Residual Analysis for Detecting Mis-modeling in fMRI

(Running title: Detecting Mis-modeling in fMRI)

J. M. Loh[1], M. A. Lindquist[2] and T. D. Wager[3]

[1]Dept of Statistics, Columbia U, New York; Email:meng@stat.columbia.edu, Tel:212-851-2139, Fax:212-851-2164

[2]Dept of Statistics, Columbia U, New York; Email:martin@stat.columbia.edu

[3]Dept of Psychology, Columbia U, New York; Email:tw2131@.columbia.edu

Abstract

The voxel-wise general linear model (GLM) approach has arguably become the dominant way to analyze functional magnetic resonance imaging (fMRI) data. The approach relies on specifying predicted patterns of signal change *a priori*. In this work we develop methods for detecting mis-modeling in the GLM framework, and derive mathematical expressions for quantifying the effects this has on bias and power. We show that even a relatively small amount of mis-modeling can result in severe power loss, and can inflate the false positive rate beyond the nominal value. Due to the massive amount of data, examining the appropriateness of the model is challenging in fMRI. We propose a simple procedure involving the residuals that can be used to identify possible voxels or regions of the brain where model misfit may be present. The key idea is that if there is model misfit − such as a mis-specification of onset, duration, or response shape − residuals will be systematically larger in mis-modeled segments of the time series. By looking at the weighted sum of consecutive residuals using a moving window, our method can pick out regions of a residual time series in which the residuals are consistently larger than expected by chance, while ignoring spurious large residuals that are expected based on the noise distribution. It may also be used more generally for identifying artifacts in fMRI time courses. We investigate the effectiveness of this method using a simulation study, and by applying it to an fMRI dataset. We develop a method and accompanying software for creating whole-brain maps showing power loss and bias due to mis-modeling. Such maps could be a valuable tool in assessing violations of statistical assumptions and informing about differences in the shape and timing of the hemodynamic response function (HRF) across the brain.

Key words: fMRI, GLM, model diagnosis, model misfit, power, residual analysis

1 Introduction

The voxel-wise general linear model (GLM) approach (Friston, Penny, Phillips, Kiebel, Hinton and Ashburner, 2002; Worsley and Friston, 1995) has arguably become the dominant way to analyze functional magnetic resonance imaging (fMRI) data. It is particularly well-suited for testing how much of the variability in an fMRI time series can be explained by a set of *a priori* specified regressors. While appealing because of its simplicity and its efficiency in analyzing massive data sets, the GLM approach is not without problems. Model mis-specification can occur for a number of reasons: incorrect design of the onset or width of the underlying neuronal activity, or incorrect specification of the function describing the hemodynamic response function (HRF) that gives rise to the observed signal. Such effects can be counterintuitive and difficult to prevent. For example, mis-modeling the response to one condition could, depending on the specifics of the design, result in a spurious difference in activation between two other conditions. Incorrect specification can lead to significant loss in power, and even minor mis-modeling can have severe effects on the analysis. Figure 1 shows an example of substantial loss in power due to seemingly minor mis-specification of the onset time and width (either neural or vascular).

Due to the massive amount of data involved in a standard fMRI experiment, examining the appropriateness of the model is challenging, and standard graphical approaches for assessing statistical assumptions (Neter, Kutner, Wasserman and Nachtsheim, 1996) are not viable options. In most fMRI experiments, the data consists of time series data sampled over M (e.g. 200-1500) time points at $N_x \times N_y \times N_z$ (e.g. $64 \times 64 \times 30$) locations, or voxels, in the brain. Each time series consists of samples of blood-oxygen-level dependent (BOLD) or cerebral perfusion responses to the sequence of stimuli presented in the experiment. In analysis with the massively univariate voxel-wise GLM, one performs a separate regression analysis on the time series at every voxel in the brain. Using residuals to check model fit is common in many statistical applications. However, for applications in fMRI, examining the residuals that are the result of $N_x \times N_y \times N_z$ regressions, each with M data points, is challenging. Recently, suggestions have been made regarding how best to perform diagnostics in the neuroimaging setting (Luo and Nichols, 2003). That work used standard diagnostic tools for linear models (e.g. Durbin-Watson, Shapiro-Wilks) to develop new methods of perform-

ing model diagnosis in the massive univariate setting. It also provided tools for summarizing the results using a series of dynamic graphical tools. In this work, we focus our attention on using the residuals to detect model mis-specification. The methods we develop here provide spatio-temporal summaries of model misfit and artifactual outliers, as well as tools for improving the design of the model.

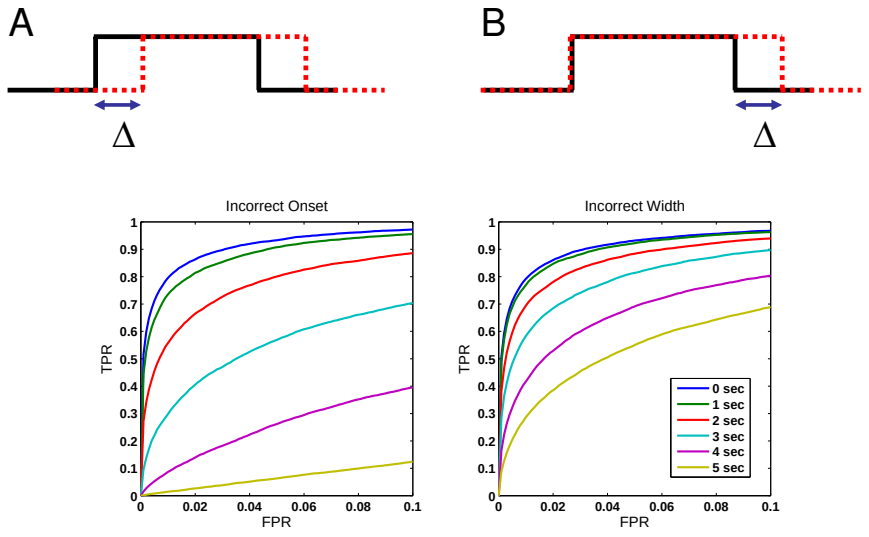

Figure 1: (A-B) Results of a simulation study. The solid and dashed lines in (A) and (B) show the modeled and true activation respectively. The difference between truth and model (delta) is allowed to vary from 0 to 5 seconds. The true activation paradigm is repeated 4 times and convolved with a canonical HRF. Noise is then added corresponding to a Cohens d of 0.5. The GLM is fit using the modeled activation pattern (solid lines) convolved with the canonical HRF. This procedure is repeated 1000 times for both delayed onset (A) and prolonged width (B). Receiver operating characteristic (ROC) curves in the bottom panels show the false positive rate (FPR) vs. the true positive rate (TPR) across statistical significance thresholds. The curves show a substantial decrease in power as a function of model mis-specification.

Specifically, we develop a procedure for identifying segments of time series data that are outliers. The key idea is that artifacts lead to blocks of large residuals. By looking at the weighted sum of consecutive residuals using a moving window, our method can pick out regions where the residuals are consistently larger while ignoring spurious large residuals. We also extend the method to construct spatial maps of model mis-fit, bias and loss in power across the brain. This will be useful for identifying regions of the brain where the statistical analysis may be inaccurate due to model mis-specification.

Where model misfit is present, the distribution of large residuals over time may be systematically related to the experimental paradigm. We develop methods for analyzing the temporal patterns of residuals in

order to study whether the mis-specification is correlated to the experimental paradigm. This tool should be useful for determining whether the model – and the hemodynamic response model in particular – results in substantial loss in power or other more pernicious problems in some brain areas. Although more sophisticated modeling approaches are available (Ciuciu, Poline, Marrelec, Idier, Pallier and Benali, 2003; Woolrich, Behrens and Smith, 2004; Lindquist and Wager, 2007), the dominant hemodynamic model among neuroscientists, psychologists, and medical researchers is a gamma or double-gamma function of fixed shape (i.e., invariant across the brain). This approach has persisted because of its relative simplicity of implementation and inference and its high power in some tasks and brain regions. However, the evoked response has been shown to vary substantially across the brain (Buckner, 2003; Wager, Hernandez, Jonides and Lindquist, 2007a; Christoff, Prabhakaran, Dorfman, Zhao, Kroger, Holyoak and Gabrieli, 2001), and methods for identifying mis-modeling and quantifying loss in power might lead to more informed model choices. Our methods for calculating power can be used to evaluate the robustness of experimental designs and models to variations in the HRF across the brain. Currently, there are no widely available tools for use by brain scientists.

In Section 2, we describe the GLM approach typically used for statistical analysis of fMRI data and derive expressions for the effects of mis-modeling on bias and power. In Section 3, we develop the residual analysis methods for detecting design mis-specification. In Section 4, we present results of a simulation study showing the effects of various types of mis-modeling on power and inferential validity, and illustrate the use of our methods to detect artifacts and mis-modeling. In Section 5.we apply the procedure to a real dataset.

2 The effects of mis-modeling in the GLM approach

In this section we provide a brief overview of the GLM procedure as applied to fMRI analysis. We also derive a theoretical framework for quantifying the effects of mis-modeling on bias and statistical power.

2.1 The GLM approach towards analyzing fMRI data

In the General Linear Model (GLM) approach, the time series, Y, is modeled as a linear combination of a number of different signal components. These components include the blood oxygenation level dependent

(BOLD) response to psychological events specified *a priori*, drift components and other nuisance parameters, which are summarized in a design matrix $\mathcal{X}$. The data can be written using the following model:

$$Y = \mathcal{X}\beta + \epsilon \quad \text{where } \epsilon \sim N(0, \mathbf{V}\sigma^2). \tag{1}$$

Here $\mathbf{V}$ is typically taken to be the covariance matrix of some autocorrelated process e.g. ARMA(1,1) or AR(p) (Bullmore, Brammer, Williams, Rabe-Hesketh, Janot, David, Mellers, Howard and Sham, 1996; Purdon, Solo, Weissko and Brown, 2001).

