cumulative periodogram with ordinary residuals (CP*), instead of BLUS residuals.

For each type of random noise structure, we created 10,000 realizations; for each realization the diagnostic statistics and corresponding P values were calculated. The performance of the statistics were measured with two criteria. First, the percentages of rejection under null hypothesis at three rejection levels (0.05, 0.01, and 0.001) under various correlation conditions are computed. Second, Q–Q plots of the logarithm of the P values were created and helpful to examine the behavior over a range of α's.

Table 4
Simulation results

α Level	DW	CP	CW-G	CW-P	SW	Outlier	CP*
			White noise				
0.05	0.0614	0.0644	0.0562	0.0508	0.0509	0.2876	0.1259
0.01	0.0146	0.0133	0.0100	0.0114	0.0104	0.0243	0.0373
0.001	0.0013	0.0016	0.0009	0.0016	0.0019	0.0005	0.0043
			AR(1) process: $\rho = 0.1$				
0.05	0.2222	0.0535	0.0557	0.0524	0.0495	0.2820	0.0540
0.01	0.0763	0.0116	0.0127	0.0113	0.0112	0.0255	0.0100
0.001	0.0151	0.0006	0.0017	0.0013	0.0012	0.0009	0.0005
			AR(1) process: $\rho = 0.2$				
0.05	0.5002	0.1466	0.0583	0.0532	0.0512	0.2778	0.1072
0.01	0.2588	0.0475	0.0138	0.0107	0.0107	0.0243	0.0311
0.001	0.0811	0.0082	0.0015	0.0010	0.0008	0.0011	0.0044
			AR(1) process: $\rho = 0.3$				
0.05	0.7745	0.3473	0.0669	0.0588	0.0506	0.2770	0.2980
0.01	0.5345	0.1531	0.0166	0.0130	0.0105	0.0244	0.1369
0.001	0.2651	0.0391	0.0026	0.0016	0.0007	0.0009	0.0356
			AR(1) process: $\rho = 0.4$				
0.05	0.9199	0.6090	0.0748	0.0605	0.0512	0.2769	0.5577
0.01	0.7905	0.3740	0.0162	0.0123	0.0113	0.0252	0.3236
0.001	0.5436	0.1507	0.0033	0.0020	0.0015	0.0006	0.1285
			AR(1) process: $\rho = 0.5$				
0.05	0.9782	0.8114	0.0862	0.0606	0.0558	0.2610	0.7813
0.01	0.9253	0.6161	0.0251	0.0135	0.0131	0.0206	0.5835
0.001	0.7846	0.3489	0.0036	0.0019	0.0016	0.0008	0.3159
			AR(12) process				
0.05	0.0498	0.0862	0.0634	0.0545	0.0527	0.2819	0.1194
0.01	0.0100	0.0217	0.0143	0.0122	0.0106	0.0249	0.0321
0.001	0.0014	0.0030	0.0011	0.0020	0.0009	0.0008	0.0031

Simulation results

Table 4 shows the estimated rejection rates of the diagnostics. The Monte Carlo standard deviations of the rejection rates are 2.18×10^{-3}, 9.95×10^{-4}, and 3.16×10^{-4} for the 0.05, 0.01, and 0.001 α levels, respectively. The *Q–Q* plots of *P* values do not reveal any behavior not captured in Table 4 and hence are omitted.

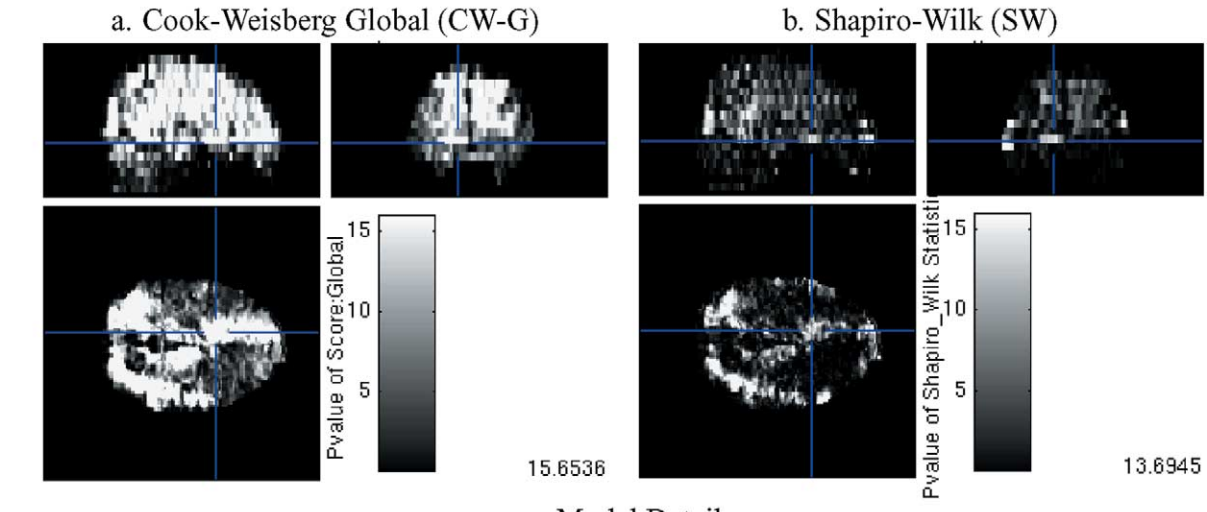
Comparisons of the autocorrelation diagnostics

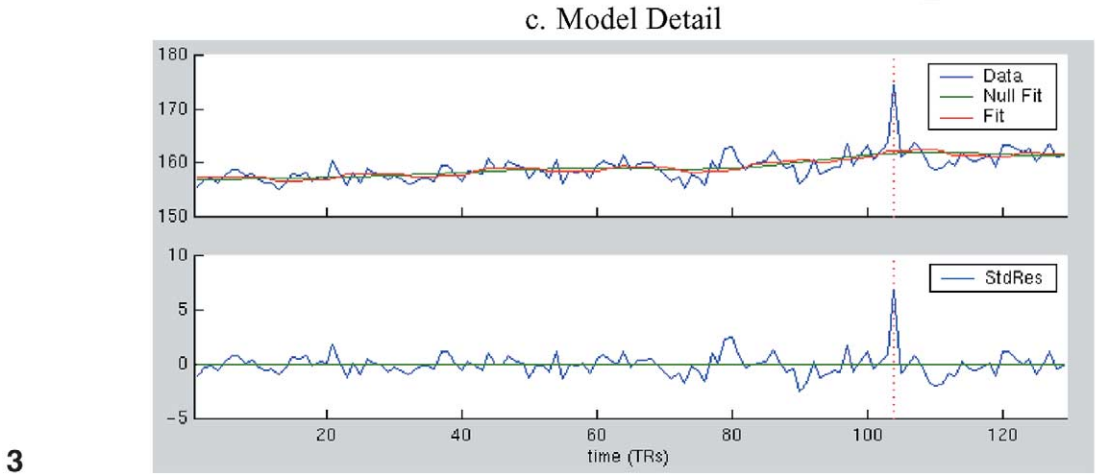
The white noise results show that estimated Type I error rates are close to nominal. Under AR(1) noise processes, as expected, the percentage of rejection increases as the correlation coefficient increases for both test statistics, except for one case ($\rho = 0.1$ for CP). Furthermore, the percentages of rejection are larger for DW statistic than that for cumulative periodogram test, which is consistent with the optimality of the DW statistic within the class of AR(1) noise. However, the cumulative periodogram test is superior to the DW test for detecting high-order autoregressive processes, as indicated in last rows of Table 4.

Performance of other diagnostics

The other purpose of the simulation study is to examine the specificity of the statistics under white noise and different error processes. Under different correlation structures, the results of SW statistic for normality are similar and most of them are within two standard deviations of the nominal α levels.

