3-10-2007

MODIFIED TEST STATISTICS BY INTER-VOXEL VARIANCE SHRINKAGE WITH AN APPLICATION TO fMRI

Shu-chih Su  
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, shsu@jhsph.edu

Brian Caffo  
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Elizabeth Garrett-Mayer  
Director of Biostatistics, Hollings Cancer Center of the Medical University of South Carolina

Susan Bassett  
The Johns Hopkins University School of Medicine, Psychiatric Neuroimaging

Suggested Citation  
Su, Shu-chih; Caffo, Brian; Garrett-Mayer, Elizabeth; and Bassett, Susan, "MODIFIED TEST STATISTICS BY INTER-VOXEL VARIANCE SHRINKAGE WITH AN APPLICATION TO fMRI" (March 2007). Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 138.  
http://biostats.bepress.com/jhubiostat/paper138

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.  
Copyright © 2011 by the authors
Modified Test Statistics by Inter-Voxel Variance shrinkage With an Application to fMRI

Shu-chih Su, Brian Caffo, Elizabeth Garrett-Mayer and Susan Spear Bassett

March 10, 2007

Abstract

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique which is commonly used to quantify changes in blood oxygenation and flow coupled to neuronal activation. One of the primary goals of fMRI studies is to identify localized brain regions where neuronal activation levels vary between groups. Single voxel t-tests have been commonly used to determine whether activation related to the protocol differs across groups. Due to the generally limited number of subjects within each study, accurate estimation of variance at each voxel is difficult. Thus, combining information across voxels in the statistical analysis of fMRI data is desirable in order to improve efficiency. Here we construct a hierarchical model and apply an Empirical Bayes framework on the analysis of group fMRI data, employing techniques used in high throughput...
genomic studies. The key idea is to shrink residual variances by combining information across voxels, and subsequently to construct an improved test statistic in lieu of the classical t-statistic. This hierarchical model results in a shrinkage of voxel-wise residual sample variances towards a common value. The shrunken estimator for voxel-specific variance components on the group analyses outperforms the classical residual error estimator in terms of mean squared error. Moreover, the shrunken test-statistic decreases false positive rate when testing differences in brain contrast maps across a wide range of simulation studies. This methodology was also applied to experimental data regarding a cognitive activation task.

**Keywords:** fMRI, GLM, group analysis, hierarchical model, image analysis, shrinkage estimation.

1 Introduction

Functional magnetic resonance imaging (fMRI) is a non-invasive technique for determining changes in blood oxygenation associated with experimental stimuli. This imaging technique has been successfully used to investigate a wide variety of neuronal functions yielding, among other things, a better understanding of a variety of brain pathologies.

In a block design fMRI experiment, a subject is placed in an MRI scanner and asked to complete a task repeated in rapid succession (Aguirre and D'Esposito, 1999) while the MRI scanner takes repeated images. Functional MRI targets the BOLD (Blood Oxygenation
Level Dependent) signal, in contrast with other MRI pulse sequencing schemes (Jackson et al., 1997). The BOLD signal is an important quantity, being an indirect measurement of neuronal activation (Heeger and Ress, 2002). In this manuscript we consider inter-subject analysis; the comparison of fMRI activation maps between subjects. Inter-subject analysis is used to find common areas of brain functions within populations or to compare differences in common areas of paradigm-related activation across populations.

Typical inter-subject fMRI analysis follows in two stages (Friston et al., 2002). In the first stage, an initial voxel (three dimensional pixel) level preprocessing and general linear model analysis is performed within each subject. Thus at this stage, subject level summaries, such as a contrast map is obtained. Further description of the first stage data preprocessing and modeling can be found in Section 3. In the second stage, subject level summaries are then compared across subjects. This two-stage approach has several benefits. For example, it approximates a random effect analysis with the goal of making population-level inter-subject conclusions. Also, the data reduction obtained by reducing the first stage fMRI time series to a single contrast map is enormous. Because of this data reduction, several second stage models can feasibly be fit, incorporating different covariates or other model structures.

Due to the high-throughput nature of the images and longitudinal quality of the fMRI technique, fMRI studies often involve an extremely large amount of data per subject. However, the high cost of imaging and limited scanner and subject availability in fMRI studies usually leads to a small number of subjects within each study. With this limitation, pre-
cise estimation of inter-subject voxel-specific variability is difficult and variance estimates obtained on a single voxel are often based on only few degrees of freedom. As a consequence, the ordinary t-statistics may not be efficient. At the other extreme, it is unlikely to be true that all voxels share equal variance, implying that a globally pooled variance estimate is not appropriate. The proposed methodology strikes a balance between these extremes by employing techniques commonly used in genomic studies. For example, Cui et al. (2005) used James-Stein shrinkage estimation (Efron and Morris, 1977; Lindley, 1962) to construct gene-specific variance estimates for modifying test statistics. A similar strategy has been applied in several microarray experiments (Baldi and Long, 2001; Smyth, 2004; Lonstedt and Speed, 2002; Storey, 2002; Wright and Simon, 2003).

There is relatively less work regarding the estimation of variability in fMRI studies. Nichols and Holmes (2001) employed a locally pooled (smoothed) variance estimate, where voxel information is combined with those of its neighbors to a locally pooled variance estimate to construct a so-called pseudo t-statistics. Weights and a neighborhood structure for the local pooling need to be specified. A similar idea is presented here that allows all voxels to provide information about both within- and between-voxel variation. Unlike the work of Nichols and Holmes (2001), we do not shrink voxels within spatial neighborhoods, in part because variogram fits suggest that residual variances for inter-subject group comparisons had little spatial correlation in our example data.

The purpose of this study is to develop methodology for constructing shrinkage variance
estimates in fMRI studies. The usual variance estimate is replaced with an empirical Bayes estimator based on a hierarchical prior distribution. The shrunken estimates are used to construct an improved signal to noise test statistic.

