Comment on *A Bayesian time course model for functional magnetic resonance imaging* by Chris Genovese

K.J. Worsley
McGill University

July 15, 2007

The challenge

It is new types of data, such as the agricultural field trials of a hundred years ago, that stimulate new research in statistics. Functional Magnetic Resonance Imaging (fMRI), and brain mapping in general, presents statisticians with new challenges that might make them the agricultural field trial of the next decade. The statistical problems are the usual ones: we are presented with observations taken in time while different 'treatments' are applied, possibly repeated on the same or different subjects ('blocks'), and we are asked to make inferences about an underlying model for the treatments.

What makes it challenging is that the observations are *3D images*, rather than univariate 'yields'. The data is highly multivariate: there are typically 100,000 to 400,000 components (voxels) to each observation, so finely sampled that the data is best regarded as coming from a continuous 3D function. Conventional multivariate statistics cannot be applied because the number of components far exceeds the number of observations, though some have tried by restricting the images to a small number of pre-selected isolated regions ('plots'), such as Broca's or Wernicke's areas.

The strengths

This paper is the first to my knowledge that attempts a fully Bayesian analysis of fMRI data. Chris Genovese has done a superb job of applying modern hierarchical Bayes models to the analysis of fMRI data. The priors are chosen not for convenience but to reflect true prior knowledge about the likely values of parameters. The advantage is that much more subtle inference can be drawn from the data, rather than simple localization via thresholding.

In defence of thresholding, the threshold value of ± 4 is probably too low. Assuming a Gaussian distribution at $\approx 15,000$ voxels, the chance of finding false activation is not more than $15000\Phi(-4) = 0.475$. To cut this down to 0.05, a threshold of $-\Phi^{-1}(0.05/15000) = 4.5$ might be better. This may remove some of the apparent false activation in Figure 4c.

Admittedly it is a little strange to answer a localization problem with a hypothesis test, but the test is designed so that detecting false activation in the unactivated part of the image is controlled (conservatively) to a pre-specified probability of say $\alpha = 0.05$. Testing for the number of connected voxels above a threshold accomplishes much the same thing (Cao, 1999).

Nevertheless, I agree that it is not easy to detect monotonicity using frequentist methods. It would of course be easy to detect a *trend* by simply regressing on a covariate taking nominal values 0, 1, 2, 3 for the four conditions, but this is not the same thing. One might be able to carry out tests for monotonic alternatives as a special type of cone alternative. Lin and Lindsay (1997) and Takemura and Kuriki (1997) have found the exact distribution of such test statistics, so it should be possible to repeat this at all voxels.

The weaknesses

Three weaknesses are addressed. The first is the issue of the computer time required to analyze a single run, requiring about one day. It should be noted that fMRI experiments, which only last a few minutes, are often repeated in several runs in the same session, several sessions on the same subject, and several subjects drawn from a population. The total computer time then becomes formidable.

The second is the noise model. At the moment, the method assumes independent temporal observations. In our data, we have detected temporal correlations ranging from 0.2 to 0.4 over a time period of 3 seconds, twice as long as here (Figure 1). Neglecting this autocorrelation causes biases in the estimated standard deviation of parameter estimates, usually resulting in smaller reported standard errors and a higher tendency to find activation where none exists.

Structural independence of parameters is mentioned as a third weakness. Neils Væver Hartvig of Aarhus has developed a stochastic geometry model for fMRI data that attempts to model activation as a sum of Gaussian densities of unknown location and scale. These are fitted using Bayesian methods with a marked point process as prior for the centres (Hartvig, 1999). A different approach has been developed by Vic Solo, Emery Brown, Patrick Purdon and Robert Weisskoff at the Massachusetts General Hospital (Solo *et al.*, 1999). Their analysis is strictly frequentist, involving a parametrized hemodynamic response function convolved with the stimulus, drift removal by regression, and an AR(1) plus white noise model for the errors. The interesting feature is the way in which the parameters are estimated by pooling the *likelihood* over adjacent voxels. The extent of the pooling is governed by a spatial kernel whose width is estimated by minimizing an estimator of the expected Kullback-Liebler information. The result is a form of regularization which spatially smooths the noise parameters but not the signal parameters. Once again, both these method are very computationally intensive, making them prohibitive for routine use.