For the Cook–Weisberg homogeneous variance test, un-





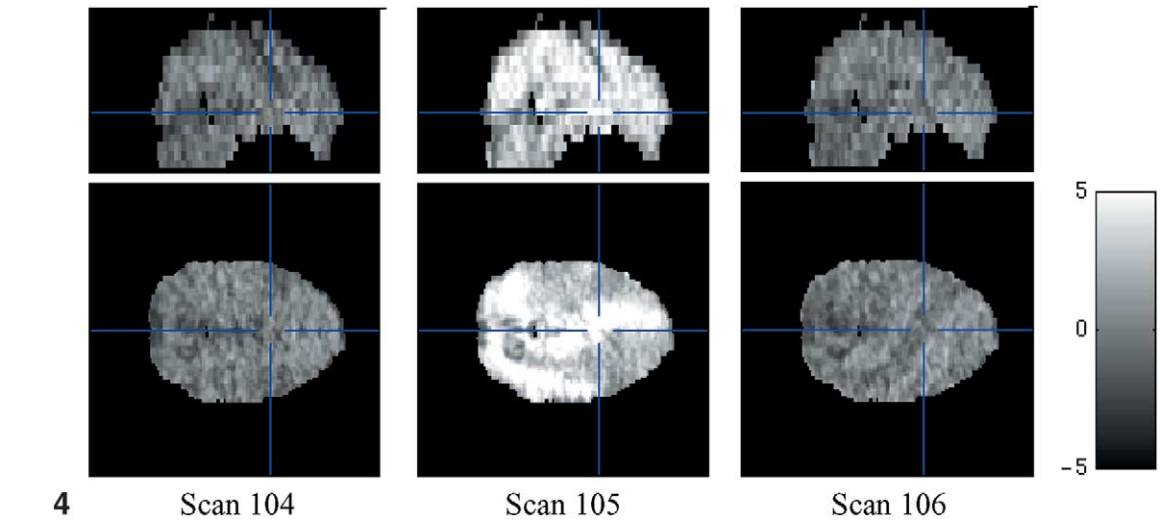
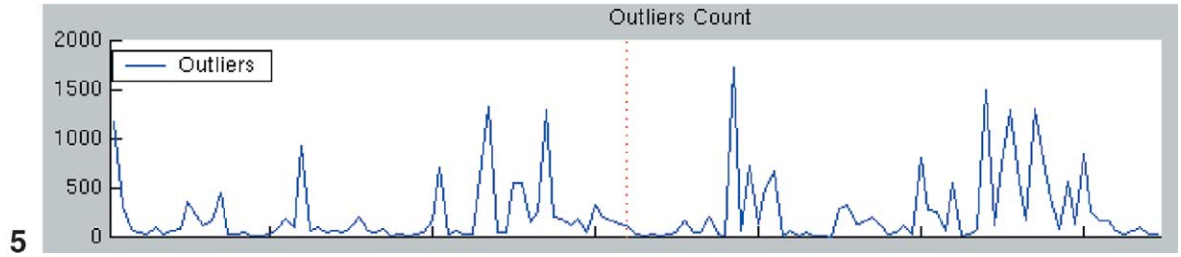
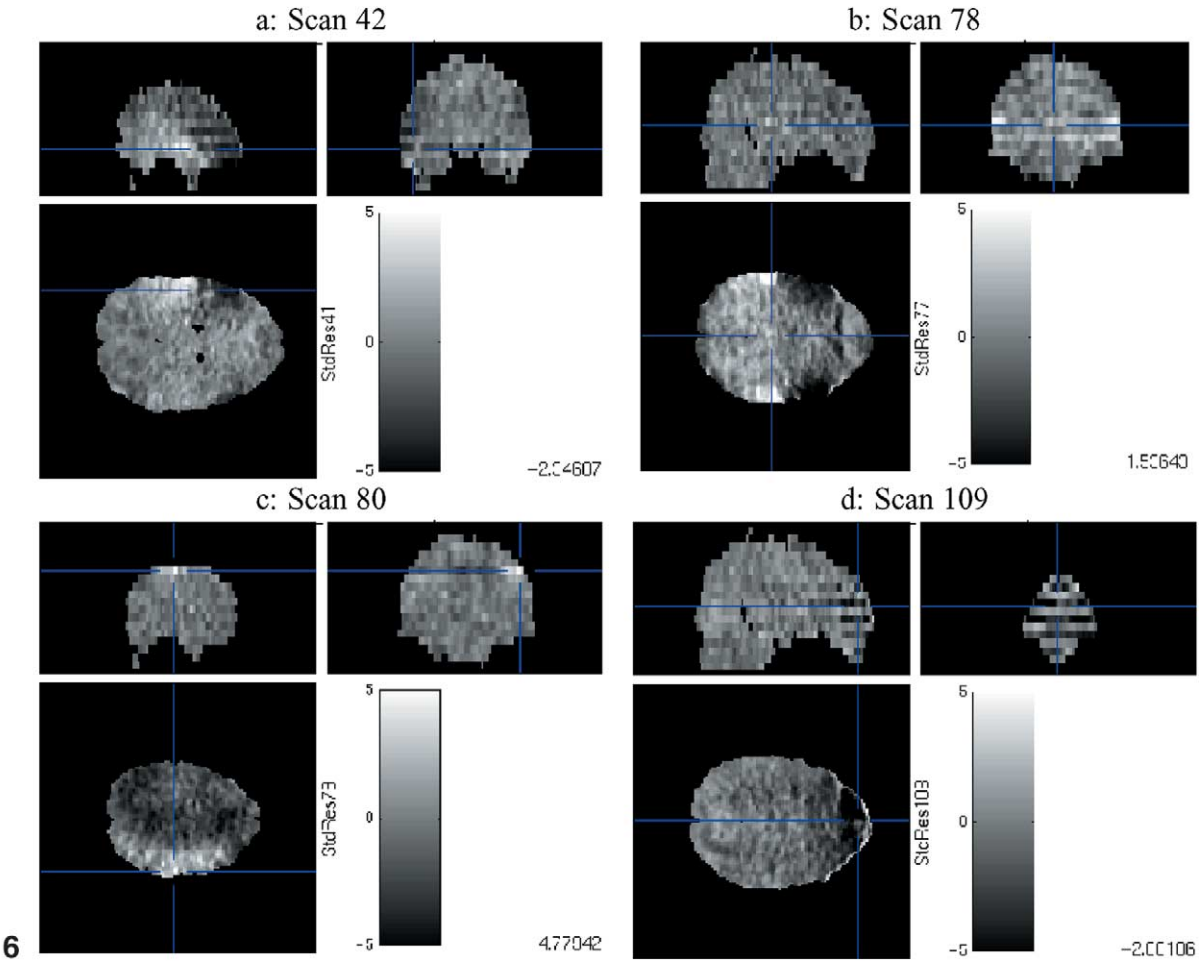


Fig. 3. Characterization of spiral artifact and identification of a problem scan. Model summaries reveal a distinct spiral pattern, here shown in the diagnostics for homogeneous variance with respect to global signal (a, CW-G), and normality (b, SW). Note that the CW-G detects problems across the volume, while SW only detected artifacts in one plane. Model detail of the fit and studentized residuals (c) for the indicated voxel shows a prominent outlier at scan 105. Note that diagnostic statistic images in this and other figures have units of $-\log_{10} P$ values for consistent visualization.

Fig. 4. Scan detail around scan 105; the studentized residual images for scans 104–106. The increased intensity of scan 105's residuals indicates a decreased intensity in the data. The spiral pattern in this scan also corresponds to that found in the CW-G and SW images (see Fig. 3).