The article is organized as follows. We first introduce the experimental data used for the demonstration of our proposed approach in Section 2. We then construct the hierarchical linear models for the analysis of designed experiments in Section 3, including the associated model distribution assumptions, the prior specifications and estimators for the hyperparameters. These methods were evaluated with simulation and experimental data in Section 4 and 5.

2 Auditory word-paired-associates learning task

The data used in this manuscript arise from an ongoing study comparing subjects at-risk for Alzheimer’s disease to matched controls in an episodic memory task (Bassett et al., 2002, 2005, 2006). In this task, while in the scanner, participating subjects were asked to remember blocks of unrelated word pairs (the encoding phase). Later, subjects were prompted with the first word of the pair and asked to think of the second word of the pair (the recall phase). Subjects participated in two 6 min and 10 s sessions, each with six trials. Each trial includes an encoding phase and a cued recalled phase. After scanning, subjects were asked to repeat as many word-pair sets as they could remember to validate that they were actively participating in the task.
MRI acquisition targeted the medial temporal, thalamus and cingulate gyrus, the regions involved in episodic memory function. The medial temporal lobe was especially considered, being the region of the brain associated with the earliest pathology in Alzheimer’s disease (Braak H, 1996). Focusing on a smaller imaging area allows the researcher to acquire higher resolution images in the same amount of time.

In the analysis that follows we consider only the 75 right handed controls. Handedness is often addressed in functional imaging analysis because of hemispheric associations (Dassonville et al., 1997; Goodglass and Quadfasal, 1954) between brain function and handedness. The data set included 37 females and 38 males, aged from 48 to 83 years old were. The control subjects had no first degree relatives affected with Alzheimer’s disease and were clinically asymptomatic. We primarily consider the encoding phase of the paradigm, compared to rest. We further focus on group comparisons in this contrast between men and women. Though not a primary aim of the study, this comparison was selected because of well known gender differences in language processing and memory (Gaab et al., 2003; Good et al., 2001).

All subjects were scanned on the same Philips 1.5 T scanner. Eighteen coronal slices were obtained with a 4.5 mm thickness and an inter-slice gap of 0.5 mm. Preprocessing of all images was conducted by the Division of Psychiatric Neuroimaging (details were described in Section 3.1) in the Johns Hopkins Department of Psychiatry using SPM. The study was approved by the Johns Hopkins Institutional Review Board and all subjects provided written
3 Model formulation

3.1 Data preprocessing and first stage modelling

In this paper we focus on second stage fMRI analysis. However, we briefly review the first stage preprocessing and modelling. During preprocessing, the raw imaging data are corrected for non-task-related variability. Specifically, image slices are timed to the fMRI paradigm and head motion is corrected through coregistration. Subsequently the imaging data are realigned, spatially normalized into a standard space, and smoothed for statistical analysis (Friston et al., 1995a). While there are many packages available to perform spatial preprocessing, we used the popular SPM program (see Frackowiak et al., 2003) in the word-paired-associates learning task data.

After preprocessing, the voxel-specific time series are regressed on a design matrix, conceptually having dummy variables set to one when the task is being given. Note that the caveat “conceptually” is necessary since the changes in blood oxygenation are delayed from the onset of the task. To account for this, a haemodynamic response function is convolved with the relevant columns of the design matrix. In addition, slowly varying temporal trend terms are included in the model to serve as a high-pass filter. See Friston et al. (1995b); Frackowiak et al. (2003); Holmes et al. (1997) for more description of the general linear model.
approach.

Contrast values, for example comparing the resting state to the task, or one task to another, are retained for second stage analysis. Comparing estimated contrasts across subjects approximates a random effect analysis (Penny et al., 2003) and is often called “random effect analysis” in the fMRI literature. The analysis would exactly correspond to a random effect model, if standard errors of the contrast maps were retained and incorporated into the second stage analysis. However, typically, the first stage variance estimates are discarded and only the contrast estimates are retained, a convention we adopt throughout. An alternative approach would analyze the normalized (estimate / standard error) contrast values in the second stage.

3.2 Second stage analysis

We consider the inter-subject analysis of contrasts maps. In our case, the contrast considers the activation phase of the task versus rest. In modeling the contrasts, we consider a linear model where estimated change in image intensity from the first stage is the outcome. Let $Y_i(v), i = 1, \ldots, n$ be the contrast values at voxel $v$ for subject $i$ and $X_i^t$ is a $p$-vector of covariates (such as gender, education, age etcetera), which does not vary by voxel. Consider the linear model

$$Y_i(v) = X_i^t \beta(v) + \epsilon_i(v),$$

(1)
where $\epsilon_i(v) \sim N\{0, \sigma^2(v)\}$ and each $\beta(v)$ is a $p$-vector of voxel-specific coefficients. Note that $\epsilon_i(v)$ encompasses both measurement error as well as biological variability of the voxel across subjects.

Our goal is to present an improved estimate of $\sigma^2(v)$ obtained by borrowing information across voxels. While improving estimation of $\sigma^2(v)$ is intrinsically of interest, we also consider the impact that improved variance estimation has on signal-to-noise statistics. In particular, consider the ordinary voxel-specific t-statistic for effect $\beta_j(v)$. That is, the ratio $\hat{\beta}_j(v)/\text{se}\{\hat{\beta}_j(v)\}$. This statistic follows Gossett’s t-distribution with $n - p$ degrees of freedom under independence and normality assumptions where $\text{se}\{\hat{\beta}_j(v)\} = s^2(v)(X^tX)^{-1}$ and $s^2(v)$ is the sum of the squared least squares residuals for voxel $v$ divided by $n - p$. We propose to replace $\text{se}(\hat{\beta}_j)$ with an estimate obtained using the ensemble of inter-voxel information. Moreover, we focus on statistics motivated by hierarchical models, that can be obtained with little computational effort.

