Diagnosis & Exploration of Massively Univariate fMRI Models

Wen-Lin Luo, Thomas E. Nichols

Dept of Biostatistics, University of Michigan, U.S.A.

May 2, 2002

Running title: Diagnosis & Exploration of fMRI Models

Address for correspondence:
Thomas E. Nichols
Department of Biostatistics
University of Michigan
1420 Washington Height,
Ann Arbor, MI48109
Phone: +1-734-936-1002
Fax: +1-734-763-2215
email: nichols@umich.edu

Abstract

The goal of this work is to establish the validity of inferences in fMRI modeling through diagnosis of linear model assumptions, and to characterize fMRI signal and artifacts through exploratory data analysis. While model diagnosis and exploration is an integral part of any statistical modeling endeavor, these aspects have been mostly neglected in functional neuroimaging. We present methods that make diagnosis and exploration of fMRI data feasible. Exploiting the spatiotemporal structure of the data, we use spatial and temporal summaries that characterize model fit and residuals. Our methods involve both statistical and graphical tools. Our statistical tools are diagnostic summary statistics with tractable null distributions, which we have validated in the context of a typical autocorrelated fMRI model. Our dynamic graphical tools allow the exploration of multiple spatial and temporal summaries, with the ability to quickly jump to spatiotemporal detail. We apply our method to a fMRI dataset, 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.

1 Introduction & Motivation

A principal use of fMRI is to localize brain regions exhibiting experimental variation. A 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 a massively univariate one (Holmes, 1994), where time series data for each voxel are independently fitted with the same model (Friston *et al.*, 1995). Images of test statistics are used to make inference on the presence of a effect at each voxel.

The primary purpose of this work is to establish the validity of inferences in fMRI 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 & 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 signal, 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 fMRI the fit and residuals are individually more amenable to thorough exploration than the full data. The fit is parameterized by the user and is readily interpretable; 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 nonhomogeneity; for example, something as simple as outlier counts per scan can quickly identify interesting scans. In fact, diagnosis and EDA are enmeshed: Diagnosis takes the form of exploration of 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 (a) image and time series summaries which characterize fit and residuals and (b) dynamic visualization tools to explore these summaries and to efficiently identify spatiotemporal regions of interest.

We use the term “spatial summaries” to refer to images that assess fit or residuals over time, and “temporal summaries” to refer to time series that assess fit or residuals over space. For the spatial summaries, we use both images of linear model parameters and images of diagnostic statistics. For example, we assess linear model assumptions like normality, homoscedasticity[1], and independence of errors with scalar diagnostic statistics; to view these diverse measures on a common scale, we create images of these statistics and images of p-values. For the temporal summaries, we use time series describing model fit and residuals, as well as time series of preprocessing parameters. For example, global intensity and outlier count per image both can capture transient acquisition problems, and head motion estimates are useful for attributing suspect scans to motion artifacts.

The dynamic visualization tools are used for simultaneously exploring multiple summary images and time series, and for quickly jumping from summary margins to the raw or residual spatiotemporal data. We use linked orthogonal viewers to explore the spatial summaries and parallel time series viewers with linked temporal cursors to facilitate study of temporal summaries. From a spatial summary, a click on a location of interest displays the temporal detail for that voxel, including time series plots of raw data, fitted model, and residuals, and traditional diagnostic plots. From a temporal summary, a click on a time point displays the spatial detail for that scan, images of studentized residual images before, during and after the time point of interest. These tools have been implemented as Statistical Parametric Mapping Diagnosis (SPMd, see Fig. 1), a toolbox for SPM99, available at `http://www.sph.umich.edu/∼nichols/SPMd`.

In this paper we assume temporal independence of the errors. While autocorrelation is a characteristic of fMRI data, signal and artifacts must be characterized before autocorrelation modeling is pursued— models that account for dependency still assume homogeneous variance and normality. Further, the autocorrelation diagnostics we describe below will capture the form and spatial heterogeneity of autocorrelation, allowing exploration of temporal dependency before it is modeled. Hence, we view our independence assumption as a working model for the temporal dependency of the noise.

There has been little previous work in fMRI model diagnosis (Razavi *et al.*, 2001; Nichols & Luo, 2001). In EDA there are many data-driven tools which have found use in fMRI, including clustering (Goutte *et al.*, 1999; Moser *et al.*, 1999), Independent Components Analysis (McKeown *et al.*, 1998) and Principal Components Analysis. 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

[1] Homogeneous variance.

statistical summaries with p-values to make inferences on the magnitude of discovered patterns (relative to a putative model).

[Figure 1 about here.]

In the next section we introduce these diagnostic summaries in spatial and temporal margins, methods for the graphical exploration, and specific strategies for model diagnosis. In the subsequent section we report on simulation studies that investigate the performance of the summaries with respect to different correlation conditions. Lastly we demonstrate our tools on a fMRI dataset. An extensive appendix contains the cumbersome definitions and null distributions of the diagnostic statistics.

2 Methods

We fit fMRI data with linear regression model at each voxel. For a given voxel this can be written as:

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

where Y is a N-vector of responses, X is a $N \times p$ matrix of p predictors, β is a p-vector of unknown, fixed parameters, and ε is a N-vector of unknown, random errors. For general linear regression model, it is assumed that ε is a vector of independent random variables with expectation $\mathsf{E}(\varepsilon) = 0$ and variance-covariance matrix $\sigma^2 I$, i.e., $N(0, \sigma^2 I)$.

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 of interest, or contrast, $c\beta$. The fitted values are $\hat{Y}$ and the residual e are:

$$e = Y - \hat{Y} \; = \; Y - X\hat{\beta}.$$

Note that even when the errors are assumed to be independent the residuals are dependent, as per $\mathsf{Cov}(e) = (I - H)\sigma^2 \neq \sigma^2 I$, where $H = X(X^\top X)^{-1} X^\top$. This has important implications for residual diagnosis which we revisit below.

Once the residuals are created, we can compute the diagnostic statistics. The two main components of our work are the spatial and temporal summary statistics of the spatiotemporal data, and the interactive visualization tools to explore those summaries. We first describe each of the summaries.

2.1 Spatial Summaries

Spatial summaries of interest include images of model parameters, to assess fit, and summaries of residuals, to assess lack-of-fit and model assumptions. We use many different summaries of the residuals, each sensitive to a potential artifact or assumption violation . The spatial statistics that we use for exploration and diagnosis are summarized in Table 1. In this section, we describe only the motivation using each of these statistics, while detailed definition of the statistics, their distribution under a null hypothesis of model fit, and how to assess them are in Appendix C. For each diagnostic statistic we create images of $-\log_{10}$ p-value; this provides a consistent metric for visualizing the diagnostic measures.

[Table 1 about here.]

2.1.1 Contrasts and Statistic Images

We use percent change contrast images and t statistic images to characterize expected signal in the data. The β coefficients are also useful, but may represent inestimable effects (Graybill, 1976) and hence we prefer contrast images. While attention is usually focused on signal-to-noise t images, the contrast image is useful for examining the profile of the signal alone; percent change is also a standard measure for the BOLD effect. However, while t images are scale-invariant and unitless, a contrast has units determined by the predictors and the contrast vector. See Appendix A for details on constructing linear models with interpretable units; in particular, so that $c\hat{\beta}/\mu \times 100\%$ is percent change, where μ is the baseline.

In addition to contrast and t images, we use F images to summarize non-scalar effects. An example of such an experimental effect is a hemodynamic response parameterized with a finite impulse response model; an example of such a nuisance effect is a set of basis drift functions. Lastly, we explicitly create a grand mean image to use as an anatomical reference. While other high resolution images may be available, there are frequently misalignment problems due to head motion or susceptibility artifacts.

2.1.2 Standard Deviation via Percent Change Threshold

The standard deviation image is the essential residual summary measure, as it is simply the root-mean-squared of the residuals. While standard deviation images are thus ideal for finding anomalous spatial structure, a particular voxel's standard deviation value is not so interpretable. Standard deviation does have units of the data (in distinction to the variance), and it lacks concrete units that, say, a contrast image has (response magnitude, all other effects held constant).

We propose characterizing variability with the half width of a $(1-\alpha)\%$ confidence interval for an effect of interest. This quantity is most easily seen as the minimum change required to obtain a significant activation at level α (either corrected or uncorrected). If an effect is expressed in units of percent change, we call this quantity the *Percent Change Threshold* (PCT); the interpretation is then the minimum percent change needed to reach significance. This measure is readily interpretable and, further, is easy to compute as it is just a scalar multiple of the standard deviation or coefficient of variation image.

Another advantage of the PCT is that it intuitively expresses the impact of standard deviation on power. For example, 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 1.5T, Moonen & Bandettini, 2000). For visualization, we find it particularly useful to window the image such that twice the modal PCT value is maximum intensity (see Appendix B); doing this gives middle gray (or middle intensity) the interpretation of typical sensitivity, and white (or maximal intensity) as less than one half typical sensitivity. See Appendix C.1 for details on PCT.

2.1.3 Independence Tests

Temporally autocorrelated noise is a common feature of fMRI data. We use two statistics to assess the independence assumption. The Durbin-Watson statistic and cumulative periodogram test are sensitive to detecting first order autoregressive noise and general non-white noise, respectively. The first test, the Durbin-Watson statistic, is a well-known way of checking for serial correlation pattern in an equally spaced sequence of residuals. The exact Durbin-Watson test is uniformly most powerful if the errors approximately follow a first-order autoregressive, AR(1), process (Durbin & Watson, 1971).

On the other hand, because fMRI noise is quite complicated with high-temporal frequency artifacts arising due to noise or physiological effects, a first order autoregressive model may be too restrictive. We prefer a test with less constraints on the correlation structure and the distribution of the data. In the spectral analysis, white noise corresponds to flat spectrum for all frequencies, and a periodogram is an estimate of the spectrum. Instead of testing the flatness of the periodogram, we test the linearity of the cumulative periodogram (Diggle, 1990). The cumulative periodogram (CP) is used because it is an ordered random sample from the uniform distribution on $(0,1)$, under the null hypothesis of white noise.

We find that direct application of the CP test is not satisfactory. In simulations with a typical fMRI model (block design experimental predictors and low frequency drift basis) and white noise, we find that the CP test rejects the white noise hypothesis in excess of the nominal level. The problem is that the residuals, as mentioned above, are not independent. The CP is appropriately sensitive to the absence of low frequency variation in the residuals, variation removed by the drift basis. We have found that the CP test based on best linear unbiased scalar covariance (BLUS) residuals appropriately controls the false positive rate, while still being sensitive to autocorrelation of the noise. See Appendix C.2 for more details on Durbin-Watson and BLUS-based cumulative periodogram test, and Section 3 for comparison of CP with BLUS and ordinary residuals.

2.1.4 Homoscedasticity Test

One assumption of linear regression is constant variance, $\text{Var}(\varepsilon_i) = \sigma^2,\ i = 1, \cdots, N$. In many cases the variances may depend on the response, on one or more of the predictors, or possibly on other factors, such as time or physical ordering. If the constant error variance assumption is not valid, then ordinary least squares should not be used. Cook and Weisberg (1983) proposed a diagnostic statistic for examining the assumption of constant variance based on the score statistic. The score test is an asymptotic statistical test that depends only on the derivative of the likelihood at the null-hypothesized parameter value (Casella, 2001).

In brief, the test models heterogeneous variance with a log-linear model. A matrix of predictors Z and parameter vector λ define a model such that $\lambda = 0$ corresponds to homoscedasticity ($\text{Var}(e_i) = \sigma^2 \exp(z_i \lambda)$). The score statistic for the hypothesis $\lambda = 0$ then furnishes a suitable diagnostic test. Remarkably, this score test takes the form of the regression sum of squares from fitting the squared residuals with Z (Cook & Weisberg, 1983). See Appendix C.3 for details on Cook-Weisberg score test.

2.1.5 Normality Test

The usual distributional assumption in regression analysis is that the errors are normally distributed. This assumption is needed to justify the hypothesis testing procedure. One technique for studying nonnormality of residuals is the normal probability plot, which is a plot of the standardized residuals against their expected values computed under the assumption of normality. If the residuals are normal, then the plot will be a straight line.

We use the Shapiro-Wilk (Shapiro & Wilk, 1965) statistic to test the null hypothesis of normally distributed residuals. There are several tests of normality, but this statistic has been shown to be superior to other normality tests, the Kolmogorov-Smirnov test in particular (Stephens, 1974). The Shapiro-Wilk statistic is essentially the square of the correlation between the standardized residuals and their expected value. See Appendix C.4 for more details on the Shapiro-Wilk test.