Typically one is interested in testing for an effect $c^T\beta$, where c is a so-called contrast vector. The contrast vector could be used to estimate signal magnitudes in response to a single type of event (c = [1]), an average over several effects (c = [1 1 ... 1]) or the difference in magnitude between two types of events (c = [1 -1]). For example, one may be interested in identifying reliable brain responses to brief periods of physical pain (a single, repeated type of event) or comparing the effects of two drugs on brain signals induced by painful stimulation (a difference between two types of events, pain with drug A and pain with drug B). Hypothesis testing is performed in the usual manner by testing individual parameters using a t-test and subsets of parameters using a partial F-test. Since the covariance matrix has to be estimated, a Satterthwaite approximation is used to calculate the effective degrees of freedom for the test statistics.

2.2 Bias and Power-loss due to Mis-modeling

In most analyses the regressor corresponding to the BOLD signal is defined to be the convolution of a boxcar function, corresponding to the experimental design, with an assumed hemodynamic response function (HRF). Often a standard canonical HRF (Friston, Penny, Phillips, Kiebel, Hinton and Ashburner, 2002), invariant across the brain, is used. However, the exact shape of the HRF is known to differ across individuals and brain locations (Aguirre, Zarahn and D'Esposito, 1998; Buckner, 2003; Wager, Hernandez, Jonides and Lindquist, 2007a; Christoff, Prabhakaran, Dorfman, Zhao, Kroger, Holyoak and Gabrieli, 2001), and thus the canonical HRF is most likely the wrong model for many brain regions.

In addition, for certain experimental conditions the exact onset and width of activation is not always known (Lindquist and Wager, 2006). For example, in task-switching experiments that compare trials in which

attention is shifted among targets with trials in which attention is maintained (Wager, Vazquez, Hernandez and Noll, 2005), onsets for neural events related to switch trials are typically modeled as beginning when the switch trial begins. However, electrophysiological evidence indicates that the shift operation may not begin until 400 ms later, depending on the task; thus, onsets are typically mis-specified. Without methods for quantitative analysis, it is unclear how much power and validity are compromised by mis-specification of onsets. Similar issues arise for specification of neural activity duration and the width of the fMRI response to be modeled. It is typical in memory literature to compare the effects of remembered vs. forgotten items and assess differences in the hippocampus and medial temporal lobe (Staresina and Davachi, 2006). The neural response for these items is, as with many other tasks, assumed to be either a nearly instantaneous response to the trial, with hysteresis in the observed signal due to the hemodynamic lag, or a linear response over a short epoch. However, data from our lab (Summerfield, Greene, Wager, Egner and Hirsch, 2006) suggest that medial temporal responses peak much later than predicted responses from either of these models, at ~ 12 s even for a 3-s stimulus presentation. The neural activity may persist for an undetermined time beyond the end of the stimulus presentation, and indeed such persistent activity may play an important role in memory formation. In these examples, tools for diagnosing how poorly the canonical model fit in these regions and how much power could have been gained by using a more flexible model would be very desirable.

Here, we give theoretical results showing the impact that mis-modeling can have on the type I and type II error of the tests performed in the GLM framework. Let Y be the fMRI time course of length M from a single voxel. Suppose further that $\mathrm{E}(Y) = \mathcal{X}\beta$ and $\mathrm{Var}(Y) = \mathbf{V}\sigma^2$, so that (1) would represent the *true* model for the GLM. However, suppose we erroneously use the design matrix $\mathbf{X}$, where $\mathcal{X} = \mathbf{X} + \mathbf{\Gamma}$ and $\mathbf{\Gamma} \neq 0$. Here $\mathbf{\Gamma}$ represents the discrepancy between the correct and modeled design matrix. With an incorrect model the estimates of β may be biased and have an inflated variance. Below we work out the bias of the estimates as well as the distribution of the residuals when we use an incorrect model in the regression.

The estimate of β using the incorrect model $\mathbf{X}$ is $\hat{\beta}_* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}Y$, with expected value

$$\mathrm{E}(\hat{\beta}_*) \quad = \quad (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\beta + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{\Gamma}\beta$$

The first term on the right hand side is β so that the bias is $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{\Gamma}\beta$. Clearly, the bias in

the estimate depends on the actual value of β as well as the amount of mis-modeling. When $\mathbf{\Gamma} = 0$, i.e. there is no mis-modeling, the bias is equal to zero, as expected.

The observed residuals under the incorrect model, given by

$$\epsilon_* = [I - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}]Y \equiv \mathbf{R}Y,$$

where $\mathbf{R}$ is the residual inducing matrix, are normally distributed with mean and variance

$$\mathrm{E}(\epsilon_*) = \mathbf{R}\mathbf{\Gamma}\beta, \quad \mathrm{Var}(\epsilon_*) = \mathbf{R}\mathbf{V}\sigma^2. \tag{2}$$

The estimate of σ^2 is

$$\hat{\sigma}^2_* \quad = \quad \frac{\epsilon_*^T\epsilon_*}{\mathrm{tr}(\mathbf{RV})}.$$

Under the correct model, $\mathbf{\Gamma} \equiv 0$ and $\mathrm{tr}(\mathbf{RV})\hat{\sigma}^2/\sigma^2$ follows a χ^2 distribution with ν degrees of freedom, with $\nu = 2\mathrm{E}(\epsilon_*^T\epsilon_*)/\mathrm{Var}(\epsilon_*^T\epsilon_*)$, determined using the Satterthwaite approximation. When there is mis-modeling, ϵ_* does not have zero mean and $\mathrm{tr}(\mathbf{RV})\hat{\sigma}^2_*/\sigma^2$ has a non-central χ^2 distribution. We can estimate ν and the non-centrality parameter δ of this distribution using the moment matching approach of the Satterthwaite approximation. In particular, we find that

$$\delta = \beta^T\mathbf{\Gamma}^T\mathbf{R}\mathbf{\Gamma}\beta \tag{3}$$

and, if δ is small, ν approximately equal to the Satterthwaite degrees of freedom. Details are left to the reader.

Often fMRI data is pre-whitened prior to analysis. Then the commonly used variance estimate is

$$\hat{\sigma}^2_* \quad = \quad \frac{1}{n-p}\epsilon_*^T\mathbf{V}^{-1}\epsilon_*.$$

Under the correct model, and the added assumption that $\mathbf{V}$ is known, $(n-p)\hat{\sigma}^2/\sigma^2$ is χ^2 with $n-p$ degrees of freedom. However, in most practical situations the covariance matrix is unknown, necessitating the use of the Satterthwaite approximation as outlined above. Note that when there is mis-modeling present, ϵ_* has a non-zero mean and the estimator has a non-central χ^2 distribution with non-centrality parameter

$$\delta = \beta^T\mathbf{\Gamma}^T\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{\Gamma}\beta. \tag{4}$$

In fMRI analyses, t or F values from the regressions of all the voxels of a brain image are computed and a map of these values, called a Statistical Parametric Map, is constructed and used to identify regions of the brain that are activated by the stimulus used in the experiment. If the model used in the regressions are incorrect, the resulting Statistical Parametric Map may not be optimal in detecting regions of activation.

In the simplest case, when testing the amplitude of a single regressor and mis-modeling only occurs in that particular regressor, under the null hypothesis of no activation, $\beta = 0$, the mean of ϵ_* is zero. Thus the type I errors of the t tests are correct even when an incorrect model is used, i.e. in voxels where $\beta = 0$, the proportion that are incorrectly classified as being activations is controlled by the specified type I error. However, power calculations require considering the amount of mis-modeling.

When there is mis-modeling in multiple regressors, however, false positive rates may not be adequately controlled at the significance level α. For example, consider a test of the amplitude of a single regressor when mis-modeling has occurred on another component of the model. In this case $\mathrm{E}\epsilon_* \neq 0$ under H_0 and the mis-modeling will therefore have effects on both the type I and type II error. The same problem occurs when testing the difference in amplitude between two regressors if only one of them is mis-modeled. Again, the mean of ϵ_* will be non-zero even under the null hypothesis of $c^T\beta = 0$. In either of these situations the t values computed have a distribution that is the ratio of a normal random variable with non-zero mean and a non-central χ^2 random variable, the doubly non-central t-distribution (Bulgren, 1971). Similarly the doubly non-central F distribution (Tang, 1938; Weibull, 1953) is of interest here when conducting F-tests.

2.3 The effect of mis-modeling on the power - a numerical study

Here, we present briefly findings of a numerical study on the loss in power when there is mis-modeling, showing the need to check for mis-modeling. We fix β to be 1 and consider specific forms of the correct and applied design matrices, $\mathcal{X}$ and X. We take V to be the identity matrix, so that we are assuming white noise with variance equal to 1. We then compare the power in the two cases when $\mathcal{X}$ and X are used as the fitted model. This is done by computing the parameter values of the noncentral and doubly noncentral t distributions, given $\beta, \mathcal{X}$ and X, and then simulating 10000 random variates. Then, for each critical value C from 1 to 3, we computed the proportion of simulated variates that are greater than C.

The specific models used for this study are shown in Figure 4, and are described in Section 4. The important point in this section is Figure 2, which essentially shows the amount of power loss due to model mis-specification. The actual amount of power loss depends on the type and degree of misfit. This ranged from 20% to as much as 80% at 0.013 significance level with the scenarios we considered. The resulting statistical parametric maps will thus not be optimal in identifying regions of activation. Our results clearly show that it is important to perform model checking and to identify if model misfit is present.