### 3.3 Variance shrinkage using a hierarchical framework

Here, we propose a hybrid approach based on Bayesian hierarchical models. The fundamental idea is to assume normality of the log of the voxel-specific residual variances, with means and variances corrected using central moments from the log of a chi-squared distribution. A second level distribution borrows strength across voxels, and is either set to a normal distribution, for ease of computing, or a mixture of normals for accurate modeling of the
inter-voxel distribution of the log-variances.

Rather than placing a hierarchical model on the linear model (1), we consider only the marginal likelihood obtained by the residual variance estimate. This drastically eases computations, and discards only a paucity of information regarding $\sigma(v)^2$. Specifically consider the consequence of the linear model, independence and normality assumptions:

$$\frac{(n-p)s^2(v)}{\sigma^2(v)} \sim \chi_{(n-p)}^2.$$  

We take a natural logarithmic transformation of the residual variance because, modeling the distribution of residual variance on the log scale is easier, more stable, and more naturally assumed to be normal than on the original scale.

For the log-scale parameters, let $\theta(v) = \log\{\sigma^2(v)\}$ be the log-scale estimand of interest. Furthermore, let $\tilde{\theta}(v) = \log\{s^2(v)\} - \psi\left(\frac{n-p}{2}\right) - \log(2) + \log(n-p)$, where $\psi$ is the digamma (derivative of the log of the complete gamma) function. The seemingly odd constant subtracted from the log residual variance is employed to make $\tilde{\theta}(v)$ an unbiased estimator of $\theta(v)$ under (2) (see Appendix A). Moreover, it can also be shown that the

$$\text{Var}\{\tilde{\theta}(v)\} = \psi'\left(\frac{n-p}{2}\right),$$

where $\psi'$ is the trigamma (derivative of the digamma) function. This equation highlights the interesting fact that the variance of the log of the empirical residual variance is a constant
and does not depend on its estimand.

The first stage of the hierarchical model assumes

\[ \tilde{\theta}(v) \mid \theta(v) \sim N\{\theta(v), \gamma^2(v)\}, \]  

(3)

where, \( \gamma^2(v) = \psi'\left(\frac{a_0 + b_2}{2}\right) \). The additional notation \( (\gamma^2) \) is necessary for generality, because imperfect registration often leads to some subjects having missing data at particular voxels; hence the value of \( n \) is voxel-specific. This is not a central issue in this study, as this problem only exists on boundary voxels, such as those near the skull or ventricles. However, the additional notation is warranted, because in related studies, such as in voxel based morphometry of gray matter, the problem can be much more severe.

The second stage model assumes a mixture of normals, which we write as

\[ \theta(v) \mid \pi, \mu, \tau \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \tau_k^2), \]  

(4)

where \( \pi, \mu \) and \( \tau \) are vectors of the \( \pi_k, \mu_k \) and \( \tau_k \) respectively. The mixture of normals (conjugate) distribution greatly eases computations, especially for the single normal case \( (K = 1) \). Simulation results suggest that a single normal is often enough to reap the benefits of shrinkage. However, we also investigate less trivial mixtures of normals, to more accurately model the inter-voxel distribution of log-variances. Our investigations have shown that only a small number of mixture components (three or fewer) are necessary in this application.
An empirical Bayes approach is adopted. Estimates for $\pi_k$, $\mu_k$ and $\tau_k$ are obtained as posterior modes from an MCMC sampler with diffuse priors (see Appendix B). The convergence of these parameter estimates is extremely rapid, owning to the tens of thousands of data points contributing to the fit of this inter-voxel distribution.

After obtaining point estimates, the conditional distribution of the log variances given the empirical ones is used for estimation. This distribution is given as

$$
\theta(v) \mid \tilde{\theta}(v) \sim \sum_{k=1}^K \pi_k^*(v)N\left\{ \mu_k^*(v), \tau_k^2(v) \right\},
$$

where

$$
\pi_k^*(v) = \frac{\pi_k \phi \left[ \{\tilde{\theta}(v) - \mu_k\} / \{\tau_k^2 + \gamma^2(v)\}^{1/2} \right] / \{\tau_k^2 + \gamma^2(v)\}^{1/2}}{\sum_{l=1}^K \pi_l \phi \left[ \{\tilde{\theta}(v) - \mu_l\} / \{\tau_l^2 + \gamma^2(v)\}^{1/2} \right] / \{\tau_l^2 + \gamma^2(v)\}^{1/2}},
$$

and

$$
\mu_k^*(v) = \frac{\tau_k^2 \tilde{\theta} + \gamma^2(v) \mu_k}{\tau_k^2 + \gamma^2(v)} \quad \text{and} \quad \tau_k^2(v) = \frac{\tau_k^2 \gamma^2(v)}{\tau_k^2 + \gamma^2(v)}.
$$

The best linear predictor of $\theta(v)$ is then $E[\theta(v) \mid \tilde{\theta}(v)] = \sum_k \pi_k^*(v)\mu_k^*(v)$.

Notice that this two-stage model is not a special case of the typical Gaussian mixture model often used for unsupervised clustering (see Hastie et al., 2001). Therefore, in addition to using the complete MCMC sampler, we propose an ad hoc procedure for fitting this mixture model that takes advantage of standard Gaussian mixture model software. In particular, we propose that one first obtains estimates for $\mu_1$ and $\tau_1$ assuming only a single component mixture ($K = 1$). Because of the simplicity of the single normal calculations, this
can be done easily using an EM algorithm or Gibbs sampler. One then uses these estimates to calculate $E[\theta(v) | \tilde{\theta}(v)]$. To refine these estimates using the mixture, we subsequently treat them as the data in a standard Gaussian mixture model algorithm. Notice that this is more correct than using the raw log empirical residual variances, as those include the $\gamma^2(v)$ extra variance component. The predictions based on this subsequent algorithm are then used instead of the estimates based on the single normal distribution.