Statistical Parametric Mapping (SPM)

Software for performing a simple statistical analysis of fMRI data has been available for several years in the SPM (Statistical Parametric Mapping) package. Developed by Karl Friston at the Functional Imaging Laboratory, University College, London, it is used in

the majority of labs worldwide, becoming the 'SAS' of the brain mapping community. It takes raw fMRI data, performs motion correction, then non-linear warping to re-align the data to a common template. The stimuli are convolved with a pre-specified hemodynamic response function to form a design matrix for a linear model. Drift is removed by low frequency filtering, and the analysis is made robust to the temporal correlation structure by high frequency filtering. Fitting is by least squares, with a standard deviation corrected for temporal correlation, modelled as an AR(1) process with a single global parameter. Activation is detected by thresholding t statistic images, or the spatial extent of adjacent voxels above a threshold, using random field theory to set the thresholds.

Most of the SPM package is written in Matlab, making it easy to write while maintaining good execution speed. There is a GUI, on-line help, e-mail assistance, well-attended annual workshops, excellent course notes, and a good book that describes the entire analysis (Frackowiak *et al.,* 1997). SPM is distributed free of charge. It is perhaps these aspects, just as much as publication in academic journals, that have ensured the success of SPM. It has captured the market for the non-statistician researcher who wants to get the experiments analyzed and the results published. It is this market that statisticians must penetrate.

A compromise

We have attempted to do so by writing our own software for the routine analysis of fMRI data. It has been in use by our researchers for the last year. It tries to combine the best aspects of all the above methods by seeking a compromise between validity, generality, simplicity and execution speed. The software consists of three main Matlab programs, available on the web (http://www.bic.mni.mcgill.ca/users/keith) that take only a few minutes of computer time to analyze each run. The method is based on that of SPM, but with local regularized AR(1) parameters, and a form of regularized random effects analysis for combining runs in the same session, sessions on the same subject, and subjects drawn from a population. Pre-processing, such as motion correction, is performed by other software.

FMRIDESIGN creates suitable design matrices for linear models for the temporal data. The time courses of the stimuli are convolved with a hemodynamic response function, taken to be the difference of two gamma functions, one to capture the main response, the second to capture the dip:

$$h(t) = (t/p_1)^{a_1} \exp((t-p_1)/b_1) - c(t/p_2)^{a_2} \exp((t-p_2)/a_2)).$$

where t is time in seconds, $p_j = a_j b_j$ is the time to the peak, and by default $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$ seconds, $c = 0.35$ (Glover, 1999). A certain amount of flexibility can be incorporated by expanding $h(t)$ as a Taylor series in an unknown scale s:

$$e^{-s} h(te^{-s}) \approx h(t) + s(-h - t\partial h(t)/\partial t).$$

We can then convolve the stimuli with $-h - t\partial h(t)/\partial t$ and add these to the model, which allows for different scales for different types of stimuli. Finally, the regressors for the linear model are formed by subsampling the convolved stimuli at the slice acquisition times.

FMRISTAT performs a statistical analysis of a single run of an fMRI data set. Drift is removed by adding polynomial regressors in the scan times, up to degree 3 by default, to

the design matrix. It is important to note that these drift terms are not convolved with the hemodynamic response function. The correlation structure is modeled as an AR(1) process; this is much easier to fit than an AR(1) plus white noise model. While there was evidence for an AR(2) model, parameter estimates and their standard errors were almost the same as for the AR(1). At each voxel, the autocorrelation parameter was estimated from the least squares residuals using the Yule-Walker equations, after a bias correction for correlation between the residuals induced by the linear model (see Appendix). Without this correction, the autocorrelation parameter was negatively biased by about -0.05. The autocorrelation parameter was then regularized by spatial smoothing with a 6mm (default) standard deviation Gaussian filter (Figure 1), then used to 'whiten' the data and the design matrix. The linear model was then re-estimated by least squares in a second run through the whitened data, to produce estimates of effects and their standard errors. This second run requires finding the pseudoinverse of a different whitened design matrix at each voxel. To reduce computations, this was evaluated only for the autocorrelation parameter rounded to the nearest second decimal, then cached.

