ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 26 (2005) 99–113

Increased sensitivity in neuroimaging analyses using robust regression

Tor D. Wager,[a,*] Matthew C. Keller,[b] Steven C. Lacey,[b] and John Jonides[b]

[a]*Department of Psychology, Columbia University, 1190 Amsterdam Avenue, New York, NY 10027, USA*
[b]*Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1109, USA*

Received 17 July 2004; revised 10 January 2005; accepted 12 January 2005
Available online 17 March 2005

Robust regression techniques are a class of estimators that are relatively insensitive to the presence of one or more outliers in the data. They are especially well suited to data that require large numbers of statistical tests and may contain outliers due to factors not of experimental interest. Both these issues apply particularly to neuroimaging data analysis. We use simulations to compare several robust techniques against ordinary least squares (OLS) regression, and we apply robust regression to second-level (group "random effects") analyses in three fMRI datasets. Our results show that robust iteratively reweighted least squares (IRLS) at the 2nd level is a computationally efficient technique that both increases statistical power and decreases false positive rates in the presence of outliers. The benefits of IRLS are apparent with small samples (n = 10) and increase with larger sample sizes (n = 40) in the typical range of group neuroimaging experiments. When no true effects are present, IRLS controls false positive rates at an appropriate level. We show that IRLS can have substantial benefits in analysis of group data and in estimating hemodynamic response shapes from time series data. We provide software to implement IRLS in group neuroimaging analyses.

Introduction

Traditional statistical inference relies on three fundamental assumptions: (1) errors are independent from one another, (2) errors are normally distributed (or have another known distributional form), and (3) error variance is constant across levels of the predicted values. Statisticians emphasize the importance of evaluating these assumptions for each analysis tested, as violations of the assumptions can produce both false positive and false negative results and undermine the interpretability of inferential statistics (e.g., P values).

The field of statistics has developed a number of diagnostic tools to check assumptions, many of them graphical (Luo and Nichols, 2003; Neter et al., 1996). However, applications that require testing of a large number of statistical models pose a problem: it is nearly impossible to check assumptions and make individual decisions about how to address potential violations in each case (but see Luo and Nichols, 2003). Neuroimaging data (e.g., PET, fMRI, SPECT) are a prototypical example of this situation, as separate regression models are typically fit for each of 30,000–100,000 voxels in the brain.

In such cases, outliers in the data can create violations of the normality and equality of variance assumptions, and they can have a disproportionately large impact on the statistical solution (see Fig. 1A). This is true particularly with large, artifact-prone datasets such as those typical in neuroimaging experiments (Langenberger and Moser, 1997; Le and Hu, 1996; Ojemann et al., 1997). Outliers are likely to exist in some proportion of the regression analyses (e.g., in some voxels). In most cases, these outliers will cause decreased power (or, equivalently, will lead to higher false negative rates), but in some cases, they will lead to higher FPRs. This unpredictability of the effects of outliers is particularly problematic because it means that a simple correction (e.g., an alpha or P value correction) is not available.

Robust regression techniques are a class of statistical tools designed to provide estimates and inferential statistics that are relatively insensitive to the presence of one or more outliers in the data (Huber, 1981; Hubert et al., 2004; Neter et al., 1996). When outlying values are present in the data, violations of distributional assumptions can lead to reduced power and increased false positive rates. Robust techniques can substantially increase power while maintaining an appropriate false positive rate (Huber, 1981; Neter et al., 1996). Robust techniques are particularly useful when a large number of regressions are tested and assumptions cannot be evaluated for each individual regression, such as with neuroimaging data. The techniques we describe here are explicitly designed to deal with outliers, and may complement other techniques, such as data filtering and incorporation of Bayesian priors, designed to increase robustness to artifacts (Ciuciu et al., 2003; Smith et al., 2002; Woolrich et al., 2004).

Robust regression and neuroimaging

In neuroimaging, analyses are conducted both individually for each subject and in a group analysis across subjects. Within a subject, a common strategy is to fit a multiple regression model to

* Corresponding author.
E-mail address: tor@psych.columbia.edu (T.D. Wager).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.01.011

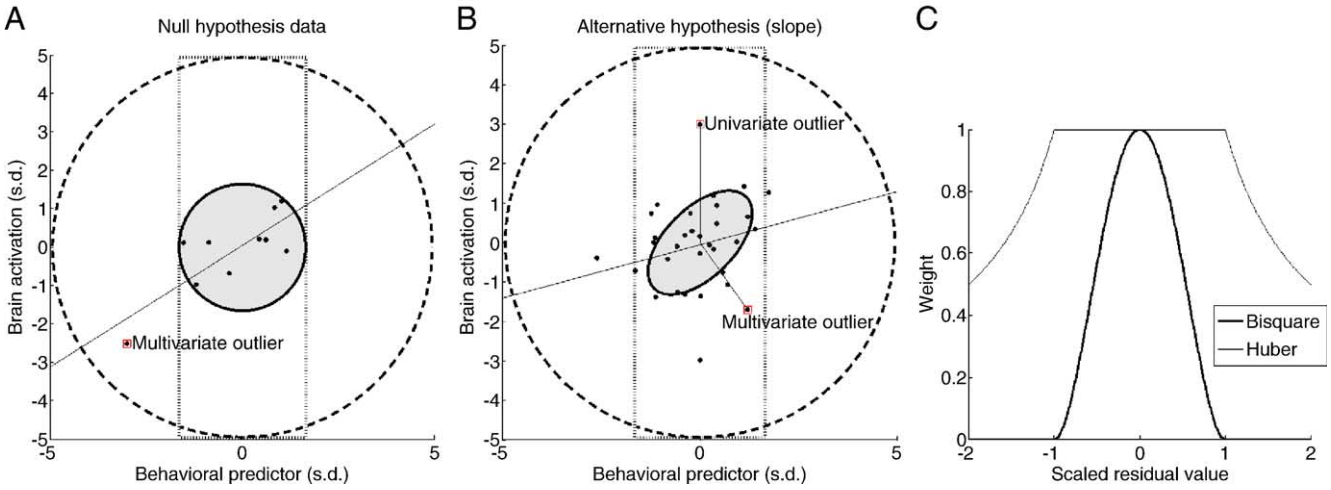

Fig. 1. (A) Example of bivariate null hypothesis (Ho) data with $n = 10$ and 10% ($n = 1$) outliers. The solid circle and dashed circle show the 95% confidence area for the noise and outlier noise distributions for multivariate (x–y) outliers, respectively. In this case, a single outlier creates a significant correlation between x and y. (B) Example of simulated data ($n = 30$) with a true effect of the predictor (alternative hypothesis, Ha, for the predictor) and a zero intercept (Ho for the intercept). The shaded region shows the 95% confidence interval for the sample mean. (C) Two weight functions used in IRLS (see legend). Huber asymptotes at 0 outside the bounds of the figure.

the time series data at each voxel (Worsley and Friston, 1995; Worsley et al., 1997). In this case, outliers (or other violations of the statistical assumptions) in the time series can substantially influence the fit of the model. Robust regression can minimize the impact of these outliers. In a typical group analysis, individual brains are first warped or anatomical regions are delineated so that voxels correspond to the same brain regions in each subject (Ashburner and Friston, 1999; Toga and Thompson, 2002). Regression parameters (or contrasts composed of a linear combination of parameters, e.g., A–B, task-control) are saved for each subject at each voxel or region of interest, and a test is performed on the parameter values, treating individual subjects' parameters as a random effect. Robust regression can be used at this level as well, minimizing the impact of outlying subjects.

Group analyses, also called "second level" or "random effects" analyses in the neuroimaging literature, can be a simple one sample t test, to test whether activation values differ from zero, or a more complex model involving repeated measures or behavioral predictors. Investigators are typically interested in (1) whether certain brain regions are activated by the task (i.e., whether contrast values differ from zero), and (2) whether behavioral scores (e.g., performance, behavioral depression scale scores, etc.) correlate with regional brain activation. These two tests correspond to tests of the intercept and slope of a simple linear regression model at the second level. Our simulations focus on this case as an illustrative example.

There are three principal reasons why robust regression techniques may be particularly important for analyzing neuroimaging data. First, as described below, there are good reasons to suspect that artifactual outliers are common in such data. Second, it is often unfeasible to check assumptions for each individual regression analysis due to the number of separate regression analyses performed (Luo and Nichols, 2003, provide a solution), and thus an efficient robust algorithm that dampens the effects of outliers would be advantageous. Finally, as noted above, robust techniques may increase statistical power (decreasing the false negative rate) and may prevent false positives in the presence of outliers or skew in the data.

Neuroimaging experiments may be more outlier-prone than many other methodologies due to the number and nature of processes that may produce artifacts, which we review briefly below. Because of the large number of comparisons that are performed in a typical "massively univariate" analysis, outliers are extremely likely to occur in some comparisons (i.e., somewhere in the brain). However, multivariate analyses (Buchel et al., 1999; Mckeown et al., 2003) are not immune to outliers. In fact, they are more influenced by outliers than univariate approaches. As the multivariate space becomes more sparsely sampled (e.g., the ratio of variables to samples grows), extreme values at some time-points can have extremely large leverages, and thus extreme influences on the overall solution.

Fig. 1A shows an example of a problematic dataset with $n = 10$ and no true effect. The data (black dots) were drawn from a null-hypothesis (Ho) distribution, shown by the shaded circle, with no correlation between the predictor (x axis) and the data (y axis). Noise from a larger-variance distribution, shown by the dashed circle, was added to one data point, marked with a square. The regression line shows a statistically significant false positive effect, caused primarily by the highly influential outlier point.

The decision to drop or downweight outliers is an important one, and the best answer depends on the nature of the data. A central issue is whether outliers are likely to arise from some process that the researcher might be interested in modeling (e.g., higher order interaction terms) or if they arise from a process that is of little theoretical interest (e.g., data collection artifacts). If from the former, outliers should not be dropped. Rather, the model should be adjusted to account for them. For example, skewed, non-normal distributions can be modeled using maximum likelihood procedures or additional predictors, such as interactions or polynomial terms, can be added to an OLS model. On the other hand, if outliers are likely to arise from processes that the researcher is uninterested in modeling, their influence should be dampened or eliminated.

Acquisition artifacts

Various kinds of acquisition artifacts are present in fMRI BOLD data, some of which are slice or region specific, and others

of which are global. Changes in gradients may produce spikes at particular time points or a range of time points. Local changes in magnetic field inhomogeneity produce artifacts specific both in space and time. The presence of such artifacts can influence an individual subject's regression parameter estimates (betas) dramatically, and thus create outliers in group analyses (i.e., random effects analyses across individual participants).

Motion artifacts

Even small movements of the head may produce large artifacts in fMRI signals. Artifacts are local in time and space. They are greater at the edges of the brain and around fluid space because magnetic field homogeneity is most sensitive to perturbations in these regions and because voxels are shifted in and out of fluid spaces with head motion (Hutton et al., 2002; Ward et al., 2002; Wu et al., 1997). In addition, they induce magnetic susceptibility changes that cannot be captured by realignment algorithms or inclusion of movement parameters in linear statistical models (Wu et al., 1997).