2.1.6 Spatial Outlier Count

Outliers are extreme observations which do not belong to the same distribution as the rest of the data. Outliers can cause trouble with any model, biasing parameters and variance estimates. With least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the squared deviations is being minimized. This could cause a misleading fit if indeed the outlying observation results from a mistake or other extraneous cause. In the context of fMRI, if one has, by chance, a few outliers in an active condition and none in a rest condition, a false positive may result. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor variable omitted from the model.

We define an outlier as an observation that has an absolute studentized residual greater than 3. We record the number of outliers for each temporal model. Voxels with large numbers of outliers may indicate artifacts or unmodeled experimental variation. See Appendix C.5 for more details on spatial outlier statistic.

2.2 Temporal Summaries

Our temporal summaries use more ad hoc measures, since we do not have an explicit spatial model to evaluate. Each element of a temporal summary assesses a single image. It is a macroscopic view at each time point used to identify any motion, physiological, or scanner artifacts or possible confounding variables. The temporal summaries that we use here include the experimental predictors, temporal outlier count, global signal, and some image preprocessing parameters such as registration shift and rotation movement parameters.

2.2.1 Experimental Predictors

In the exploration of fit and residuals, the experimental design is of obvious importance. Any pattern evident in a spatial plot should be referenced with predictors from the design matrix to see if it has an experimental cause. Hence we include experimental predictors as temporal summaries, though they are quite data-independent.

An example where this could be useful is if a voxel exhibits temporally autocorrelated noise. In this case the residual time series plot should be referenced to the experimental predictor to see if experimental, but out-of-phase variation is responsible.

2.2.2 Global Signal

Because neuroimaging experiments often test hypotheses regarding local changes in neuronal activity, variations in signal that are common to the entire brain volume, global signal, have been considered as nuisance effects to be eliminated.

Aguirre *et al.* (1998) suggested that, to allow for the proper interpretation of neuroimaging results, the degree of correlation of the global signal with the experimental paradigm should be reported in any study to understand the role of global signal. In particular, if the global signal can be explained by experimental predictors, the global is a confound; otherwise it is simply a nuisance variable. To facilitate this, we regress the global on the design matrix X, display the fitted line, and report the significance of this regression.

Note, while we prefer the mode instead of the mean to measure typical intracerebral intensities, we use the mean for creating time series of the global signal. For diagnosis/exploration purposes, the mean's sensitivity to tail behavior is advantageous.

2.2.3 Temporal Outlier Count

We use outlier count per image at each time point as a measure sensitive to shot noise or transient acquisition problems. The definition of outlier is the same as that in the Spatial Outlier Count (Appendix C.5), but we instead sum outliers over space. This outlier count localizes problems in time, and leads the analyst to scans with possible artifacts.

We cannot assign p-values to temporal outlier counts without specifying a spatial covariance structure. We do, however, note the expected number of outliers, q times the number of voxels in the image, where q is the probability of outlier (See Appendix C.5).

2.2.4 Image Preprocessing Parameters

Before linear models are fitted there are many preprocessing methods that applied to the data image-by-image . Each of these preprocessing steps can fail, or, if successful, may capture aspects of artifacts or anomalies. Hence time series of preprocessing parameters are valuable temporal summaries for diagnosing fMRI linear models.

The most important parameter time series are movement parameters. Transient movements can induce changes in B_0 (Birn *et al.*, 1998) that are uncorrectable by standard registration methods. Hence when localizing artifacts in time we view time series of shifts and rotations, focusing on times when sudden shifts or rotations occur.

There are many other preprocessing parameters not considered in this work which could be useful: The image registration goodness-of-fit metric can indicate failures of the rigid body registration model; EPI deghosting parameters estimated scan-by-scan and multishot phase corrections can indicate acquisition problems; and respiratory and cardiac time series can also inform possible physiological effects.

2.3 Visualization

A key aspect to our method is the use of dynamic graphical tools to find interesting spatiotemporal regions in the fit and residuals. In lieu of examining every voxel of every β and residual image, we use spatial and temporal summaries to identify exceptional scans and voxels. Investigating individual scans and voxels will then lead to further localization of the feature of interest. We use four viewers (Fig. 1) to efficiently explore the summaries and, as guided by the summaries, the fit and residuals.

Temporal Summary Viewer Temporal summaries are presented using parallel time series plots (Fig. 1-a). A temporal cursor is present on all temporal plots (temporal summaries and temporal detail), and clicking in any temporal plot moves the cursor to this point. This facilitates comparing the impact of transient effects, for example, to check if a sudden change in global intensity is related to subject movement or any other temporal measure. Right-clicking on a temporal summary brings up the spatial detail (see below).

Spatial Summary Viewer Spatial summaries are displayed with linked orthogonal slice viewers (Fig. 1-b). Additionally, each orthogonal viewer prints the intensity of the selected voxel. While orthogonal viewers do not show all data at once (as with arrays of 2D axial slices), they are ideal for close inspection of 3D detail. As usual, clicking on any point makes the 3 orthogonal views display slices that intersect at that point. All spatial views are updated simultaneously; in this way, multiple summaries can quickly be compared. For example, after clicking on an area of activation in the t-image, a quick glance of the other summaries will reveal if there are problems with autocorrelation, heterogeneous variance or nonnormality. Each orthogonal viewer can optionally be replaced with a maximum intensity projection (MIP) image; this is useful when hunting for structure in a very sparse volume. We use a dynamic MIP of a rotating volume to visualize complicated spatial structures. Right-clicking on a spatial summary brings up temporal detail for the selected voxel (see below).

Temporal Detail Viewer For a selected voxel, the temporal detail shows time series and residual plots (Fig. 1-c). One time series plot shows raw data with fitted values, another shows the residuals. A panel of residual plots show the standard diagnostic plots corresponding to the diagnostic summary statistics in the spatial summary viewer. For example, if the Durbin-Watson and Shapiro-Wilk images detected problems, one would view a lagged residual (e_i vs. e_{i+1}) and QQ-normal plot. As with the temporal summary viewer, right-clicking on a time-series plot brings up the spatial detail.

Spatial Detail Viewer Spatial detail is depicted with a sequence of studentized residual images using linked orthogonal slice viewers (Fig. 1-d). Use of studentized residuals is important, since spatial structure in the variance can obscure artifactual patterns in the unscaled residuals. If five residual images are viewed and scan i is selected, the studentized residual images for scans $i - 2$ through $i + 2$ are displayed; the mean functional image is also displayed for co-localization. The spatial detail is important for characterizing both the spatial and temporal extent of a problem.

2.4 Spatiotemporal Diagnosis Strategies

[Table 2 about here.]

In previous sections, we have described summaries of fit and statistics sensitive to possible artifacts and violations of assumptions, and dynamic graphical tools to efficiently search interesting regions in both spatial and temporal aspects. It is not immediately clear 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 (a) check assumptions and (b) explore expected and unexpected variability, (c) address problems found (see Table 2). In short, we search summaries and detail, and perform exploration and diagnosis of noise before exploration of signal.

Step 1: Explore Temporal Summaries

We use the temporal summaries to find scans affected by artifacts and acquisition problems. 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 temporal outlier count, and the motion parameters. Usually, these jumps correspond to head movements and acquisition artifacts. Third, we check the relationships between different temporal 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 Spatial Summaries

The spatial summaries are next used to do both diagnosis and exploration. For diagnostic purposes, we check the regression assumptions; homoscedasticity with the Cook-Weisberg score test and spatial outlier count, independence with the Durbin-Watson and cumulative periodogram tests, and normality with the Shapiro-Wilk test. We pay special attention to regions with both significant diagnostic statistics and anticipated experimental effects. We also note other locations with significant diagnostic statistics. These regions may contain information about model fitting or possible artifacts.

For exploratory purposes, we search images of signal and noise, focusing on the noise first. We study variability with the PCT image; we identify the typical gray matter PCT value, and then look for regions PCT values in great excess, say, twice that value. 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 temporal 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 percent change, t or F images. We localize interesting signals 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 temporal detail.

Step 3: Explore Temporal Detail

For a given voxel we examine the temporal detail plots, doing so interactively with spatial summaries to characterize the sources of the significant experimental or diagnostic statistics. From the time series plots of data with fit and residuals, we can not only assess the goodness-of-fit of the model to the signal, but also identify unmodeled signals. Also, from the time series plot of residuals, we note possible outlier scans. We reference these with the temporal summary outlier count and characterize their spatial extent with the spatial 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 score statistic, we use a residual plot versus predictor variable to verify that systematic heteroscedasticity and not an outlier is responsible.

Step 4: Explore Spatial Detail

As guided by the temporal detail or temporal summaries, we use a sequence of studentized residual images to find the spatiotemporal extent of problems identified in previous steps. Fixed and spinning MIP images are helpful to give an overall indication of the problems. When examining the temporal series of residual images, we note the spatiotemporal extent of the problem. 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 & Lewis, 1994), like gross movements or spike noise. Another approach to outliers is Windzorization (Arm, 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. 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 unmodeled signal from the temporal 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).

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 signal has assumptions clearly satisfied, while a questionable region has some significant diagnostics but exploration of the fit and residuals has shown the signal to be believable. Artifactual activation is clearly due to outliers or acquisition artifacts which could not be remedyd. In regions with no significant activation, it is also important to report the character of the regions with significant diagnostic statistics. The source of these significant results may be related to the unmodeled signal or artifactual variation and may be the source of new neuroscientific hypotheses or problems for MR physicists to solve.

3 Simulation Studies

In this section we examine the performance of the spatial summary diagnostic statistics using simulated datasets. All of our spatial summary statistics have been well-studied under 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. 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.

3.1 Simulation Methods

Time series data was simulated from a 84-observation model, corresponding to a publicly available dataset (`http://fil.ion.ucl.ac.uk/spm/data`); we used such a short-length time series to characterize the small-sample limitations of our diagnostic statistics. The simulated data were comprised of the sum of two series: One was the fixed response effect including nine covariates corresponding to intercept (1), experimental condition (1), and drift terms (7); 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-twelve autoregressive process. The parameters of these covariates and the twelve AR parameters were obtained from a real dataset. 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), 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 (see Appendix C.2).

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 depends on two criteria. First, the percentage of rejection under null hypothesis at three rejection levels (0.05, 0.01, and 0.001) for each statistics under various correlation conditions are computed and listed in Table 3. On the other hand, under the null hypothesis, the distribution of the p-values should be uniform. Therefore, Q-Q plots of the logarithm of the p-values were created and helpful to examine the behavior of these spatial summary statistics.

[Table 3 about here.]

3.2 Simulation Results

Table 3 shows the estimated rejection rates of the diagnostics. The Monte Carlo standard deviations of the rejection rates are $2.18{\times}10^{-3}$, $9.95{\times}10^{-4}$, and $3.16{\times}10^{-4}$ 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 the table, 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, cumulative periodogram test is superior to the DW test for detecting high order autoregressive process, as indicated in last rows of Table 3.

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 test are similar and most of them are within two standard deviations of the nominal α-levels.

As for Cook-Weisberg score test, under 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.

4 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 motion correction, where the subject was asked to speak aloud.

4.1 Experiment

The study employed a block design of word generation task, with rest and activation conditions. The stimulation paradigm consisted of 6 cycles of rest/activation, with a final rest condition; there were 20 scans per cycle. During the activation, subject was asked to generate a word that starts with each letter of the alphabet starting from “A”. Functional data was acquired on a 1.5T GE Signa magnet. A sequence of 130 EPI images were collected with a TR of 3,000ms, a TE of 40ms. Images were 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://fil.ion.ucl.ac.uk/spm`). Time series at each voxel were scaled by the global signal and fitted voxelwise with a linear regression model with a covariate consisting of the convolution of box-car design with a canonical hemodynamic response function, and a 6-elements 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 evaluate the data and model as outlined in Section 2.4.

4.2 Results

We start with temporal and spatial summaries, and then explore temporal and spatial detail as guided by the summaries. Inspection of the temporal 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 time series has a general downward trend and has several negative dips. The outlier count has several spikes, one image in particular having over 70% outliers (scan 105). 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.

[Figure 2 about here.]

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

[Figure 3 about here.]

[Figure 4 about here.]

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 over-scaled. Hence instead of omitting a scan, we alter the model by removing the global scaling.