MULTISTAT combines the output from FMRISTAT across runs within sessions, sessions within subjects, then finally subjects within a population in a hierarchical random effects analysis. Runs within a session are combined with another linear model for the run effects (as data), weighted inversely by the square of their standard errors from FMRISTAT. The degrees of freedom of the standard errors are quite large (> 100), but the number of runs in a session is quite small (< 10), so a straightforward random effects analysis would leave too few residual degrees of freedom. Instead, a regularized random effects analysis was performed by first estimating the ratio of the random effects variance σ^2_{random} to the fixed effects variance σ^2_{fixed}, then regularizing this ratio by spatial smoothing with a Gaussian kernel with a standard deviation of $w_{\text{ratio}} = 6.4$mm (default) (Figure 2). The residual variance was then estimated by

$$\sigma^2_{\text{residual}} = \sigma^2_{\text{fixed}} \text{smooth}(\sigma^2_{\text{random}}/\sigma^2_{\text{fixed}})$$

to achieve higher degrees of freedom at the expensive of introducing a small bias. The effective degrees of freedom of the smoothed variance ratio was estimated by assuming that the spatial correlation function of the fMRI data was a Gaussian function with standard deviation w_{data} (2.5mm by default):

$$\nu_{\text{ratio}} = \nu_{\text{random}}(2(w_{\text{ratio}}/w_{\text{data}})^2 + 1)^{3/2},$$

where ν_{random} is the degrees of freedom of the second linear model for the runs. The final effective degrees of freedom of the residuals, ν_{residual}, is estimated by

$$1/\nu_{\text{residual}} = 1/\nu_{\text{ratio}} + 1/\nu_{\text{fixed}},$$

where ν_{fixed} is the residual degrees of freedom of the fixed effects analysis, equal to the sum of the degrees of freedom of the preceding separate FMRISTAT analyses. Thus the w_{ratio} parameter acts as a convenient way of providing an analysis mid-way between a random effects and a fixed effects analysis; setting $w_{\text{ratio}} = 0$ (no smoothing) produces a random effects analysis; setting $w_{\text{ratio}} = \infty$, which smooths the variance ratio to one everywhere, produces a fixed effects analysis. In practice, we choose w_{ratio} to produce a final ν_{residual}

which is at least 100, so that errors in its estimation do not greatly affect the distribution of test statistics. Sessions within subjects are combined with another MULTISTAT analysis, and subjects are combined within a population in a final MULTISTAT analysis.

TSTAT_THRESHOLD provides a threshold for the resulting t statistic image chosen so that the probability of finding activation in unactivated regions is controlled conservatively to a prespecified level, say $\alpha = 0.05$ (Figure 3). The method uses the minimum given by a Bonferroni correction and random field theory, which approximates the probability of exceeding the threshold by the expected Euler characteristic of the set of voxels above the threshold (Adler, 1981, 1999; Worsley, 1995; Worsley *et al.*, 1996).

Statistics and brain mapping

My hope is that more statisticians might get involved with brain mapping data. New types of data are being created all the time. Besides fMRI, there is the older PET methodology which can target brain chemistry rather than simply activation. Non-linear models are needed to fit compartment models that describe the transfer and uptake of tracer-labelled neurotransmitters. John Aston at Imperial College has started a careful statistical analysis of this type of data (Aston *et al.*, 1999). Magnetic Transfer Ratios can detect changes in multiple sclerosis (MS) lesions before the lesions actually appear (Pike *et al.*, 1999). The appearance of MS lesions themselves represent fascinating challenges for applying survival analysis to image data, where there is a hazard function at every point in the 3D image. Finally, there is interest in functional connectivity, defined as high correlation between distant brain voxels (Cao and Worsley, 1999a).