5



6

Fig. 5. Scan summary of the outlier count without global scaling. All other scan summaries are the same as before (see Fig. 2).
Fig. 6. Acquisition artifacts reveals in the studentized residuals (scan detail). These scans are identified as having large numbers of outliers.

der the white noise, the rejection rates are nominal for all α levels. For the AR(1) noise processes, both CW-G and CW-P tend to give estimated Type I errors that are higher than the nominal α level. As the AR(1) process correlation increases, the CW-G shows increasing Type I error at the three nominal α levels, while CW-P is better, not showing appreciable anticonservativeness until $\rho \geq 0.3$. Under the AR(12) process, the Type I errors are all slightly greater than the α levels.

Due to the discreteness of the outlier count, the rejection rates are far from the nominal $\alpha = 0.05$ and $\alpha = 0.01$. However, comparison of the rejection rates across noise processes shows that the rejection rates are quite stable, suggesting that the outlier count is quite resilient with respect to autocorrelation.

The rejection rates for CP* under the white noise shows the problem with using ordinary residuals with the cumulative periodogram test: For all three α levels, the CP* rejection rates are about twice the CP rates. Moreover, for almost all dependent noise simulations, the CP rejection rates exceeded the CP* rates. Hence these results suggest that the cumulative periodogram test using BLUS residual is both more specific and more sensitive than using ordinary residuals.

In summary, these simulation results argue that our autocorrelation diagnostics are specific and sensitive, that our

normality and outlier statistics retain specificity under autocorrelation. The Cook–Weisberg score tests for heteroscedasticity show some excess false positives when autocorrelation is strong. We conclude that if a CW statistic is large when strong autocorrelation is detected, an appropriate residual plot should be checked to confirm the presence heterogeneous variance.

Real data analysis

In this section, we demonstrate our methods and their ability to localize subtle artifacts and to understand their causes. We use data from a study of new motion correction methods, where the subject was asked to speak aloud.

Experiment

The study employed a block design of a word generation task. The stimulation paradigm consisted of six cycles of rest/active, with a final rest condition; there were 20 scans per cycle. During the active condition, the subject was asked to generate a word that starts with each letter of the alphabet starting from "A." Functional data were acquired on a 1.5-T GE Signa magnet. A sequence of 130 EPI images was collected with a TR of 3000 ms and a TE of 40 ms. Images consisted of $128 \times 128 \times 20$ voxels, with voxel dimensions of $1.88 \times 1.88 \times 7$ mm. The first scan was discarded to allow for T1 stabilization.

Images were corrected for slice timing effects and subject head motion using SPM99 (http://www.fil.ion.ucl.ac.uk/spm). While some recommend against the use of the global signal (R. Henson, personal communication), we used the conventional scaling approach. After global scaling, the data at each each voxel were fitted with a general linear model; covariates consisted of the convolution of design box-cars with a canonical hemodynamic response function and a six-element discrete cosine transform basis set to account for drift. Summary statistics described above were computed for diagnosis, including a t image based on rest and activation contrast and a grand mean image for comparison and localization. We evaluated the data and model as outlined previously.

Results

We start with scan and model summaries and then explore model and scan detail as guided by the summaries. Inspection of the scan summaries reveals no systemic problems (Fig. 2), and in particular there is no significant correlation between the global signal and experimental condition ($P = 0.7181$). The global signal has a general downward trend and has several negative dips. The outlier count has several spikes, scan 105 in particular having over 70% outliers (0.03% is nominal). Significantly, the dips in the global signal correspond to spikes in the outlier count. The movement parameters display some transient movements, but these do not correspond with outlier or global events; the magnitude of estimated movement is modest.

Of the model summaries, the homoscedasticity Cook–Weisberg score tests (CW) and the normality Shapiro–Wilk (SW) test are the most notable, with a dramatic spiral pattern (Figs. 3a and b). This pattern is limited to one slice on the CW score test computed with respect to the experimental predictor (CW-E) and the SW test, but extends over the whole brain for the CW score test computed with respect to the global signal (CW-G). We examine the model detail for a voxel $(-11, -30, -20)$ in the slice with this artifact (Fig. 3c) and find that the data are nominal except for an outlier at scan 105. This leads us to view the scan detail about scan 105 (Fig. 4). There is an global hyperintensity exhibited in the residuals at scan 105, with the spiral artifact clearly evident.

Having identified this corrupted observation, one course of action would be to remove scan 105. However, we are more concerned of this as an artifact of global normalization. Standardizing by global intensity presumes that perturbations captured by the global are common to all voxels. However, the large residuals all over scan 105 and local spiral pattern are consistent with a single-plane hypointensity artifact: A local reduction in T2* magnitude causes a dip in global intensity, which results in the whole volume being overscaled. Hence instead of omitting a scan, we alter the model by removing the global scaling.

Global scaling eliminated

The scan summaries are the same after removing the global scaling, except for the outlier count (Fig. 5). The outlier plot is improved, but many scans have considerably more than the expected outlier rate of 0.03% or 145 per scan. Checking the scan detail (studentized residuals) for scan 105 reveals that the volume as a whole is nominal, while one plane is, as before, corrupted. In fact, the scan detail for most outlier-spike images shows similar acquisition artifacts either confined to a single plane or to every other plane (see Fig. 6 for examples.) In general these dramatic artifacts are not evident by inspection of the raw images. However, examination of temporally differenced raw images does reveal similar patterns, implying that these patterns are not attributable to the particular model we use.

The model summaries still reveal problems. The DW and CP images detect regions with periodic variation corresponding to about one-quarter cycle off from the experimental paradigm (Fig. 7). The regions exhibiting this temporal pattern are principally in the primary visual cortex and in the cerebellum, though this pattern is also found throughout the posterior surface of the brain and even in third ventricle. Hence we note this temporal pattern as artifactual and probably vocalization-related.

The CW-E image has a pronounced hyperintensity in the frontal pole (Fig. 8a). Exploration of model detail localizes the heterogeneous variance to the last epoch (Fig. 8c), and

residuals for one of these scans (109) reveals a pattern of signal loss also in the frontal pole (Fig. 6d).

The SW image identifies bilateral regions as non-normal (Fig. 9a). The diagnostic plot (Fig. 9b) and fit and residual plots in the model detail viewer (Fig. 9c) reveal this as a problem of negative outliers. The outliers tend to fall at the end of each epoch and are perhaps related to swallowing.

The artifacts identified in the CW-E and SW images are troubling and we want to remedy these problems by removing corrupted scans. One possibility would be to investigate the scan detail of each outlier spike and to establish whether an artifact is responsible; scans with artifacts would be removed. However, a less labor-intensive solution is to simply remove scans with large numbers of outliers. We remove all scans with more than four times the number of outliers expected under the null hypothesis (1.1% or 530 voxels) and those belonging to the last epoch (due to the problem in the frontal pole, see Fig. 8c). The 34 scans that meet this criterion include almost all of the artifactual scans detected above.

Corrupted scans removed

The 95-scan analysis has much improved diagnostics. The outlier plots are reduced in magnitude and only have one notable spike. Maximum intensity projections of the model summaries before and after problem scans are removed are shown in Fig. 10. The CW and SW images are now mostly uniform, while the DW image and CP image (not shown) exhibit hyperintensities only in vascular and edge voxels. In fact, based on a 0.05-FDR-thresholded images (not shown), the autocorrelation is only significant in gray matter voxels. This suggests a physiological source of autocorrelation that should be addressed in the final modeling of this data set.

The problems in the frontal pole and bilateral frontal regions are much reduced. Inspection of model detail at the few hyperintensities in the SW image (Fig. 9b) identifies about six additional scans with artifacts. As none of these artifacts are as severe as those previously identified, and since any removal of outliers invariably leads to creation of new outliers, we chose not to remove any other scans.