Global Scaling Eliminated The temporal summaries are the same after removing the global scaling, except for the outlier count (Fig. 5). The outlier time series is improved, but many scans have considerably more than the expected 0.3% outlier rate. Checking the spatial 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 spatial detail for most outlier-spike scans shows either similar acquisition artifacts 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. Hence these patterns are not attributable to the particular model we use.

[Figure 5 about here.]

[Figure 6 about here.]

The spatial summaries still reveal problems. The DW and CP images detect regions with periodic variation corresponding to about 1/4 cycle off from the experimental paradigm (Fig. 7). The regions exhibiting this temporal pattern are primarily 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.

[Figure 7 about here.]

The CW-E image has a pronounced hyperintensity in the frontal pole (Fig. 8-a). Exploration of temporal detail localizes this to the last epoch (Fig. 8-c), and spatial detail characterized the pattern of signal loss (See Fig. 6-d).

[Figure 8 about here.]

The SW image identifies bilateral regions as non-normal (Fig. 9-a). The diagnostic plot (Fig. 9-b) and temporal detail (Fig. 9-c) 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.

[Figure 9 about here.]

The artifacts identified in the CW-E and SW images are troubling and we wish to remedy these problems by removing corrupted scans. One possibility would be to investigate the spatial 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 4 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. 8-c). 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 time series are reduced in magnitude and only have one notable spike. Maximum intensity projections of the spatial summaries before and after problem scans are removed are shown in Figure 10. The CW and SW images are now mostly uniform, while the DW image and CP image (not shown) exhibit hyperintensities in vascular and edge voxels. In fact, based on a 0.05-FDR-thresholded images (not shown), the autocorrelation is only negligible in white matter voxels. This suggests a physiological source of autocorrelation which should be addressed in the final modeling of this dataset.

The problems in the frontal pole and bilateral frontal regions are much reduced. Inspection of temporal detail at the few hyperintensities in the SW image (Fig. 9-b) 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 choose not to remove any other scans.

[Figure 10 about here.]

With a largely artifact-free dataset, we continue our exploration of the spatial summaries of nuisance variability and noise. The F image of the drift basis coefficients reveals no unusual patterns (not shown), mainly identifies slow monotonic drifts at the posterior surface of the brain. The PCT image is used to characterize noise; the modal value of 0.47% ($\alpha = 0.05$ FDR-corrected, 0.30% $\alpha = 0.05$ uncorrected), means that in a typical voxel changes as small as 0.47% can be detected. We window the PCT so that the maximal intensity corresponds to twice the modal value, to accentuate highly variable areas. Of immediate concern is the increased PCT in the bilateral motor and left dorsal lateral frontal areas (Fig. 11-a), the very regions of expected activation. Inspection of temporal detail suggests that while a few scans are affected by acquisition artifacts, none are severe. Instead, we note that voxels in these regions all exhibit experimental variation that rises early relative to the model, by about 1/8-cycle (Fig. 11-c). 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; if we have instead found experimental variation 1/4-cycle out of phase, we would be more concerned about vocalization artifacts.

[Figure 11 about here.]

We next examine spatial summaries of the signal with percent change and t images. There are focal signals in the bilateral sensory-motor cortices, bilateral auditory cortices, and bilateral cerebellum, and diffuse signal 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 temporal detail for each foci (not shown), we confirm that artifactual sources are not responsible for the effects; however, for each voxel examined there is the 1/8-cycle phase error evident (Fig. 11-c). 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 12-a). While this is easily identified as susceptibility-related, it is a demonstration of the merit of examining unthresholded images of signal.

[Figure 12 about here.]

In summary, the application of our diagnostic tools finds violation of the independence assumption, due to physiology and out-of-phase experimental variation, and violation of homoscedasticity assumption, due largely to artifacts. We remedy these problems by eliminating global scaling and removing scans with serious artifacts. The resulting reduced dataset 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 due to an 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 (or locally estimated) hemodynamic 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; other solutions include Windzorization and data imputation. 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 data even after thorough diagnosis and exploration.

5 Discussion

While it is straightforward to apply the general linear model to fMRI data, and seemingly easy to make inferences on activations based, 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 dataset consists of 1,000 images and 100,000 voxels.

We have proposed methods for evaluating assumptions on massively univariate linear models and exploring fMRI data. The key aspects are spatial and temporal summaries sensitive to anticipated effects and model violations, and interactive visualization tools that allow efficient exploration of the 3D parameter images and 4D 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. From the dataset in this paper, the proposed diagnostic tools illustrate their importance for these two purposes: On the first point, we found it necessary to eliminate global scaling and a collection of bad scans; On the second, we found 1/4-cycle out-of-phase variability distributed throughout the brain, and 1/8-cycle out-of-phase activations in anticipated regions. Thus, we have refined the quality of our data and learned about a shift in this subject's hemodynamic response, neither which could have accomplished solely with inspection of thresholded t images.

While we have focused on single subject fMRI in this work, virtually all of these tools and statistical summaries are relevant for second level fMRI, PET and SPECT. The only significant differences are the reduced concern about autocorrelation and reduced power of diagnostic statistics due to smaller sample size.

The principal contributions of this work are both statistical and computational. We have identified and characterized diagnostic statistics relevant for linear modeling of fMRI, focusing in particular on impact of autocorrelation. In addition to defining all of these measures in a consistent manner, we have presented practical methods for calculating p-values. Computationally, we have specified and created a system for the efficient exploration of massive datasets. Using linked, dynamic viewers, all the various summary measures can be rapidly integrated and understood. Lastly, we have given practical recommendations on how to diagnose and explore typical fMRI datasets.

The principal directions for future work is the diagnosis of models of temporal dependence, models of spatial dependence, and multivariate exploration of the residuals. We have focused solely on temporally independent linear models to utilize the rich diagnostics literature which assumes white noise. Assessing models of temporal dependence may require model-specific methods and lack the wide range of diagnostic measures. Inference using Gaussian Random Fields requires assumptions on the spatial dependence structure, and methods are needed to assess the validity of continuity and stationarity assumptions. Finally, our approach using spatial and temporal residual summaries may miss extended spatiotemporal patterns. Methods such as ICA applied to the standardized or BLUS residuals may provide valuable tools for this purpose.

6 Conclusion

In this work, we have developed a general framework for diagnosis of linear models fit to fMRI data. Using spatial and temporal summaries and dynamic viewers, we have shown how to swiftly localize rare anomalies and artifacts in large 4D datasets.

7 Acknowledgements

The authors wish to 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.

References

1998. *Encyclopedia of Biostatistics*. Vol. 6. John Wiley & Sons. Pages 4588–4590.

Aguirre, G, Zarahn, E, & D'Esposito, M. 1998. A critique of the use of the kolmogorow-smirnov (KS) statistic for the analysis of BOLD fMRI data. *Magnetic Resonance in Medicine*, **39**, 500–505.

Barnett, V., & Lewis, T. 1994. *Outliers in statistical data (Third edition)*. John Wiley & Sons.

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. *Magnetic Resonance in Medicine*, **40**, 55–60.

Casella, G. 2001. *Statistical inference (Second edition)*. Wadsworth Publishing Company.

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.

Draper, N. R., & Smith, H. 1998. *Applied regression analysis (Third edition)*. John Wiley & Sons.

Durbin, J., & Watson, G. S. 1950. Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, **37**(3/4), 409–428.

Durbin, J., & Watson, G. S. 1951. Testing for Serial Correlation in Least Squares Regression: II. *Biometrika*, **38**(1/2), 159–177.

Durbin, J., & Watson, G. S. 1971. Testing for serial correlation in least squares regression, III. *Biometrika*, **58**, 1–19.

Friston, KJ, Holmes, AP, Worsley, KJ, Poline, J-B, & Frackowiak, RSJ. 1995. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, **2**, 189–210.

Genovese, CR, Lazar, N, & Nichols, TE. 2001. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage, to appear*.

Goutte, C, Toft, P, Rostrup, E, Nielsen, FA, & Hansen, LK. 1999. On Clustering fMRI Time Series. *NeuroImage*, **9**, 298–310.

Graybill, Franklin A. 1976. *Theory and Application of the Linear Model*. Duxbury Press.

Hartigan, J. A. 1977. Distribution Problems in Clustering. *Pages 45–72 of: Classification and Clustering*. Academic (New York; London).

Hoaglin, David C., Mosteller, Frederick, & Tukey, John W. (Ed). 1983. *Understanding Robust and Exploratory Data Analysis*. Wiley.

Holmes, AP. 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`.

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. *Human Brain Mapping*, **6**, 160–188.

Moonen, C.T.W., & Bandettini, P.A. (eds). 2000. *Functional MRI*. Springer.

Moser, E, Baumgartner, R, Barth, M, & Windischberger, C. 1999. Explorative signal processing in functional MR imaging. *International Journal of Imaging Systems and Technology*, **10**, 166–176.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. 1996. *Applied linear statistical models (Fourth edition)*. Richard D. Irwin Inc.

Nichols, T.E., & Luo, W.L. 2001. Data Exploration Through Model Diagnosis. *NeuroImage*, **13**(6 (2/2)), S208. Proceedings of Seventh Annual Meeting of the Organization for Human Brain Mapping, June 10-14, 2001, Brighton, England.

Razavi, M, Grabowski, TJ, Mehta, S, & Bolinger, L. 2001. The source of residual temporal autocorrelation in fMRI time series. *NeuroImage*, **13**(6 (2/2)), S228. Proceedings of Seventh Annual Meeting of the Organization for Human Brain Mapping, June 10-14, 2001, Brighton, England.

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

Ryan, T. P. 1997. *Modern regression methods*. Wiley-Interscience.

Schlittgen, R. 1989. Tests for white noise in the frequency domain. *Computational Statistics [Formerly: Computational Statistics Quarterly]*, **4**, 281–288.

Scott, David W. 1992. *Multivariate density estimation. Theory, practice, and visualization*. John Wiley & Sons.

Shapiro, S. S., & Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Shapiro, S. S., Wilk, M. B., & Chen, Mrs. H. J. 1968. A comparative study of various tests for normality. *Journal of the American Statistical Association*, **63**, 1343–1372.

Smirnov, N. 1948. Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics*, **19**(2), 279–281.

Stephens, M. A. 1970. Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society, Series B, Methodological*, **32**, 115–122.

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730–737.

Theil, H. 1971. *Principles of Econometrics*. John Wiley & Sons.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley.

A Percent change and interpretable linear models

Percent change is a standard measure of BOLD signal. For the linear model defined in Section 2, a ratio of a contrast and the baseline mean $c\beta/\mu \times 100\%$ will not have units of percent change unless the predictors and the contrast vector have been appropriately scaled. In this appendix we give the requirements on X and c in order for $c\beta/\mu \times 100\%$ to have percent change units and for $c\beta 100\%$ to that interpretation approximately.

To create images of percent change based on the general linear model, $c\beta$ must have units of the data. In particular, the predictors must have unit scaling and the contrast vector must preserve that scaling. Unit scaling of predictor X_j means that a unit change in β_j will cause a unit change of the predicted signal in $\hat{Y}$. A binary effect coded 0/1 has this effect, but a coding of -1/1 does not. In fMRI, this amounts to requiring the baseline-to-plateau range of each experimental predictor to be one. To preserve β's units, it is straightforward to show that the contrast vector must have an absolute sum of unity ($\sum_j |c_j| = 1$). If the X and c are so scaled, $c\beta$ will have units of the data and $c\beta/\mu$ will have units of fractional change from baseline.

Percent change is not the only reason to scaling design matrices and contrast vectors in this manner. Another motivation is simply to make the linear model as interpretable as possible. To have meaningful contrast and parameter images one needs to have data with meaningful units. Scaling the data to have approximate baseline mean of 100 ensures that the β's and $c\beta$ have approximate interpretation of percent change (*if* the design matrix and contrast vectors have been scaled as per above).

Standardizing the baseline mean in fMRI has typically been done by scaling the entire dataset such that the baseline mean image has "global" intensity of 100. (Note that this is different from global scaling of each image.) Global is usually defined as the arithmetic mean of all intracerebral voxels. However, the mean is very sensitive to hyperintensity outliers and the segmentation of brain from nonbrain. In particular, if a simple intensity threshold is used and 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; it requires no topological operations on the image data and is easily coded and quickly computed.

To summarize, we assert that models should be constructed to be as interpretable as possible. The fruit of this endeavor is 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.