This ad hoc procedure has several notable benefits. First, in general, the single normal fit is usually of interest, so would be performed regardless. Secondly, the procedure can utilize existing Gaussian mixture model algorithms. Thirdly, simulations studies suggest that, unless the mixture distribution is multimodal, this algorithm does a reasonable job of accounting for mild departures from normality in the inter-voxel log-variance distribution.

The shrunken estimates for $\theta(v)$ are used to obtain estimates for $\sigma^2(v)$ to be used in constructing t-statistics. Since

$$\exp\{E[\theta(v) | \tilde{\theta}(v)]\} \neq E[\exp\{\theta(v)\} | \tilde{\theta}(v)],$$

one has to chose between the two. We chose the later in the single normal case. This leads to the best linear predictor of $\theta(v)$ which is equal to $\sum_k \pi^*_k(v) \mu^*_k(v)$ (see Appendix C). In the true mixture case ($k \geq 2$) where simulation studies suggested that $\exp\{E[\theta(v) | \tilde{\theta}(v)]\}$ leads to a lower MSE than $E[\exp\{\theta(v)\} | \tilde{\theta}(v)]$. Thus $\exp\{E[\theta(v) | \tilde{\theta}(v)]\}$ is suggested in this case. These estimates replace $s^2(v)$ in the denominator of t-statistics used to create.
statistical maps. The resulting statistics do not follow Gossett’s t-distribution. However, simulation results suggest an improvement in performance for ranking significant voxels.

4 A simulation study

We performed a simulation study to evaluate the proposed methodology. Data sets were generated from the hierarchical model stated in Section 3, starting at the second stage model, under varying parameter scenarios. We generated 100 simulated data sets with 5,000 voxels per subject. The linear model (1) used for simulation at each voxel contained a group effect only. That is, $X_i$ contained an intercept term and an indicator variable.

At each simulation, a random log $\sigma_v^2$ value was drawn from from either a single or three component mixture of normals. Then $\epsilon_i(v)$ values were then sampled randomly from a normal distribution with mean 0 and variances $\sigma_v^2$. We assumed there was differential activation ($\beta_1(v) \neq 0$) at 10% of the voxels. This was accomplished by either simulating all of the $\beta_1(v)$ from a standard normal distribution and declaring the top 10% in absolute value as differentially activated or fixing 500 of the $\beta_1(v)$ to be 0 and simulating the remainder from a uniform distribution with range $-4$ to $4$. This procedure was replicated to create 100 simulated data sets.

Several measures of performance were considered. We emphasize the average mean squared error (simply labeled MSE), which is defined as the estimated voxel-specific mean squared error averaged over the 5,000 voxels. This was considered on the natural and log
scales. In addition, we considered areas under the receiver operating characteristic (AUC) (Hanley and McNeil, 1982; Hilden, 1991) for the modified t-statistic as well as false and true positive rates for a given cutoff.

4.1 One mixture component

First, we consider the performance when a single normal is used for simulation and fitting. The equation in (4) now is modified as

\[
\theta(v) \mid \mu_1, \tau_1 \sim N(\mu_1, \tau_1^2).
\]

We set \( \mu_1 \) to be \((-14, -5, -1, 0.01, 0.5, 2)\) and varied the coefficient of variation (or CV, \( \frac{\tau_1}{\mu_1} \)) to be \((1, 0.1, 0.05)\). Both sets of values were motivated by the verbal-paired-associates data. Note that the values of \( \mu_1 \) depend on the units of the contrast measurements in the second stage fMRI analysis, which depend on arbitrary scaling factors used in either stage. We also considered varying numbers of subjects \( (n) \) to investigate the interplay between the number of subjects and the benefits of shrinkage estimation.

When varying the coefficient of variation (Table 1) for a given \( \mu_1 \), the shrunken variance estimators on the log scale lead to a consistently lower MSE (ratios ranged from 0.1% to 94.2%). This improvement is most significant when the coefficient of variation decreases while \( \mu_1 \) gets closer to zero. When examining the estimates on the natural scale, generally
the shrunken estimates have a lower MSE, except for the case where $\mu_1 = 2$ and CV = 1. When the sample size is varied (Table 2) for a given $\mu_1$, on the log scale, the constructed shrunken variance estimators lead to a lower MSE for all specified sample sizes, with more improvement as the sample size decreases (see Figure 1) and $\mu_1$ approached zero. A similar pattern was observed on the natural scale.

Consider comparing the performance of the modified t-statistics (Tables 3 and 4). Generally the modified t-statistic has a greater AUC than the classical t-statistic, indicating a better ability to discriminate between zero and non-zero differences in activation. The increase in AUC varies from less than 0.0001% to 4%, depending on the sample size and the coefficient of variation. However, one must take into account that the AUC are quite large, hence the room for improvement is small.

As expected, as the sample size decreased, the AUC improvement by the modified t-statistic increased. In addition, the false positive rate for the modified t-statistic decreased by a range of 0.002% to 0.04% when compared to the classical t-statistic. There is trade-off between the true positive and the false positive, the classical t-statistic has a slightly higher true positive rate than the modified t-statistic. However, we emphasize that the modified t-statistic has higher AUC than the classical t-statistics.
4.2 Multiple mixture components

We now consider log variances simulated from a three component mixture of normals for the inter-voxel log variance distribution. The mean and the variance of the distribution used for simulation was based on estimated values from the verbal-paired-associates data. However, we also varied the parameters to account for possible distributions of the true log residual variances. Figure 2 shows the three different scenarios for the hypothesized mixture normal prior (Table 5). The first scenario resembles the distribution of the log residual variances had a marked skewness and slight bend in the right tail, which was motivated by the experimental data. In the second, the distribution of the log residual variances had heavy tails, while in the third the distribution of the log residual variances is bimodal. One hundred simulated data sets were generated for each scenario.