Physiological noise artifacts

Heartbeat and breathing both induce pulsatile motion in the brain, which creates artifacts in the time series directly, by moving brain tissue with respect to the sampling grid, and indirectly, by inducing magnetic susceptibility artifacts (Frank et al., 1993, 2001; Kruger and Glover, 2001). Troublingly, these artifacts may often correlated to some degree with the task design. Although some spurious correlation can be expected by chance, many cognitive and emotional states produce changes in respiration. Task-correlated physiological artifact can create outliers in individual subjects' regression parameter estimates, and the magnitude of these effects varies widely across participants, exacerbating the problem in individual differences analyses at the group level. If there is a systematic physiological noise-induced bias in one individual participant, they are likely to be an outlier in the group of participants, and robust regression could prove beneficial at the group level.

Normalization and anatomical variability artifacts

Group analyses are often performed by warping or normalizing each participant's brain to a reference template, and thereafter assuming that each voxel covers the same anatomical brain tissue and functional brain region for each participant. If this process fails for a particular brain region within even one subject—or functional localization is different for that subject—the parameter estimates for that subject can become outliers in group analysis.

Behavioral outliers

This final category of outliers is very important, because behavioral (X) outliers exert high leverage on the parameter estimates (subject activation contrast scores) throughout the whole brain. This kind of outlier may be caused by error or inaccuracy in behavioral measurement, or because a participant is drawn from a different population from other participants.

The present study

A successful application of robust regression to neuroimaging should demonstrate that (1) the technique is more sensitive in brain regions that are known to show true positive responses, (2) the technique improves the reliability of estimates, and (3) FPRs are reduced in regions known *not* to show true responses.

We address each of these in a simulation comparing several methods of robust techniques with OLS, using parameters and sample sizes similar to those encountered in imaging studies. We then apply the first two of these criteria in three experiments of real fMRI data. In each experiment, we have a priori expectations for regions that should be active, and we perform brain-wise and region-of-interest (ROI) analyses comparing robust IRLS and ordinary least squares (OLS) in those regions.

Experiment 1 is a cognitive task that requires both left- and right-handed responses in a single-trial event-related fMRI design. We compare sensitivity of OLS and IRLS to contralateral and ipsilateral primary motor (M1) responses in a group of subjects ("random-effects" analysis). Experiment 2 compares OLS and IRLS random-effects analyses of brain-wise responses to anticipation of pain. Experiment 3 employs a visual-motor paradigm with long inter-trial intervals (ITIs, 30 s). In this experiment, we explore the effects of using IRLS at the individual subject level.

Methods

Linear modeling framework

Before turning to robust regression techniques, we briefly review the general linear model (GLM). GLM finds the combination of predictors, each scaled by some value (β_i), that best fits the data. In algebraic terms, the GLM projects data (y) in an n-dimensional space (n independent data observations) onto a k-dimensional model subspace (k predictors). This framework is described by the equation:

$$y = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{1}$$

where $\mathbf{X}$ is the $n \times k$ model matrix whose columns contain values for the predictors, $\boldsymbol{\beta}$ is a $k \times 1$ vector containing the regression parameter estimates, y is an $n \times 1$ vector containing the observed data, and ε is an $n \times 1$ vector of unexplained error values. The most common GLM technique, ordinary least squares (OLS), defines "best fit" as the β_I that minimize the sum of squared deviations from the predicted values. This is equivalent to finding the vector $\boldsymbol{\beta}$ of length k, which minimizes

$$\left(\vec{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \left(\vec{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{2}$$

The algebraic solution to this problem is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y. \tag{3}$$

Dropping outliers

The simplest procedure for removing the influence of outliers is to drop them from the analysis altogether. One way to locate outliers is to identify data points that are far from the mean for each variable individually (univariate outliers). For example, data points more than three standard deviations from the mean of each variable could be sequentially located and removed. The weakness of this approach is that, if data are multivariate, highly influential outliers can be missed. Fig. 1B shows an example of a univariate outlier caused by adding noise to the y value of the

point. The noise added was drawn from a wider univariate distribution than the rest of the data, whose 95% confidence limits are shown by the top and bottom ends of the dotted square. Fig. 1B also shows a multivariate outlier, or a data point that is not an outlier for either variable individually, but that is an outlier when the two variables are considered together. This outlier was created by adding noise whose distribution is shows by the dashed circle. Such data points are outliers in the sense that they stand apart from the pattern of the majority of the data and have a disproportionate pull on the regression line. Importantly, we know these points are outliers in this dataset because we created them that way—we know the process by which they were generated—but that does not guarantee that they are the most outlying values in the dataset. In practice, we have to guess at which points might be outliers based on where they fall compared to the rest of the data. Knowing the true outlier status of points is important for validating techniques, which makes simulation a useful tool.

A method of locating both univariate and multivariate outliers is to employ multidimensional distance measures, such as Mahalanobis distance. Like standardized scores, Mahalanobis distances take into account a point's distance from the mean of some variable relative to that variable's variation, but unlike standardized scores, Mahalanobis distances also account for the covariance between the variables. Data points that stand apart from the data cloud have higher Mahalanobis distances than data points that lie within the data cloud. The Mahalanobis distance for a particular observation (subject in a group analysis) i is:

$$M_i = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T S^{-1} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \quad (4)$$

where $\mathbf{X}_i$ is the $p \times 1$ vector of scores on all p variables (both predictors and dependent variables) for the ith subject, $\bar{X}$ is the $p \times 1$ vector of means for each variable, and S^{-1} is the inverse of the $p \times p$ sample covariance matrix. A disadvantage of dropping points that have a high Mahalanobis distance is that it can lead to an overestimation of the degree of relationship between variables and thus, as we show below, can inflate FPRs.

Iteratively reweighted least squares (IRLS)

The GLM framework can be generalized to include additional information about error variance and covariance by incorporating a weighting matrix ($\mathbf{W}$) into the model estimation. The diagonals of $\mathbf{W}$ contain information about the estimated variance of the distribution from with each individual data point was drawn, and the off-diagonals contain information about the estimated relationship (covariance) between the distributions from which any two data points were drawn. In the traditional GLM framework, the error variances associated with each data observation are assumed to be equal, and the data are assumed to be independent. Thus, the $\mathbf{W}$ matrix in traditional GLM is the identity matrix.

The generalization of the GLM to account for correlated variables and unequal variances is a simple extension of the GLM. Let $\mathbf{Q}$ be the matrix square root of $\mathbf{W}$, that is, a matrix such that $\mathbf{W} = \mathbf{Q}^T\mathbf{Q}$. If we multiply each side of Eq. (1) by $\mathbf{Q}$, applying $\mathbf{Q}$ to both $\mathbf{X}$ and y, we obtain

$$\mathbf{Q}y = \mathbf{Q}\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\varepsilon \quad (5)$$

The best linear estimator of $\boldsymbol{\beta}$ in terms of $\mathbf{Q}y$ in the least-square sense is

$$\begin{aligned}\hat{\beta} &= \left((\mathbf{Q}\mathbf{X})^T\mathbf{Q}\mathbf{X}\right)^{-1}(\mathbf{Q}\mathbf{X})^T\mathbf{Q}y \\ &= \left(\mathbf{X}^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Q}^T\mathbf{Q}y \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}y \quad (6)\end{aligned}$$

In weighted least squares, the reciprocal of estimates of the variability of each data point ($1/\sigma_i^2$) are placed on the diagonals of $\mathbf{W}$, and the off-diagonals are zero (signifying independence of errors). When the model is estimated, the observations with the highest error variance are given the least weight in determining the regression parameters.

An important extension of weighted least squares is the iterative reweighted least squares (IRLS) algorithm. IRLS is particularly useful in massively univariate settings because it does not require a priori knowledge of the variability of data points and because it can be accomplished using an automated and efficient algorithm. The IRLS procedure works as follows:

(1) The researcher chooses a weighting scheme that down-weights residuals as they become larger according to some function. The bisquare and Huber are two common weighting schemes, and they are shown in Fig. 1C. The bisquare reduces the influence of all points as their residuals grow, while the Huber down-weights only those points that pass some threshold (defined by a tuning constant). Bisquare weighting schemes tend to down-weight more aggressively, and thus have slightly more power in the context of outliers but also slightly inflated FPRs relative to the Huber.

(2) The algorithm initially performs OLS regression (i.e., where the $\mathbf{W}$ matrix is an identity matrix) to obtain initial $\hat{\boldsymbol{\beta}}$. In addition, an adjustment factor for the residuals is calculated as 1/sqrt(1−$\boldsymbol{h}$), where $\boldsymbol{h}$ is a vector containing leverages for each observation. Leverages are the diagonals of the "hat matrix" $\mathbf{H}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The higher the leverage, the more influence the observation has on the regression plane—which means its residual value will be an underestimate of the true value, because it pulls the regression plane closer to it. Another way of saying this is that even if errors are normally distributed with equal variance, $\boldsymbol{\epsilon} \sim N(0,\sigma^2)$, residuals will have variance inversely proportional to the leverage, $\hat{\boldsymbol{\epsilon}} \sim N(0, I-H)$. Thus, high-leverage points tend to have low residual values, and they will tend to dominate the robust regression fit if not accounted for. This adjustment factor will be used to correct residual values below. Note that only the x (predictor) component of Mahalanobis distances contribute to high leverage, as leverage is purely a function of the design matrix $\mathbf{X}$.

(3) Residuals ($\hat{\boldsymbol{\epsilon}}$) are obtained by subtracting the fitted response from the data according to the formula $\hat{\boldsymbol{\epsilon}} = y - \mathbf{X}\hat{\boldsymbol{\beta}}$. Residuals are standardized by their median absolute deviation (MAD), an outlier-robust estimation of spread. They are then multiplied by the adjustment factor from the previous step, which inflates the error estimates for high-leverage points (so that they are subsequently downweighted). Finally, the residuals are multiplied by the weighting function, which generally downweights very high error values disproportionately, to obtain new weights for the observations. These weights are placed on the diagonal of the $\mathbf{W}$ matrix. The lower the weight given to an observation, the less influence it has on the subsequent regression fit.

(4) The regression is rerun using the new **W** matrix, according to Eq. (6). Steps 3 and 4 are iterated until the fit statistics converge. Convergence in the Matlab 6.5 (Mathworks, Natwick, MA) algorithm we used in this paper is defined as a change in successive iterations less than the square root of the maximum numerical precision available (e.g., 1×10^{-8}).

Inferential statistics for IRLS procedures have not been fully developed. Results for large samples have been obtained (Hoaglin et al., 1985), but these results may not apply when outliers are present. Therefore, it has been suggested that bootstrapping is the preferred method if inference is desired when using IRLS (Neter et al., 1996), but such an approach defeats the primary advantage of its computational efficiency. An alternative has been presented by DuMouchel and O'Brien (1989) based on principles in Huber (1981), which shrinks the robust error variance toward the OLS error variance if the latter is greater. One purpose of the present paper is to use simulations to understand the validity of inferential statistics from IRLS using this method under conditions commonly encountered in neuroimaging studies.