There is also great interest in brain *structure* rather than brain *function*. It is thought that different brain regions might shrink in response to disease (e.g. Alzheimers or MS) or grow in response to age (growth curves might be useful here) or in response to a stimulus (pianists have larger motor cortex for finger control). The simplest type of structural data is binary masks, where a value of 1 is assigned to the structure of interest (e.g. cortex) and 0 elsewhere. Jonathan Taylor has analyzed this type of data using logistic regression at every voxel with a quasi-binomial likelihood (Taylor *et al.*, 1998; Paus *et al.*, 1999). The outer surface of brain structures can now be extracted using sophisticated methods that shrink-wrap a triangular mesh around the cortex. Surface normal displacements can be used for statistical analysis, but now the data sits on a 2D curved manifold, rather than a uniform grid in 3D space, which presents fascinating challenges for random field theory (Worsley *et al.*, 1996). Vector deformations required to warp the data to an 'atlas' brain encode shape information. Jin Cao (Bell Labs) and myself have treated this as trivariate Gaussian data, and used random fields of Hotelling's T^2 statistics to look for shape differences (Cao and Worsley, 1999b). Paul Thompson at UCLA has used this type of data to look for brain growth in children and adolescents (Thompson *et al.*, 1999). Finally, diffusion tensor imaging now produces 3×3 symmetric matrices at each voxel that indicate the direction and degree of alignment of white matter tracts (Pierpaoli *et al.*, 1996).

Why are so few statisticians working in this area? A few years ago, there were less than 10. I have two reasons. The first is that the amount of data is forbidding. A single fMRI run generates about 100MB of data; this is repeated on the same subject, and over several subjects to generate several gigabytes of data. Statisticians are more used to comfortable

data sets of 100 numbers or less. This type of data is awkward to transmit, it often comes in strange formats, and requires special tools to manipulate. Analysis has to be done in C or Matlab; conventional statistics packages usually cannot cope. Visualizing the results often requires special 3D imaging tools.

The second is that statisticians *must* actually work side by side, in the same lab, as the neuroscientists. As Bill Eddy has said, brain mapping is a 'team sport', involving psychologists who design the experiments, technicians and physicists who operate the scanner, computer specialists who handle the data, and statisticians who should (I hope!) carry out the analysis. To make any progress, you have to be on site. This type of work cannot be done at a distance from the comfort of an office in a Department of Statistics. I would like to make a plea for statisticians to get involved, to get their hands 'dirty' with the data, as Fisher did in the fields of Rothamstead. Chris Genovese has clearly done this, and reaped the rewards.

Appendix

For an AR(p) model, the bias can be reduced by equating sample autocovariances to their expectations, as follows. Let $\mathbf{e} = \mathbf{RY}$ be the vector of least squares residuals, where $\mathbf{Y}$ is the data vector at a single voxel, $\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{X}$ is the design matrix. Let $\mathbf{D}_r$ be a matrix of zeros with ones on the rth upper off-diagonal, so that the expectation of the lag r autocovariance, $0 \leq r \leq p$, is

$$\mathrm{E}(\mathbf{e}'\mathbf{D}_r\mathbf{e}) = \mathrm{tr}(\mathbf{R}\mathbf{D}_r\mathbf{R}\mathbf{V}) \approx v_0\mathrm{tr}(\mathbf{R}\mathbf{D}_r) + \sum_{j=1}^{p} v_j \mathrm{tr}(\mathbf{R}\mathbf{D}_r\mathbf{R}(\mathbf{D}_j + \mathbf{D}_j')),$$

where $\mathbf{V} = \mathrm{Var}(\mathbf{Y})$ is approximated by v_j on the jth off-diagonal, $0 \leq j \leq p$, and zero elsewhere. Equating $\mathbf{e}'\mathbf{D}_r\mathbf{e}$ to its approximate expectation for $0 \leq r \leq p$ and solving for v_j, $0 \leq j \leq p$, gives estimates of the autocovariances. Inserting these into the Yule-Walker equations gives estimates of the AR(p) parameters which were shown by simulation to have much reduced bias.

Additional references

Adler, R.J. (1981). *The Geometry of Random Fields.* Wiley, New York.

Adler, R.J. (2000). On excursion sets, tube formulae, and maxima of random fields. *Annals of Applied Probability*, in press.