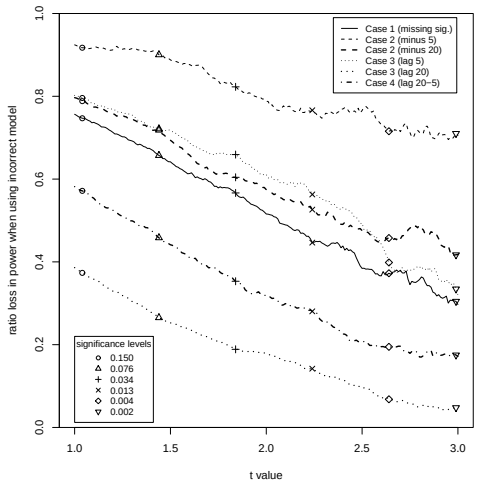

Figure 2: This figure shows the results of a numerical study on the effect of mis-modeling on the power of rejecting the null hypothesis of $\beta = 0$. With assumed values of the effect β, covariance mattrix V and correct and incorrect models $\mathcal{X}$ and X, the expressions in Section 2.2 were used to obtain values of the parameters of the test distributions. The plot shows the ratio of the power, as a function of the critical t value.

3 Methods

It is clear from the previous section that correctly modeling the data in the GLM framework is crucial. Even minor modeling errors can severely impact the efficiency and validity of the ensuing statistical analysis. Here, we propose a simple procedure to identify possible voxels, or regions of the brain, where model misfit may be present. The key idea is that if there is model misfit − such as a mis-specification of onset, duration, or response shape − residuals will be systematically larger in mis-modeled segments of the time series. By looking at the weighted sum of consecutive residuals using a moving window, our method can pick out regions

where the residuals are consistently larger than expected by chance, while ignoring spurious large residuals due to the noise distribution.

Suppose $r_i, i = 1, \ldots, M$ are the de-noised residuals obtained from a GLM analysis of the time series from one voxel. Let us define a window with bandwidth w, where preferably $2w << M$. Define $Y_w(t)$ by

$$Y_w(t) \quad = \quad \sum_{i=-w}^{w} K(i) r_{t+i}, \tag{5}$$

where K is a kernel such that $\sum_{j=-w}^{w} K(|j|)^2 = 1$. This last condition ensures that $Y_w(\cdot)$ has the same variance as the residuals. Under the null hypothesis that the model is correct, $\mathrm{E}(r_j) = 0$, so for any t, we have $\mathrm{E}(Y_w(t)) = 0$. For any fixed window bandwidth w and location t, $Y_w(t)$ is a statistic for testing whether the mean of residuals in the window is 0. Thus the statistic

$$S_w = \max_t Y_w(t)$$

measures the strongest evidence against the null hypothesis. The value of t that yields this maximum indicates the most likely location of mis-modeling within the time course.

If the kernel K is constant over the window, then Y_w is related to the likelihood ratio test statistic. We call this kernel the uniform kernel. Using this kernel may yield the highest power in detecting model misfit. The significance of the identified mis-modeling can be obtained via Monte Carlo simulation. Specifically, the same statistic is computed for, say, 999 simulated sets of residuals, yielding reference values $S^*_{w,j}, j = 1, \ldots, 999$. The rank S_w relative to the reference values $S^*_{w,j}$ provides an estimate of the true p value, with small p values suggesting model misfit. Note that because these p values are based on the distribution of the maximum statistic they are intrinsically controlled for family-wise error. On a Pentium 4 3.8 GHz computer, computing the reference values took 30 seconds. Note that if the choice of kernel K and bandwidth w have been made prior to the experiment, the reference values can be computed beforehand. Then, during the actual analysis, only the comparison of S_w relative to the reference values needs to be done.

Since the Monte Carlo method for determining significance may be too computationally intensive in certain situations, it could be desirable to use alternative methods that can quickly provide an estimate of the p-value. One way to do this is to use a Gaussian kernel in (5). Then the resulting Y_w's will form a Gaussian random process and the desired p-value can be estimated using results that have been derived for

the maxima of Gaussian processes. In fMRI the expected Euler characteristic has been employed with great success to identify threshold values (Worsley, Marrett, Neelin and Evans, 1992; Worsley, Marrett, Neelin, Vandal, Friston, and Evans, 1996b) in Gaussian random fields, to correct for multiple testing across a large number of voxels. Here we focus on correction for searches over time. Using a result in Worsley, Marrett, Neelin, Vandal, Friston, and Evans (1996b), the estimated p-value is given by

$$P(S_w \geq t) \approx \rho_0(t) + \tau\rho_1(t),$$

where $\tau = M/w$ and $\rho_i(t)$ are Euler characteristic densities in the ith dimension depending on the threshold t. Expressions for $\rho_i(t)$ can be found in Worsley, Marrett, Neelin, Vandal, Friston, and Evans (1996b).

Various other approximations for the maxima of Gaussian processes and Gaussian random fields are also available. These include Hotelling's volume-of-tube formula (Hotelling, 1939), the Poisson clumping heuristic (Aldous, 1989) and Sidak and Slepian inequalities (Slepian, 1962; Sidak, 1967, 1968). We mention these methods, but do not pursue them further here, as Adler (2000) suggests that using the Euler characteristic yields as good an approximation as any of the other methods. In summary, using the Gaussian kernel in (5) together with the expected Euler characteristic to estimate significance can provide a substantial increase in speed, at the possible expense of some power, compared with the Monte Carlo approach.

It should be noted that in the neuroimaging setting, the procedure of estimating the p-value of S_w will typically be repeated over a large number of different voxels. In this situation the need to correct the p-values for searches over space, as well as over time, will arise. This can easily be done using equivalent results regarding the expected Euler characteristic for a 4D Gaussian random field (3 spatial directions and 1 temporal). Here we assume that M is the maximum value of S_w across all voxels in the search volume V. The p-value that controls the family-wise error rate over all voxels is then given by

$$P(M \geq t) \approx \sum_{d=0}^{3} R_d(V)(\tau\rho_{d+1}(t) + \rho_d(t)),$$

where $R_d(V)$ represents the resel (resolution elements) count which depends on certain d-dimensional features of the search volume. Again, $\rho_i(t)$ represent the Euler characteristic densities described above.

Regardless of the kernel used, clearly, the power of the method for fixed w depends on both the type and amount of mis-specification. For example, if the (incorrect) fitted model differs in signal width from the

correct model, using a value of w that matches the difference in width will yield higher power than some other value of w. However, it is often not clear what the amount of mis-specification is. A variation of the kernel weighted residual sum approach is to use a range of bandwidths and compute the statistic

$$S = \max_{w \in W} S_w,$$

where W is a set of consecutive integer values. The p-value of this statistic can be obtained by Monte Carlo or by using the expected Euler characteristic if the Gaussian kernel is used. When a uniform kernel is used, the procedure is equivalent to using a scan statistic (Naus, 1965; Kulldorff, 1997). With the Gaussian kernel, the method for estimating the p-value, given above, needs to account for search over different bandwidths. Results for the Euler characteristic densities required in this case can be found in Worsley, Marrett, Neelin and Evans (1996a). In Section 4 we compare the varying and fixed bandwidth approaches.

Once model mis-specification has been identified, there are two additional issues that are of interest. The first is whether the mis-modeling is repeating systematically with the experimental stimulus. This would indicate either a systematic error in the timing of the neuronal stimulus or an error in the assumed shape of the HRF. The other issue is the amount of bias and power loss caused by the mis-specification. Generating bias and power loss maps would be useful for identifying regions of the brain where the usual GLM analysis may not be accurate. These are described in the next two sections.

3.1 Detecting systematic mis-modeling

When evidence of significant mis-modeling is detected in a certain voxel, the next step is to determine if the mis-modeling repeats systematically for each repetition of the experimental stimulus. If this is the case we expect clusters of residuals to be unusually large in a way that correlates with the experimental paradigm.

A simple approach towards detecting the systematic reoccurrence of clusters of large residuals is to fit a linear model using the residual analysis results and a finite impulse response (FIR) basis set. This model contains one free parameter for every time-point following stimulation in every cognitive event type modeled (Glover, 1999; Goutte, Nielsen and Hansen, 2000; Ollinger, Shulman and Corbetta, 2001). This allows us to estimate the shape of the stimulus dependent mis-modeling. As the FIR basis set makes minimal assumptions

about the shape of the mis-modeling, this provides a computationally simple, but effective indication of how the model error correlates with the experimental paradigm. The model can then be modified by changing the shape of the HRF, or alternatively, by changing the estimates of the onset or the neuronal activity duration.