The estimation of $\mu_k$, $\tau_k$ and $\pi_k$ were obtained by the ad hoc algorithm. To consider the impact of misspecification, we included results from the single mixture component. These were both compared to the classical residual variance estimator.

The evaluation of all three variance estimation methods are presented in Table 6. As expected, the mixture of normals model has the lowest MSE over all parameter settings, especially in the first and third simulation scenarios. As such, this validates the ad hoc algorithm’s ability to appropriately model more intricate inter-voxel distributions.

On the natural scale, the two shunken estimators had higher mean squared errors in the second scenario. The modified t-statistics (using either a single normal or mixture of normals)
generally performed better than the standard t-statistic, with larger AUCs and lower false positives rate (Table 7). Again, for the true positive rate, the modified t-statistics did not show an improvement over the classical t-statistic. In general, the difference between the shrunken variance estimates and the classical variance estimates were prominent in MSE and AUC. However, the improvement on efficiency and effect discrimination is not uniform across all circumstances. Moreover, even under a mixture of normals assumption, the single normal shrunken estimates are still comparable to the mixture of normals shrunken estimates. Thus, unless the variances themselves are of intrinsic interest, a single normal mixture appears sufficient. In summary, all of the experimental results suggest a potential improvement in t-statistic performance based on some degree of shrinkage. In contrast, elaborate modelling of the inter-voxel distribution of variances only seems worthwhile when the variances themselves are of interest.

5 Experimental results

In this section, we apply this methodology to the fMRI experiment outlined in Section 2. The effect of interest was the difference in the encoding versus rest contrast maps between males and females. The contrast maps had dimension $79 \times 95 \times 68$. Recall the imaging area was reduced to focus on a coronal band through the medial temporal lobe.

We constructed t-statistic maps to identify brain regions where the contrasts of activation differs across genders. First, we demonstrate the validity of the inter-voxel distributional
assumptions. Figure 3 shows the kernel density estimates of $\tilde{\theta}(v)$ and the logs$^2(v)$ obtained after trimming the upper and lower 1% which suggested $\theta(v)$ might follow a distribution with more than one mixture component.

The fitted distributions, obtained using a single normal mixture component and a three component mixture of normals fit by the full model as well as by the ad hoc approach. The three component mixture models both seem to fit the empirical variance distribution quite well.

Maps of the classical and modified t-statistics, again using a single mixture and both methods with three mixture components, are shown in Figures 4 and 5. It shows the t-maps for detecting true differences between groups according to the shrunken t-statistics and the classical t-statistics where we threshold the absolute value of t-statistics at 3.01. That is, we only highlight the region where there is statistical evidence to show that the contrasts of activation differs across genders. We see that the map for the classical t-statistic is more diffuse, which might be due to noise and, according to the simulation study, might include more false positives. However, the t-statistic maps were similar for all of the methods due to the peculiarly large number of subjects in this study.

5.1 Performance on a subset of cognitive activation task data

In this section, we consider a subset of twenty subjects from the cognitive activation task, to highlight and discuss differences in the classical and modified t-statistics. Note that most
fMRI studies contain on the order of twenty or fewer subjects. Therefore, we compare the fitted variance and t-statistics from the subset data to those from the full data set. Moreover, by selecting a subset we match on important covariates. The selected 20 subjects from the original data set were matched on their age, IQ and education. Figures 6 and 7 show the resulting t-maps. They suggest that t-statistics with shrunken estimators lead to smaller regions as the ordinary residual variance is less stable and large values for the classical t-statistics might be due to small residual variance estimates. Though the evidenced changes are somewhat small, we must bear in mind that interpretation of fMRI results depends heavily on comparing threshold maps with areas of known anatomical function. Therefore, differences of even small number of voxels can lead to drastic differences in interpretation.

6 Discussion

In this manuscript we developed a flexible method for estimating voxel-specific variances in fMRI experiments by combining inter-voxel information, an idea that is used extensively in the genomics literature, but less so in fMRI. We employed this Empirical Bayes methodology to provide a statistically rigorous way of improving residual variance estimate. The performance of the methodology was evaluated by simulation and implemented on experimental data. The results show that shrinkage estimates of variance components are generally more efficient and robust, especially on the log scale, for small sample sizes or when there is a small variance for the inter-voxel distribution of the log variances.
After transformation to natural scale, the improvement is less uniform, especially when the variability between voxels is high. However, in general, the modified t-statistic has more power than the classical one and leads to a higher AUC and lower false positive rate.

This manuscript demonstrates the trade-offs and useful mechanisms for shrinking variance components. We relegate to future work a discussion and comparison of optimal thresholding of the modified t-statistics. In particular, as these statistics no longer follow Gossett’s t-distribution, popular methods for thresholding statistical maps (Worsley et al., 1996) do not immediately apply. In addition, multiplicity concerns (Shaffer, 1995) would also need to be addressed.

A limitation of this approach is that it does not utilize any spatial information in the variance. However, variogram fits suggested such spatial correlation was small in our verbal-paired-associates task. Moreover, adjusting for any spatial correlation would introduce a great deal of additional computational burden. Similarly, accounting for the spatial correlation in the effect estimates would also be interest. However, to a large degree, spatial smoothing in the first stage capitalizes on this information.