In summary, robust regression techniques that use the generalized least squares (GLS) framework, such as IRLS, incorporate a weighting matrix into the statistical estimation so that some points have less influence on estimates of the regression parameters. An additional advantage is that the same GLS framework is used to account for autocorrelation (e.g., in fMRI time series) and covariance of parameter estimates, which are incorporated in recent versions of statistical software for neuroimaging analysis (e.g., SPM2, FSL). Thus, robust estimation can easily be incorporated into a sophisticated individual subject analysis. However, robust group analysis can proceed whether or not IRLS is incorporated into the individual subjects' analysis, by entering from any imaging statistics software into a robust 'random-effects' analysis.

Other techniques

Other common robust techniques include least median of squares (LMS) regression, least absolute deviations (LAD) regression, least trimmed squares (LTS) regression, and minimum covariance determinant (MCD) estimation. LMS and LAD are early robust techniques and minimize alternative error measures to the sums of squared deviation measure used in OLS. LMS minimizes the median of the squared residuals rather than the sum of the square residuals, and therefore outlying values have very little influence on the fit. LTS regression minimizes a central portion (specified by the analyst, with a default of 75% of the observations) of the squared deviations. For LTS, LMS, and LAD, bootstrapping is commonly used if inference is desired (Neter et al., 1996). This is a large drawback in massively univariate settings because of the time necessary to bootstrap large numbers of analyses.

MCD is a more recent technique that determines multivariate outliers with a specified breakdown point using an iterative strategy (Hubert, 2001). The idea of MCD is to minimize the determinant of the covariance matrix of the central f points. Thus, the technique essentially finds the set of f points that are most multidimensionally co-planar. The primary disadvantage to MCD in the context of numerous analyses is that it can be hundreds of times slower than IRLS. Our initial simulations compared MCD and IRLS, but we found that the false positive rate in MCD was inflated substantially with small numbers of observations (which may be expected; see (Hubert et al., 2004).

Simulation: comparing regression methods

We performed regressions on simulated data to determine the power and false positive rates across a range of parameters relevant to neuroimaging experiments.

Data generation

We generated data to simulate participants' contrast values (both slopes and intercepts) in one brain region, although the data can be generalized to any bivariate data setting. The intercept parameter corresponds to the one-sample t test on brain activity in a voxel across subjects. The slope parameter corresponds to the simple regression of activation contrast on a behavioral index (e.g., performance).

Data were pairs of random vectors ($\boldsymbol{x}$ and $\boldsymbol{y}$) drawn from standard normal distributions ($\bar{x} = 0$, $\sigma^2 = 1$), or $N(0, 1)$. We adjusted these vectors in two ways in order to obtain FPR and power statistics for both slope and intercept estimation. To simulate "activation" alternative hypothesis (H_a) data, we added a constant to the data, so that it was distributed $N(.5, 1)$. We also set the covariance between x and y to be either 0 (for H_0 data) or 0.5 (for H_a data). Thus, for H_a data, the correlation explained 25% of the variance in y (Cohen's $d = 0.25$) and the intercept explained 50% of the variance in y (Cohen's $d = 0.5$). These numbers were chosen as reasonable approximations of typical effect sizes in imaging data analysis. We varied the number of observations (n) per sample between 5 and 40. For each combination of parameters, we regressed y on x in the general linear model 2000 times.

FPRs for the slope parameter were estimated by the proportion of 10,000 H_0 datasets that showed significant positive or negative regression slopes at $P < 0.05$, two-sided. Similarly, FPRs for the intercept parameter were estimated by the proportion of 10,000 H_0 datasets that showed significant non-zero intercepts at $P < 0.05$. Power rates for both slopes and intercepts were estimated as the proportion of 10,000 H_a datasets whose parameters were significant in the expected (positive) direction.

Outlier generation

We simulated the effects of both univariate and multivariate outliers in the data. To introduce univariate outliers in y values, we added additional Gaussian noise with a larger standard deviation ($\sigma = 3$) to a specified proportion ($q = 0.1$) of the y values. We performed additional simulations using a range of different sizes of outliers ($\sigma = 3$ to 10) and proportions of outliers ($q = 0.05$ to 0.20), but these did not change the pattern of results and so we do not report on these additional analyses. To introduce outliers in the multivariate setting, we drew outliers from a bivariate normal distribution with greater variance ($\sigma = 3$) in x and y.

Regression models

In our simulations, we compare five types of regression model. The first is ordinary least squares (OLS). We also modeled two simple approaches to robust regression: dropping data points that are likely to be outliers and then running OLS on the remaining data. The univariate trimming approach (Univar) excludes data points that are more than 3 standard deviations from the mean on y. As Mahalanobis distances approximately follow a χ^2 distribution (Neter et al., 1996; Rocke and Woodruff, 1996), multivariate outlier trimming (Mahal) consisted of removing observations with Mahalanobis values greater than 5.99 ($P < 0.05$ on the χ^2 distribution for $df = 2$).

The final two robust regression methods we employed were robust IRLS models using either the bisquare or the Huber weighting functions.

Experimental paradigms

Experiment 1: cognitive/motor activity

In Experiment 1, eleven participants performed a cognitive task in which they were required to respond with a single button-press using either the right or the left thumb. Participants were asked to match a center stimulus with a corresponding flanker. On each trial, two flankers, a yellow square and a blue square, appeared at the sides of the screen. A central stimulus—either a yellow or blue square, one of two abstract shapes, or a shape overlaying a yellow or blue square—appeared at the same time. When colored squares only appeared (Color), participants located the like-colored flanker and pressed the corresponding thumb. Participants learned that one shape was the "blue" shape, and one was the "yellow" shape (both shapes appeared in white). When shapes appeared (Shape), the participants matched shape form to the flanker with the corresponding color association. When the center stimulus consisted of shapes over colored squares (Shape Mixed), participants matched shape to flanker color, and ignored the central colored square. The central colored square could be congruent with the correct response, in the sense that it matched the color of the square on the correct side of the screen (Congruent) or it could be incongruent (Incongruent). Thus, there were eight trial types: Color only, Shape only, Shape Congruent, and Shape Incongruent crossed with R and L motor response.

40-s long blocks of tasks (Color, Shape, and Shape Mixed) were cued and presented in pseudorandom order. Each block consisted of four trials spaced 10-s apart, and the trial orders were counterbalanced up to 2 trials back, to ensure even transitional probabilities among trials and prevent trial dependencies from causing artifacts in estimated HRF shapes.

Our analysis in the context of the present report focuses on primary motor cortex. Contiguous voxels showing strong lateralized (L thumb > R thumb in the R hemisphere, and vice versa for the L hemisphere) responses in a group random-effects analysis were masked with the gray-matter precentral sulcus region from the ICBM labeled template (Kochunov et al., 2002; Mazziotta et al., 2001). Data were extracted from R and L motor ROIs. Our analysis examines individual subject parameter estimates for each of the eight trial types for outliers, and compares OLS and IRLS random-effects analysis for each ROI.

Spiral-out gradient echo images were collected on a GE 3T fMRI scanner (Noll et al., 1995) with 3.75 × 3.75 × 7 mm voxels, TR = 1 s, TE = 25 ms, flip angle = 90, FOV = 24 cm. 16 slices provided near whole-brain coverage.

Experiment 2: pain

Participants (n = 23) were given visual cues signaling upcoming thermal pain stimulation on the left forearm. After a variable anticipation period (1–16 s, mean = 9 s), a 20-s thermal stimulus was delivered. The stimulus was calibrated to be moderately painful for each participant, and included a 1.5 s ramp-up period, a 17-s plateau period, and a 1.5-s ramp-down period. Participants were cued to report the intensity of pain after a variable interval following stimulation offset.

Regressors for anticipation and pain were dummy coded predictors of responses to the 4 s following the anticipation cue and the last 10 s of painful stimulation, corresponding to the subjectively most painful portion of the stimulus. Additional anticipation, pain, and pain reporting periods were modeled simultaneously with these; they were not collinear with the predictors we report here, and they are not discussed further in this report.

Spiral-out gradient echo images were collected on a GE 3T fMRI scanner (Noll et al., 1995) with 3.75 × 3.75 × 5 mm voxels, TR = 1.5 s, TE = 25 ms, flip angle = 90, FOV = 24 cm. 25 slices provided near whole-brain coverage.

Experiment 3: visual/motor HRF estimation

In this experiment 10 participants performed two visual-motor tasks in which they observed contrast-reversing checkerboards (16 Hz) and made manual button-press responses. In the first task (localizer), 16-s blocks of unilateral contrast-reversing checkerboards were presented in alternating left (L) and right (R) order. Subjects made motor responses continuously with the index and middle fingers of the ipsilateral hand. Results were used to localize primary visual and motor cortices in each hemisphere for each subject (contiguous voxels on the precentral gyrus or in the calcarine fissure at $t > 3.5$), and data from the main task (below) was analyzed in each of these four ROIs.

The main task consisted of eight functional runs of visual/motor stimuli. In each run, a series of brief contrast-reversing checkerboard stimuli (250 ms stimulus duration) separated by a 1-s stimulus-onset asynchrony (SOA) were observed by participants. Stimuli were as close to full-field as the stimulus-presentation system allowed (approximately 30 degrees of visual angle). Participants were instructed to press the index and middle fingers of both hands together each time they saw a checkerboard stimulus. Each stimulus series consisted of either 1, 2, 5, 6, 10, or 11 stimuli separated by 1 s, followed by 30 s of rest. The order of the 6 series types (i.e., series of 1, 2, 5, 6, 10, or 11 stimuli) was counterbalanced across runs, and 2 trials of each train length were presented in each run, for a total of 16 trials per series length for each participant.

Spiral-out gradient echo images were collected on a GE 3T fMRI scanner (Noll et al., 1995). Seven oblique slices were collected through visual and motor cortex, 3.12 × 3.12 × 5 mm voxels, TR = 0.5 s, TE = 25 ms, flip angle = 90, FOV = 20 cm. We extracted hemodynamic response estimates (HRFs) for each stimulus series length in each participant in each of the four ROIs using an unsmoothed finite impulse response (FIR) model. Our first analysis compared linear model fits of the canonical SPM HRF (Friston et al., 1995) composed of two gamma functions, to ROI timecourses using OLS and IRLS. Our second analysis compared FIR-derived HRF estimates derived with OLS and IRLS using a split-half (odd runs vs. even runs) reliability metric for consistency across HRF shapes within stimulus length, participant, and brain region.

Comparison of t values

To compare t scores for ordinary and robust voxel-wise estimates, we used a z test to compare z-transformed differences between t scores for IRLS and OLS at each voxel. An estimate of P values was performed by first transforming differences of t scores to z scores according to the formula:

$$\zeta = \frac{t_{\mathrm{IRLS}} - t_{\mathrm{OLS}}}{\sqrt{2\sigma^2_{t(df)}}} \tag{7}$$

where $\sigma^2_{t(df)}$ is the variance of the t distribution with $n-1$ degrees of freedom. For purposes of testing and display, we thresholded images at $\zeta > 1.64$, corresponding to $P < 0.1$. Simulations confirmed that ζ is asymptotically approximately normally distributed with large n, but is somewhat conservative with low n (P values were 1.7× too high in simulations with 10 df, but correct with $n = 40$.) We chose this method over P value based t to z transformations because of the numerical inaccuracy of the latter in the tails of the distribution (e.g., that employed by SPM; http://www.fil.ion.ucl.ac.uk/spm/). Importantly, Eq. (7) strictly holds only for independent t statistics; however, our t scores were calculated on the same data, and are therefore dependent. For positive dependence, as we have here, the variance of the difference between t values will be overestimated, and the test is quite overconservative. However, we do not intend to use this to establish a strict inferential test for when robust methods are "significantly more reliable," but rather as a guideline for interpreting results.