Aston, J.A.D., Worsley, K.J., Bleicher, C., Ma, Y., Gunn, R.N., Evans, A.C.& Dagher, A. (1999). Statistical methods for measuring neurotransmitter release with Positron Emission Tomography. *NeuroImage*, **9**:S208.

Cao, J. (1999). The size of the connected components of excursion sets of χ^2, t and F fields. *Advances in Applied Probability*, **31**:577-593.

Cao, J. & Worsley, K.J. (1999a). The geometry of correlation fields, with an application to functional connectivity of the brain. *Annals of Applied Probability*, **9**:1021-1057.

Cao, J. & Worsley, K.J. (1999b). The detection of local shape changes via the geometry of Hotelling's T^2 fields. *Annals of Statistics*, **27**:925-942.

Collins, D.L., Zijdenbos, A.P. & Evans, A.C. (1998). Improved automatic cerebral structure segmentation. *NeuroImage*, **7:**S707.

Collins, D.L., Holmes, C.J., Peters, T.M. & Evans, A.C. (1996). Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, **3:**190-208.

Frackowiak, R.S.J., Friston, K.J., Frith, C.D., Dolan, R. & Mazziotta, J.C. (Eds.) (1997). *Human Brain Function.* Academic Press, New York.

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

Hartvig, N.V. (1999). A stochastic geometry model for fMRI data. *Research report No. 410, Department of Theoretical Statistics, University of Aarhus,* submitted for publication.

Lin, Y. & Lindsay, B.G. (1997). Projections on cones, chi-bar squared distributions and Weyl's formula. *Statistics & Probability Letters*, **32:**367-376.

MacDonald, D., Avis, A. & Evans, A.C. (1998). Automatic segmentation of cortical surfaces from MRI with partial volume correction. *NeuroImage*, **7:**S703.

Paus, T., Zijdenbos, A., Worsley, K.J., Collins, D.L., Blumenthal, J., Giedd, J.N., Rapoport, J.L. and Evans, A.C. (1999). Structural maturation of neural pathways in children and adolescents: In vivo study. *Science*, **283**:1908-1911.

Pierpaoli, C., Jezzard, P., Basser, P.J., Barnett, A. & Di Chirico, G. (1996). Diffusion tensor MR imaging of the human brain. *Radiology*, **201:**637-648.

Pike, G.B., De Stefano, N., Narayanan, S., Worsley, K.J., Pelletier, D., Francis, G.S., Antel, J.P. & Arnold, D.L. (1999). Imaging pre-lesional pathology in multiple sclerosis using magnetization transfer. *Radiology*, in press.

Solo, V., Purdon, P., Brown, E., and Weisskoff, R. (1999). A signal estimation approach to functional MRI. *IEEE Transactions on Medical Imaging*, submitted.

Takemura, A. & Kuriki, S. (1997). Weights of $\bar{\chi}^2$ distribution for smooth or piecewise cone alternatives. *Annals of Statistics*, **25:**2368-2387.

Taylor, J., Worsley, K.J., Zijdenbos, A.P., Paus, T. & Evans, A.C. (1998). Detecting anatomical changes using logistic regression of structure masks. *NeuroImage*, **7**:S753.

Thompson, P.M., Giedd, J.N., Woods, R.P., MacDonald, D., Evans, A.C. & Toga, A.W. (1999). Growth patterns in the developing human brain detected using continuum-mechanical tensor mapping. *Nature*, submitted.

Worsley, K.J. (1995). Boundary corrections for the expected Euler characteristic of excursion sets of random fields, with an application to astrophysics. *Advances in Applied Probability*, **27**:943-959.

Worsley, K.J., MacDonald, D., Cao, J., Shafie, Kh. & Evans, A.C. (1996). Statistical analysis of cortical surfaces. *NeuroImage*, **3**:S108.