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 consistent estimation of mode[2]. 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 don't 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 over-estimate 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 percentile to work on all images we have considered. Our threshold estimate is thus

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

where n is the number of voxels in the image. 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 non-tail 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 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. Matlab code for this method is available at `http://www.sph.umich.edu/~nichols/PCT`.

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

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

optimal (and consistent) histogram bin width for estimating the mode is order $n^{1/5}$; we use bin widths equal to $1.595 \times$ IQR$n^{-1/5}$ (Scott, 1992, pg 100), 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 datasets. 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 Figure 13).

[Figure 13 about here.]

C Linear Model Diagnosis Statistics

In this appendix, we describe the definition, distribution under null hypothesis of model fit (when possible), and how to assess the statistic for each spatial summary statistic.

C.1 Percent Change Threshold

Definition The percent change threshold (PCT) is the half-width of a $(1-\alpha)\%$ confidence interval for an effect of interest. For t statistics, the half-width of a confidence interval is the product of the standard deviation of an estimate times a t critical value (Neter *et al.*, 1996). In the model defined in equation (2), the estimate of interest is the percent change contrast $c\hat{\beta}/\mu$, where μ is the baseline. The estimate's standard deviation is

$$\sqrt{c(X^\top X)^{-1}c^\top}\hat{\sigma}/\mu.$$

The effect $c\hat{\beta}$ must have units of the data, which depends on the predictors and contrast being scaled as described in the Methods above. The critical value can be an uncorrected or corrected threshold. We recommend a corrected threshold, such as obtained by using the False Discovery Rate (Genovese *et al.*, 2001).

Assessment When used with a corrected threshold, the PCT image depicts the sensitivity for detecting changes when searching over the whole brain For characterizing artifacts, however, it can be useful to use an uncorrected threshold to measure sensitivity independent of the number of voxels examined. Hence we recommend viewing both corrected and uncorrected PCT by using a dual-axis color bar, one side showing uncorrected, the other showing corrected values.

For a technical note on PCT and Matlab software to create PCT images, see `http://www.sph.umich.edu/~nichols/PCT`.

C.2 Independence Tests

C.2.1 Durbin-Watson Statistic

Definition The Durbin-Watson statistic (Durbin & Watson, 1950) tests an assumption of independence against an alternative of a first order autoregressive model. The null hypothesis of zero autocorrelation, $\mathcal{H}_0 : \rho = 0$, is tested via the statistic

$$D = \frac{\sum_{i=2}^{N}(e_i - e_{i-1})^2}{\sum_{i=1}^{N}(e_i)^2} = \frac{e^\top A e}{e^\top e}$$

where A is the real symmetric matrix

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \cdots & \cdots & \cdots & \cdots \\ 0 & -1 & 2 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{bmatrix}$$

Note that the Durbin-Watson statistic accounts for the dependencies in the residuals, as per their orthogonality to the column space of X.

Distribution The distribution of D lies between 0 and 4 and is symmetric about 2. The mean and variance of D are given by (Durbin & Watson, 1950):

$$\begin{aligned} \mathsf{E}(D) &= \frac{P}{N-p-1} \\ \mathsf{Var}(D) &= \frac{2}{(N-p-1)(N-p+1)}[Q - P\mathsf{E}(D)] \end{aligned}$$

where

$$\begin{aligned} P &= \text{tr}A - \text{tr}\{X^\top AX(X^\top X)^{-1}\}, \\ Q &= \text{tr}A^2 - 2\text{tr}\{X^\top A^2 X(X^\top X)^{-1}\} + \text{tr}[\{X^\top AX(X^\top X)^{-1}\}^2] \end{aligned}$$

The distribution of D depends on X, and is difficult to calculate routinely. Therefore, the distribution of D is approximated by a Beta distribution (Durbin & Watson, 1951), and this should be sufficiently accurate if the number of degrees of freedom is large enough, say greater than 40, a criterion usually met with fMRI experiments. For positive autocorrelation, $\frac{1}{4}D$ is distributed in the Beta distribution, $\beta_{m,n}$, where m and n are found by the equations:

$$\begin{aligned} m + n &= \frac{\mathsf{E}(D)(4 - \mathsf{E}(D))}{\mathsf{Var}(D)} - 1 \\ n &= \frac{1}{4}(m+n)\mathsf{E}(D) \end{aligned}$$

Small values of D indicate positive autocorrelation; this property, combined with the finite range of D make the statistic unsatisfactory for visualization. An image of $\frac{m(4-D)}{nD}$, which is distributed as $F_{2n,2m}$, is much more amenable to visual inspection.

Assessment Values of Durbin-Watson statistic near zero indicate positive autocorrelation of the random errors, and values near 2 indicate no correlation. For large N, the Durbin-Watson statistic approximates $2(1-\hat{\rho})$, where $\hat{\rho}$ is the estimated serial correlation coefficient. As described above, for the purpose of visualization, we use both the transformed F statistic and corresponding p-value approximated by Beta distribution to assess the significance of autocorrelation.

C.2.2 Cumulative Periodogram Test

Definition The general definition of periodogram is

$$I(w_k) = N \left| \sum_{i=0}^{N-1} y_i \exp(-jw_k i) \right|^2$$

where N is the length of time, and w_k is the Fourier frequencies, $w_k = 2\pi k/N, \; k = 1, \cdots, N$. Define quantities $C_k = \sum_{i=1}^{k} I(w_i) \; : \; k = 1, \cdots, M$ where M is the largest integer strictly less than $N/2$, and $B_k = C_k/C_M$. A plot of B_k against k/M', where $M' = M - 1$, is called the cumulative periodogram (Diggle, 1990). The cumulative periodogram should increase approximately linearly from 0 to 1 as k runs from 1 to M' under the null hypothesis of white noise. A statistic which measures the departure from linearity using cumulative periodogram is defined as following (Schlittgen, 1989):

$$C = \max_{k=1,\ldots,M'} (\max(|B_k - k/M|, |B_k - (k-1)/M|)).$$

The cumulative periodogram test is the Kolmogorov-Smirnov statistic to test the hypothesis that the B_k are an ordered random sample from the uniform distribution on $(0, 1)$.

However, modifications are necessary when the periodogram is computed from the least-squares residuals. The variance-covariance matrix of the residuals is $\mathsf{Var}(e) = \sigma^2(I - H)$ where $H = X(X^\top X)^{-1}X^\top$. Thus even when the errors in a regression model are serially uncorrelated, the calculated residuals are correlated by virtue of the singularity of their distribution, corresponding to $e^\top X = 0$, their orthogonality to the column space of X. Residuals are typically negatively correlated, and may have a peculiar spectral density depending on H. The cumulative periodogram test is sensitive to these dependencies and will falsely detect dependent errors in excess of nominal levels. To remove this correlation from the residuals, we instead use Best Linear Unbiased residuals with Scalar (diagonal) covariance matrix, or BLUS residuals (Theil, 1971). BLUS residuals are unbiased in the sense that their expectation is zero, the same as the unobservable errors (ε); they are best in that their distance from the true errors is minimized in expectation.

The BLUS residuals are defined as follows: First, partition the observation matrix $[Y \; X]$ into two submatrices consisting of p and $N - p$ rows, respectively. The basic equation and its least square estimator then become

$$\begin{array}{r} p \text{ rows} \\ N-p \text{ rows} \end{array} \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix} = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \beta + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \end{bmatrix} = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \hat{\beta} + \begin{bmatrix} e_0 \\ e_1 \end{bmatrix}$$

Then the $N - p$ element vector

$$\hat{e}_1 = e_1 - X_1 X_0^{-1} \left[\sum_{r=1}^{R} \frac{d_r}{1 + d_r} g_r g_r^\top \right] e_0$$

is the BLUS residual vector, where $d_1^2, \cdots, d_R^2$ are the eigenvalues of the matrix $X_0(X^\top X)^{-1}X_0^\top$, and $g_1^2, \cdots, g_R^2$ are eigenvectors (corresponding to the R eigenvalues that are less than one) of the matrix. The BLUS residuals ($\hat{e}_1$) can be seen as the ordinary residuals (e_1) adjusted to have a diagonal covariance matrix. While any p rows can used to form the base of the transformation, we are interested in detecting autocorrelation and so we do not want to introduce any discontinuities. We use the first p observations as the base, though we could also use the last p.

Distribution The approximate p-value corresponding the test statistic is calculated according to the equation (Smirnov, 1948):

$$\mathsf{P}(C' \le c') = 1 - 2\sum_{\nu=1}^{\infty}(-1)^{\nu-1}e^{-\nu^2 c'^2} = (2\pi)^{\frac{1}{2}}c'^{-1}\sum_{\nu=1}^{\infty} e^{-(2\nu-1)^2\pi^2/8c'^2}$$

where C' is the transformed test statistic, $C' = C(\sqrt{M'} + 0.12 + 0.11/\sqrt{M'})$ (Stephens, 1970). The approximate p-value is obtained with a truncated series; we find that as few as 30 terms are sufficiently accurate.

Assessment Cumulative periodogram statistic measures the maximal departure of the observed normalized cumulative periodogram from increasing linearly from 0 to 1. It is more intuitive to interpret this statistic. The larger this statistic is, the more the observed curve is different from linear, and thus the periodogram of the data is not flat. Higher test statistic results in smaller p-value and rejection of the null hypothesis of white noise.

C.3 Homoscedasticity Test

Definition The Cook-Weisberg (1983) score statistic tests an assumption of homogeneous variance against an alternative of heteroscedasticity parameterized by a log-linear model. Assume that the random errors are independent normal variates with mean 0 and

$$\mathsf{Var}(\varepsilon_i) = \sigma^2 \exp(z_i\lambda)$$

where z_i is a known $1{\times}p'$ vector of covariates and λ is a p'-vector of unknown parameters. Other forms of variance models are also possible; see (Cook & Weisberg, 1983).

Homoscedasticity is thus expressed as $\mathcal{H}_0 : \lambda = 0$. A score statistic is used to test $\mathcal{H}_0$; a score test is the slope of the log-likelihood at the null hypothesis value, normalized to unit variance (Casella, 2001). Let U be a vector with elements $e_i^2/\hat{\sigma}^2$, where $\hat{\sigma}^2 = \sum_{i=1}^{N} e_i^2/N$. Let Z be the $N{\times}p'$ matrix of covariates, and $\bar{Z}$ be the column-centered version of Z, $Z - \mathbf{1}\mathbf{1}^\top Z/N$, where $\mathbf{1}$ is a $N{\times}1$ vector of ones. Then the score test statistic for the hypothesis $\lambda = 0$ is

$$S = \frac{1}{2}U^\top \bar{Z}(\bar{Z}^\top \bar{Z})^{-1}\bar{Z}^\top U$$

Computationally, S is equivalent to one-half of the sum of squares for the regression of U on Z. Therefore, the score test can be easily obtained using a general regression analysis.

Distribution Under the hypothesis $\lambda = 0$, the asymptotic distribution of score statistic is central chi-squared with p' degree of freedom.

Assessment The homoscedasticity test is based on the score test assessing the dependence of residual on presumed variables. As mentioned above, this score test is equal to one half of the sum of squared error from the regression of the residuals on possible predictors. Therefore, high score statistic causes smaller p-value. In particular, large score test means that the residuals are independent on the predictors, and the residuals are homogeneous with respect to the predictors.

C.4 Normality Test

Definition Many statistics have been proposed for testing a sample for normality. Extensive comparisons have been made on the relative sensitivity of normality tests, as applied to samples from a variety of nonnormal populations (Shapiro *et al.*, 1968). The results reveals that the Shapiro-Wilk W statistic provides a superior omnibus indicator of non-normality, judged over the various symmetric, asymmetric, short- and long-tailed alternatives. In particular, the Shapiro-Wilk W out-performs the Kolmogorov-Smirnov statistic (Stephens, 1974).

The Shapiro-Wilk (1965) statistic tests the null hypothesis that the residuals follow a Normal distribution. It is based on the regression of ordered residuals on corresponding expected normal order statistics is linear. Let $s^\top = (s_1, ..., s_N)$ denote the vector of expected values of standard normal order statistics, $s_1 \leq s_2 \leq \cdots \leq s_N$, and let $V = (v_{ij})$ be the corresponding $N \times N$ covariance matrix, $v_{ij} = \mathsf{Cov}(s_i, s_j)$..