With a largely artifact-free data set, we continue our exploration of the model summaries of nuisance variability and noise. The F image of the drift basis coefficients reveals no unusual patterns (not shown) and mainly identifies slow monotonic drifts at the posterior surface of the brain. The PCT image has a mode of 0.47% (for $\alpha = 0.05$ FDR-corrected, or a mode of 0.30% for $\alpha = 0.05$ uncorrected), meaning that in a typical voxel changes as small as 0.47% can be detected. Of concern is the increased PCT in the bilateral motor and left dorsal lateral frontal areas (Fig. 11a), the very regions of expected activation. Inspection of model detail suggests that, while a few scans are affected by acquisition artifacts, no problems are severe. Instead, we note that voxels in these regions all exhibit experimental variation that rises early relative to the model, by about one-eighth of a cycle (Fig. 11c). This could be due to experimental timing errors or simply poor fit of the canonical hemodynamic response for this subject. Hence the increased variability in these regions is likely due to model misspecification.

We next examine images of model summaries of the signal with percent change and t images. There are focal changes in the bilateral sensory–motor cortices, bilateral auditory cortices, and bilateral cerebellum and diffuse changes in left prefrontal regions (Fig. 12). The signals are of expected change magnitude, the local maxima ranging between 3 and 5.5%. By examining the model detail for each foci (not shown), we confirm that artifactual sources are not responsible for the effects; however, for each voxel examined the one-eighth-cycle phase error is evident (Fig. 11c). The spatial extent of voxels with this phase error is consistent with the overlap between regions of activation and regions of hypervariability in the PCT, suggesting that lack of fit is responsible for the increased residual variability.

Finally, there are broad patterns of positive and negative changes about orbitofrontal regions (Fig. 12a). While this is easily identified as susceptibility-related, it is a demonstration of the merit of examining unthresholded images of estimated signal.

In summary, the application of our diagnostic tools finds violation of the independence assumption, owing to physiology and out-of-phase experimental variation, and violation of homoscedasticity assumption, owing largely to artifacts. We remedy these problems by eliminating global scaling and removing scans with serious artifacts. The resulting reduced data set is satisfactory, except for typical fMRI autocorrelation in gray matter and vascular regions. In regions of activation we find no extensive violations of assumptions, aside from model misspecification owing to a phase error in the predictor. The reduced dataset is now ready for a final model fitting, in particular, with a model that uses a shorter hemodynamic delay (or one that allows for variable delay) and one that accounts for intrinsic temporal autocorrelation.

There are several limitations and qualifications to this demonstration. First, this analysis does not constitute a study of global scaling. Rather we have demonstrated how careful study of the data can lead to selecting an appropriate model. Also, we do not advocate a routine deletion of scans based on outlier counts. The origin of outlier spikes should be explored and understood; we have only removed scans when an obvious acquisition artifact is identified. Further, removing scans is just one possible remediation, and we could have instead Windzorized. Finally, we note that vocalization can create a confounding of signal and artifact (Birn et al., 1998), a situation that is to be avoided and that troubles the interpretation of this data even after thorough diagnosis and exploration.

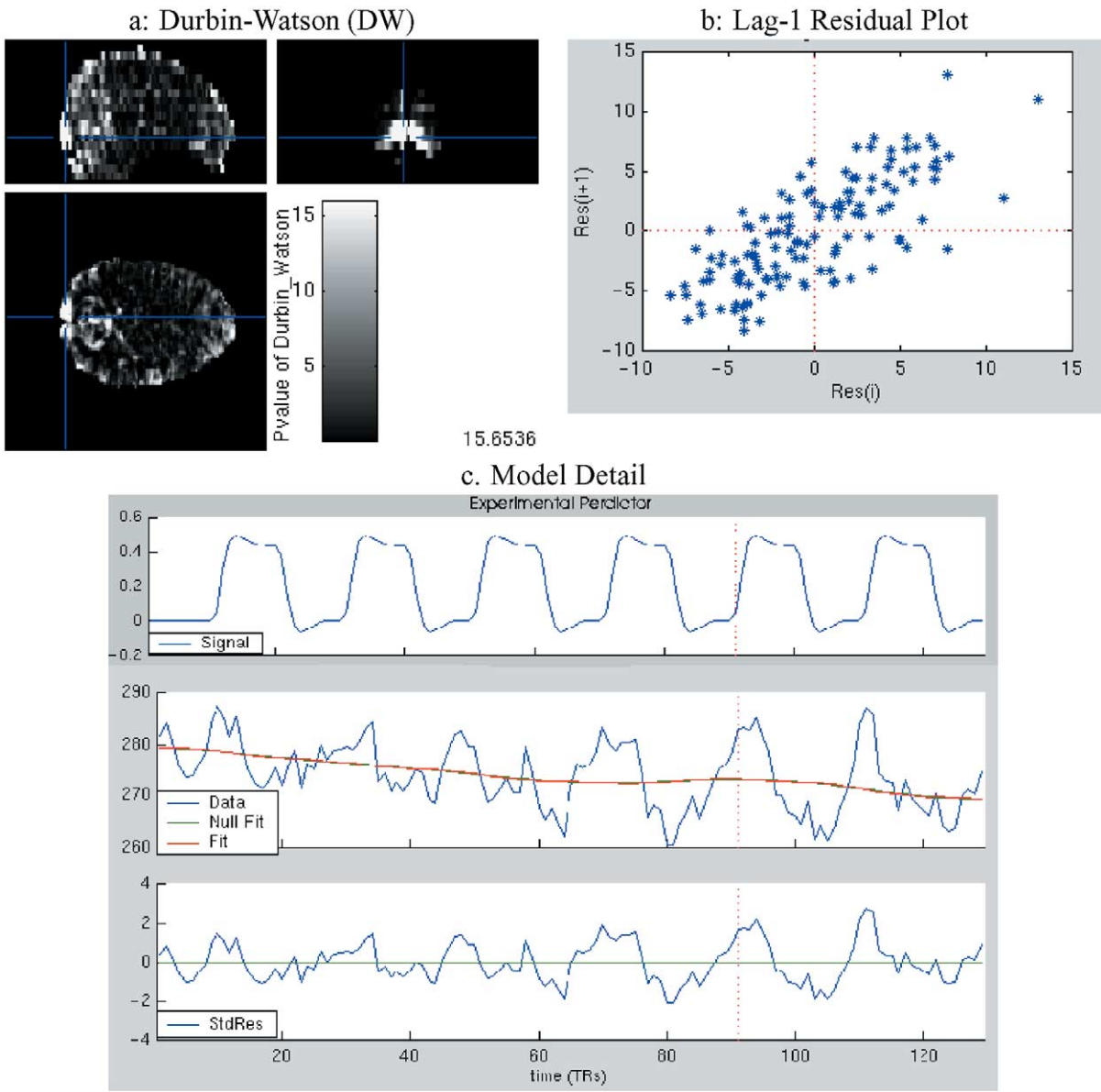


Fig. 7. Out-of-phase variation found with autocorrelation model summaries. Diagnostics for both zero autocorrelation (a, DW) and white noise (CP, not shown) indicate problems in posterior regions. The lag-1 residual plot (b) confirms the autocorrelation found by DW, and model detail (c) identifies the cause of the dependence, a one-quarter-cycle out-of-phase signal.

Discussion

While it is straightforward to apply the general linear model to neuroimaging data, and seemingly easy to make inferences on activations, the validity of statistical inference depends on model assumptions. One can have no confidence about their inferences unless these assumptions have been checked. Further, systematic exploration of the data is essential to understand the form of expected and unexpected variation. However, both diagnosis and exploration are formidable tasks when a single data set consists of 1,000 images and 100,000 voxels.