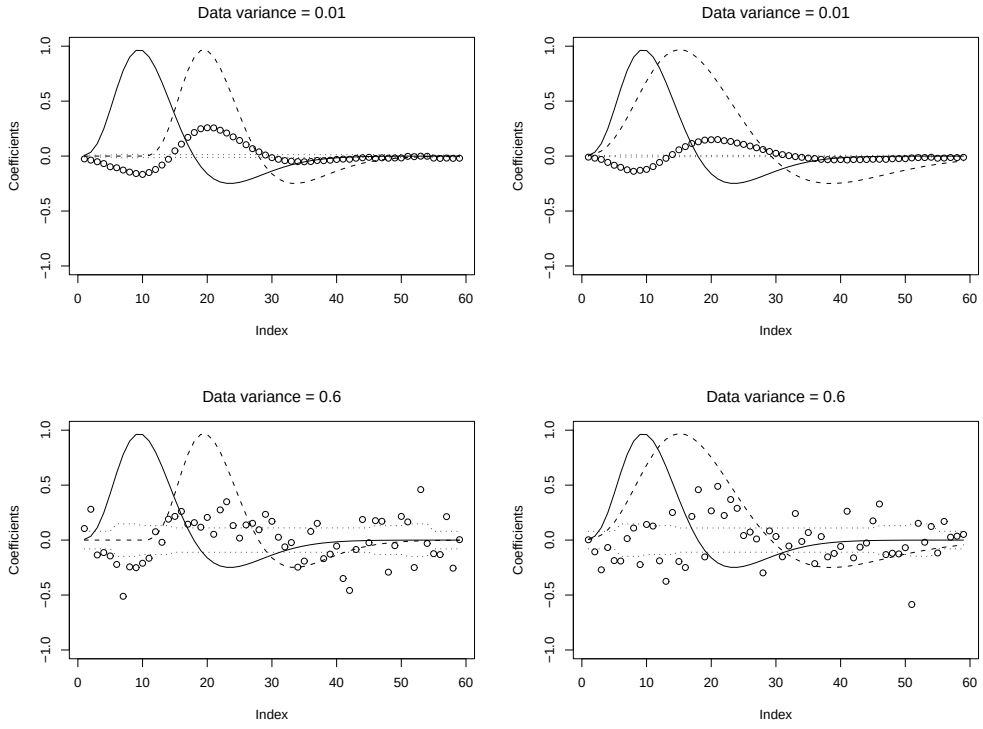

Figure 3: Plots of coefficients obtained from regressing the averaged residuals from a fitted model on the FIR basis set. The dashed and solid curves are respectively the true hemodynamic response function (HRF) and the HRF used in the model. The significant coefficients identify the locations where the two HRFs differ. The dotted lines indicate critical values for the 0.05 significance level, accounting for multiple comparisons with the Bonferonni method.

Here, we use a numerical study to illustrate the effectiveness of using a FIR basis set to identify locations of systematic mis-modeling. We consider the fixed experimental paradigm x of 4 identical stimuli of width 45, separated by intervals of 100. With a particular HRF (SPMs double gamma function) taken to be the truth, we simulated a dataset by convolving this HRF with x, adding a constant and normally distributed white noise. This data is then regressed on a model based on x but convolved with a different HRF. We considered two cases. In the first case, the true HRF has a lag relative to the modeled HRF, while in the second case, the true HRF is wider than the modeled HRF. The true and modeled HRFs are shown in Figure 3. We also considered two values for the variance of the noise, 0.01 and 0.6.

The residuals from the regression are averaged over a window of bandwidth 2 at each time point, ignoring the time points at the ends. The averaged residuals are then regressed against the FIR basis set. Our results are shown in Figure 3. The figure shows plots of the coefficients, indicated by points on the plots. The true and modeled HRFs are represented by the dashed and solid lines respectively. The critical values for the coefficients are indicated by the dotted lines. These critical values are for 0.05 significance level after accounting for multiple comparisons using the Bonferonni correction method.

The top two plots, for the cases when the noise variance is low, clearly show that this method of using the FIR basis set can identify the locations where the modeled and true HRFs differ. The bottom two plots show the effects of increased noise variance. Although the coefficients are now more scattered, especially at the end, the trough and peak showing the locations of the error of the modeled HRF are still evident.

We note that there are methods that are robust against errors in onset or width of the assumed HRF. For example, first and second time derivatives of the assumed HRF can be included as regressors (Liao, Worsley, Poline, Duncan and Evans, 2002; Friman, Borga, Lundberg and Knutsson, 2003; Calhoun, Stevens, Pearlson and Kiehl, 2004; Worsley and Taylor, 2006). These methods are useful tools for statistical analysis. The procedure described here deal with the issues from a different angle and thus are complementary to these other methods. Specifically, instead of accommodating possible errors in the HRFs, our method is useful for identifying the kind of error that is present. Our hope is that this will allow scientists to improve on the HRF that is used and provide guidance in determining whether a more flexible model is appropriate.

3.2 Bias and Power-loss maps

Once significant mis-modeling is detected in a particular voxel, it is of great interest to determine the amount of bias and power loss that can be directly attributed to this mis-specification. In this section we discuss methods for detecting regions of the brain where there is an enhanced risk of bias and power loss. These results will help us evaluate the validity and accuracy of the statistical maps that we obtain through the GLM analysis. To obtain these results, we tie together the theoretical results derived in Section 2 with the empirical results obtained through application of the residual analysis method described in Section 3.

In Section 2 we found that the bias due to mis-modeling is $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{\Gamma}\beta$. Hence, to estimate

the bias we first need estimates of $\mathbf{\Gamma}$ and β. Rather than estimate these parameters individually, a joint estimate of $\mathbf{\Gamma}\beta$ is given by (2) and assuming that the mean of the residuals are well estimated by the observed smoothed residuals, i.e.

$$\mathrm{E}(\epsilon_w) = \mathbf{Y}_w,$$

where $\mathbf{Y}_w = (Y_w(1), \cdots Y_w(N))$. Putting these results together we obtain

$$\widehat{\mathbf{\Gamma}\beta} = \mathbf{R}^{-1}\mathbf{Y}_w.$$

Using the plug-in-principle we can obtain an estimate of the bias. Furthermore we can use this estimate to calculate the power loss in the voxel due to mis-specification. This is done by calculating the power to detect significant activation both in the situation when mis-modeling is present and when it is absent, i.e. when $\mathbf{\Gamma} = 0$. The difference in these measures, indicates the loss in power that is attributable to mis-modeling. Here the non-centrality parameter shown in (3) can be written as $\hat{\delta} = \mathbf{Y}_w^T\mathbf{R}^{-1}\mathbf{Y}_w$.

Once the bias and power loss have been calculated for each voxel, they can be summarized in a map and presented together with the traditional statistical parametric maps included in the output of an experiment. If one is solely interested in determining which regions are affected by power loss, rather than obtaining exact measures of the power-loss, one can construct an image of the non-centrality parameter derived in (3). If this parameter is equal to 0 then there will be no power loss due to mis-modeling. However, if this parameter is significantly greater than zero the difference in power can potentially be significant. As the loss in power will be proportional to the size of the parameter, one can simply create a map of the non-centrality parameter across the brain in order to provide a way to identify hot spots where there appears to be a loss in power due to mis-modeling. We construct bias and power loss maps in our data example in Section 5 .

4 Simulation study of the residual analysis procedure

We performed an extensive simulation study to explore the performance of the residual analysis method described in Section 3. We considered five cases of model mis-specification. For cases 1 to 4, the correct model was based on a time series x_0 of length $M = 225$ with two signals to represent the true response. The incorrect model for case 1 (x_1) contains a missing signal. In case 2, the incorrect model has signals of

incorrect width, specifically, 5, 10, 15 and 20 units narrower, while for case 3 the signals have time lags of 5, 10, 15 and 20 (x_{2a} to x_{2d}, and x_{3a} to x_{3d} respectively). In case 4, the time lags are different for the two signals (models x_{4a}, x_{4b} and x_{4c} with lags 10 and 5, 20 and 5 and 20 and 10 respectively). Lastly, in model 5, the correct model has one signal with greater height (x_5) while the incorrect model is x_0, with two signals of the same (smaller) height. This is an example of unmodeled parameter modulation. These models are shown in Figure 4. We also ran an additional simulation with the true and fitted models both equal to x_0.

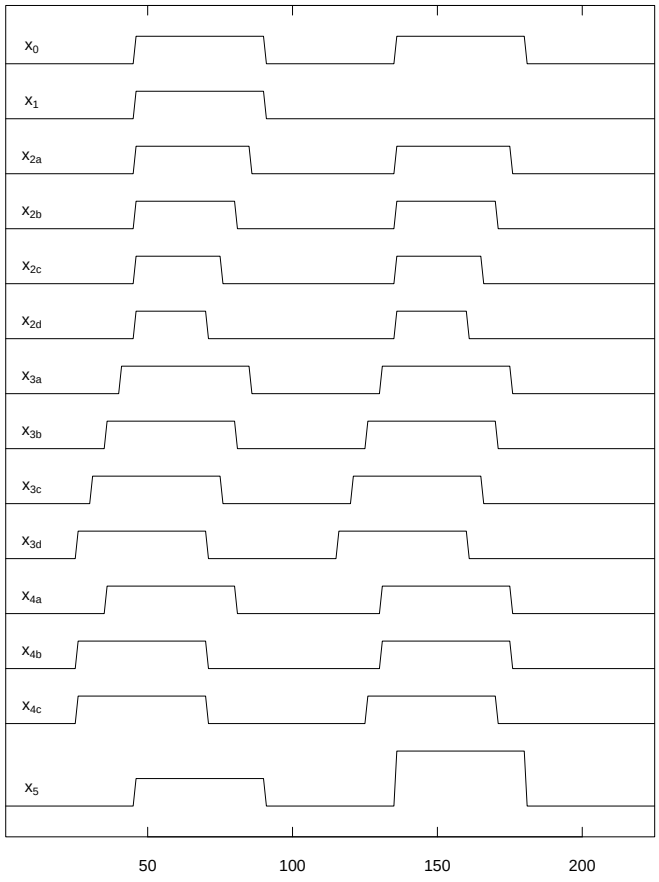

Figure 4: These models, convolved with the HRF function, are used in the simulation study of the performance of the residual analysis method of identifying model misfit. They are also used in the numerical study of Section 2.3, examining the effect on power of model mis-specification.

Each experiment was run 1000 times. For each run, the simulated data, y, is obtained by first convolving the correct model with an HRF, and then adding a constant and normally distributed noise of variance σ^2 to

the convolved function. This data y is then regressed against the chosen model, giving residuals $r_1, \ldots, r_{225}$. The residuals are then standardized to have the same variance by dividing residual r_i by $(1-h_{ii})\hat{\sigma}$ where h_{ii} is the i-th diagonal element of the hat matrix of the regression. The statistic S_w is then computed and the significance of the statistic estimated. We considered two kernels, uniform and Gaussian, with bandwidths $w = 2$ and 7. For the Gaussian kernel, we took the bandwidth to be three times the standard deviation. We used two versions of the Gaussian kernel: one (Gauss 1) with variance equal to that of the uniform kernel, so that the bandwidth is $\sqrt{3}w$, and the other with standard deviation equal to $w/3$ (Gauss 2). To study the effect of signal to noise ratio, we also considered a range of values of σ^2, specifically, $\sigma^2 = 0.3, 0.6, 1, 2$ and 4.