Suppose $e^\top = (e_1, \cdots, e_N)$ is a random sample of residuals , ordered $e_{(1)} < e_{(2)} < \cdots < e_{(N)}$. Then the normality test statistic is

$$W = \frac{[\sum_1^N a_i e_{(i)}]^2}{\sum_1^N (e_i - \bar{e})^2}.$$

The a_i are coefficients such that $\sum_1^N a_i e_{(i)}$ is the best linear unbiased estimate of the slope of $e_{(i)}$ on their expected values, s_i, the standard normal order statistics.

However, since the elements of V are known only up to samples of size 20, Shapiro and Wilk (1965) offer a satisfactory approximation for a which improves with increasing sample size N. By definition,

$$a^\top = \frac{s^\top V^{-1}}{(s^\top V^{-1} V^{-1} s^\top)^{\frac{1}{2}}} = \frac{s^\top V^{-1}}{K}$$

is such that $a^\top a = 1$. Let $a^* = s^\top V^{-1}$, then $K^2 = a^{*\top} a^*$. The suggested approximation is

$$\hat{a}_i^* = \begin{cases} 2s_i, & i = 2, 3, ..., N-1, \\ \left(\frac{\hat{a}_1^2}{1-2\hat{a}_1^2} \sum_{i=2}^{N-1} \hat{a}_i^{*2}\right)^{\frac{1}{2}}, & i = 1 = N, \end{cases}$$

where for $N > 20$,

$$\hat{a}_1^2 = \hat{a}_N^2 = \frac{\Gamma(\frac{1}{2}(N+1))}{\sqrt{2}\Gamma(\frac{1}{2}N+1)}.$$

Note $a_i = -a_{N-i+1}$. By Stirling's formula, $\hat{a}_1^2$ and $\hat{a}_N^2$ may be approximated and simplified (Royston, 1982)

$$\hat{a}_1^2 = \hat{a}_N^2 = \left[\frac{6N+7}{6N+13}\right] \left(\frac{\exp(1)}{N+2} \left[\frac{N+1}{N+2}\right]^{N-2}\right)^{\frac{1}{2}}.$$

Distribution Shapiro-Wilk statistic is scale and origin invariant with maximum and minimum to be 1 and $Na_1^2/(N-1)$, respectively. The distribution of W does not depend on the sample size N, under the null hypothesis of normal distribution, and is independent of mean and standard deviation. The asymptotic distribution of W is not normal, and for sample size between 20 and 2000, Royston (1982) proposed the following approximate normalizing transformation for W:

$$h = (1-W)^\tau \quad \text{and} \quad t = (h - \mu_h)/\sigma_h$$

where t is a standard normal deviate, τ is the power transformation parameter, and μ_h and σ_h are the mean and s.d. of h. The quantities τ, μ_h and σ_h are all polynomial functions of N, given in Royston (1982, pg.199). This equation allows one to calculate the significance level of the original W.

Assessment Normality is rejected if W is too small. Using Royston's approximation (Royston, 1982) for significance level of W, large values of t (as opposed to small values of W) indicate departure from normality. The approximation works well for sample size less than 2000. If the sample size is greater than 2000, it is suggested to use Kolmogorov test. Since the fMRI experiment usually has sample size less than 2000, Shapiro-Wilk statistic is applied in this work.

C.5 Spatial Outlier Count

Definition Residual outliers can be identified from residual plots against X or $\hat{Y}$, as well as from box plots, stem-and-leaf plots, and dot plots of the residuals (Neter *et al.*, 1996). A general definition of outlier is that it is at least three or four standard deviations from the center of the data (Ryan, 1997). However, the variance of e_i is not necessarily constant, $\mathsf{Var}(e_i) = \sigma^2(1 - H_{ii})$, even if homoscedasticity assumption is true. The studentized residuals are residuals with standardized variance. It is defined as $e_i/(\sigma\sqrt{1 - H_{ii}})$, and has mean zero and variance 1. We defined a residual to be outlier if the absolute studentized residual is greater than 3. The number of outliers at each voxel are summed over time:

$$L = \# \left\{ \left| \frac{e_i}{\hat{\sigma}\sqrt{1 - H_{ii}}} \right| > 3 \right\}, \tag{2}$$

where $\hat{\sigma}$ is the residual standard deviation.

Distribution If the residuals are Normal and independently and identically distributed, then L would follow a Binomial distribution with success probability $q = 2\Phi(-3) = 0.3\%$. We use this result, though the dependence of the residuals makes this hold only approximately.

Assessment The assessment of the image of outlier count is intuitive, which the value represents the number of outliers at each voxel. Large number of outlier count may be an indicator of misfitting of the regression model or certain artifacts in the data.

List of Figures

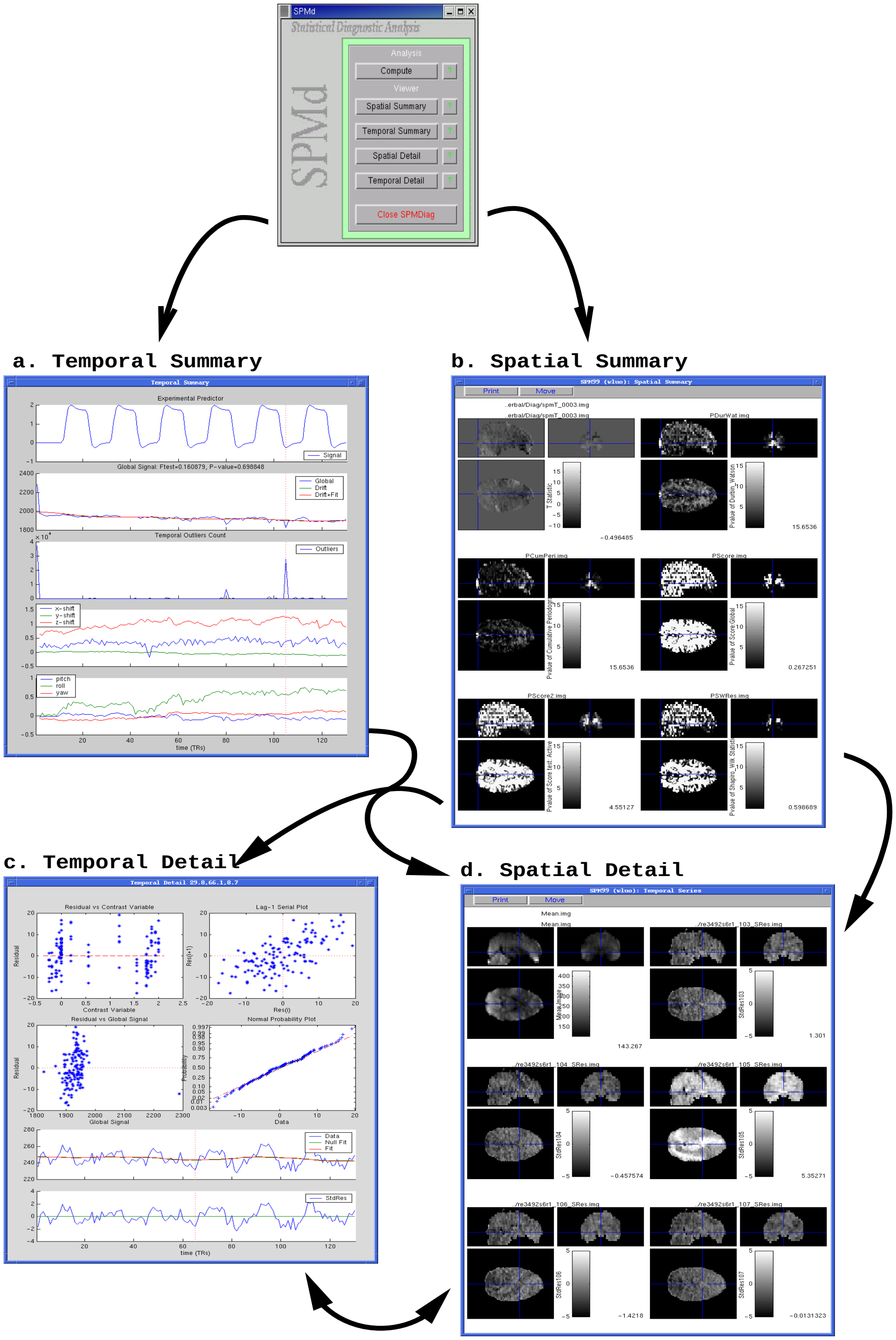


Figure 1: Statistical Parametric Mapping Diagnosis (SPMd), a toolbox for SPM99. SPMd provides efficient tools for the diagnosis and exploration of fMRI data. As guided by spatial and temporal summaries, spatial and temporal detail are used to localize artifacts and anomalies.

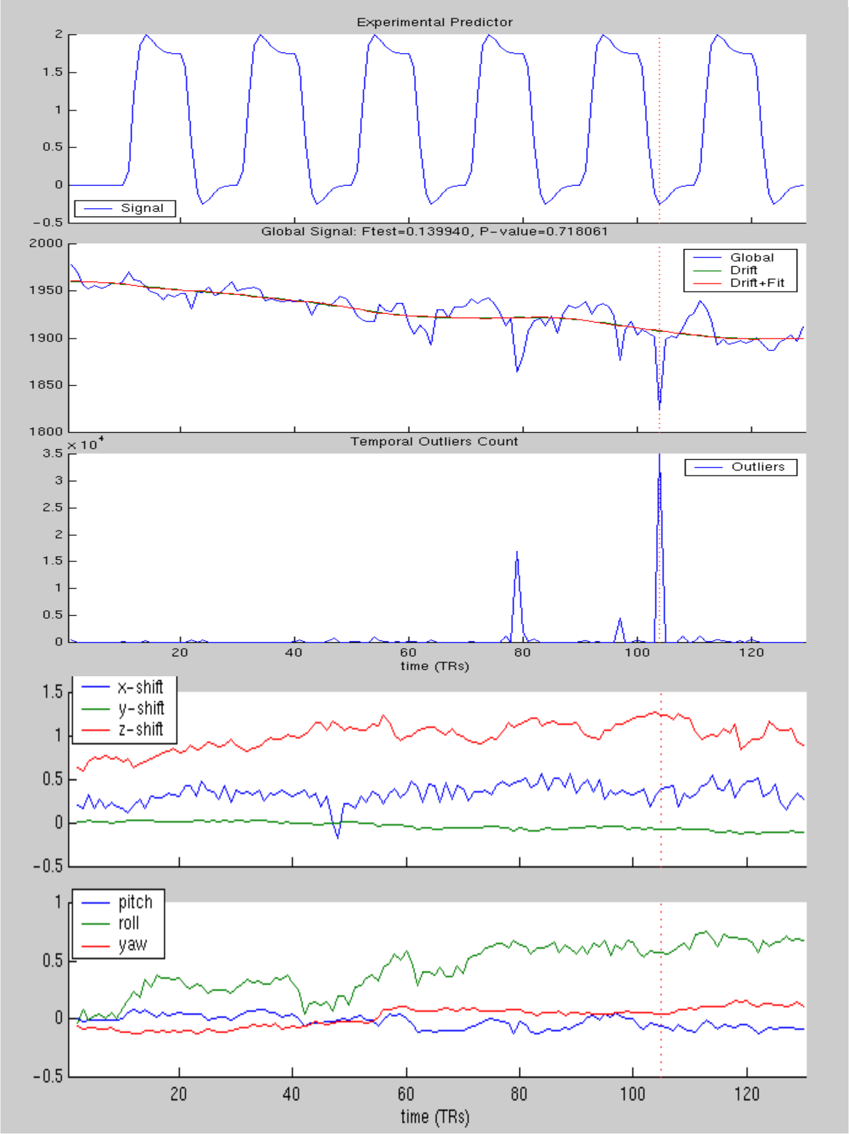
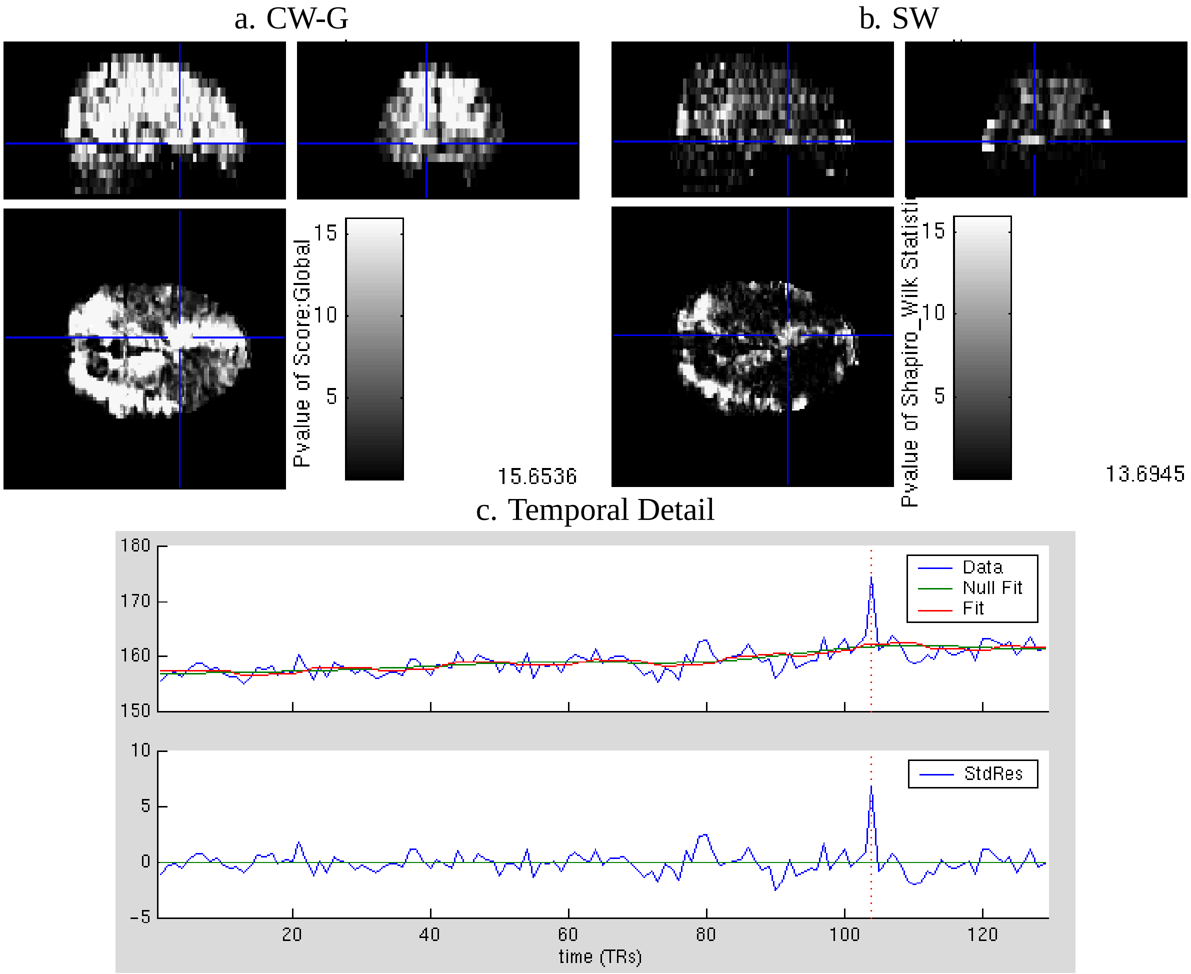