In conclusion, we emphasize that, although the methods were developed and illustrated for fMRI analysis, they are potentially useful in other areas, such as high throughput genomic studies. Of particular interest are the flexible and easily implemented method for fitting a mixture distribution to the log variances.
A Log chi squared results

Let \( s^2(v) = \frac{\text{SSE}}{n-p} \) where SSE is the sum of the squared least squares residuals for voxel \( v \).

Assume \((n - p)s^2(v)\) is independent of \( \sigma^2(v) \). Then we have

\[
\frac{(n-p)s^2(v)}{\sigma^2(v)} \sim \chi^2_{(n-p)}.
\]

Taking the natural logarithmic transformation on both sides yields,

\[
\log\{s^2(v)\} + \log(n - p) - \log\{\sigma^2(v)\} \sim \log\{\chi^2_{(n-p)}\},
\]

and therefore, using standard results for the log of a chi-squared (see the moment generating function given below),

\[
E[\log\{s^2(v)\}] = E[\log\{\chi^2_{(n-p)}\}] + \log\{\sigma^2(v)\} - \log(n - p).
\]

Let \( \tilde{\theta} = \log\{s^2(v)\} - \psi\left(\frac{n-p}{2}\right) - \log(2) + \log(n - p) \). Then \( \tilde{\theta}(v) \) is an unbiased estimator for \( \log\{\sigma^2(v)\} \). Further notice that

\[
\text{Var}\left[\log\{s^2(v)\}\right] = \text{Var}\left[\log\{\chi^2_{(n-p)}\}\right].
\]
Thus the derivation of $\text{Var}[\log(s^2(v))]$ is equivalent to the derivation of $\text{Var}[\log(\chi_{(n-p)}^2)]$.

Consider the moment generating function below:

$$E \left[ \exp \left\{ t \log(\chi_{n-p}^2) \right\} \right] = E \left[ \chi_{n-p}^{2t} \right] = \frac{\Gamma\left(\frac{n-p}{2} + t\right)}{\Gamma\left(\frac{n-p}{2}\right)} 2^t.$$ 

Therefore, the cumulant generating function is

$$K(t) = \log \left\{ E \left[ \exp\left\{ \log(\chi_{n-p}^2) \right\} \right] \right\} = \log \left\{ \Gamma\left(\frac{n-p}{2} + t\right) \right\} - \log \left\{ \Gamma\left(\frac{n-p}{2}\right) \right\} + t \log(2).$$

Hence we obtain that:

$$E \left[ \log \left\{ \chi_{(n-p)}^2 \right\} \right] = K'(0) = \psi\left(\frac{n-p}{2}\right) + \log(2)$$

and

$$\text{Var} \left[ \log \left\{ \chi_{(n-p)}^2 \right\} \right] = K''(0) = \psi'(\frac{n-p}{2}).$$

**B  Gibbs Sampler full conditionals**

**B.1 Single normal, $K = 1$**

Consider the $K = 1$ case. We assume that $\mu_1 \sim N(0, 10,000)$ and $\sigma_1^2 \sim IG(\alpha, \beta)$ is a diffuse prior; for example setting $\alpha$ and $\beta$ both equal to $10^{-l}$ for large $l$. (We investigated several
values of $l$). The full conditional distributions are:

$$
\begin{align*}
\theta(v) | \tilde{\theta}(v), \mu_1, \tau_1^2 & \sim N \left[ \frac{\mu_1}{\tau_1^2} + \frac{\tilde{\theta}(v)}{\gamma^2(v)} \gamma^2(v), \frac{1}{\tau_1^2} + \frac{1}{\gamma^2(v)} \right]^{-1}, \frac{1}{\tau_1^2} + \frac{1}{\gamma^2(v)} \right]^{-1} \\
\mu_1 | \tilde{\theta}(v), \theta(v), \tau_1^2 & \sim N \left\{ \frac{1}{V} \Sigma \theta(v), \frac{\tau_1^2}{V} \right\} \\
\tau_1^2 | \tilde{\theta}(v), \theta(v), \mu_1 & \sim IG \left[ \alpha + \frac{V}{2}, \beta + \frac{1}{2} \Sigma \{\theta(v) - \mu_1 \}^2 \right]
\end{align*}
$$

B.2 Multiple mixture components, $K \geq 2$

Assume that $\theta(v) | \eta(v) = k \sim N(\mu_k, \tau_k^2)$. Let $P(\eta(v) = k) = \pi_k$. Let $\Theta = (\mu, \tau^2, \pi)$ be the vector of parameters to be estimated. We assume that $\mu_k \sim N(0, \tau_0^2)$, where we set $\tau_0^2$ to be a large number (usually 10,000), $\tau_k^2 \sim IG(a, b)$ as in the previous subsection and $\pi_1, \ldots, \pi_K \sim Dirichlet(\alpha, \ldots, \alpha)$ where $\alpha$ was set to 1.

Let $n_k = \sum_v I(\eta(v) = k)$. The full conditional distributions for the Gibbs Sampler are as follows. The full conditional for $\tilde{\theta}(v)$ is

$$
N \left[ \mu_{\eta(v)} \gamma^2(v) + \tilde{\theta}(v) \tau_{\eta(v)}^2, \frac{1}{\gamma^2(v)} + \frac{1}{\gamma^2(v)} \right]^{-1} 
$$

The full conditional for $\eta(v)$ is Multinomial\{1, $p_1(v), \ldots, p_K(v)$\} where

$$
p_k(v) = \frac{\phi\{(\theta(v) - \mu_k)/\tau_k\} \pi_k}{\sum_{k'} \phi\{(\theta(v) - \mu_{k'}/\tau_{k'}) \pi_{k'} \}.}
$$
The full conditional for \( \pi \) is \( \text{Dir}(n_1 + \alpha, \ldots, n_K + \alpha) \). The full conditional for \( \mu_k \) is

\[
N \left[ \frac{\tau_0^2 \sum_v \theta(v) I\{\eta(v) = k\}}{n_k \tau^2 + \tau_k^2}, \frac{\tau_k^2 \tau_0^2}{n_k \tau_0^2 + \tau_k^2} \right]
\]

where \( I \) is an indicator function. Finally the full conditional for \( \tau_k^2 \) is

\[
IG \left[ a + \frac{n_k}{2}, b + \frac{1}{2} \sum_v \{\theta(v) - \mu_k\}^2 I\{\eta(v) = k\} \right].
\]