For each simulated run, we estimated the p value using Monte Carlo simulation and, with the Gaussian kernels, using the expected Euler characteristic as well. Thus for each experiment (i.e. each chosen model) we have 1000 p values corresponding to the 1000 runs. We examined the average of these p values, as well as the proportion of these p values that are 0.05 or less.

4.1 Simulation Results

The results are summarized in Table 1, which shows, for each fitted (incorrect) model and value of σ^2, the median p value (over 1000 runs) of the test. The columns labeled "MC" and "Euler" correspond to p values found using Monte Carlo and the expected Euler characteristic respectively.

We find that if the correct model is used the median p value (not shown) is close to 0.5. With an incorrect model, we find that the p value is very small when σ^2 is small, showing that model misfit is detected with strong significance. As expected, the median p values generally increase as σ^2 is increased. The strength of the detection varies with the kind and amount of model misfit. When the model misfit is due to lags or differences in signal width, the greater the lag or width difference, the greater the significance of the misfit detection. We also performed a set of simulations without convolving x_0 with the HRF. We do not present these results, but note that in this case, the median p values are consistently smaller than the values in Table 1 when there is model misfit. The differences from the values in Table 1 depend on the amount of misfit, with greater differences when the mis-specification is small. This is expected, since the HRF smears out the signals, so the mis-specification becomes less detectable when the difference between the true and

$w = 2$		Uniform	Gauss 1		Gauss 2		$w = 7$		Uniform	Gauss 1		Gauss 2	
Case	σ^2	MC	MC	Euler	MC	Euler	Case	σ^2	MC	MC	Euler	MC	Euler
1	0.6	0.06	0.13	0.14	0.26	0.46	1	0.6	0.004	0.004	0.004	0.02	0.02
	2	0.31	0.34	0.47	0.43	0.78		2	0.16	0.17	0.17	0.22	0.25
2a	0.6	0.39	0.40	0.58	0.49	0.91	2a	0.6	0.38	0.38	0.42	0.33	0.40
	2	0.50	0.49	0.76	0.53	1.03		2	0.58	0.53	0.71	0.53	0.73
2d	0.6	0.10	0.16	0.18	0.28	0.50	2d	0.6	0.01	0.01	0.01	0.04	0.04
	2	0.33	0.36	0.50	0.44	0.80		2	0.17	0.20	0.20	0.26	0.30
3a	0.6	0.29	0.33	0.46	0.44	0.80	3a	0.6	0.41	0.23	0.23	0.24	0.27
	2	0.46	0.46	0.68	0.52	0.97		2	0.51	0.49	0.61	0.47	0.60
3d	0.6	0.11	0.18	0.20	0.32	0.57	3d	0.6	0.08	0.01	0.02	0.05	0.06
	2	0.28	0.32	0.42	0.44	0.79		2	0.15	0.15	0.15	0.21	0.23
4a	0.6	0.16	0.24	0.28	0.35	0.62	4a	0.6	0.07	0.06	0.06	0.09	0.09
	2	0.39	0.41	0.60	0.47	0.87		2	0.34	0.37	0.40	0.37	0.45
4c	0.6	0.12	0.18	0.20	0.32	0.57	4c	0.6	0.02	0.02	0.02	0.05	0.06
	2	0.31	0.34	0.47	0.44	0.79		2	0.18	0.20	0.19	0.24	0.28
5	0.6	0.42	0.43	0.63	0.49	0.92	5	0.6	0.38	0.40	0.45	0.39	0.49
	2	0.48	0.49	0.76	0.53	1.00		2	0.57	0.54	0.73	0.53	0.72

Table 1: Table showing the median p value over 1000 independent runs of the model misfit test, for cases 1 to 5 of model misfit (see Figure 4), and for $\sigma^2 = 0.6$ and 2. The p values are obtained with an approximation using the expected Euler characteristic (Euler) and with Monte Carlo (MC) simulation. The p values obtained when an incorrect model is used should be low, indicating detection of mis-modeling.

incorrect fitted model is small. We also applied the procedure using varying bandwidths. The median p values obtained by Monte Carlo simulation are shown in Table 2.

Figures 5 and 6 show, for a representative sample of the simulation experiments, plots of the proportion of Monte Carlo p values that are 0.05 or less when an incorrect model is used, as a function of σ^2. This is essentially the power of the test under various alternatives. The figures correspond to the uniform and Gaussian kernels with the same variance, respectively.

The clearest feature in each of the plots is the drop in power as σ^2 is increased. This is, of course, expected. The power of the test depends on the type and amount of mis-modeling. In cases where the model mis-specification is small (such as x_{2a} and x_{3a}, a mis-specification in width by 5 and a lag difference of 5 respectively - the middle two plots on the left column of Figure 5), the power is low. This is not surprising, since if the difference between the correct and incorrect model is small, this difference is easily masked by the HRF and the errors. For example, for model x_{3a}, the power for the scan statistic method is about 0.4 when $\sigma^2 = 0.3$, and decreases quickly as the error variance is increased. When the width mis-specification or lag is larger, the power of the test is much larger and the drop in power, as the error variance is increased, is more gradual. This is also the case with x_1, the model mis-specification of a missing signal.

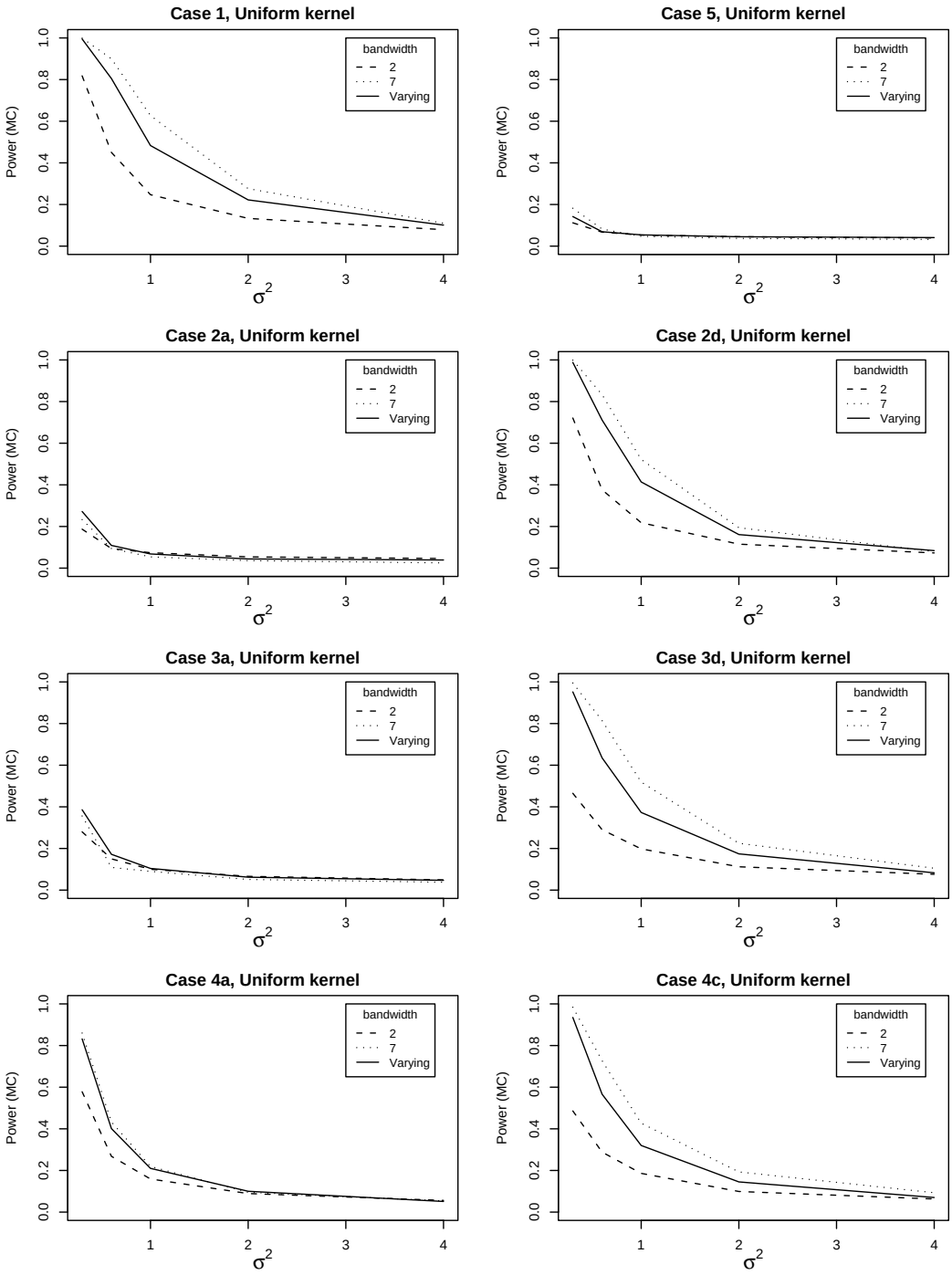

Figure 5: Plots showing the power of the residual analysis procedure with uniform kernel to detect model misfit for various incorrect fitted models (see Figure 4). The power is found from the proportion of p values, numerically obtained by Monte Carlo (MC) simulation, that are 0.05 or less. For varying bandwidths, we used $w = 2$ to 7.