Figure 2: Temporal summaries of the full 129-scan dataset.



a. CW-G

b. SW

c. Temporal Detail

Figure 3: Spiral pattern in the diagnostic images. Spatial summary images of Cook and Weisberg's score test with respect to the global signal (a. CW-G), and Shapiro-Wilk statistic (b. SW). Note the CW-G detects problems across the volume, while SW is only sensitive to artifacts in one plane. c. Temporal detail of the fit and studentized residuals at a voxel with high intensity in the CW-G and SW images.

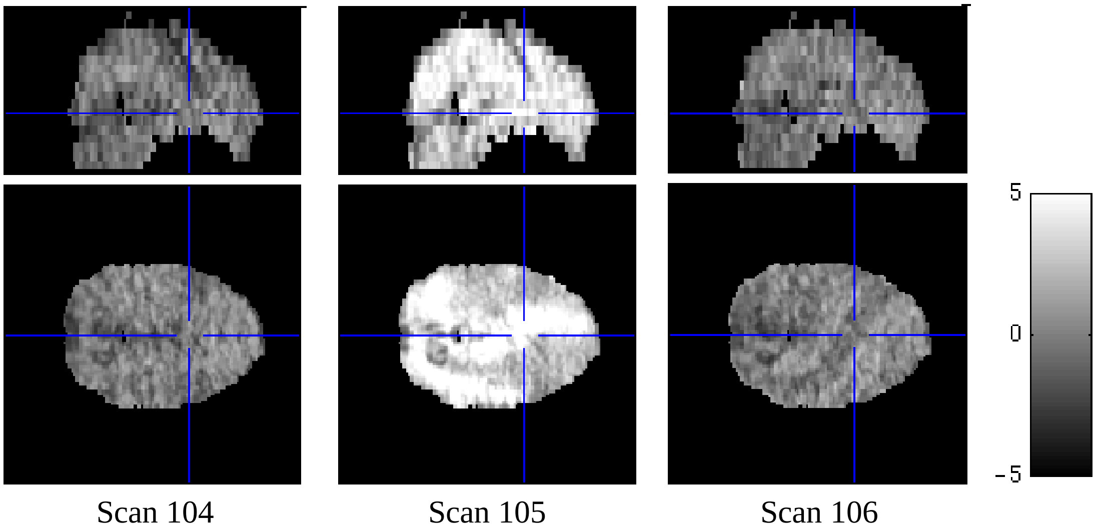


Figure 4: Spatial detail around scan 105, studentized residuals for scans 104–106. The intensity of scan 105 is much higher than other scans. The spiral pattern in scan 105 also corresponds to the similar pattern in the CW-G and SW images. See Figure 3.

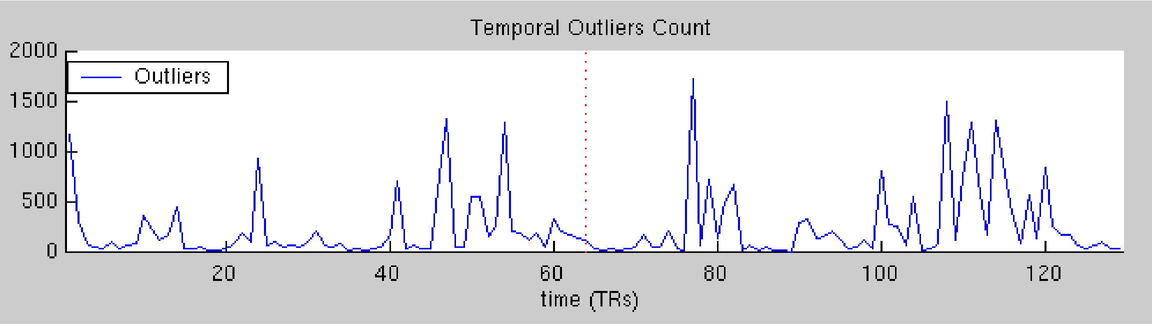
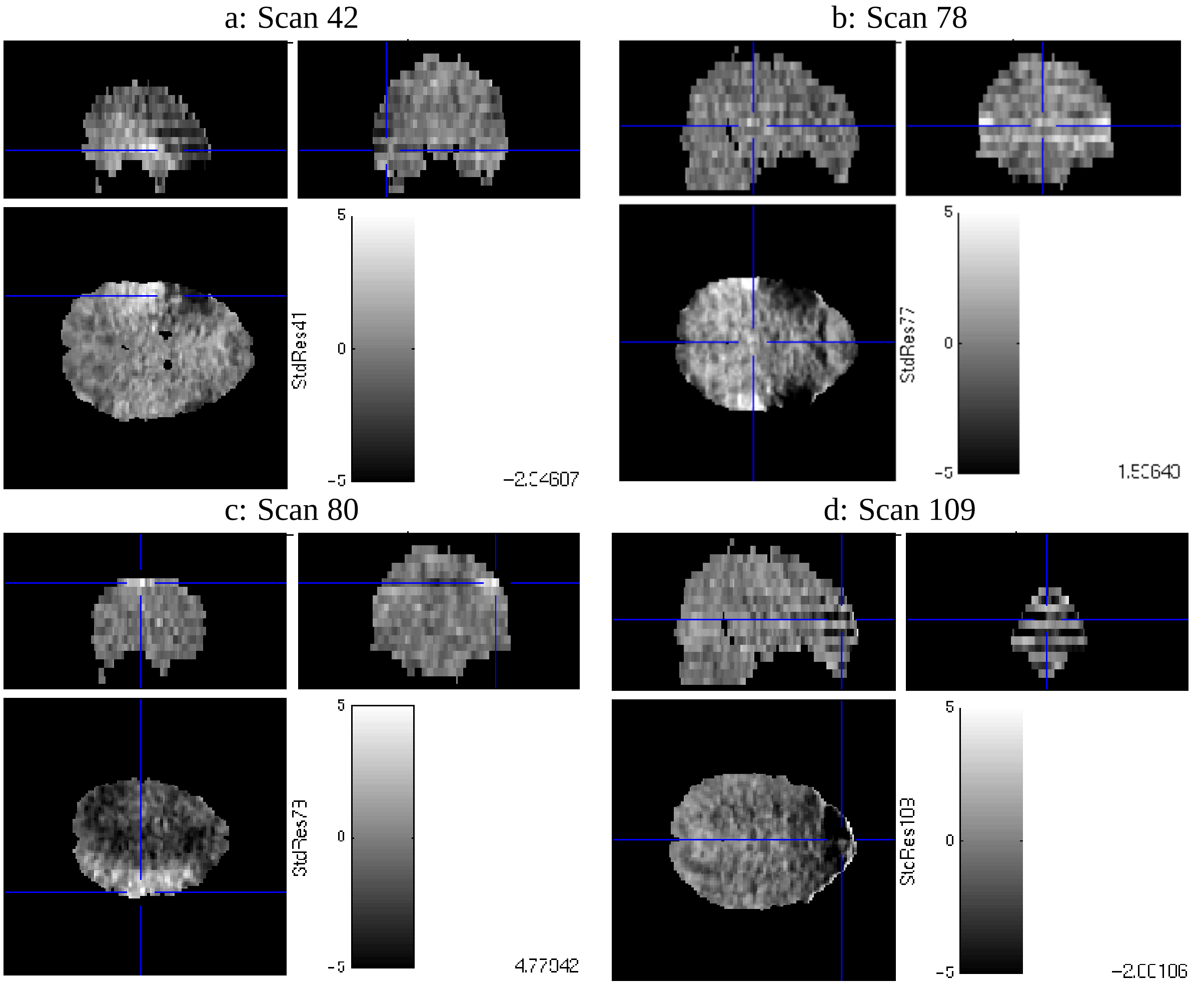


Figure 5: Temporal summary of outlier count without global scaling. All other plots are the same as before (See Fig. 2).