### C Converting shrunken estimators to the natural scale

Recall that \( \theta(v) \mid \tilde{\theta}(v) \sim \sum_{k=1}^K \pi_k^*(v) N \{\mu_k^*(v), \tau_k^*(v)\} \). Note that this implies

\[
\exp\{\theta(v)\} \mid \tilde{\theta}(v) \sim \sum_{k=1}^K \pi_k^*(v) \ast \text{Log Normal} \{\mu_k^*(v), \tau_k^*(v)\}.
\]

Therefore,

\[
E[\exp\{\theta(v)\} \mid \tilde{\theta}(v)] = \sum_{k=1}^K \pi_k^*(v) \exp\{\mu_k^*(v) + \tau_k^*(v)/2\}.
\]
Figure 1: MSE Ratio comparisons when varying the sample size on the log scale (left hand side) and natural scale (right hand side). The constructed shrunken variance estimators lead to a lower MSE for all specified sample sizes, with more improvement as the sample size decreases and $\mu$ approaches zero.
Figure 2: Scenarios for the distribution of mixture normals on the log Residual variances: scenario I resembles a distribution with a marked skewness and slight bend in the right tail; scenario II resembles a distribution with heavy tails; scenario III resembles a bimodal distribution.
Figure 3: Kernel density estimates: (1) logsigmasq.trim is the $\log^2(v)$ after trimming; (2) $\tilde{\theta}(v)$ is the unbiased estimate for $\log\sigma^2$; (3) theta.drawn is the theta drawn from Gibbs sampler when assuming one mixture component; (4) normal.gibbs represents the shrunken variance estimates on the log scale based on the one mixture component assumption; (5) mix.gibbs represents the shrunken variance estimates on log scale by Gibbs sampler based on a three mixture components assumption; (6) mix.gibsem represents the shrunken variance estimates on the log scale by the ad hoc approach based on the three mixture components assumption.
Figure 4: T-map comparisons, (a) classical t-statistic, (b) single normal modified t-statistic, (c) mixture normal modified t-statistic obtained by the ad hoc algorithm, (d) mixture of normals modified t-statistic obtained by the Gibbs sampler.
Figure 5: T-map by slice comparison, (a) classical t-statistic, (b) single normal modified t-statistic, (c) mixture normal modified t-statistic obtained by the ad hoc algorithm, (d) mixture of normals modified t-statistic obtained by the Gibbs sampler.
Figure 6: T-map comparison: subset, (a) classical t-statistic, (b) single normal modified t-statistic, (c) mixture normal modified t-statistic obtained by the ad hoc algorithm, (d) mixture of normals modified t-statistic obtained by the Gibbs sampler.
Figure 7: T-map by slice comparison: subset, (a) classical t-statistic, (b) single normal modified t-statistic, (c) mixture normal modified t-statistic obtained by the ad hoc algorithm, (d) mixture of normals modified t-statistic obtained by the Gibbs sampler.
## E Tables