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

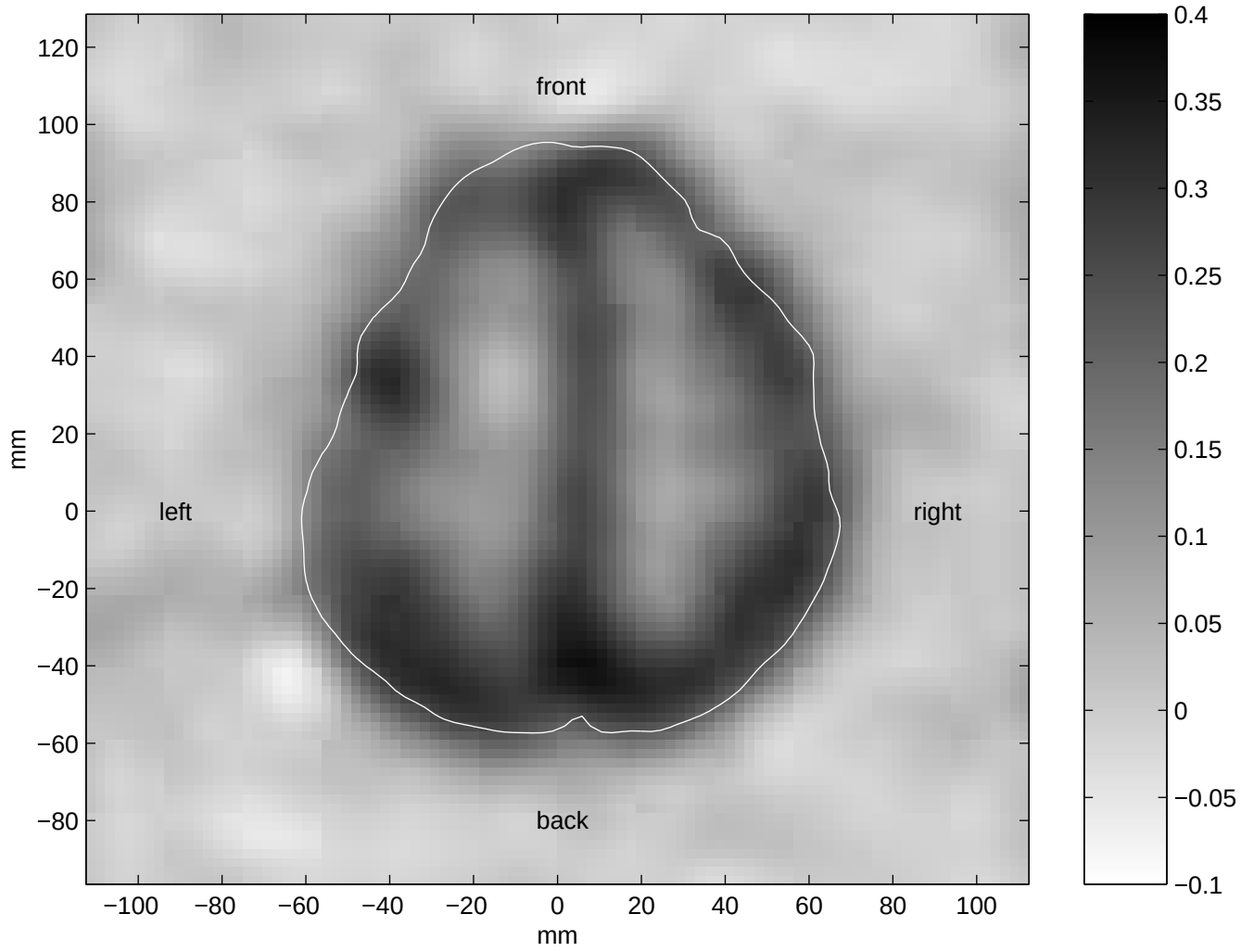

Figure 1: The autocorrelation at a lag of 3 seconds in one horizontal slice, after bias correction and spatial smoothing. The data come from an fMRI experiment to detect pain perception. 120 scans (128×128 2.35mm pixels per slice, 13 slices 7mm apart) were taken every 3 seconds while a painful heat stimulus (3 scans) was alternated with a control warm stimulus (3 scans) each seperated by a baseline (3 scans), repeated 10 times. Note that autocorrelations are 0.2-0.4 in cortical areas.