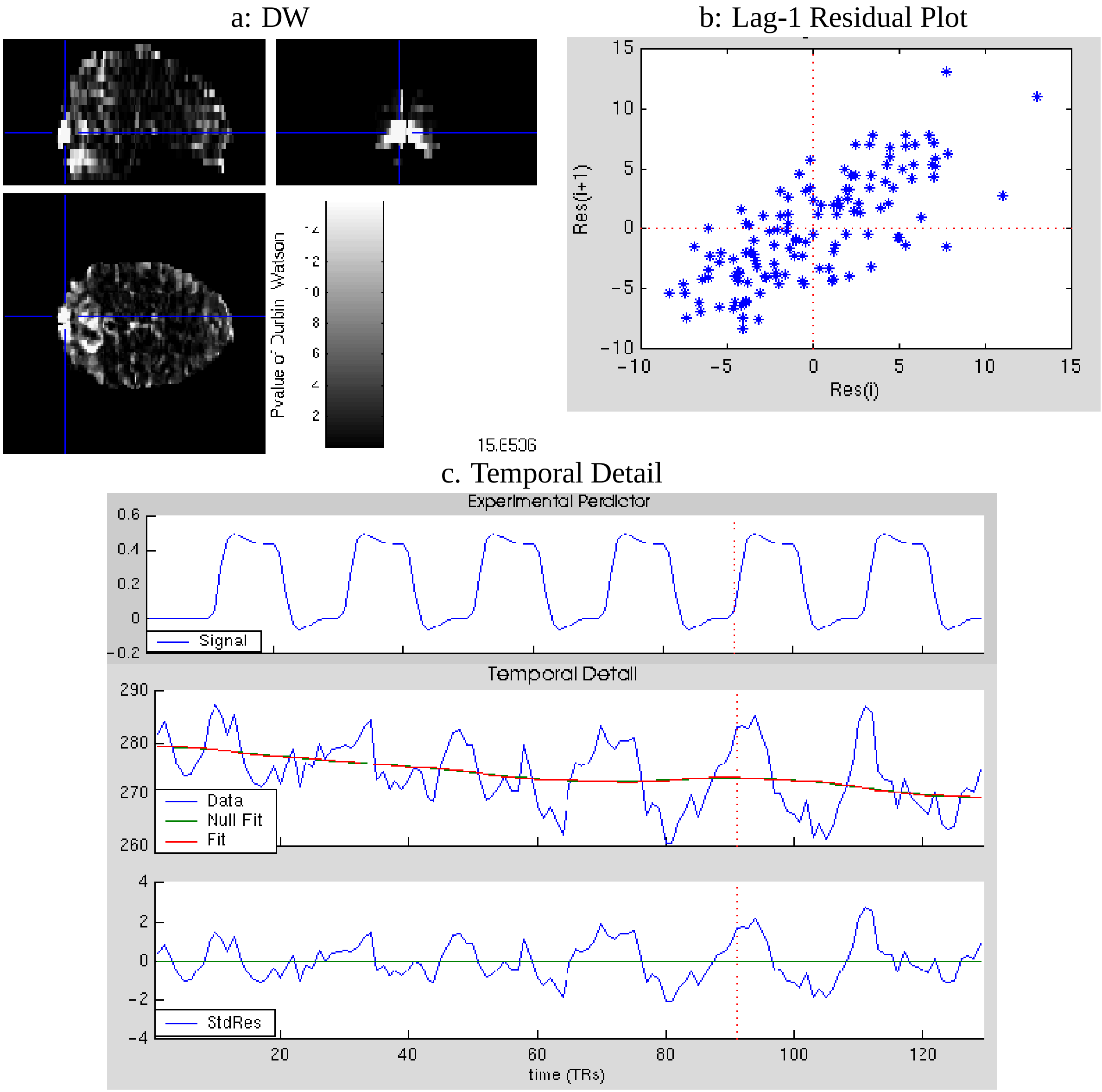

a: Scan 42

b: Scan 78

c: Scan 80

d: Scan 109

Figure 6: Acquisition artifacts revealed in spatial detail. These studentized residual images are identified as having high outlier counts.

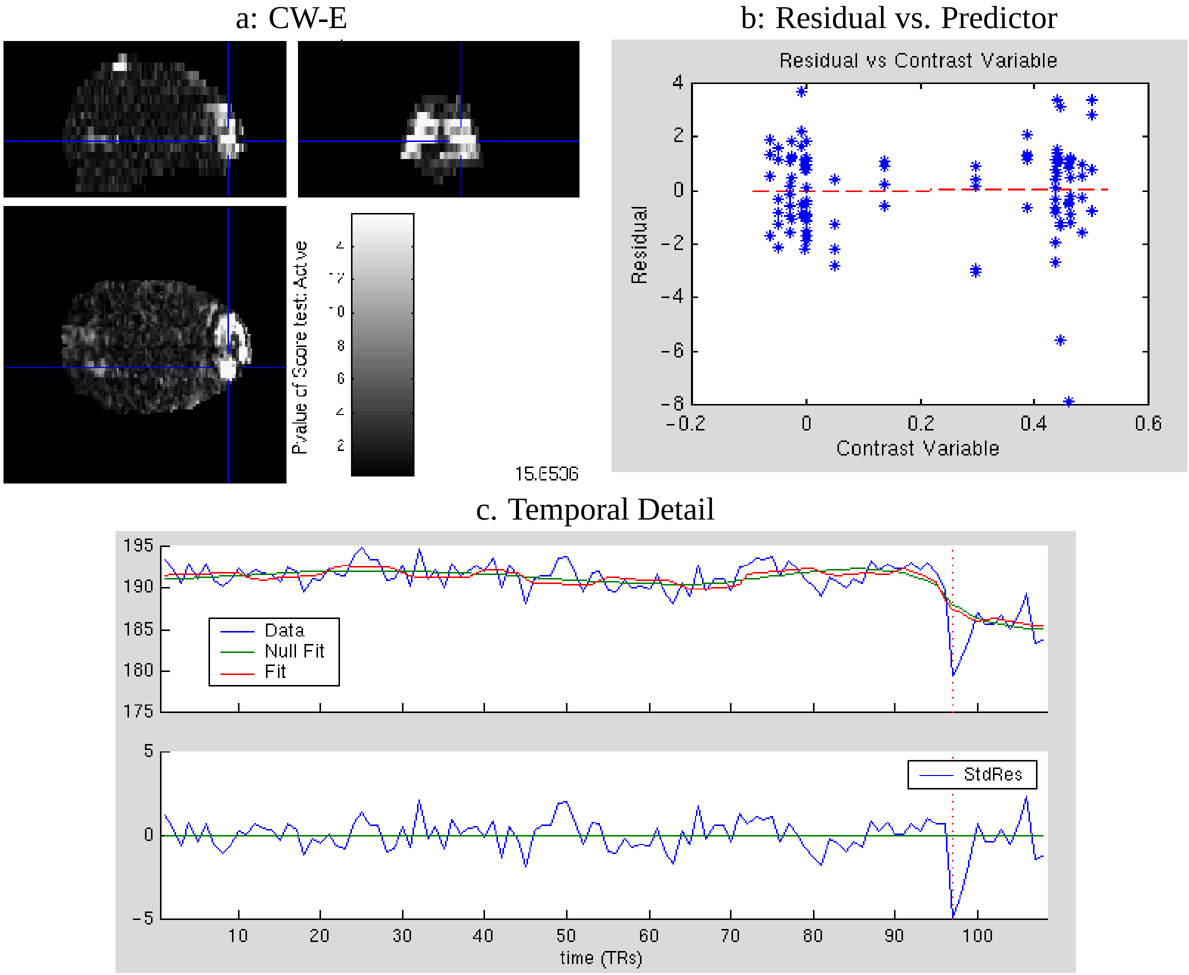


a: DW

b: Lag-1 Residual Plot

c. Temporal Detail

Figure 7: Temporal detail in region with significant DW and CP statistics. The time series plots show the problem of out-of-phase signal, which results in a periodic pattern in the residual time series and autocorrelation in the residual plot. a: Durbin-Watson statistic; b: Lag-1 residual plot; c: Temporal detail of the fit and residuals.

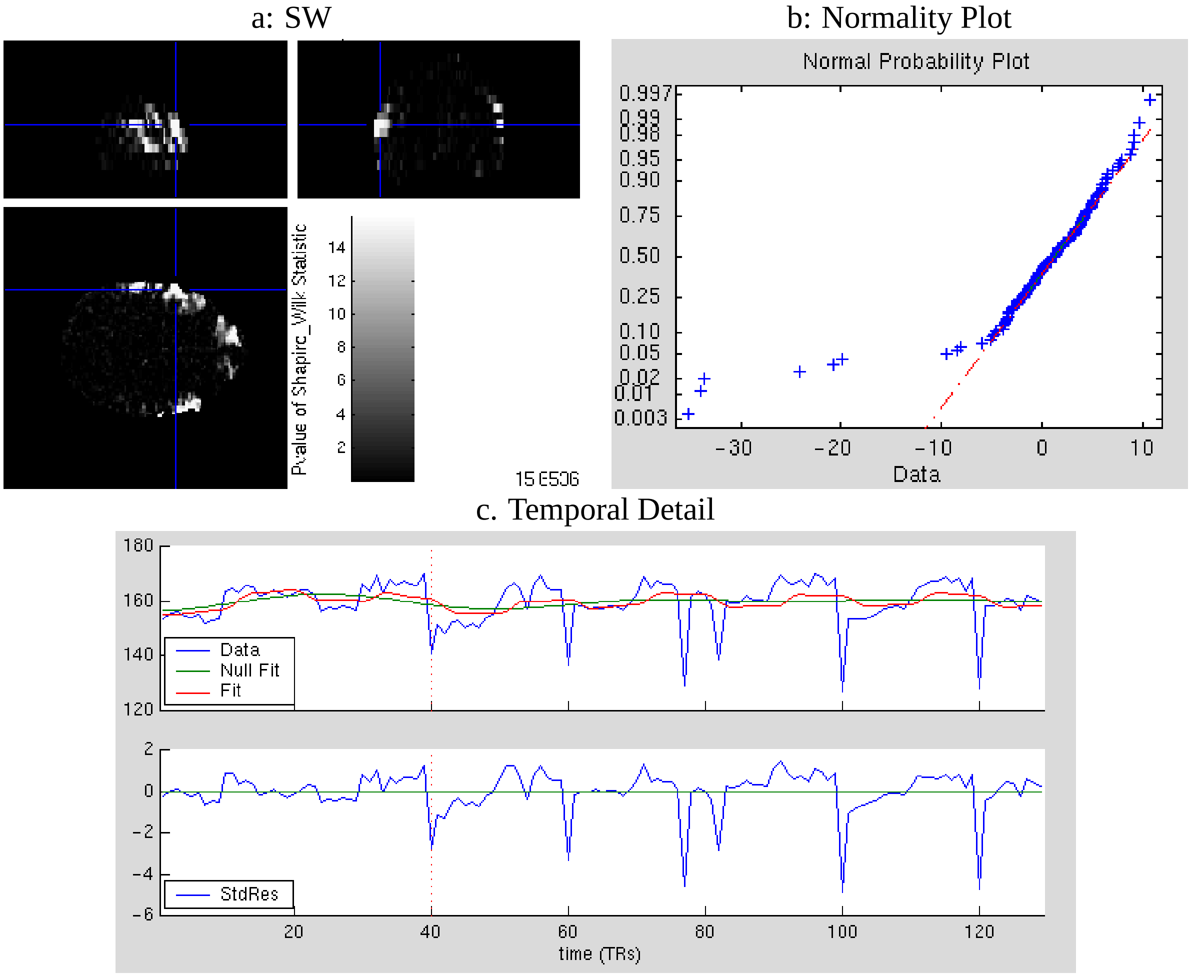


a: CW-E

b: Residual vs. Predictor

c. Temporal Detail

Figure 8: Result of homoscedasticity test. The hyperintensity in the frontal region is a result of the outlier scans (see Fig. 6-d), where images also exhibit signal loss in the same region. a: Image of CW-E; b. Diagnostic plot of residual vs. predictor; c. Temporal detail of residuals in voxel at the frontal region.