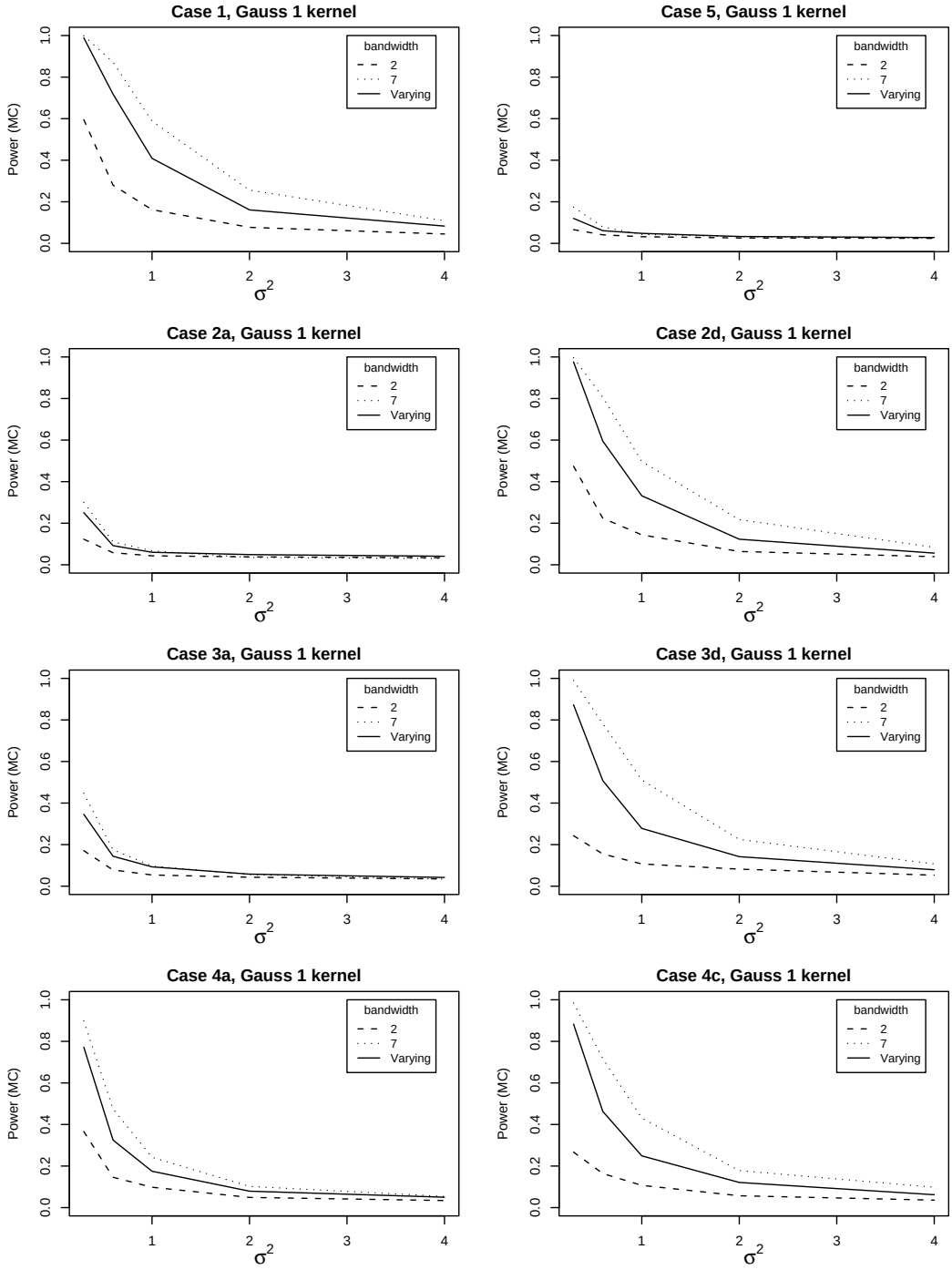

Figure 6: Plots showing the power of the residual analysis procedure with the Gaussian kernel of the same variance as the corresponding uniform kernel of bandwidth w. to detect model misfit for various incorrect fitted models (see Figure 4). The power is found from the proportion of p values, numerically obtained by Monte Carlo (MC) simulation, that are 0.05 or less. For varying bandwidths, we used Gaussian kernels with variances equal to that of the uniform kernel with bandwidths $w = 2$ to 7.

$w = 2$ to 7		Uniform	Gauss 1	Gauss 2
Case	σ^2	MC	MC	MC
1	0.6	0.01	0.01	0.07
	2	0.22	0.25	0.36
2a	0.6	0.35	0.35	0.42
	2	0.52	0.50	0.52
2d	0.6	0.01	0.03	0.11
	2	0.24	0.29	0.37
3a	0.6	0.26	0.29	0.35
	2	0.46	0.46	0.50
3d	0.6	0.03	0.05	0.14
	2	0.21	0.26	0.34
4a	0.6	0.07	0.13	0.21
	2	0.36	0.39	0.45
4c	0.6	0.04	0.06	0.15
	2	0.24	0.29	0.35
5	0.6	0.39	0.41	0.46
	2	0.51	0.49	0.52

Table 2: Table showing the median p value over 1000 independent runs of the model misfit test, for cases 1 to 5 of model misfit (see Figure 4), and for $\sigma^2 = 0.6$ and 2, where the residual analysis procedure with varying bandwidths is used. The p values are obtained by Monte Carlo (MC) simulation.

The effect of window bandwidth on power depends on the particular mis-specification. For x_1, the missing signal is of width 25. Thus the power is much greater for $w = 7$ than with $w = 2$ (top left plot of Figures 5 and 6). This is also the case with signal width or lag mispecification of 20 (middle plots on the right column of Figures 5 and 6). When, for example, the lag mis-specification is 5, the difference between using $w = 2$ and $w = 7$ is smaller. The power of the varying bandwidth procedure is in most cases between the power achieved by the fixed windows. By searching for model misfit (via consecutive large residuals) over a range of bandwidths , power is retained over varied forms of model misfit. On the other hand, by doing so, it results in less power than when the optimal bandwidth is used.

We also examined the proportion of p values that are 0.05 or less when the correct model is used (not shown), i.e. the type I error of the tests. We find in all cases that this proportion is close to 0.05.

Although curves generally have the same shape, regardless of whether the uniform or Gaussian kernel is used, it appears that there is slightly less power with the Gaussian kernel. This was true for both the Gaussian kernels we considered. This is not conclusive, however, as there are many more bandwidths for the Gaussian kernel that we did not consider in our simulations.

4.2 Power of the test

We studied the power of our residual analysis procedure to detect model misfit. Specifically, for the model mis-specifications considered, we constructed ROC curves showing the power of the test as a function of type I error. Figures 7 and 8 show our results with the uniform and Gaussian kernel respectively, for models $x_{2a}, \ldots, x_{2d}$ and $x_{3a}, \ldots, x_{3d}$ which shows the main features that are present in the other models. We applied the procedure using varying bandwidths and windows with fixed bandwidth $w = 2$ and 7.

We find that the ROC curves for the procedure using bandwidth of 2 were the lowest in all cases. At the smallest mis-specifications, the ROC curves are roughly equal. However, as the amount of the mis-specification is increased, the procedure with $w = 7$ yielded increasingly higher ROC curves. In general, the ROC curves for the varying bandwidths procedure lie between those for $w = 2$ and $w = 7$.

Thus, we find that, if the degree or extent of mis-specification is not known, the varying bandwidth procedure serves as a reasonable robust procedure that can identify a wide variety of mis-specifications. However, if there is additional information about the type of mis-specification, additional power (and speed) may be achieved by using the procedure with the appropriate fixed bandwidth.

5 Experimental Design

The experimental data collected at the University of Michigan consisted of a visual paradigm conducted on a single subject, in accordance with Institutional Review Board guidelines. It consisted of a blocked alternation of 11 s of full-field contrast-reversing checkerboards (16 Hz) with 30 s of open-eye fixation baseline. Blocks of stimulation were presented on an in-scanner LCD screen (IFIS, Psychology Software Tools). Spiral-out gradient echo images (Noll, Cohen, Meyer and Schneider, 1995) were collected on a GE 3T fMRI scanner. Seven oblique slices were collected through visual and motor cortex, 3.12 x 3.12 x 5 mm voxels, TR = 0.5 s, TE = 25 ms, flip angle = 90, FOV = 20 cm, 410 images. Data from all images were corrected for slice-acquisition timing differences using 4-point sinc interpolation (Oppenheim, Schafer and Buck, 1999) and corrected for head movement using 6-parameter affine registration (Woods, Grafton, Holmes, Cherry and Mazziotta, 1998) prior to analysis.