Results

Simulation: comparing regression methods

We compared FPR and experimental power for five regression techniques: OLS, IRLS with bisquare and Huber weighting functions, univariate outlier removal (Univ), and multivariate outlier removal (Mahal). Fig. 2 shows the results of simulations for all n (5, 10, 15, 25, 40) at $q = 0.1$, $m = 3$, and $t = 3$ for the intercept term (i.e., detecting activations in a random effects analysis), where q is the proportion of outliers, m is the standard deviation of the outlier noise distribution, and t is the threshold for univariate outlier removal. The first column (2A) shows results for normally distributed data with no outliers. The second column (2B) shows results with univariate (y) outliers. The third column (2C) shows results with multivariate outliers. The x axis of each graph increases with increasing n. The y axis of the top panel shows observed FPRs at $\alpha = 0.05$; thus, FPRs above 0.05 are inflated. The y axis of the bottom panel shows observed power given the effect sizes described earlier. Fig. 3 shows the same data for the slope term.

With well-behaved data with no outliers, FPR was essentially appropriately controlled for all methods except for a slight inflation for the Mahal method (green triangles), which increases with n (top panel, Fig. 2A) and a very slight increase in FPR for IRLS with low n. When no outliers are present, the power is virtually equivalent for all methods (bottom panel, Fig. 2A). The same patterns of FPRs and power can be observed for the regression slope (Fig. 3A). These results are important because they signify that, compared to OLS, there are no disadvantages for three of the four robust techniques in FPRs or in power when outliers are absent.

If univariate outliers are present, FPRs remains appropriate and essentially unchanged for all methods, again showing a slight increase (anticonservative) for Mahal relative to other methods (Fig. 2B) and a slight conservativeness for IRLS and OLS. However, power is dramatically reduced for OLS regression (blue circles). (No difference is observed for $n = 5$ because no outliers are present for $n = 5$ and $q = 0.1$.) The power increases with IRLS are striking in that they are not accompanied by concomitant increases in FPR; power can always be achieved at the expense of validity. Univ (cyan stars) is the most powerful method and Mahal is slightly less powerful. The IRLS techniques (red squares and black dashed lines) offer a substantial improvement over OLS. The benefit of using IRLS increases with larger n and converges with Mahal at $n = 40$. The reason for this is that larger sample sizes afford a better estimate of which points are truly outliers, and allow the IRLS weighting to work with greater efficiency. The same pattern holds for slope estimates (Fig. 3B). While Univ may seem appealing according to these simulations, it is important to remember that the true outlier standard deviation ($m = 3$) and the cutoff for removing outliers are the same, which could artificially inflate power with this method.

If multivariate outliers are present (e.g., outliers in brain activation and in behavioral covariates), FPR is controlled roughly appropriately for the intercept (Fig. 2C), with a slight increase for Mahal. Observed power is also highest for Mahal, with Univ and IRLS offering similar improvement over OLS, with the benefit increasing as n increases. The benefit of Mahal here can be understood intuitively by recalling that multivariate outliers are often not univariate outliers, and so are not detected.

The results for the slope parameter (Fig. 3C) were quite different than the results for the intercept parameter. No technique except Mahal came close to controlling the FPR. For every technique except Mahal, FPR was ~0.125 for $n = 5$, higher than the nominal 0.05 rate, and generally increased with n. IRLS and Univ techniques attenuate this problem, but do not correct it. Power was highest for bisquare IRLS, slightly lower for Mahal and Huber, and substantially lower for OLS and Univ. This result shows that outliers in the brain-behavioral (bivariate) space can exert a powerful, meaningful influence on brain-behavior correlation estimates, and FPR is not properly controlled if there are high-leverage, outlying points in the behavioral data. Most importantly, the default regression method, OLS, performs the worst in terms of both power and FPRs, and thus use of some kind of robust technique seems necessary when multivariate outliers are likely to exist in the data.

We observed the same pattern of results across variations in q and m. However, if outliers are trimmed at $t = 2$ standard deviations rather than 3, the FPR increased substantially for outlier removal methods, suggesting that $t = 2$ is not an appropriate threshold. This is an important point because heavier trimming may seem intuitively appealing to some investigators. Overall, the simulations suggest that IRLS techniques offer a compromise between OLS and all-or-none outlier removal.

Though our simulations compare power and FPR at a nominal alpha threshold of 0.05, we expect the results to hold for much lower alpha levels (e.g., $P < 0.001$) and for corrected P values; however, we do not simulate at these levels here because the simulation accuracy decreases as the cutoff point becomes more extreme (thus, very many more iterations would need to be run).

Applications to fMRI data

The simulations performed above seem to indicate that robust techniques may be important in limiting false positives and in maximizing power when outliers are present, but that they perform about as well as OLS when outliers are absent. It is important to understand how robust techniques perform on real data. We explore IRLS techniques as applied to real fMRI data in three experiments that have well-defined a priori regions of interest within which to investigate results. For brevity, we restrict the

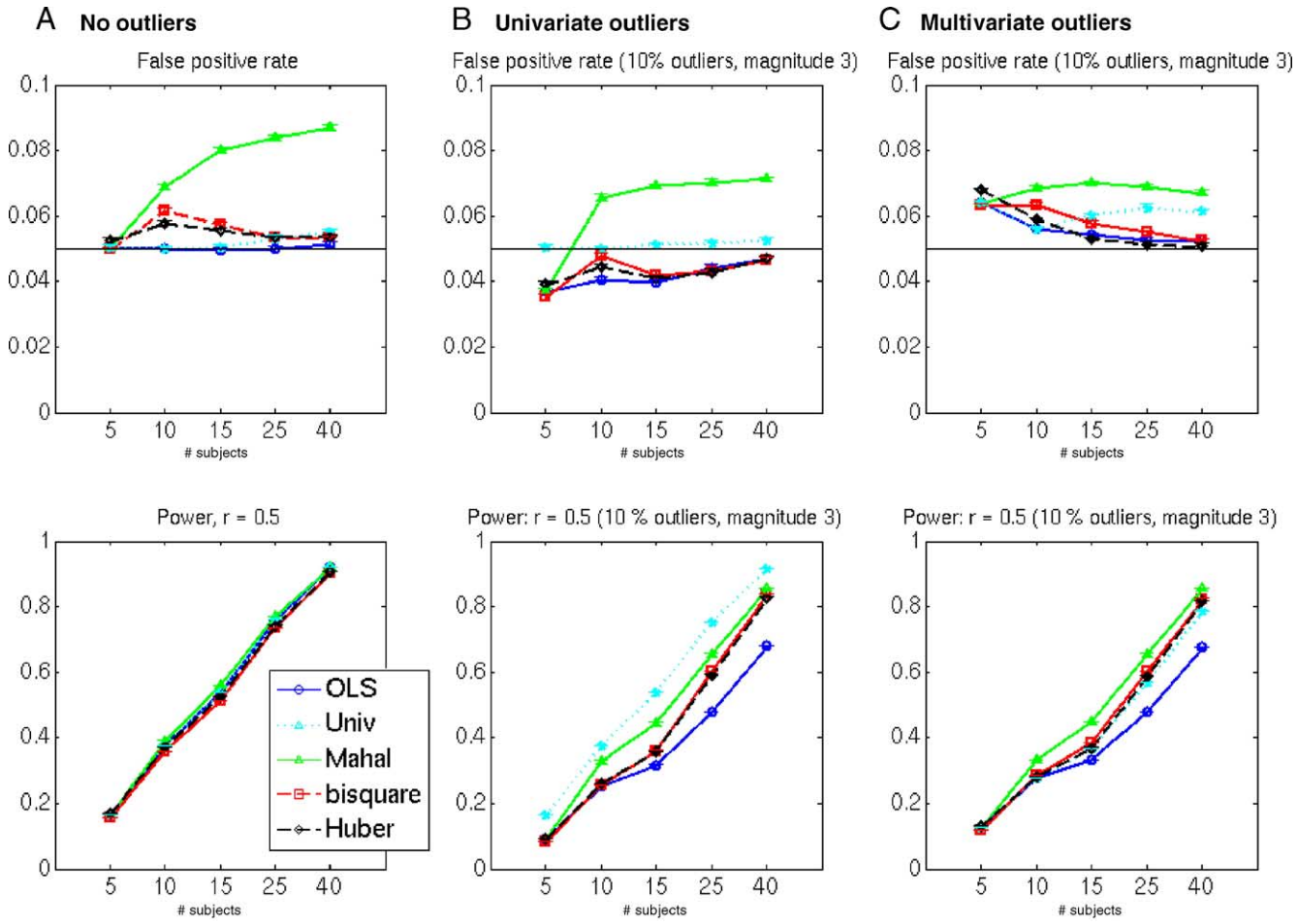

Fig. 2. Simulated false positive rates and power under different outlier conditions, and with different regression methods, for estimation of the regression intercept (overall activation in an imaging context). (A) The left column shows results for normally distributed noise without (top) and with (bottom) a true signal. (B) The middle column shows results with univariate outliers in the data. (C) The right column shows results with multivariate outliers in both the predictor and the data, as in Fig. 1A. Univariate outlier removal offers the best if univariate outliers are present, but is poor in the presence of multivariate outliers. Multivariate outlier removal is best for multivariate outliers, but increase false positive rates under normal conditions. The IRLS techniques offer an intermediate solution, providing gains in power that increase with sample size (n, x axis) while controlling false positive rates. Error bars show standard errors of the means across 10 replications of the simulations.

comparison below to OLS and two robust techniques—IRLS using the bisquare weighting function (referred to simply as IRLS below) and the dropping of univariate outliers that are ± 3 standard deviations from the mean on the dependent variables (Univ). We focus on IRLS using the bisquare weighting scheme because it had almost the exact same FPR as the Huber weighting scheme but had slightly more power in the simulations, and we focus on the dropping of univariate outliers because it tended to have lower FPR but comparable power to the dropping of outliers based on Mahalanobis distance.

Experiment 1: motor responses

Experiment 1 compared IRLS and Univ with OLS in two regions that are expected to show strong motor responses. By isolating regions of interest where we have a strong a priori expectation of activity (Figs. 4A and B, first panel), we can infer that higher t scores reflect increased power rather than increased FPR. In four conditions in this experiment (1–4 on the x axis of Fig. 4; see Methods), responses were made with the right thumb. In the remaining four conditions (5–8), responses were made with the left thumb. Significant activation in right and left motor cortices was found for both contralateral and ipsilateral responses, with much higher t scores for contralateral responses, as expected (Figs. 4A and B, second and fourth panels).