We have proposed methods for evaluating assumptions on massively univariate linear models and exploring neuroimaging data. The key aspects are model and scan summaries sensitive to anticipated effects and model violations, and interactive visualization tools that allow efficient exploration of the 3-D parameter images and 4-D residuals. We have demonstrated how these tools can be used to rapidly identify rare anomalies in over 10^7 elements of data.

Diagnostic tools have two important usages. First, the diagnostic methods can be used to suggest appropriate remedial action to the analysis of the model. Second, they may result in the recognition of important phenomena that might otherwise be unnoticed. The study of our fMRI data set illustrates both of these roles: On the first point, we found it necessary to eliminate global scaling and a collection of bad scans; on the second, we found one-quarter-

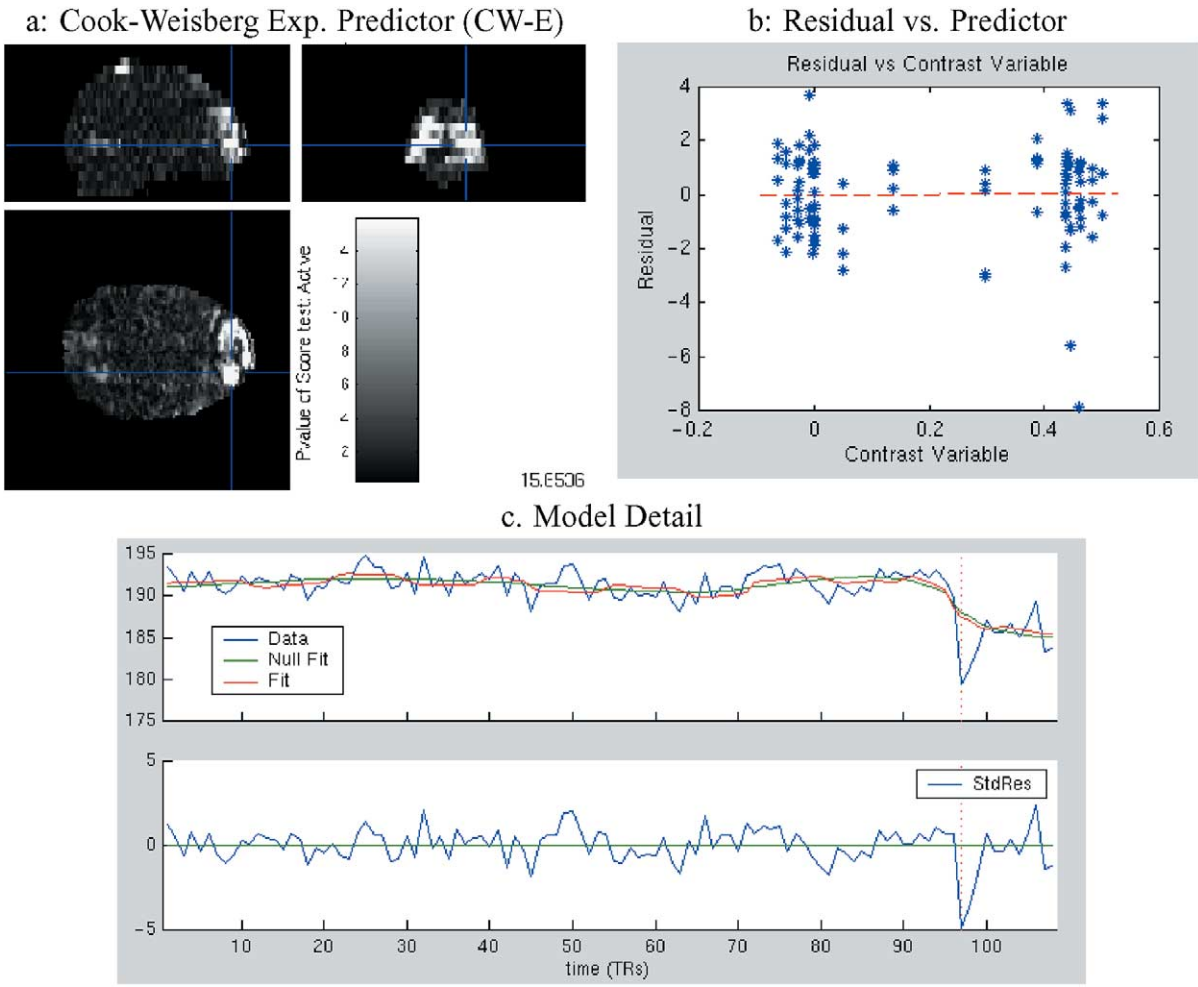


Fig. 8. Transient artifacts found with homoscedasticity model summary. The diagnostic image for homogeneous variance with respect to the experimental predictor (a, CW-E) has a hyperintensity in the frontal region. The residual versus predictor plot (b) shows this is due to a few outliers during an active condition. The model detail (c) identifies the last epoch as the source of the problem. A residual image from this epoch, shown in Fig. 6d, shows signal loss in the same frontal region.

cycle out-of-phase variability distributed throughout the brain and one-eighth-cycle out-of-phase activations in regions of anticipated activation. Thus, we have improved the quality of our model and learned about a shift in this subject's hemodynamic response, neither of which could have been accomplished solely with inspection of thresholded *t* images.

While we only considered a single-subject fMRI example in this work, these tools and statistical summaries are equally relevant PET, VBM, and second-level fMRI. The essential differences are the reduced power of diagnostic statistics, owing to smaller sample size and a lessened concern about autocorrelation.

The principal contributions of this work are both statistical and computational–graphical. We have identified and characterized diagnostic statistics relevant for linear modeling of neuroimaging data, focusing in particular on impact of autocorrelation. Computationally, we have specified and created a system for the efficient exploration of massive data sets. Using linked, dynamic viewers, all the various summary measures can be rapidly compared and understood. Finally, we have given practical recommendations on how to diagnose and explore neuroimaging data sets.

The principal direction for future work is the exploration of temporal dependence, models of spatial dependence, and multivariate exploration of the residuals. A temporal convariance model is needed to decorrelate fMRI data; in future work we will address how one selects such a model for temporal dependence. Gaussian random field methods make assumptions on the spatial dependence, and we are developing methods to assess these assumptions, stationarity in particular. Finally, our approach using spatial and temporal residual summaries may miss extended spatiotemporal patterns. Methods such as PCA or ICA applied to the studentized or BLUS residuals may provide valuable tools for this purpose.

Conclusion

In this work, we have developed a general framework for the diagnosis of linear models fit to fMRI data. Using model