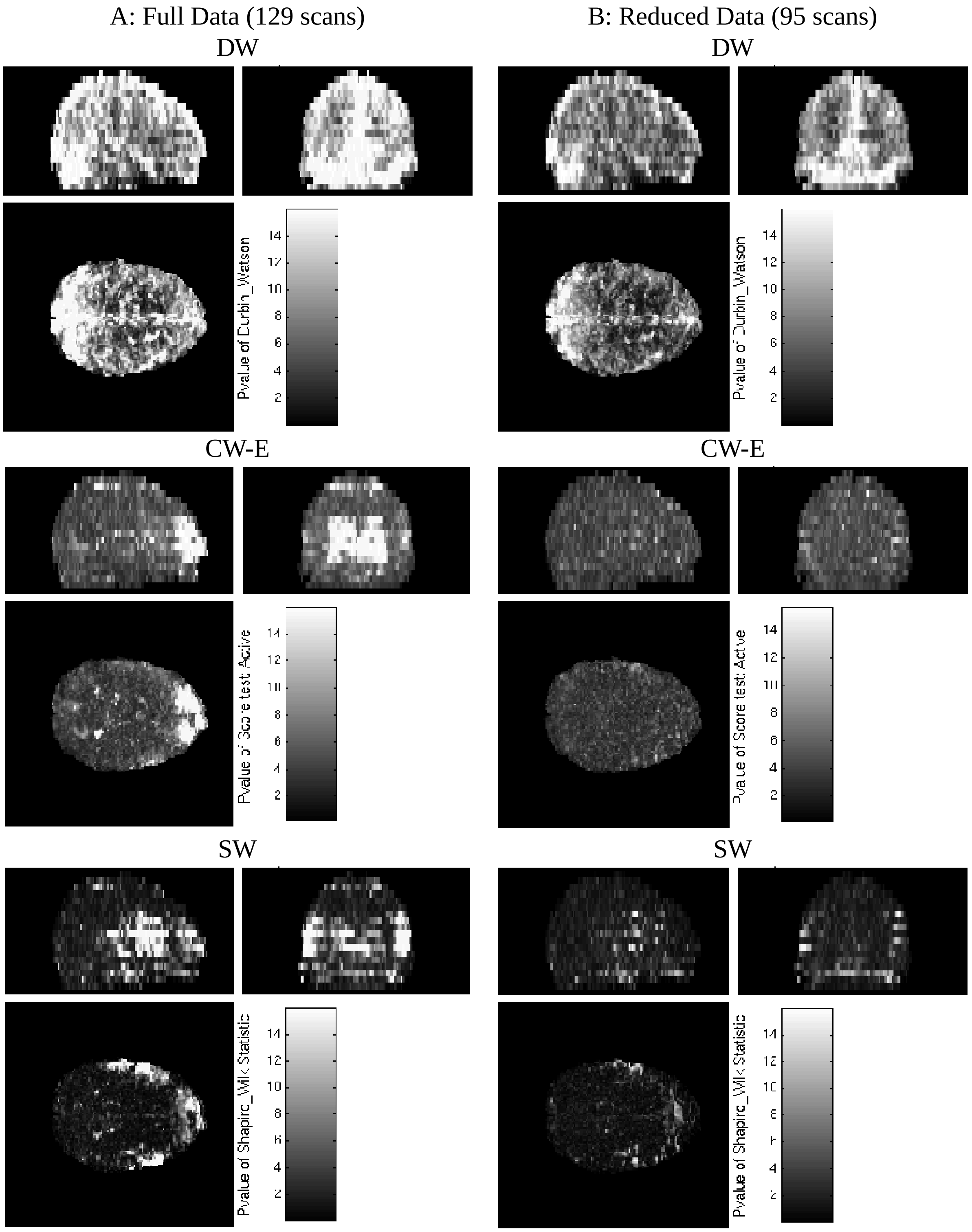


a: SW

b: Normality Plot

c. Temporal Detail

Figure 9: Result of normality test. The significant SW statistic is mainly in bilateral frontal regions, and caused in part by experimentally related outliers. a: Image of SW statistic; b: Normality plot of residuals; c: Temporal detail at the voxel with significant normality test.


A: Full Data (129 scans)

B: Reduced Data (95 scans)

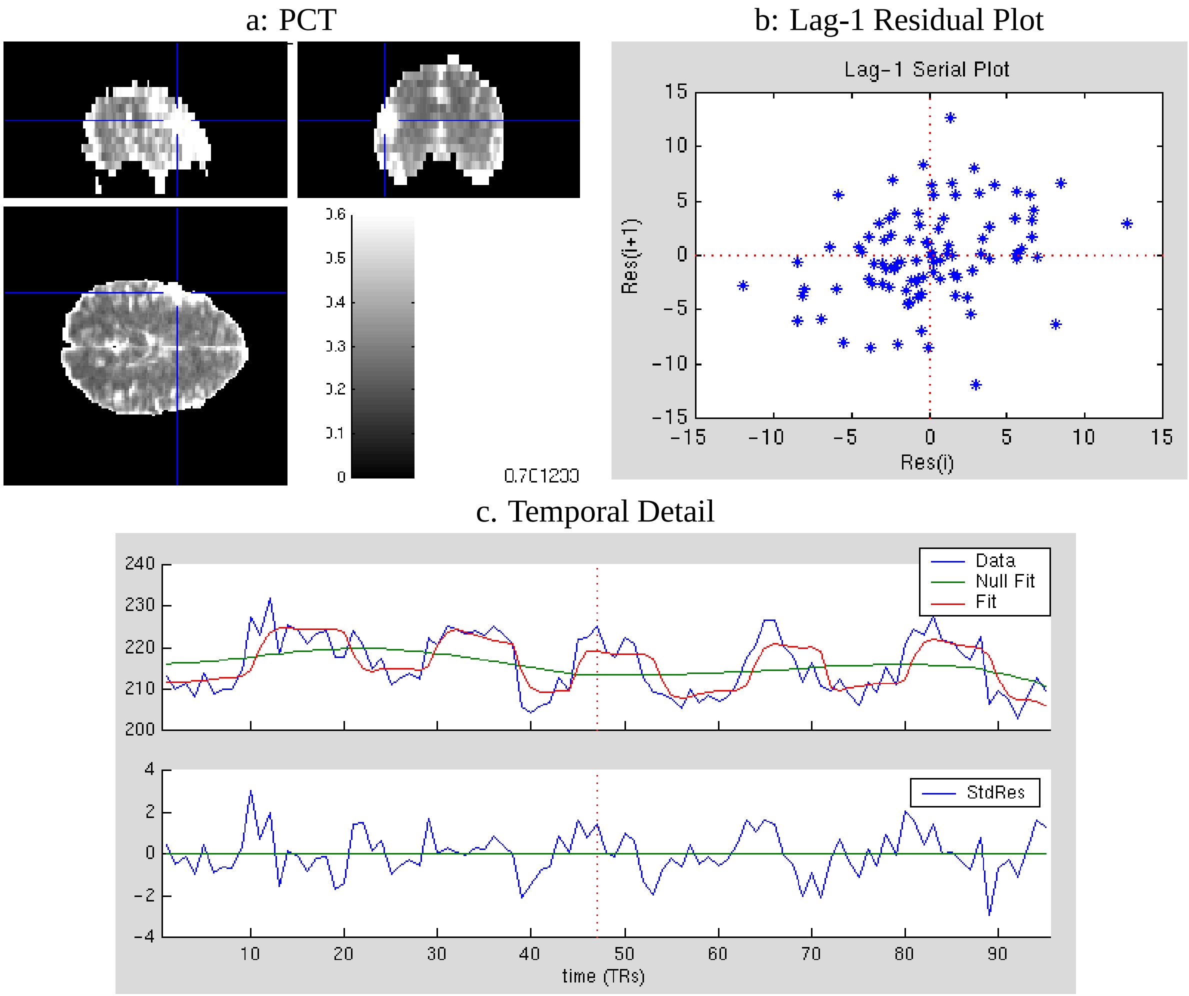
DW

CW-E

SW

Pvalue of Durbin_Watson
2
4
6
8
10
12
14
Pvalue of Score test: Active
Pvalue of Shapiro_Wilk Statistic

Figure 10: Comparisons of maximum intensity projection of DW, CW-E, and SW in full data analysis and in reduced data analysis. The images from reduced data are much uniform than those from the full data.


a: PCT

b: Lag-1 Residual Plot

c. Temporal Detail

Figure 11: Exploration of PCT image. The intensity of PCT is much higher in the bilateral motor and left dorsal lateral frontal areas. These regions have experimental signal but are shown 1/8 out of phase with predictor. a: Image of PCT; b. Lag-1 residual plot; c. Temporal detail of residuals in voxel with high PCT.

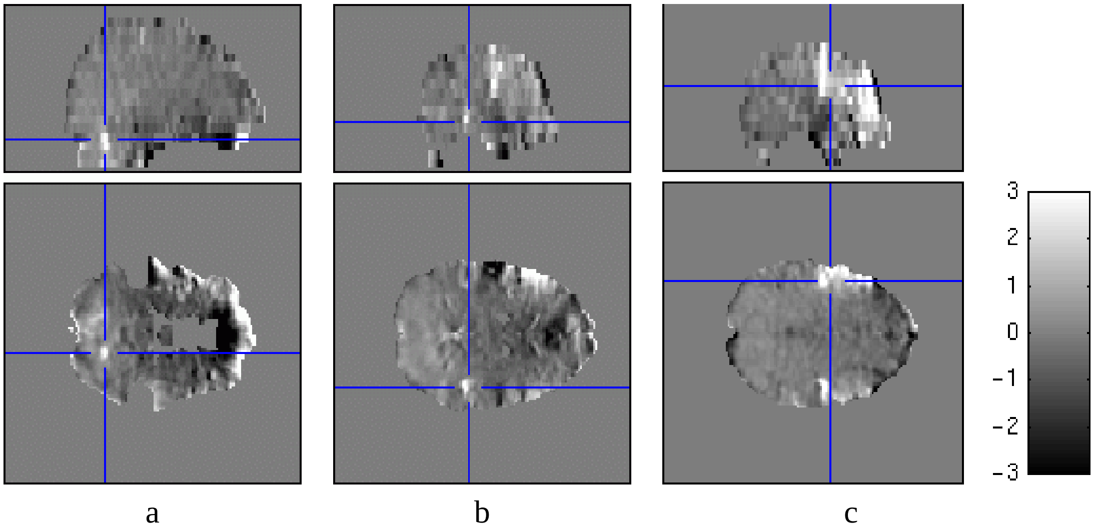


Figure 12: Percent change images of activation. a: Bilateral, though mostly right cerebellum, and artifactual orbitofrontal changes; b: Bilateral, though mostly right auditory cortex; c: Bilateral motor cortices and left dorsolateral prefrontal cortices.

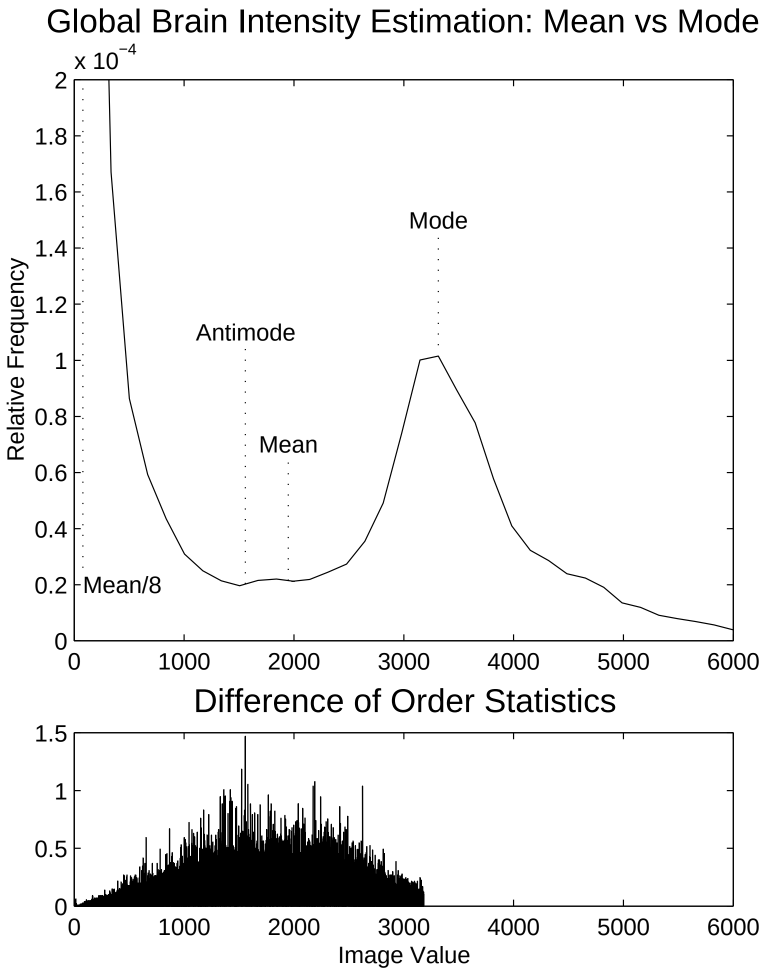


Figure 13: Our mode estimate compared to SPM99's mean estimate. Top plot shows the distribution of intensities of the baseline image of the dataset in the paper. The mean is determined by SPM99, where all voxels greater than 1/8 of mean image intensity are considered. The bottom figure shows how the 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.

List of Tables

Table 1: Spatial Summaries

Statistic	Assesses	Null Dist$^{\text{n}}$
Contrasts	Signal	t^*
Standard Deviation/PCT**	Artifacts	
Durbin-Watson	Cor(ε_i,$\varepsilon_{i+1}) = 0$	Beta
Cumulative Periodogram	Var$(\varepsilon) = \sigma^2 I$	Uniform
Score Test	Var$(\varepsilon_i) = \sigma^2$	χ^2
Shapiro-Wilk	Normality	Normal***
Outlier Count	Artifacts	Binomial

* After standardization

** Percent Change Threshold

*** After transformation

Table 2: Spatiotemporal Diagnosis Strategies

	Step	Action
1	Explore Temporal Summaries	Check for systemic problems Check for transient problems Check for relationships between summaries
2	Explore Spatial Summaries	Check for violations of assumptions Explore noise, nuisance variability Explore experimental signal
3	Explore Temporal Detail	Check for unmodeled signals Note possible problem scans Check specificity of significant diagnostic statistics
4	Explore Spatial Detail	Check temporal 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 & artifactual variation

Table 3: 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