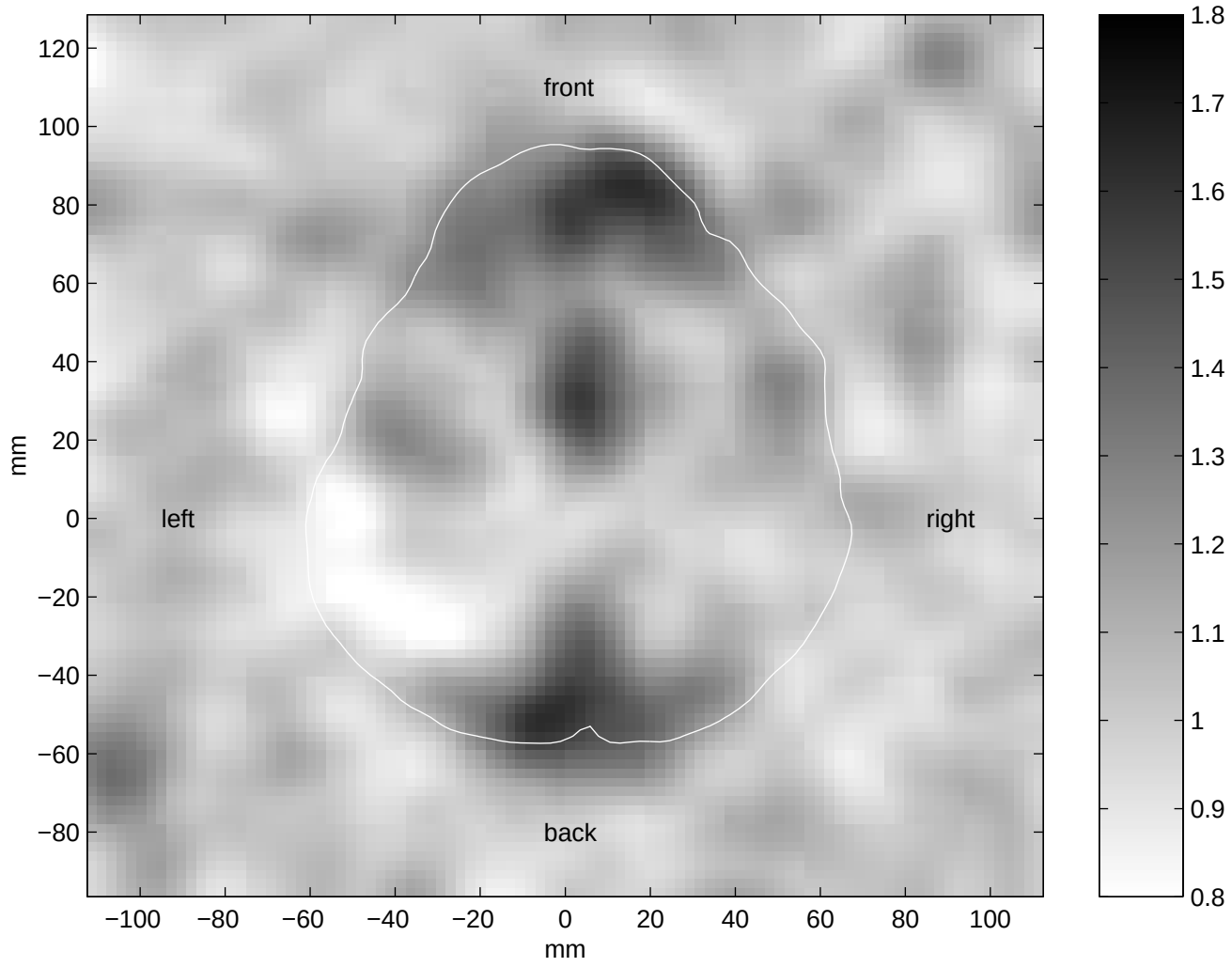

Figure 2: The ratio of random effects standard deviation to fixed effects standard deviation in one horizontal slice after spatial smoothing, for 4 runs of the experiment described in Figure 1. Note that most of the brain shows no evidence of a random effect (ratio ≈ 1), but the random effect due to subjects within a population can be much higher (ratio ≈ 3, not shown).