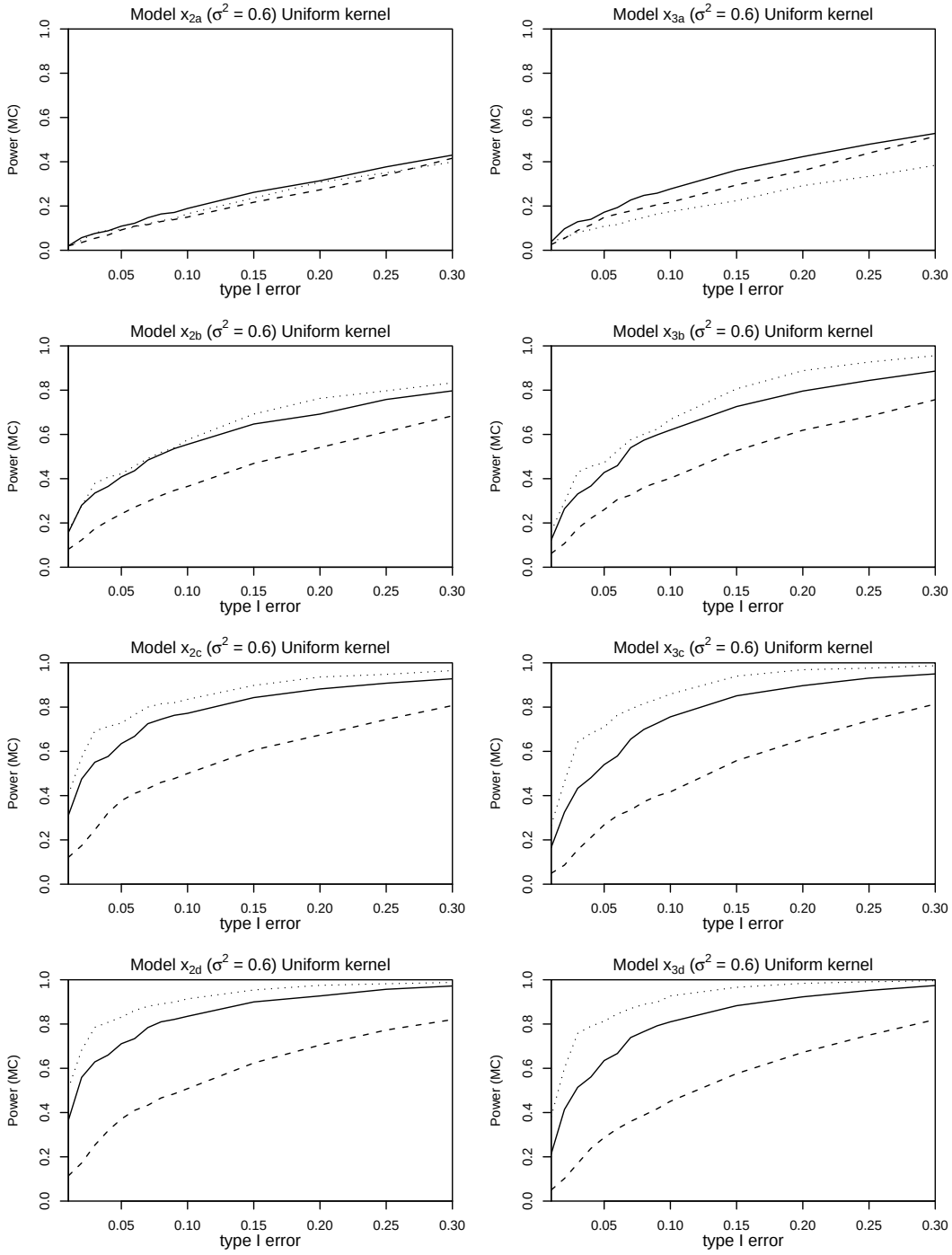

Figure 7: ROC plots showing the power of the residual analysis procedure with uniform kernel to detect model misfit as a function of the type I error for mis-specification of signal width (left column) and lag (right column). The amount of mis-specification is respectively 5, 10, 15 and 20 units going from the top plot to the bottom plot. The different line types correspond to the procedure applied with varying bandwidths (solid line), and fixed bandwidths with $w = 2$ (dashed line) and $w = 7$ (dotted line).

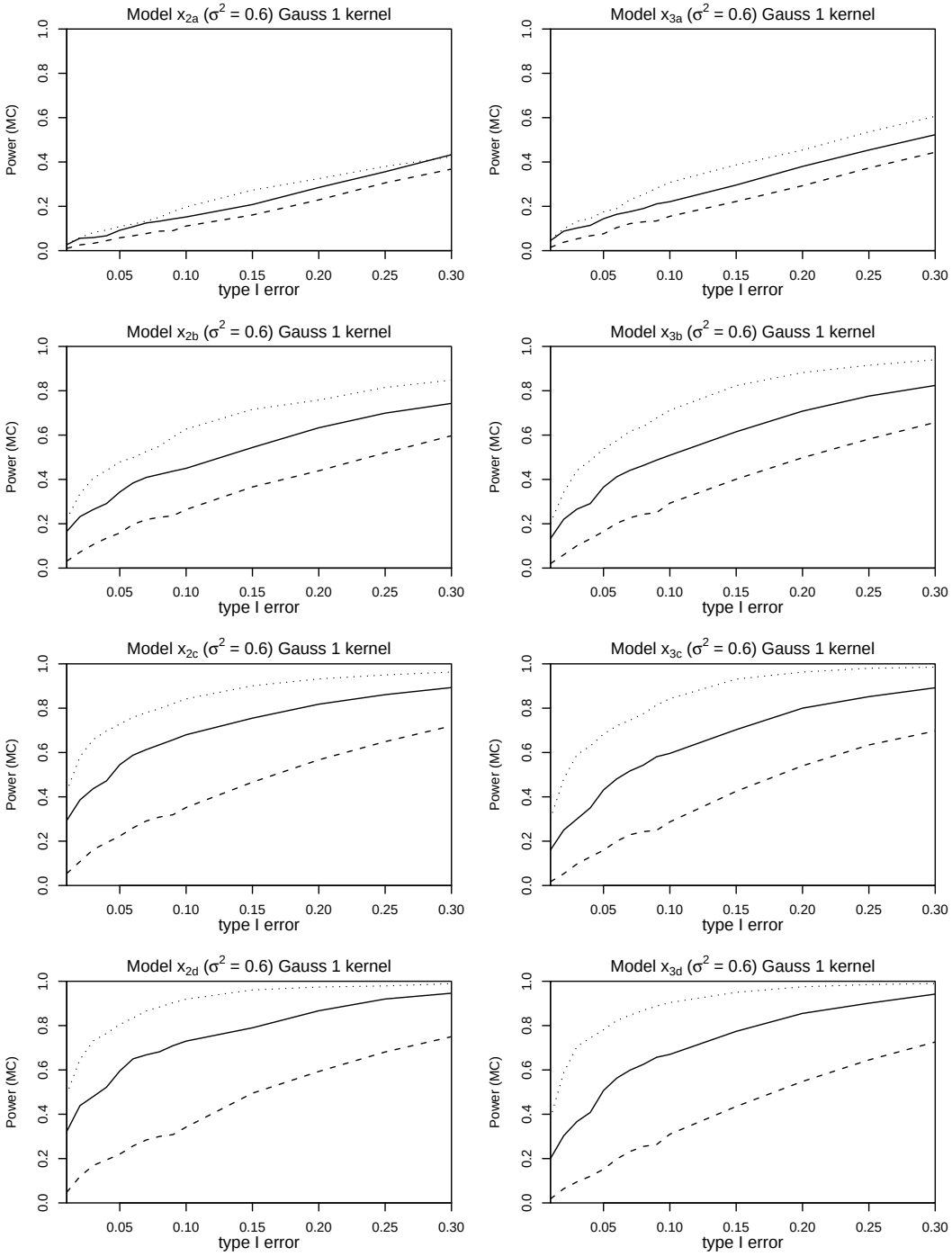

Figure 8: ROC plots showing the power of the residual analysis procedure with Gaussian kernel to detect model misfit as a function of the type I error for mis-specification of signal width (left column) and lag (right column). The Gaussian kernel used has the same variance as that of the uniform kernel with bandwidth w. The amount of mis-specification is respectively 5, 10, 15 and 20 units going from the top plot to the bottom plot. The different line types correspond to the procedure applied with varying bandwidths (solid line), and fixed bandwidths with $w = 2$ (dashed line) and $w = 7$ (dotted line).

The data was analyzed using a standard GLM procedure, where the design matrix consisted of three regressors associated with a quadratic drift term and one regressor corresponding to the expected BOLD response. This regressor was obtained by convolving a boxcar function corresponding to the experimental design with SPMs canonical HRF. We performed this analysis five times, purposefully mis-modeling the activation onset of the boxcar design so the difference in modeled and true onset time took the values -2, -1, 0, 1 and 2 seconds. (Figure 9 top). For each case, our residual analysis approach using the uniform kernel with bandwidth of 5 was applied to the residuals to detect evidence of significant mis-modeling. Maps of the estimated bias and power-loss due to mis-modeling were computed for each case.