In the second panel of Fig. 4, mean BOLD contrast is shown for each condition, with individual participants' values in each condition shown by symbols. In a number of cases, some individual participants were potential univariate outliers ($z > 1.96$, circled in red). The third panel shows t scores for OLS (black bars), IRLS (gray bars), and Univ (white bars). Because none of the potential outliers reached $z \geq 3$, Univ and OLS results are identical.

In most cases, IRLS resulted in a small reduction in t scores, as down-weighting observations effectively reduces the variance of the predictors and the data (this kind of variance is advantageous). Another way to think of this might be as a reduction in the effective degrees of freedom in the robust test. However, in some cases, IRLS resulted in a significant improvement in reliability ($P < 0.05$ by z test). This happened when one or more outliers fell below the central mass of observations, thus increasing the standard error and decreasing the estimated response. Outliers above the central mass increase both the estimate and the standard error, resulting in little net difference in t values. Notably, this case is one that should,

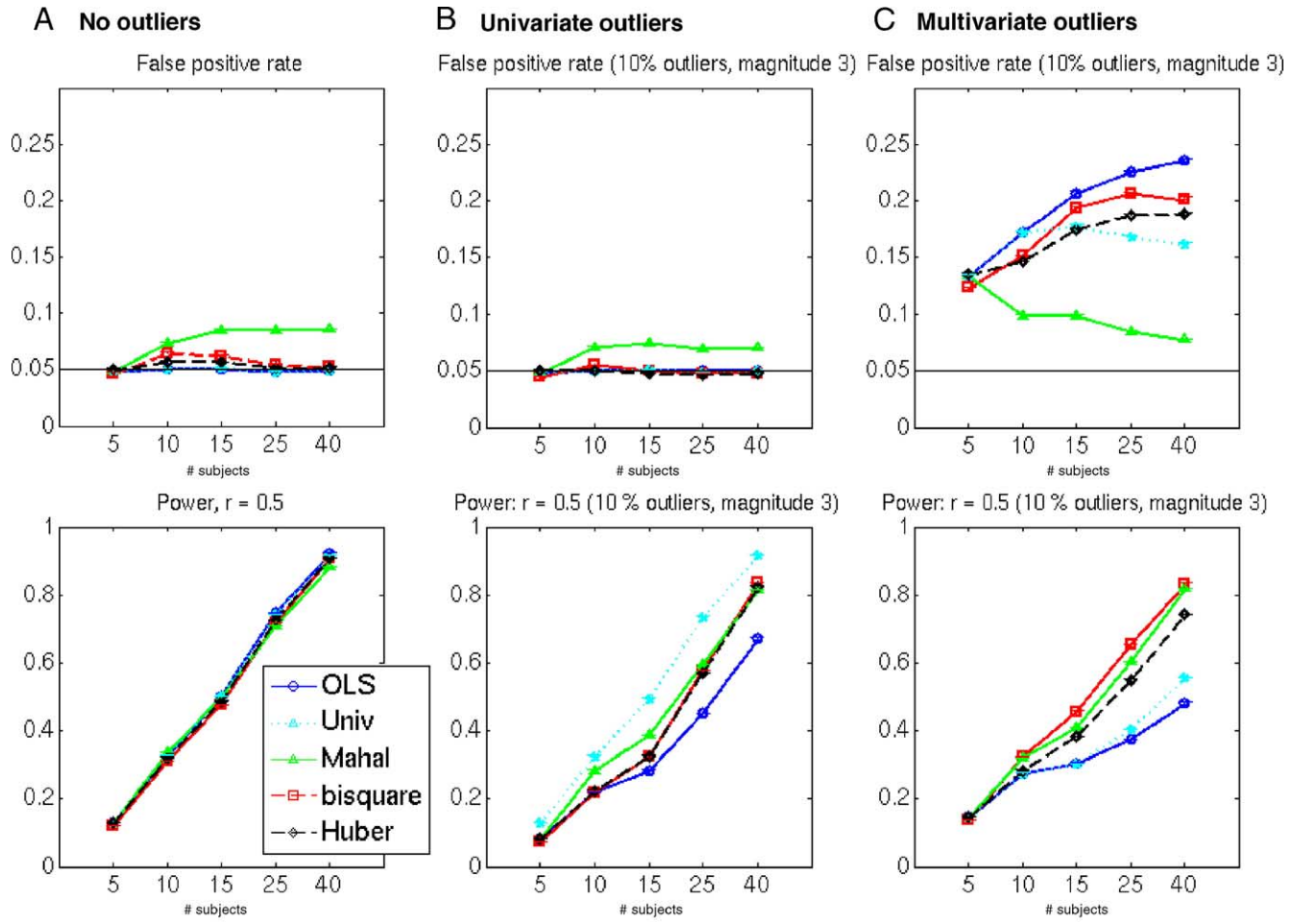

Fig. 3. Simulated false positive rates and power for estimation of the regression slope (correlation with behavior in an imaging context). A–C are as in Fig. 2. False positive rates are high in the presence of multivariate outliers, but are reduced somewhat by robust methods. Univariate outlier removal is best if only univariate outliers exist, but is poor if there are multivariate outliers. IRLS offers increased in the presence of both univariate and multivariate outliers, with improved or equivalent control of false positive rates.

according to the simulations, offer the least advantage of IRLS: We were estimating intercept terms with $n = 10$ (small sample size).

Experiment 2: pain

Fig. 5A shows responses to a foveal visual cue signaling upcoming pain. The left and middle columns show random-effects activations ($n = 22$, $P < 0.001$) for OLS (left column) and IRLS (middle column). T-maps are superimposed on slices of a canonical brain, and slices were picked that showed the maximal difference between OLS and IRLS, unbiased with respect to the direction of the difference. Red-yellow indicates activation and green-blue indicates deactivation.

The right hemisphere shows reduced activation with OLS activation, whereas the expected bilateral activation is detected with IRLS. In addition, right anterior prefrontal cortex and ventral striatum is detected with IRLS, and this activation is expected from previous studies of anticipation of pain and aversive events (Becerra et al., 2001; Jensen et al., 2003; Ploghaus et al., 1999; Porro et al., 2002).

The right panel shows z scores for the difference in t scores for IRLS-OLS. Red indicates significantly more positive t values for IRLS, and green indicates significantly more negative t values for IRLS. Activated regions discussed above showed significant increases in reliability for IRLS. Posterior cingulate, deactivated with both OLS and IRLS, is significantly more deactivated with IRLS. This deactivation is expected, as this region is commonly deactivated in attention-demanding tasks (Gusnard and Raichle, 2001; Raichle et al., 2001). These results illustrate that IRLS can produce qualitative changes in activated regions—depending, of course, on the threshold used—in biologically meaningful brain regions.

Fig. 5B shows responses during pain. As expected from previous studies, bilateral SII (superior posterior insula), anterior insula, anterior cingulate, and anterior PFC were activated, and posterior cingulate was deactivated (Becerra et al., 2001; Davis et al., 1998; Peyron et al., 2002; Ploghaus et al., 1999; Schneider et al., 2001). Activations in right (contralateral to stimulation) SII, anterior insula, and anterior PFC showed more reliable activation with IRLS, and posterior cingulate bilaterally showed more reliable deactivation.

Experiment 3: visual/motor responses

This experiment investigated the effects of using robust regression on fitting time series data at the individual participant level. Our prediction was that when outliers were present in the time series, robust IRLS would result in higher t values and greater

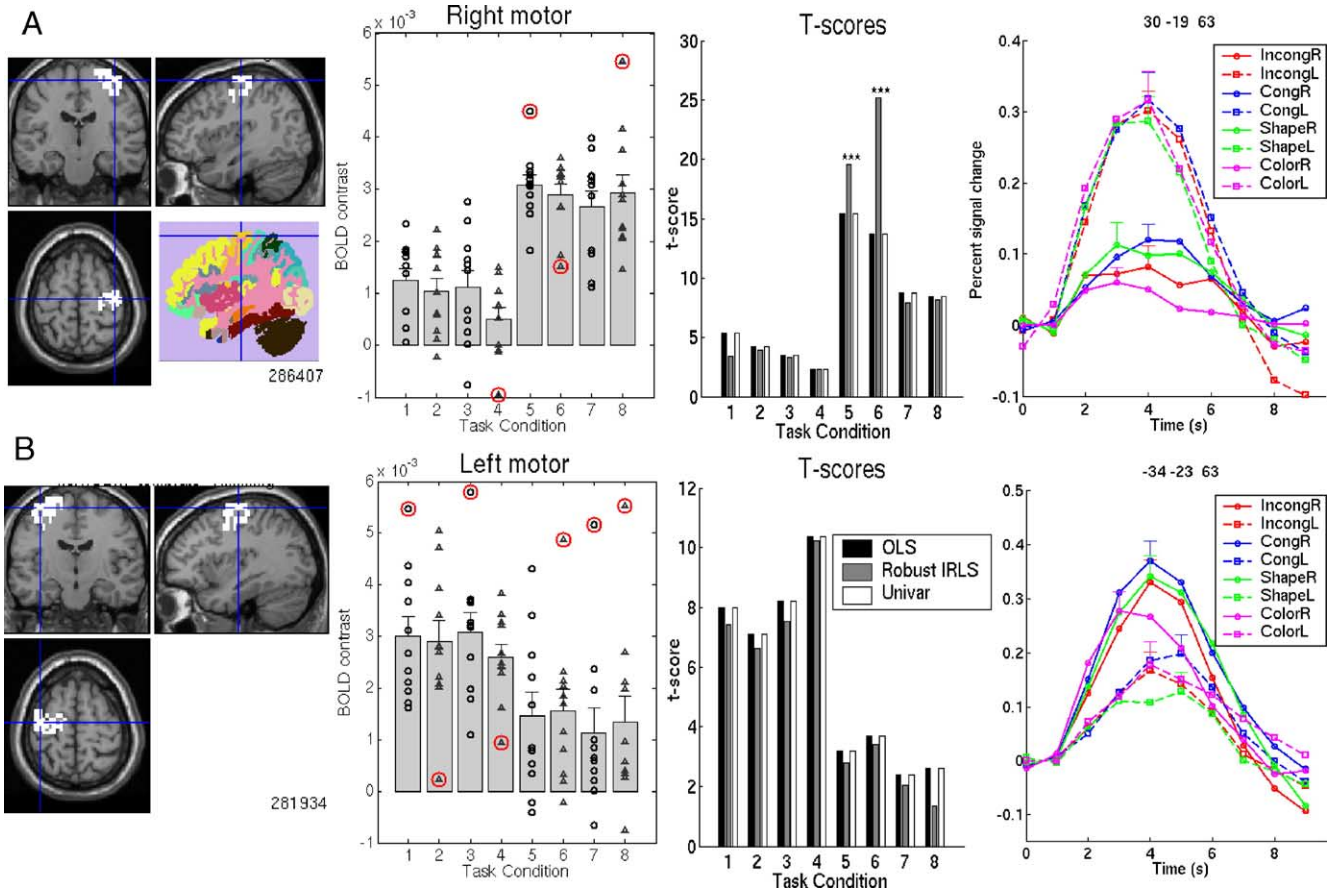

Fig. 4. Experiment 1. (A) Right motor ROI. Panel 2 shows contrast values with standard error bars for each trial type. Individual subjects are plotted with symbols, and those with z scores >1.96 (potential outliers) are marked in red. Conditions 1–4 required ipsilateral (R) responses, and conditions 5–8 required contralateral (L) responses. Panel 3 shows t values for ordinary (OLS), robust (IRLS), and univariate outlier removal (Univar). ***$P < 0.001$ (two tailed) in a z test between the robust method and OLS. Panel 4 shows L motor responses (dashed lines, conditions 1–4 in Panel 2) and R motor responses (solid lines, conditions 5–8). (B) Results for left motor cortex, as in A.

reliability of HRFs for individual participants. Higher t values can be used as a measure of sensitivity only in a priori ROIs in which we can be confident of a true positive activation; thus, we used individually defined, primary sensory and motor ROIs that are known to be activated by sensorimotor tasks. The estimated HRFs, shown in Fig. 6, were highly consistent across participants (Fig. 6A), and extremely consistent across ROIs (Fig. 6B), and are very similar to the canonical SPM HRF.