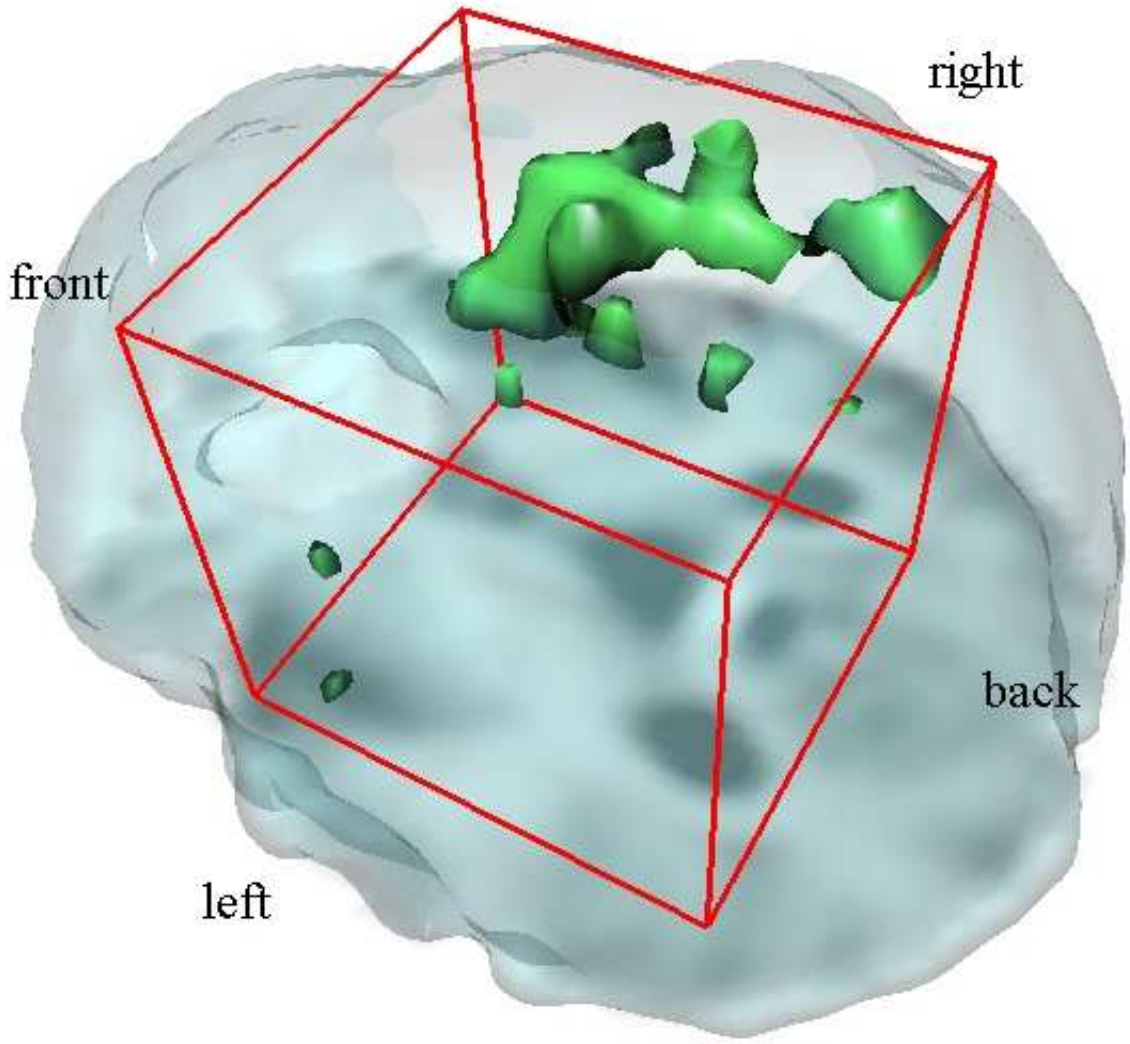

Figure 3: The t statistic from the data described in Figure 2 for detecting a difference between the hot and control (warm) stimulus. The effective degrees of freedom is $\nu_{\text{residual}} = 112$, and the image is thresholded at 4.86, chosen to control the probability of detecting activation in non-activated ares to 0.05.