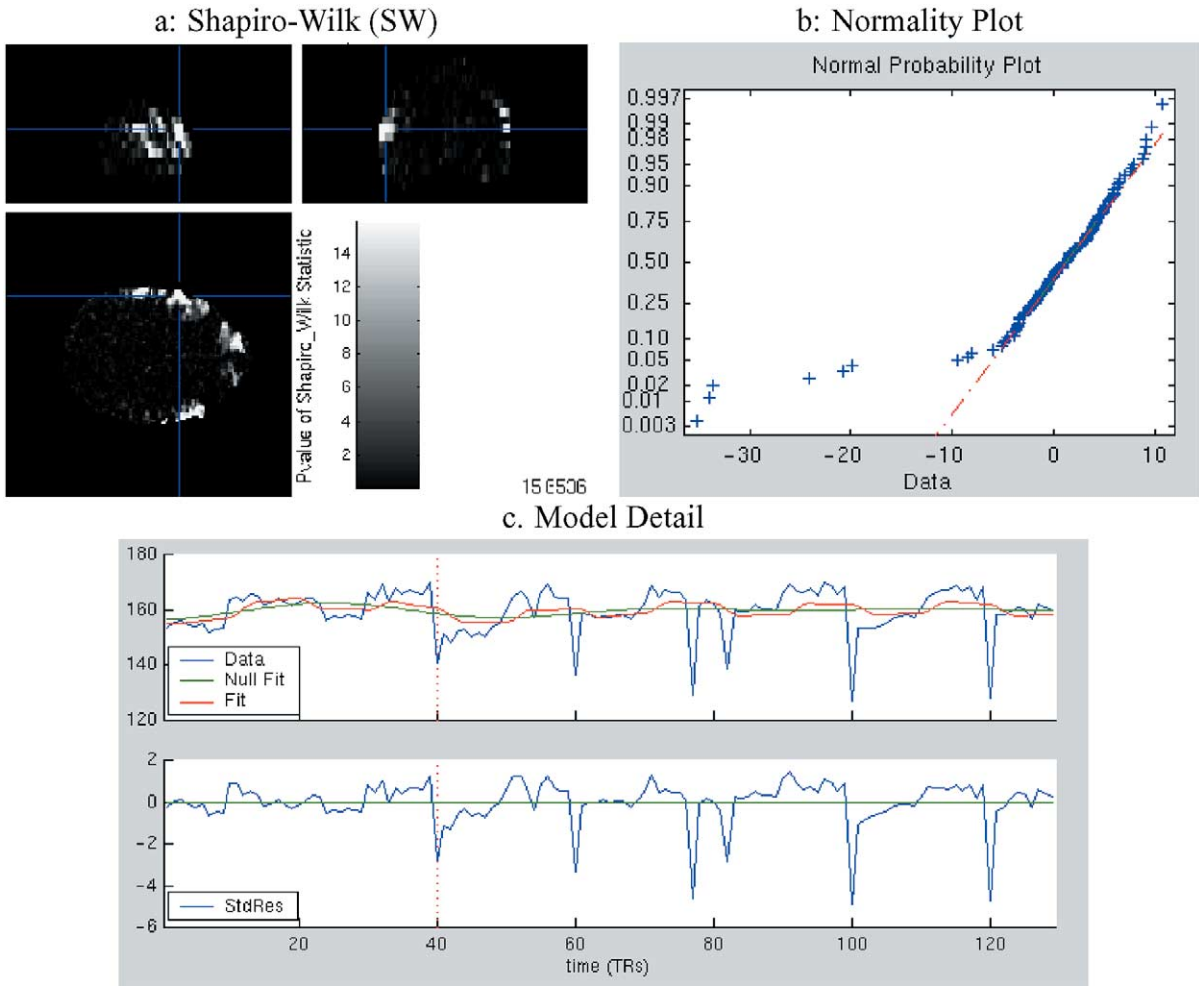


Fig. 9. Periodic artifacts found with the normality model summary. The normality diagnostic (a, SW) is mostly nominal, except in in bilateral frontal regions. The *Q–Q* plot of one of these nonnormal voxels (b) shows that many negative outliers are responsible, and the model detail (c) shows these outliers are regularly spaced with roughly the experimental periodicity.

and scan summaries and dynamic linked viewers, we have shown how to swiftly localize rare anomalies and artifacts in large 4-D data sets.

Acknowledgments

The authors thank Brian Ripley and Philip Howrey for valuable conversations about this work. We also thank Boklye Kim for using the example data and John Ashburner for creating the orthogonal viewer tool used extensively in our software.

Appendix A

Percent change and interpretable linear models

This appendix shows how to create interpretable linear models, from which it is easy to create percent change contrast images. Neuroimaging linear models are traditionally not constructed to yield maximally interpretable parameters. For example, the standard two-sample t test, implemented with a zero–one design matrix and $c = [-1\ 1]$, yields a $c\beta$ with half-magnitude units. Here we show how to scale X and c such that $c\beta$ has the same units as the data and that $c\beta/\mu \times 100$ has percent change units; if the data are intensity normalized, $c\beta$ can approximate percent change.

The two key requirements are that (1) each column, X_j corresponding to an experimental effect represent a unit change in the data and (2) contrasts have absolute sum of unity. Condition 1 ensures that a unit change in β_j corresponds to a unit change in $\hat{Y}$, and condition 2 ensures that these units are preserved. While condition 2 is easy to enforce in software, condition 1 requires care while constructing predictors.

Categorical predictors will satisfy condition 1 when the "off" to "on" difference is unity, such as coding dummy variables with 0 and 1 (but not -1 and 1). In block-design fMRI, this requires the that baseline-to-activation be scaled to unity, and in event-related fMRI with isolated or equally spaced events, the baseline to peak height can be scaled to

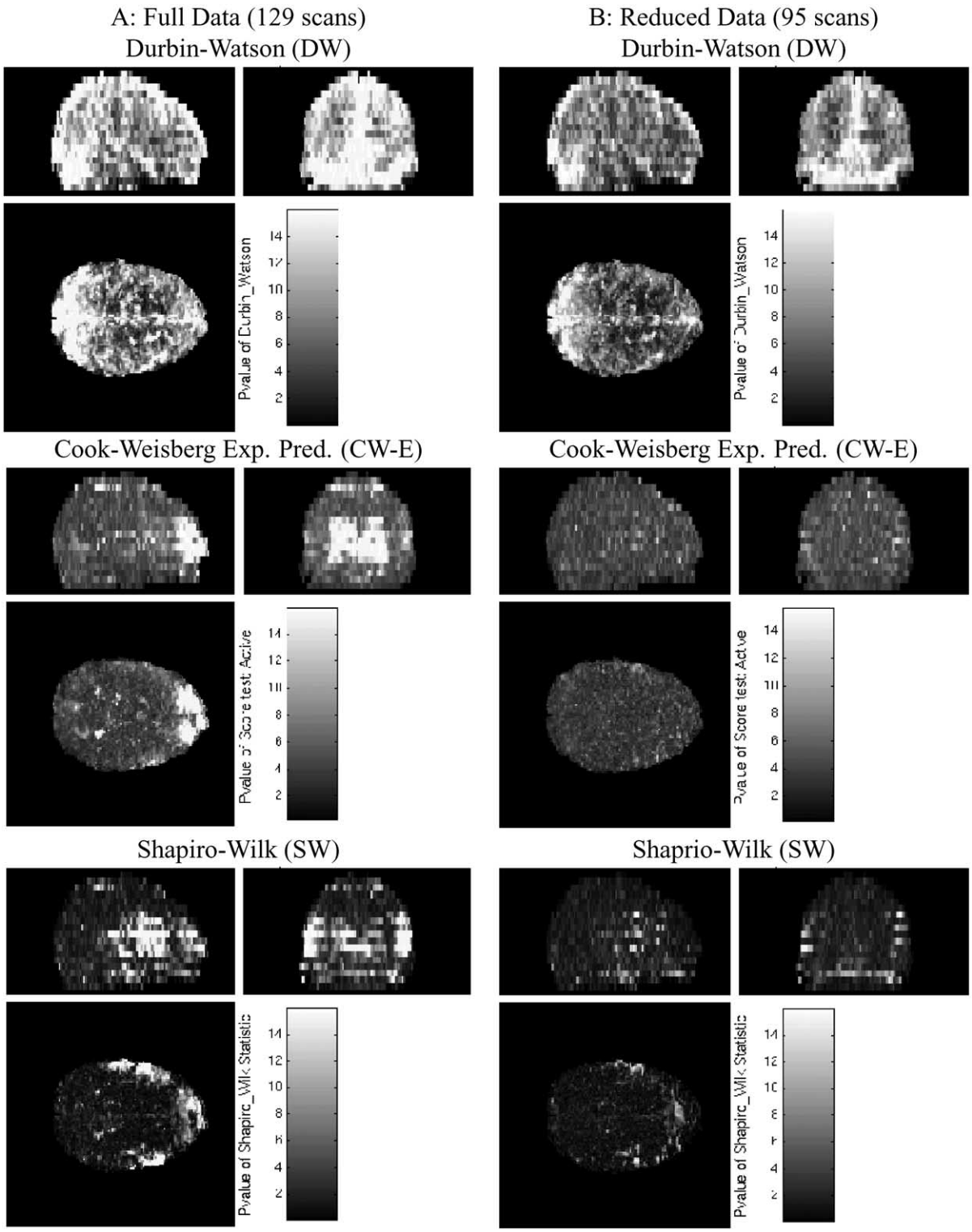


Fig. 10. Comparisons of maximum intensity projections of DW, CW-E, and SW in full data analysis (A) and in reduced data analysis (B). The images from reduced data are much more uniform than those from the full data.

one. Jittered event-related fMRI predictors are more problematic since there is no unique baseline-to-peak range. As a practical solution we recommend scaling to unity the range or trimmed range (e.g., 5th to 95th percentile) of predictor values. See the related technical report (http://www.sph.umich.edu/~nichols/PCT) for more detail.