Fig. 7 shows results for one region (left visual cortex) comparing IRLS to OLS results using the canonical SPM HRF. Other regions, not shown for space reasons, produced similar results. Black bars indicate OLS t values for each stimulus type for each participant, white bars indicate IRLS t values, and dark gray bars show the overlap. Thus, white bars appearing above dark gray bars indicate higher t values for IRLS, and black bars appearing above gray bars indicate higher t values for OLS. Two participants (P4 and P12) showed very strong benefits of IRLS estimation across conditions, indicating that activation to visual stimulation in left visual cortex was much more reliable with IRLS estimation. Three participants (P1, P5, and P11) showed smaller but appreciable improvements with IRLS. (Each unit change in t value signifies an approximately 10-fold decrease in P value, as a rough guide). No participants showed systematic benefits for OLS.

Second, we used a finite impulse response model (FIR) in the GLM framework to determine HRFs for each participant in each region × stimulus type combination. For each HRF, odd runs were used to calculate one HRF, and even runs were used to calculate another, with both OLS and IRLS estimation methods. We then correlated the two HRF estimates for odd and even runs to derive a split-half reliability, and compared reliability for OLS and IRLS. Across participants, robust IRLS resulted in significantly higher reliability scores for short stimulus trains (1 stimulus, $t(10) = 2.07$, $P < 0.05$; 3 stimuli, $t(10) = 2.33$, $P < 0.05$), and never resulted in significantly lower reliability scores. Fig. 8 summarizes these reliability estimates. The blue bars show a histogram of reliability differences between IRLS and OLS for all stimulus types. Red bars show reliability differences for one-stimulus trains only.

Most reliability scores were quite similar for IRLS and OLS, as shown by the top left inset. In the inset, the left and right panels show HRF estimates for one participant in one brain region. Solid lines are from odd runs, and dashed lines are from even runs. In this inset, HRFs look very similar for both regression methods.

However, for some cases, outliers in the time series exert strong influences on HRF estimates, resulting in a lower reliability score for OLS estimates. The bottom and right insets of Fig. 8 show two examples, corresponding to values in the difference histogram of approximately 0.58 and 1.02, respectively. Values can exceed 1

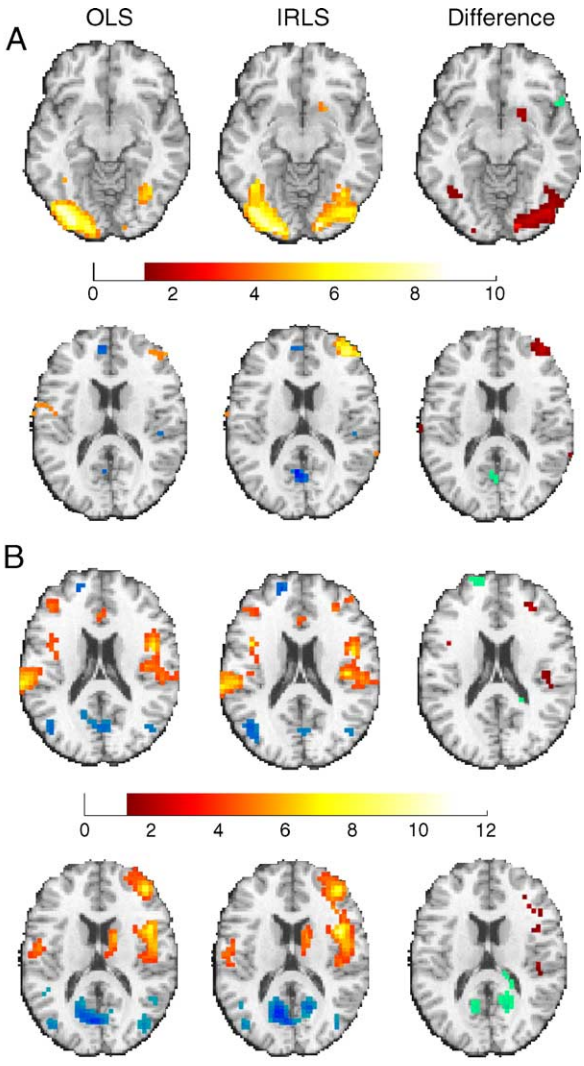

Fig. 5. Experiment 2. Results from ordinary least squares (OLS, left column) and robust IRLS (middle column) during anticipation and experience of pain relative to a fixation baseline. The first two columns show t maps thresholded at $P < 0.001$, with positive activations in red/yellow and deactivations in green/blue. The third column shows z scores for the difference in reliability for IRLS-OLS, thresholded at $P < 0.10$ (two-tailed). Voxels in which IRLS produced significantly greater statistical significance by a z test are shown in red; those in which OLS produced more significant results are shown in green. Slices were selected in an unbiased manner by choosing the slice with the greatest overall difference between OLS and IRLS, irrespective of direction. (A) Tow row: early anticipation of pain, after receiving a warning cue. IRLS revealed more reliable activation in the right occipital cortex, showing bilateral activation, whereas OLS activation was largely left-lateralized, and greater reliability in ventral striatum, expected during pain anticipation. Second row: late anticipation. Right anterior prefrontal cortex (aPFC) shows more positive results with IRLS, and posterior cingulate shows more reliable deactivation. (B) Top row: thermal pain (20 s duration) applied to left forearm, early (0–10 s) phase. IRLS shows more reliable activity in contralateral SII and anterior PFC. Second row: pain, late (10–20 s) phase. IRLS shows more reliable activity in posterior (SII) and anterior insula and PFC, and more reliable deactivation in posterior cingulate. Deactivation is estimated to be unilateral with OLS and bilateral with IRLS.

slightly if the reliability estimate for IRLS is near 1, but the reliability estimate for OLS is negative (true reliability should never be negative, but estimates can be negative). For example, a reliability estimate of 0.99 for IRLS and −0.05 for OLS would yield a difference of 1.04; this occurred for two cases. As the inset figures show, IRLS minimized artifactual spikes in HRF estimates in these cases. IRLS never resulted in substantially lower reliability scores (the maximum decrease was less than $r = 0.10$ over 640 estimates).

Discussion

The results from both simulations and experimental data demonstrate that robust estimation methods can offer substantial benefits in neuroimaging analyses. Robust techniques are well suited for cases in which artifactual outliers may exist in the data, and automated robust techniques, such as IRLS, offer substantial improvements when each regression analysis cannot be individually checked for violations of assumptions. Both situations are true of neuroimaging data. Outliers in time series data can be caused by gradient changes, transient susceptibility, and motion-related effects in neuroimaging, and outliers in group data can be caused by time series outliers and by anatomical and functional variability among individuals.

IRLS, false positives, and power

IRLS is a computationally efficient robust estimation technique that can be employed when investigators need to run a large number of separate regression analyses. The primary advantage of IRLS, demonstrated in both simulation data and in actual experimental data, is that it substantially increases power in the presence of outliers. The advantage of IRLS is evident at all sample sizes and increases as sample sizes increase. Our simulations demonstrate that even with a relatively large sample size ($n = 40$ is currently considered large for neuroimaging experiments), a small number of outliers can create large reductions in experimental power. Robust estimators can effectively minimize problems created by influential outliers.

In our simulations, IRLS controlled FPRs at the same level as OLS or better. A potential misconception is that IRLS capitalizes on chance by down-weighting observations that do not fit the model and thereby increases FPRs. However, IRLS essentially down-weights observations that do not fit the central mass of the data, not those that do not fit the hypothesis. Thus, under the null hypothesis, IRLS is equally likely to down-weight data that favor the alternative hypothesis in either positive or negative tails of the distribution, resulting in a zero net bias. However, particularly with small samples, a large proportion of the observations can be co-planar by chance, resulting in down-weighting of observations favoring the null hypothesis and broadening of the tails of the distribution (more false positives in both positive and negative directions; Hubert et al., 2004). We observed this problem when we used MCD outlier estimation; however, IRLS may be more effective because it (a) offers a ‘softer’ weighting approach than trimming, decreasing the impact of chance co-planarity, and (b) uses an error variance estimate that is adjusted towards the OLS value, if the OLS error variance is greater. Our simulations bear out the validity of P values obtained with this technique. Thus, IRLS tends not to increase FPR in either the presence or absence of

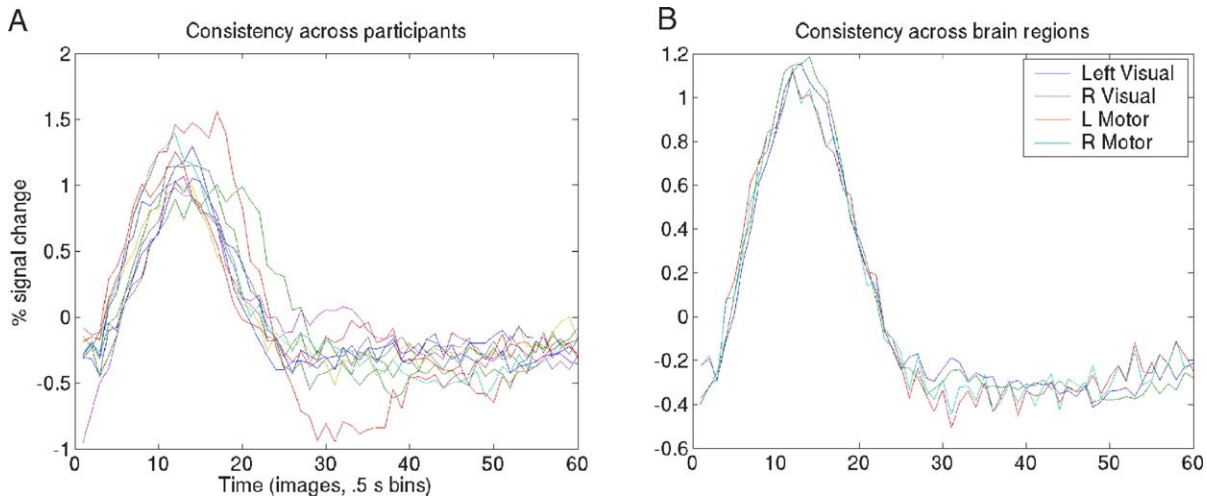

Fig. 6. (A) Hemodynamic response function (HRF) estimates for each participant, averaged across four visual and motor regions of interest, for a series of 3 visual/motor responses spaced 1 s apart. (B) HRF estimates for each region, averaged across participants. Other series lengths showed similar consistency, although consistency was noticeably lower for 1-stimulus series.

outliers, which means that *P* values obtained from IRLS can be used with confidence and can be interpreted in the same way that they are interpreted when using OLS.