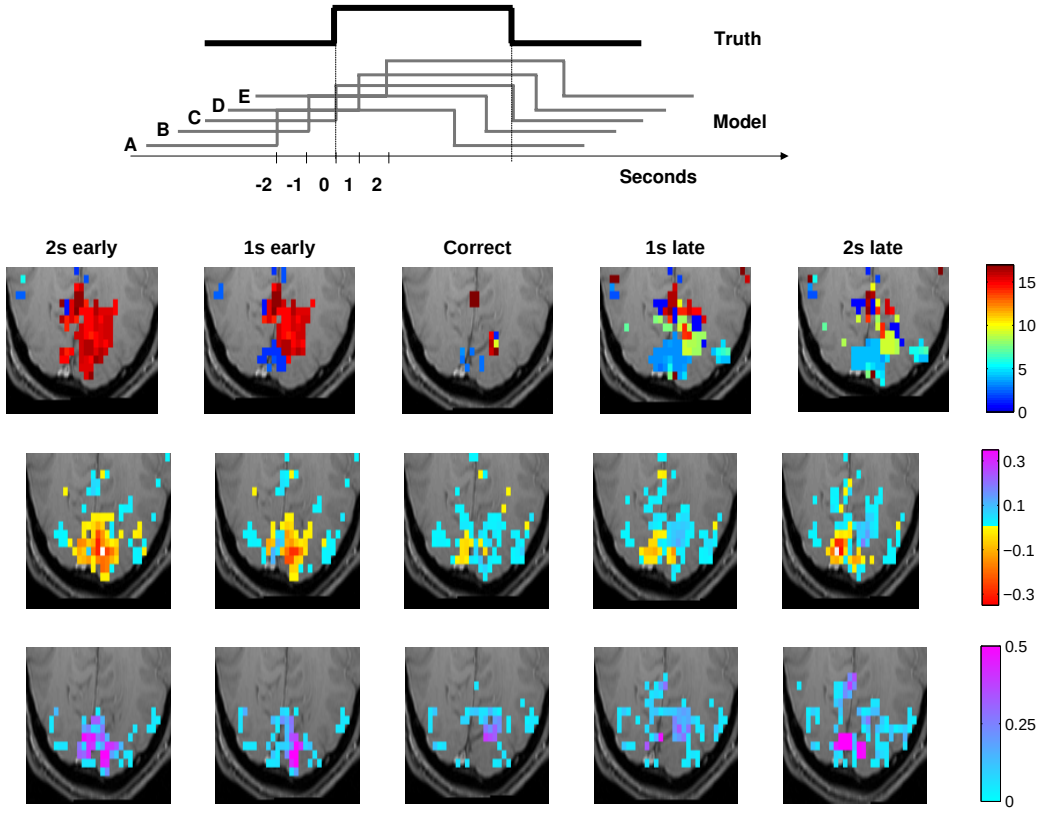

Figure 9: (Top) Experimental data analyzed using a standard GLM where the activation onset is purposefully mis-modeled so the difference in modeled and true onset time took values -2, -1, 0, 1 and 2 seconds. (First row of images A-E) The location in time of the statistic S for voxels where significant mis-modeling is detected. Blue (red) indicates voxels that show a cluster of mis-modeled points toward the beginning (end) of the visual stimuli. Results are consistent with the erroneous model formulation. Bias (second row) and power-loss (third row) maps show an increase in bias and decrease in power as the amount of mis-modeling increases.

5.1 Results

The location in time of the statistic S is shown for voxels with significant mis-modeling in Figures 9A-E (first row). Blue (red) indicates voxels that show a cluster of mis-modeled points toward the beginning (end) of the visual stimuli. The results are consistent with the erroneous model in each case. Clearly, the greater the amount of mis-modeling, the more voxels show significant deviations from iid Normal residuals, demonstrating that the residual analysis approach can detect model mis-specification. In addition, when the correct model is used, there is a minimal amount of significant deviations.

In addition, in each case we include maps of the bias and power-loss due to mis-modeling (Figures 9 A-E, bottom two rows). The maps tell a similar story and indicate the regions of the brain where mis-modeling has the greatest impact. They allow us to judge the validity of the statistical parametric maps that are typically used to summarize the results of a GLM analysis, identifying regions that should be further studied.

6 Discussion and Conclusion

In this paper, we derive expressions for bias and power loss due to systematic mis-specification of a GLM model. We introduce a procedure for detecting deviations in fMRI time series residuals. Using these two ideas, we can construct whole-brain bias and power loss maps due to systematic mis-modeling. We apply these methods both to simulated and real fMRI data.

A key idea we explicate in our theory and simulations is that mis-modeling (design mis-specification) can result in bias in addition to loss in power. Positive bias inflates the Type I error rate beyond the nominal α level, so that p-values for the test are inaccurate. For example, a statistical parametric map thresholded at $p < .001$ may actually only control the false positive rate at, say, $p < .004$. In recent work (Wager, Lindquist and Kaplan, 2007b), we estimate that roughly 10-20% of reported activations in neuroimaging literature are false-positives, posing a serious problem for the accumulation of knowledge in the field. The procedures for generating bias maps over the brain developed here can be used to detect regions of the brain in which inflation of the false positive rate is likely.

Lack of sensitivity is also an important issue, since lack of activation across studies in a particular task is

generally taken to imply that the region is not important for the task (though this inference is not strictly valid). For example, inconsistencies activation detection across studies have spurred debates about the effects of mental imagery in V1 (Kosslyn and Thompson, 2003) and the role of the amygdala in anxiety disorders (Etkin and Wager, 2007). We find that even relatively minor model mis-specification can result in substantial power loss. In light of our results, it seems important for studies that use a single canonical HRF or a highly constrained basis set to construct maps of bias and power loss, so that regions with low sensitivity or increased false positive rates may be identified.

The methods developed here are useful in other ways as well. fMRI data may contain artifacts from many sources, such as head movement, physiological noise and intermittent gradient failures. Our approach can identify deviations in the residuals due to these and other sources. As regression is extremely sensitive to outliers, appropriate identification and removal of outliers may help to substantially increase power.

7 Acknowledgement

We would like to thank Doug Noll and Luis Hernandez for kindly providing the data used in this paper.

References

Adler, R. J. (2000). On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability* **10**, 1–74.

Aguirre, G. K., Zarahn, E. and D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *NeuroImage* **8**, 360–369.

Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*. Springer, New York.

Buckner, R. L. (2003). The hemodynamic inverse problem: making inferences about neural activity from measured MRI signals. *Proceedings of the National Academy of Sciences U S A* **100**, 2177–2179.

Bulgren, W. G. (1971). On representations of the doubly non-central F distribution. *Journal of the American Statistical Association* **66**, 184–186.

Bullmore, E. T., Brammer, M. J., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A. S., Mellers, J. D. C., Howard, R. and Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine* **35**, 261–277.

Calhoun, V. D., Stevens, M. C., Pearlson, G. D. and Kiehl, K. A. (2004). fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *NeuroImage* **22**, 252–257.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J. and Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage* **14**, 1136–1149.

Ciuciu, P., Poline, J.-B., Marrelec, G., Idier, J., Pallier, C. and Benali, H. (2003). Unsupervised robust non-parametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE Transactions on Medical Imaging* **22**, 1235–1251.

Etkin, A. and Wager, T. D. (2007). Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder and specific phobia. Submitted *American Journal of Psychiatry* .

Friman, O., Borga, M., Lundberg, P. and Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage* **19**, 837–845.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G. and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage* **16**, 465–483.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* **9**, 416–429.

Goutte, C., Nielsen, F. A. and Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Transactions on Medical Imaging* **19**, 1188–1201.

Hotelling, H. (1939). Tubes and spheres in n-spheres and a class of statistical problems. *American Journal of Mathematics* **61**, 440–460.

Kosslyn, S. M. and Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery. *Psychological Bulletin* **129**, 723–746.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]* **26**, 1481–1496.

Liao, C., Worsley, K. J., Poline, J.-B., Duncan, G. H. and Evans, A. C. (2002). Estimating the delay of the response in fMRI data. *NeuroImage* **16**, 593–606.

Lindquist, M. A. and Wager, T. D. (2006). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human Brain Mapping* DOI: 10.1002/hbm.20310.

Lindquist, M. A. and Wager, T. D. (2007). Modeling state-related fMRI activity using change-point theory. *NeuroImage* **35**, 1125–1141.

Luo, W.-L. and Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage* **19**, 1014–1032.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on the line. *Journal of the American Statistical Association* **60**, 532–538.

Neter, J., Kutner, M. H., Wasserman, W. and Nachtsheim, C. J. (1996). *Applied Linear Statistical Models*. McGraw-Hill/Irwin.

Noll, D. C., Cohen, J. D., Meyer, C. H. and Schneider, W. (1995). Spiral K-space MR imaging of cortical activation. *Journal of Magnetic Resonance Imaging* **5**, 49–56.

Ollinger, J. M., Shulman, G. L. and Corbetta, M. (2001). Separating processes within a trial in event-related functional MRI. *NeuroImage* **13**, 210–217.

Oppenheim, A. V., Schafer, R. W. and Buck, J. R. (1999). *Discrete-Time Signal Processing*. 2nd edition, Prentice-Hall.

Purdon, P. L., Solo, V., Weissko, R. M. and Brown, E. (2001). Locally regularized spatiotemporal modeling and model comparison for functional MRI. *NeuroImage* **14**, 912–923.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626–633.

Sidak, Z. (1968). On multivariate normal probabilities of rectangles: Their dependence on correlations. *The Annals of Mathematical Statistics* **39**, 1425–1434.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal* **42**, 463–501.

Staresina, B. P. and Davachi, L. (2006). Differential encoding mechanisms for subsequent associative recognition and free recall. *Journal of Neuroscience* **26**, 9162–9172.

Summerfield, C., Greene, M., Wager, T. D., Egner, T. and Hirsch, J. (2006). Neocortical connectivity during episodic memory formation. *PLoS Biology* **4**.

Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs* **2**, 126–137.

Wager, T. D., Hernandez, L., Jonides, J. and Lindquist, M. A. (2007a). Elements of functional neuroimaging. In *The Handbook of Psychophysiology* (J. Cacioppo, L. Tassinary and G. Berntson, eds.), 3rd edition, Cambridge University Press.

Wager, T. D., Lindquist, M. A. and Kaplan, L. (2007b). Meta-analysis of functional neuroimaging data: Current and future directions. Submitted to *Social Cognitive and Affective Neuroscience* .

Wager, T. D., Vazquez, A., Hernandez, L. and Noll, D. C. (2005). Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage* **25**, 206–218.

Weibull, M. (1953). The distributions of t- and F- statistics and of the correlation and regression coefficients

in stratified samples from normal populations with different means. *Skandinavisk Aktuarie Tidskrift* **36**, 9–106 (Supplement).

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R. and Mazziotta, J. C. (1998). Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography* **22**, 139–152.

Woolrich, M. W., Behrens, T. E. and Smith, S. M. (2004). Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage* **21**, 1748–1761.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fMRI time-series revisited-again. *NeuroImage* **2**, 173–181.

Worsley, K. J., Marrett, S., Neelin, P. and Evans, A. C. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism* **12**, 900–918.

Worsley, K. J., Marrett, S., Neelin, P. and Evans, A. C. (1996a). Searching scale space for activation in PET images. *Human Brain Mapping* **4**, 74–90.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. (1996b). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* **4**, 58–73.

Worsley, K. J. and Taylor, J. E. (2006). Detecting fMRI activation allowing for unknown latency of the hemodynamic response. *NeuroImage* **29**, 649–654.