For $c\beta$ to approximate $c\beta/\mu \times 100$, μ must be approximately 100. When it is reasonable to approximate the

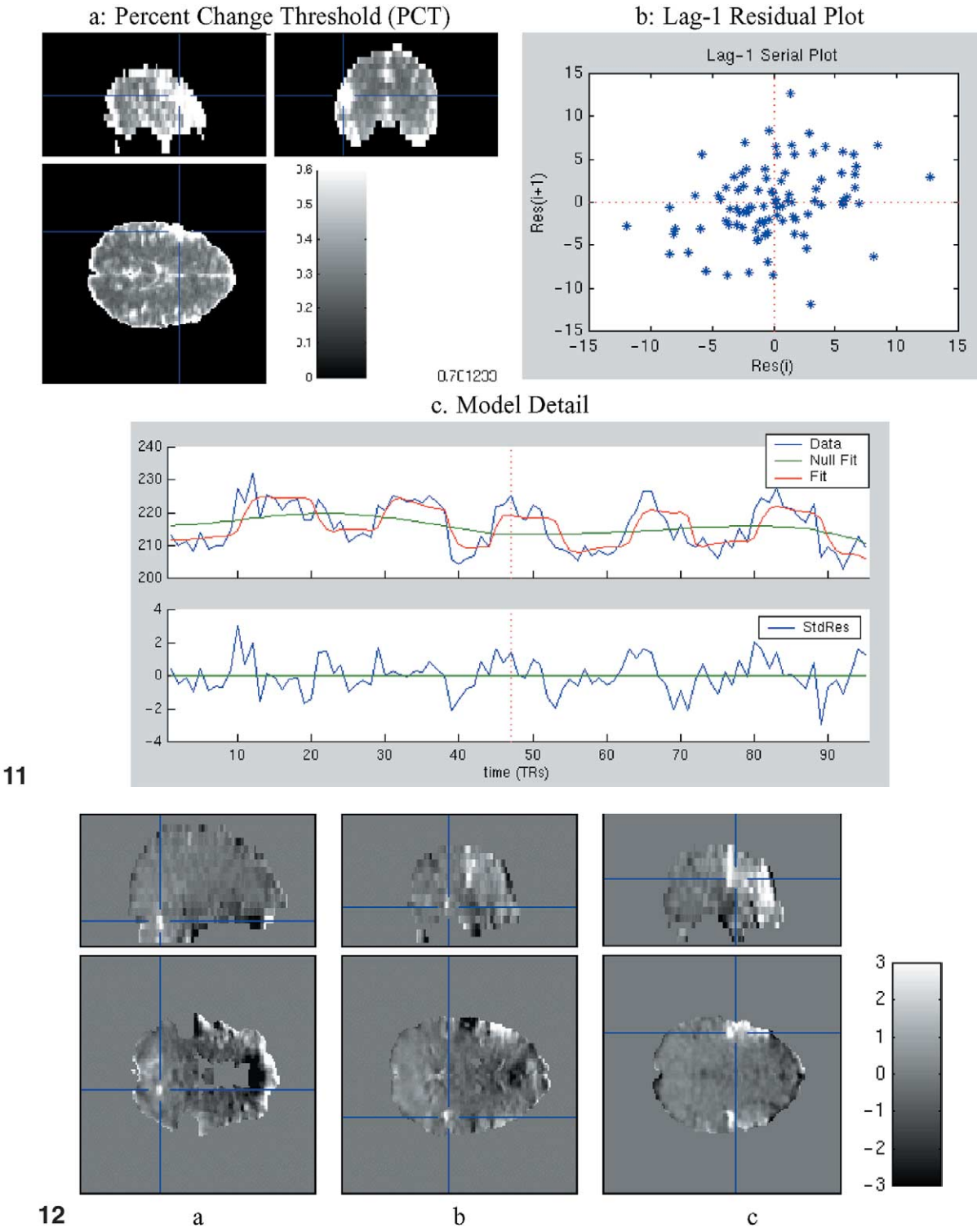


Fig. 11. Results of exploring the PCT image. The standard deviation, as visualized with the minimum significant percent change (a, PCT), is much higher in the bilateral motor and left dorsal lateral frontal areas. The model detail (c) suggests that the source of this inflated variance due to experimental signal which is one-eighth of a cycle out of phase. This phase error induced some autocorrelation in the residuals (b) but not enough to stand out in either the DW or CP image (not shown).

Fig. 12. Results of exploring percent change images of activation. (a) Bilateral, though mostly right cerebellum, and artifactual orbitofrontal changes; (b) bilateral, though mostly right auditory cortex; (c) bilateral motor cortices and left dorsolateral prefrontal cortices.

baseline (grand mean) image with a constant (e.g., in fMRI), all data are scaled such that the baseline images' "global" intensity is 100. But we find that the usual global estimate, the arithmetic mean of all intracerebral voxels, is unsatisfactory. The mean is sensitive to hyperintensity outliers and the segmentation of brain from nonbrain. If a simple intensity threshold is used to segment brain and the threshold is set too low, the global average can be far below typical brain intensities. We propose that the mode is a more accurate global measure than the mean, as the mode is very robust with respect to brain threshold. In Appendix B we give a method to estimate the intracerebral mode.

To summarize, we assert that models should be constructed to be as interpretable as possible. The rewards of this endeavor are that, if the voxel grand means are around 100, the predictors have unit scaling, and the contrast vector has an absolute sum of unity, then the contrast image will have approximate interpretation of percent change. Ratioing such a contrast image with a grand mean image will produce percent change exactly.

Appendix B

Estimation of the intracerebral modal intensity

This appendix describes the estimation of the mode of intracerebral voxel intensities. This method consists of estimating a brain–nonbrain threshold and then estimating the mode of the distribution of brain voxel intensities. While there is an extensive literature on mode estimation using kernel density estimation (see, e.g., Scott, 1992), we simply use a histogram estimate with appropriate bin widths for a consistent estimator[1] of the mode. Our approach uses no topological operations on the image data and is easily coded and quickly computed.

Estimating brain–nonbrain threshold with the antimode

We estimate a brain–nonbrain threshold using the distribution of all voxel intensities. Our threshold is the location of minimum density between the background and gray matter modes; call this the antimode. Let $f(x)$ be the distribution of intensities in the brain image. Hartigan (1977) shows that a consistent estimator of the antimode is the location of the maximally separated order statistic between modes. Since we do not know the location of modes, we instead just search over the whole density excluding the tails; the tails must be excluded as the global minimum of $f(x)$ will be found there. A crude overestimate of the tails is sufficient, since the antimode estimate will only be perturbed if we include tails with less density than the antimode or exclude the actual location of the antimode. We have found the 10th and 90th percentiles to work on all images we have considered. Our threshold estimate is thus

$$T = \left\{ \frac{1}{2} \left(x_{(k+1)} + x_{(k)} \right) : k = \underset{0.1n<i<0.9n}{\operatorname{argmax}} \left(x_{(i+1)} - x_{(i)} \right) \right\}, \tag{1}$$

where n is the number of voxels in the image and $x_{(k)}$ is the k-th order statistic. If k is not unique, we take an average of the locations.