The only exception to the non-inflation of FPRs when using IRLS occurs when multivariate outliers are present in the data, such as might occur when response variables (e.g., fMRI data) are regressed on continuous predictors (e.g., behavioral performance data). In this case, FPRs with all techniques, including IRLS, can be much higher than expected. Our recommendation in this case is to carefully check for the presence of outliers in behavioral data, as these problems occur when data are outliers in the behavioral predictor and in brain data.

The simulations show that FPR and power do not always directly trade off across regression techniques—a technique may

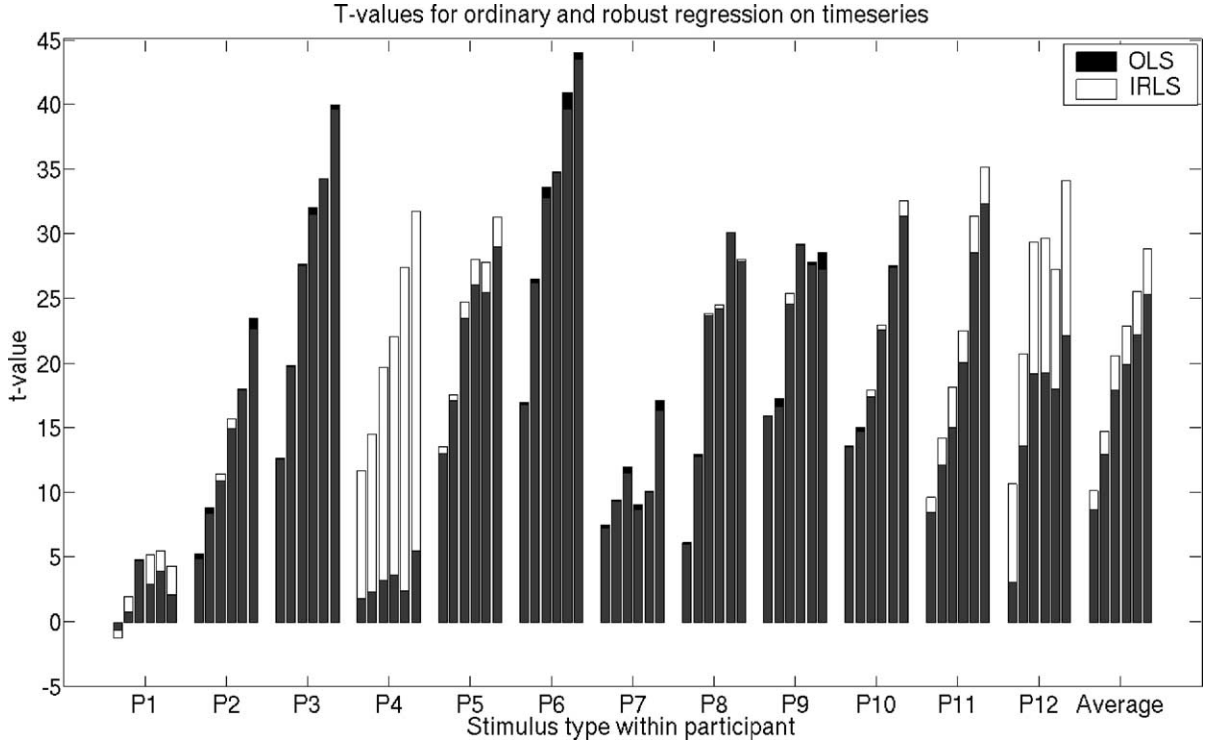

Fig. 7. Experiment 3. Results comparing ordinary least squares (OLS) to robust iteratively reweighted least squares (IRLS) performed on individual participants' time series. The data were taken from a left-hemisphere visual region on the cuneate gyrus defined in each individual participant based on an independent task. Bars show the average *t* value across participants, and error bars show the standard error across participants. Analysis was conducted on high-pass filtered, voxel-averaged time series data. Linear models constructed by convolving the canonical HRF with stimulus onset functions (e.g., a typical SPM model) were fit to time series data. The stimuli were contrast-reversing checkerboards (16 Hz, 250 ms duration) occurring every 1 s in trains of 1, 2, 5, 6, 10, or 11 consecutive stimuli, separated by 30 s of rest. Robust IRLS shows an advantage over OLS for each of the six independent contrasts.

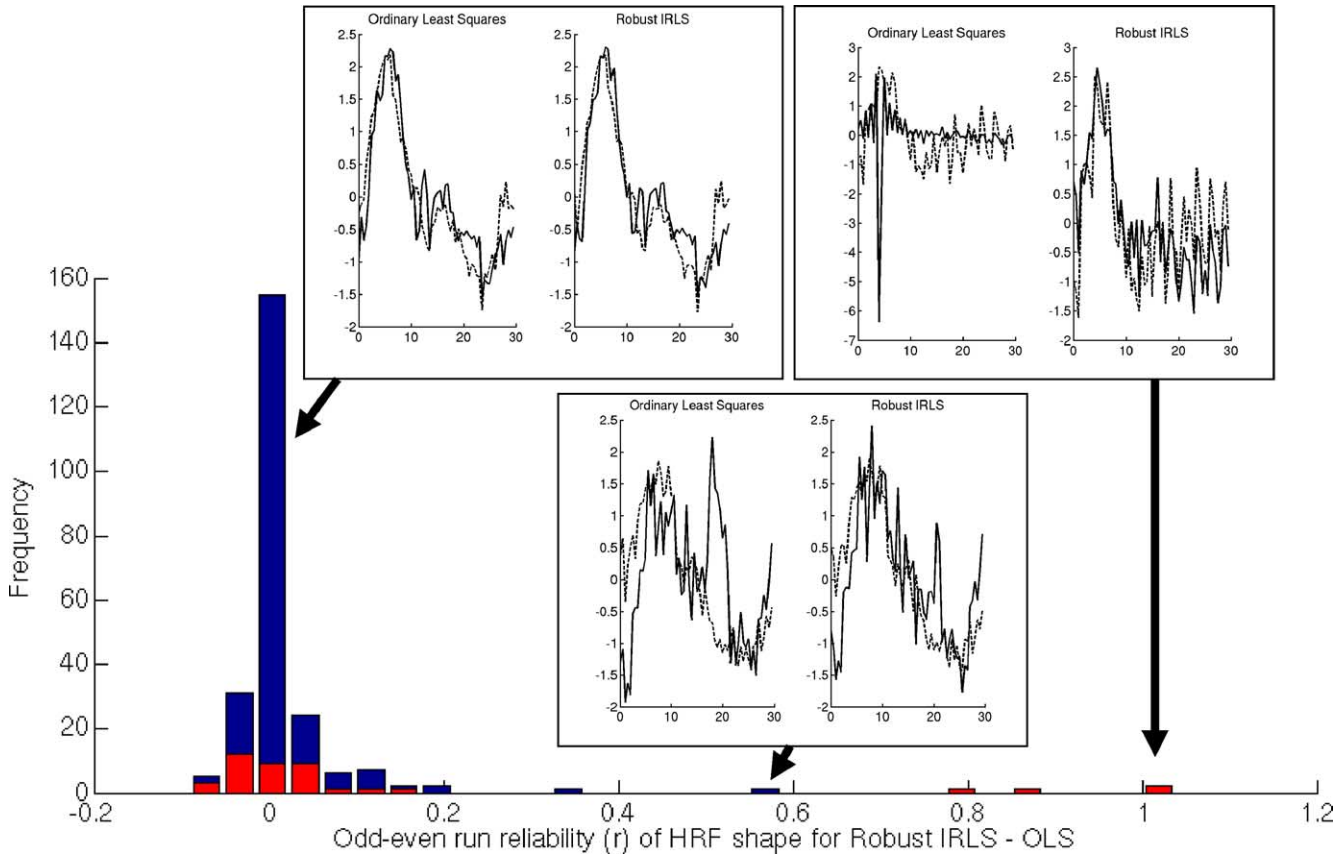

Fig. 8. Frequency histogram showing the difference in HRF estimation reliability for IRLS-OLS. Reliability for each analysis was computed as the correlation between estimated HRF shapes for odd and even runs. Positive reliability differences indicate better reliability for IRLS. Frequencies for all 6 stimulation lengths combined are shown in blue, and those for 1 s stimulation only are overlaid in red. Inset 1 shows typical HRF estimates for one subject, in which OLS and robust estimates were very similar. Solid lines are from odd runs (1, 3, 5), and dashed lines are from even runs (2, 4, 6). Insets 2 and 3 show cases in which IRLS produced much higher reliability scores due to transient artifacts that influence the OLS estimates.

have both low FPR and high power. This idea may be counter-intuitive, in that within a given technique, lowering the alpha level always increases power. However, this same relationship does not hold when comparing across different statistical techniques, as the validity and sensitivity of a technique are functions of different aspects of the fit between technique and data.

To drop, or not to drop?

The simulation results show that power is maximized when outliers are dropped according to the same criteria used in creating outliers. Dropping data points based on Mahalanobis distance maximizes power in the presence of multivariate outliers and dropping data points based on z scores maximizes power in the presence of univariate outliers. However, if the distribution from which outliers are drawn does not match the technique used to drop outliers, dropping outliers are much less powerful.

More troublingly, outlier removal techniques can substantially elevate FPRs. For overall activation (when considering the intercept term), Mahal led to unacceptably high FPR levels. When considering the regression slope, dropping univariate outliers produced comparable FPRs to the other techniques assessed and thus offers no advantage over these techniques. However, dropping data points based on Mahalanobis distance produced mixed results. It was the only technique to substantially inflate FPR in the presence of univariate outliers, but was also the only technique that produced FPRs close to the nominal value of 0.05 in the presence of multivariate outliers.

The mixed results of FPRs for Mahal are inherent to the way that the technique removes points that do not fit the pattern of the data. When no true pattern exists (in H_0 datasets), this technique tends to capitalize on chance patterns and thus inflates FPRs. This represents a major drawback to removal of points based on Mahalanobis distance. However, when multivariate outliers are present in the data, other techniques tend to be overly influenced by multivariate outliers (thus inflating FPRs), while Mahal recognizes these points as aberrant and removes them.

An alternative to dropping outliers is to Windsorize them, or to adjust values whose absolute distance from the mean is above a certain limit (e.g., 3 standard deviations) down to the limit. While this is a potentially useful technique in the univariate family, with many of the same advantages and disadvantages as Univ, we do not explore it further here.

Robust technique recommendations

One important problem with removing or down-weighting observations that appear to be outliers is that, as extreme values, those points are the most informative ones in the sample and have the most ability to support or disconfirm a model. On the other

hand, many brain researchers are interested in characteristics of typical representatives of a population, and individual outlying scores may arise from a number of theoretically uninteresting factors. Individuals may have abnormal vascular responses (due, e.g., to hypertension), abnormal BOLD responses due to high hematocrit levels, sleepiness, breath-holding, or use of caffeine or other drugs. How should we deal with observations that appear to come from a different distribution than the others?