While this works well on continuous-valued image (e.g., a floating point mean image), it does not work with a discrete-valued image (e.g., an integer T2* image). The problem is that the distance between order statistics will be 0 or 1 except at the very extreme tails. Hence if the image is discrete we then revert to a simpler histogram method. We construct a histogram based on all nontail data (10th to 90th percentile) and use the location of the minimum bin as the antimode estimate. To construct the histogram we use the bin width rule for the mode (described below). We have found that this serves as a robust estimate of a brain–nonbrain threshold.

Whether through the inter-order-statistic distance or the

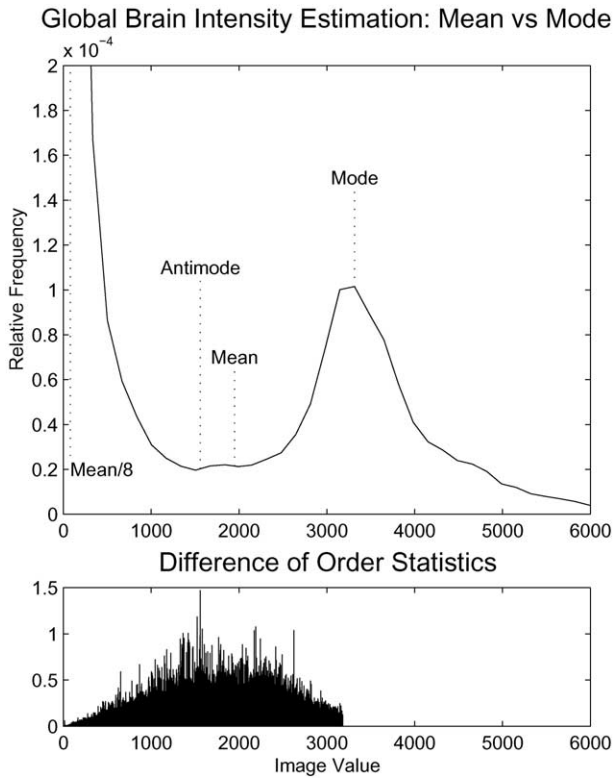


Fig. 13. Our mode estimate compared to SPM99 and SPM2's mean estimate. Top plot shows the distribution of intensities of the baseline image of the data set in this paper. The mean is determined by SPM, where all voxels greater than one-eighth of mean image intensity are considered. The bottom figure shows how the antimode is determined: the differences of order statistics are plotted versus the order statistics; the location of the greatest gap between order statistics is the estimate of the antimode.

[1] With more and more data, a consistent estimator converges to the true value in probability.

histogram approach, this antimode estimate is only used to eliminate the lower mode of background voxels and hence does not need to be highly accurate.

Estimating global brain intensity with the mode

We estimate the mode of the brain voxel intensities using a type of histogram estimate for simplicity and computational efficiency. The optimal (and consistent) histogram bin width for estimating the mode is order $n^{1/5}$; we use bin widths equal to $1.595 \times \mathrm{IQR} n^{-1/5}$ (Scott, 1992, p100), where IQR is the interquartile range of the brain voxels. While this rule is based on independent normal data, it has performed quite well on many PET and fMRI data sets. The mode estimate is the location of the maximal histogram bin.

While we do not argue that this mode estimate is optimal in the sense of mean squared error, we have found it to be robust and sufficiently accurate for the purposes of this work. In particular, we have found it more accurate than the simple estimator used in SPM (see Fig. 13). Matlab code for this method is available at http://www.sph.umich.edu/~nichols/PCT.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The inferential impact of global signal covariates in functional neuroimaging analyses. NeuroImage 8, 302–306.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, third ed. Wiley, New York.

Birn, R.M., Bandettini, P.A., Cox, R.W., Jesmanowicz, A., Shaker, R., 1998. Magnetic field changes in the human brain due to swallowing or speaking. Magn. Reson. Med. 40, 55–60.

Burock, M.A., Dale, A.M., 2000. Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient and unbiased approach. Hum. Brain Map. 11, 249–260.

Casella, G., 2001. Statistical Inference, second ed. Wadsworth, Belmont, CA.

Cook, R.D., Weisberg, S., 1983. Diagnostics for heteroscedasticity in regression. Biometrika 70, 1–10.

Diggle, P.J., 1990. Time Series: A Biostatistical Introduction. Oxford University Press, Oxford.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley, New York.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-B., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Map. 2, 189–210.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.A., Hansen, L.K., 1999. On clustering fMRI time series. NeuroImage 9, 298–310.

Hartigan, J.A. 1977. Distribution problems in clustering, in: Classification and Clustering, Academic Press, New York/London, pp. 45–72.

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), 1983. Understanding Robust and Exploratory Data Analysis, Wiley, New York.

Holmes, A.P. 1994. Statistical Issues in Functional Brain Mapping, Ph.D. thesis. University of Glasgow. Available from: http://www.fil.ion.ucl.ac.uk/spm/papers/APH_thesis.

Holmes, A.P., Friston, K.J., 1999. Generalisability, random effects and population inference. NeuroImage 7 (4 (2/3)), S754.

Kherif, F., Poline, J.-B., Flandin, G., Benali, H., Simon, O., Dehaene, S., Worsley, K.J., 2002. Multivariate model specification for fMRI data. NeuroImage 16 (4), 1068–1083.

Luo, W.-L., Nichols, T.E. 2002. Diagnosis and Exploration of Massively Univariate fMRI Models, technical report. Department of Biostatistics, University of Michigan.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T-P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Map. 6, 160–188.

Moonen, C.T.W., Bandettini, P.A. (eds.), 2000. Functional MRI. Springer Verlag, New York.

Moser, E., Baumgartner, R., Barth, M., Windischberger, C., 1999. Explorative signal processing in functional MR imaging. Int. J. Imag. Syst. Technol. 10, 166–176.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models, fourth ed. Irwin, Chicago, IL.

Nichols, T.E., Luo, W.L., 2001. Data exploration through model diagnosis. NeuroImage 13 (6(2/2)), S208.

Petersson, K.M., Nichols, T.E., Poline, J.-B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. Phil. Trans. R. Soc. Lond. B 354, 1261–1281.

Razavi, M., Grabowski, T.J., Mehta, S., Bolinger, L., 2001. The source of residual temporal autocorrelation in fMRI time series. NeuroImage 13 (6(2/2)), S228.

Royston, J.P., 1982. An extension of Shapiro and Wilk's *W* test for normality to large samples. Appl. Stat. 31, 115–124.

Ryan, T.P., 1997. Modern Regression Methods. Wiley-Interscience, New York.

Schlittgen, R., 1989. Tests for white noise in the frequency domain. Comput. Stat. 4, 281–288.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York.

Shapiro, S.S., Wilk, M.B., Chen, H.J., 1968. A comparative study of various tests for normality. J. Am. Stat. Assoc. 63, 1343–1372.

Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. 19 (2), 279–281.

Stephens, M.A., 1970. Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. J. R. Stat. Soc. Ser. B Methodol. 32, 115–122.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. 69, 730–737.

Theil, H., 1971. Principles of Econometrics. J Wiley, New York.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Wilcox, R. 1998. in: Trimming and winsorization, Encyclopedia of Biostatistics, Vol. 6. Wiley, New York, pp. 4588–4590.

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S., 2001. Temporal autocorrelation in univariate linear modeling of fMRI data. NeuroImage 14, 1370–1386.