OLS weights each observation equally, but because squared deviations are minimized, data points that are most uncharacteristic of the rest of the data have the most influence. This is highly undesirable when artifactual outliers are likely to exist in the data. On the other hand, dropping outliers completely removes all influence of those observations that do not fit the pattern of the rest of the data, and thus can inflate FPRs. The IRLS techniques offer an intermediate solution, down-weighting influential points without removing them altogether. The bisquare weighting function tends to offer more power than the Huber function at the cost of very slightly inflated FPRs (although they are still below the OLS FPRs). In addition, low IRLS weights can be used as a way to earmark individual participants for careful checking—for example, review of data for artifacts, normalization errors, and other problems.

Perhaps most importantly, IRLS leads to substantially greater power than OLS when outliers are present, but shows no diminution in power when outliers are absent. We conclude that IRLS is a good overall compromise between OLS and outlier removal, that *P* values from IRLS appear trustworthy, and that IRLS can help uncover otherwise hidden patterns in the data when outliers are likely to be present. We suggest that IRLS is an underutilized technique that offers few disadvantages but important advantages in the analysis of neuroimaging data.

Extensions of robust regression: nonparametric, hierarchical, and multivariate techniques

Robust techniques can also be appropriate for use in multivariate analyses, in which a number of components are extracted from a larger set of variables (Lin et al., 2003; McKeown et al., 1998; Peltier et al., 2003). In these cases, it is often difficult to check for multivariate outliers and multivariate normality. Furthermore, multivariate analyses are exquisitely sensitive to outliers—much more so than the one-sample *t* test—as they operate on covariance values among continuous variables. Although several groups have developed robust versions of multivariate techniques (Hubert and Vanden Branden, 2003; Hubert and Verboven, 2003; Hubert et al., 2002), we focus here on the former, "massively univariate" approach. And, although we focus on fMRI and PET, other kinds of methods such as ERP could benefit from employing robust techniques if artifactual outliers are expected at some time points.

Also, although we focus here on ordinary statistical techniques, even nonparametric techniques such as permutation tests (Hayasaka et al., 2004; Nichols and Holmes, 2002) are not immune to the effects of outliers, which influence the distribution of statistics used for thresholding and can result in substantial decreases in power.

As a final note, IRLS offers another important advantage over OLS in neuroimaging analysis. The IRLS technique utilizes a weight matrix (**W**) to minimize the influence of outliers. The same weight matrix is also used to provide several generalizations of the GLM that might be relevant to other neuroimaging applications. Weights on the off-diagonals can be used to pre-whiten the data and model before estimation if error terms have a known (or estimated) covariance structure. This particularly applies to fMRI time series models. Autoregressive models employed in several statistical programs incorporate constrained estimates of temporal autocorrelation into W. Hierarchical models (FSL) use the same framework, carrying error variance estimates forward to higher levels as weights in W. Thus, IRLS can be incorporated into existing neuroimaging analysis software (e.g., SPM, fMRI, BrainVoyager, AFNI) with relative ease.

Acknowledgments

We would like to thank Martin Lindquist for his helpful advice. This research was supported by grant MH60655 to the University of Michigan (John Jonides, P.I.). Software is available from: http://www.columbia.edu/cu/psychology/tor/.

References

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. 7 (4), 254–266.

Becerra, L., Breiter, H.C., Wise, R., Gonzalez, R.G., Borsook, D., 2001. Reward circuitry activation by noxious thermal stimuli. Neuron 32 (5), 927–946.

Buchel, C., Coull, J.T., Friston, K.J., 1999. The predictive value of changes in effective connectivity for human learning. Science 283 (5407), 1538–1541.

Ciuciu, P., Poline, J.-B., Marrelec, G., Idier, J., Pallier, C., Benali, H., 2003. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. IEEE Trans. Med. Imag. 22 (10), 1235–1251.

Davis, K.D., Kwan, C.L., Crawley, A.P., Mikulis, D.J., 1998. Event-related fMRI of pain: entering a new era in imaging pain. NeuroReport 9 (13), 3019–3023.

DuMouchel, W.H., O'Brien, F.L., 1989. Integrating a robust option into a multiple regression computing environment. Paper presented at the Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface, American Statistical Association.

Frank, L.R., Buxton, R.B., Kerber, C.W., 1993. Pulsatile flow artifacts in 3D magnetic resonance imaging. Magn. Reson. Med. 30 (3), 296–304.

Frank, L.R., Buxton, R.B., Wong, E.C., 2001. Estimation of respiration-induced noise fluctuations from undersampled multislice fMRI data. Magn. Reson. Med. 45 (4), 635–644.

Friston, K.J., Frith, C.D., Turner, R., Frackowiak, R.S., 1995. Characterizing evoked hemodynamics with fMRI. NeuroImage 2 (2), 157–165.

Gusnard, D.A., Raichle, M.E., 2001. Searching for a baseline: functional imaging and the resting human brain. Nat. Rev., Neurosci. 2 (10), 685–694.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage 22 (2), 676–687.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1985. Exploring Data Tables, Trends, and Shapes. Wiley, New York.

Holland, P.W.W.R.E., 1977. Robust regression using iteratively reweighted least-squares. Commun. Stat., Theory Methods A6, 813–827.

Huber, P.J., 1981. Robust Statistics. Wiley-Interscience, New York.

Hubert, M., 2001. Multivariate outlier detection and robust covariance matrix estimation—Discussion. Technometrics 43 (3), 303–306.

Hubert, M., Vanden Branden, K., 2003. Robust methods for partial least squares regression. J. Chemom. 17 (10), 537–549.

Hubert, M., Verboven, S., 2003. A robust PCR method for high-dimensional regressors. J. Chemom. 17 (8–9), 438–452.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemom. Intell. Lab. Syst. 60 (1–2), 101–111.

Hubert, M., Rosseeuw, P.J., Val Aelst, S., 2004. Robustness. In: Sundt, B.T., Teugels, J. (Eds.), Encyclopedia of Actuarial Sciences. Wiley, New York.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fMRI: a quantitative evaluation. NeuroImage 16 (1), 217–240.

Jensen, J., McIntosh, A.R., Crawley, A.P., Mikulis, D.J., Remington, G., Kapur, S., 2003. Direct activation of the ventral striatum in anticipation of aversive stimuli. Neuron 40 (6), 1251–1257.

Kochunov, P., Lancaster, J., Thompson, P., Toga, A.W., Brewer, P., Hardies, J., et al., 2002. An optimized individual target brain in the Talairach coordinate system. NeuroImage 17 (2), 922–927.

Kruger, G., Glover, G.H., 2001. Physiological noise in oxygenation-sensitive magnetic resonance imaging. Magn. Reson. Med. 46 (4), 631–637.

Langenberger, K.W., Moser, E., 1997. Nonlinear motion artifact reduction in event-triggered gradient-echo fMRI. Magn. Reson. Imag. 15 (2), 163–167.

Le, T.H., Hu, X., 1996. Retrospective estimation and correction of physiological artifacts in fMRI by direct extraction of physiological activity from MR data. Magn. Reson. Med. 35 (3), 290–298.

Lin, F.H., McIntosh, A.R., Agnew, J.A., Eden, G.F., Zeffiro, T.A., Belliveau, J.W., 2003. Multivariate analysis of neuronal interactions in the generalized partial least squares framework: simulations and empirical studies. NeuroImage 20 (2), 625–642.

Luo, W.L., Nichols, T.E., 2003. Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage 19 (3), 1014–1032.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al., 2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. London, B Biol. Sci. 356 (1412), 1293–1322.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., et al., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp. 6 (3), 160–188.

McKeown, M.J., Hansen, L.K., Sejnowsk, T.J., 2003. Independent component analysis of functional MRI: what is signal and what is noise? Curr. Opin. Neurobiol. 13 (5), 620–629.

Neter, J., Kutner, M.H., Wasserman, W., Nachtsheim, C.J., 1996. Applied Linear Statistical Models, fourth ed. McGraw-Hill/Irwin.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.

Noll, D.C., Cohen, J.D., Meyer, C.H., Schneider, W., 1995. Spiral K-space MR imaging of cortical activation. J. Magn. Reson. Imaging 5 (1), 49–56.

Ojemann, J.G., Akbudak, E., Snyder, A.Z., McKinstry, R.C., Raichle, M.E., Conturo, T.E., 1997. Anatomic localization and quantitative analysis of gradient refocused echo-planar fMRI susceptibility artifacts. NeuroImage 6 (3), 156–167.

Peltier, S.J., Polk, T.A., Noll, D.C., 2003. Detecting low-frequency functional connectivity in fMRI using a self-organizing map (SOM) algorithm. Hum. Brain Mapp. 20 (4), 220–226.

Peyron, R., Frot, M., Schneider, F., Garcia-Larrea, L., Mertens, P., Barral, F.G., et al., 2002. Role of operculoinsular cortices in human pain processing: converging evidence from PET, fMRI, dipole modeling, and intracerebral recordings of evoked potentials. NeuroImage 17 (3), 1336–1346.

Ploghaus, A., Tracey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., et al., 1999. Dissociating pain from its anticipation in the human brain. Science 284 (5422), 1979–1981.

Porro, C.A., Baraldi, P., Pagnoni, G., Serafini, M., Facchin, P., Maieron, M., et al., 2002. Does anticipation of pain affect cortical nociceptive systems? J. Neurosci. 22 (8), 3206–3214.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proc. Natl. Acad. Sci. U. S. A. 98 (2), 676–682.

Rocke, D.M., Woodruff, D., 1996. Identification of outliers in multivariate data. J. Am. Stat. Assoc. 91, 1047–1061.

Schneider, F., Habel, U., Holthusen, H., Kessler, C., Posse, S., Muller-Gartner, H.W., et al., 2001. Subjective ratings of pain correlate with subcortical-limbic blood flow: an fMRI study. Neuropsychobiology 43 (3), 175–185.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., et al., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17 (1), 479–489.

Street, J.O., Carroll, R.J., Ruppert, D., 1988. A note on computing robust regression estimates via iteratively reweighted least squares. Am. Stat. 42, 152–154.

Toga, A.W., Thompson, P.M., 2002. New approaches in brain morphometry. Am. J. Geriatr. Psychiatry 10 (1), 13–23.

Ward, H.A., Riederer, S.J., Jack Jr., C.R., 2002. Real-time autoshimming for echo planar timecourse imaging. Magn. Reson. Med. 48 (5), 771–780.

Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M., 2004. Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Trans. Med. Imag. 23 (2), 213–231.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—Again. NeuroImage 2 (3), 173–181.

Worsley, K.J., Poline, J.B., Friston, K.J., Evans, A.C., 1997. Characterizing the response of PET and fMRI data using multivariate linear models. NeuroImage 6 (4), 305–319.

Wu, D.H., Lewin, J.S., Duerk, J.L., 1997. Inadequacy of motion correction algorithms in functional MRI: role of susceptibility-induced artifacts. J. Magn. Reson. Imaging 7 (2), 365–370.