<table>
<thead>
<tr>
<th></th>
<th>Log Scale</th>
<th>CV=0.05</th>
<th>CV=0.1</th>
<th>CV=1</th>
<th>Natural Scale</th>
<th>CV=0.05</th>
<th>CV=0.1</th>
<th>CV=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1=-14$</td>
<td>59.7</td>
<td>82.5</td>
<td>94.2</td>
<td></td>
<td>60.9</td>
<td>70.4</td>
<td>73.3</td>
<td></td>
</tr>
<tr>
<td>$\mu_1=-5$</td>
<td>17.0</td>
<td>44.3</td>
<td>93.3</td>
<td></td>
<td>19.8</td>
<td>46.9</td>
<td>96.3</td>
<td></td>
</tr>
<tr>
<td>$\mu_1=-1$</td>
<td>0.9</td>
<td>3.3</td>
<td>73.4</td>
<td></td>
<td>1.0</td>
<td>3.9</td>
<td>68.9</td>
<td></td>
</tr>
<tr>
<td>$\mu_1=0.01$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>$\mu_1=0.5$</td>
<td>0.2</td>
<td>0.9</td>
<td>44.2</td>
<td></td>
<td>0.3</td>
<td>1.0</td>
<td>47.3</td>
<td></td>
</tr>
<tr>
<td>$\mu_1=2$</td>
<td>3.3</td>
<td>11.8</td>
<td>88.0</td>
<td></td>
<td>3.9</td>
<td>13.9</td>
<td>119.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: MSE ratios for the single mixture component shrunken residual variance estimates versus the classical residual variance estimate. Here the coefficient of variation (CV) is varied at three levels and results reported for the natural and log scales. MSE ratio(%) = \[ \frac{\text{MSE for Shrunken}}{\text{MSE for classical}} \times 100\% . \]
Table 2: MSE ratios for single mixture component shrunken residual variance estimate versus the classical residual variance estimates when varying the sample size. MSE ratio(%) = \( \frac{\text{MSE for Shrunken}}{\text{MSE for classical}} \times 100\% \).
<table>
<thead>
<tr>
<th>CV</th>
<th>$\mu_1$</th>
<th>$t_{\text{modified}}$</th>
<th>$t_{\text{classical}}$</th>
<th>increase by (%)</th>
<th>True Positive Rate $t_{\text{modified}}$</th>
<th>True Positive Rate $t_{\text{classical}}$</th>
<th>False Positive Rate $t_{\text{modified}}$</th>
<th>False Positive Rate $t_{\text{classical}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>$-14$</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.0016</td>
<td>0.9994</td>
<td>0.9994</td>
<td>0.0268</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>$-5$</td>
<td>0.9899</td>
<td>0.9892</td>
<td>0.0659</td>
<td>0.9707</td>
<td>0.9725</td>
<td>0.0228</td>
<td>0.0499</td>
</tr>
<tr>
<td></td>
<td>$-1$</td>
<td>0.9208</td>
<td>0.9152</td>
<td>0.6192</td>
<td>0.7749</td>
<td>0.7815</td>
<td>0.0214</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.8746</td>
<td>0.8655</td>
<td>1.0478</td>
<td>0.6342</td>
<td>0.6482</td>
<td>0.0214</td>
<td>0.0501</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.8380</td>
<td>0.8266</td>
<td>1.3720</td>
<td>0.5297</td>
<td>0.5450</td>
<td>0.0213</td>
<td>0.0504</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6820</td>
<td>0.6720</td>
<td>1.4896</td>
<td>0.1760</td>
<td>0.2194</td>
<td>0.0214</td>
<td>0.0498</td>
</tr>
<tr>
<td>0.1</td>
<td>$-14$</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.0012</td>
<td>0.9996</td>
<td>0.9997</td>
<td>0.0285</td>
<td>0.0499</td>
</tr>
<tr>
<td></td>
<td>$-5$</td>
<td>0.9892</td>
<td>0.9885</td>
<td>0.0722</td>
<td>0.9684</td>
<td>0.9697</td>
<td>0.0248</td>
<td>0.0495</td>
</tr>
<tr>
<td></td>
<td>$-1$</td>
<td>0.9237</td>
<td>0.9184</td>
<td>0.5831</td>
<td>0.7800</td>
<td>0.7885</td>
<td>0.021</td>
<td>0.0495</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.8723</td>
<td>0.8629</td>
<td>1.0865</td>
<td>0.6330</td>
<td>0.6430</td>
<td>0.0211</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.8400</td>
<td>0.8283</td>
<td>1.4067</td>
<td>0.5373</td>
<td>0.5508</td>
<td>0.0210</td>
<td>0.0504</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6826</td>
<td>0.6727</td>
<td>1.4624</td>
<td>0.1722</td>
<td>0.2191</td>
<td>0.0223</td>
<td>0.0495</td>
</tr>
<tr>
<td>1</td>
<td>$-14$</td>
<td>0.9274</td>
<td>0.9267</td>
<td>0.0794</td>
<td>0.8439</td>
<td>0.8507</td>
<td>0.0292</td>
<td>0.0498</td>
</tr>
<tr>
<td></td>
<td>$-5$</td>
<td>0.9323</td>
<td>0.9302</td>
<td>0.2202</td>
<td>0.8273</td>
<td>0.8385</td>
<td>0.0290</td>
<td>0.0498</td>
</tr>
<tr>
<td></td>
<td>$-1$</td>
<td>0.9126</td>
<td>0.9065</td>
<td>0.6627</td>
<td>0.7451</td>
<td>0.7611</td>
<td>0.0273</td>
<td>0.0498</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.8727</td>
<td>0.8636</td>
<td>1.0549</td>
<td>0.6311</td>
<td>0.6435</td>
<td>0.0212</td>
<td>0.0501</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.8353</td>
<td>0.8247</td>
<td>1.2912</td>
<td>0.5160</td>
<td>0.5423</td>
<td>0.0255</td>
<td>0.0497</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6986</td>
<td>0.6933</td>
<td>0.7656</td>
<td>0.2687</td>
<td>0.3195</td>
<td>0.0293</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

Table 3: Single mixture component shrunken and classical t-statistics comparison when varying the coefficient of variation.
Table 4: Single mixture component modified and the classical t-statistic comparison when varying the sample size.
<table>
<thead>
<tr>
<th>Scenarios</th>
<th>$\mu_k$</th>
<th>$\tau_k$</th>
<th>$\pi_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (-14.3, -13.48, -12.34)</td>
<td>(0.26, 0.48, 0.38)</td>
<td>(0.48, 0.43, 0.09)</td>
<td></td>
</tr>
<tr>
<td>II (-13.90, -14.50, -13.50)</td>
<td>(2.10, 0.80, 0.60)</td>
<td>(0.50, 0.30, 0.20)</td>
<td></td>
</tr>
<tr>
<td>III (-14.80, -13.80, -13.20)</td>
<td>(0.49, 0.24, 0.36)</td>
<td>(0.40, 0.12, 0.48)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Scenarios considered for three mixture normal component model.
<table>
<thead>
<tr>
<th>Log scale</th>
<th>$\hat{\sigma}_v^2$ (N)</th>
<th>$\hat{\sigma}_v^2$ (M)</th>
<th>$s_v^2$</th>
<th>$\frac{\hat{\sigma}_v^2}{s_v^2}$ (%)</th>
<th>$\frac{\sigma_v^2}{s_v^2}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2.63</td>
<td>2.48</td>
<td>2.79</td>
<td>94.14</td>
<td>88.96</td>
</tr>
<tr>
<td>II</td>
<td>2.75</td>
<td>2.74</td>
<td>2.80</td>
<td>98.30</td>
<td>98.02</td>
</tr>
<tr>
<td>III</td>
<td>2.68</td>
<td>2.59</td>
<td>2.81</td>
<td>95.69</td>
<td>92.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Natural Scale</th>
<th>$\hat{\sigma}_v^2$ (N)</th>
<th>$\hat{\sigma}_v^2$ (M)</th>
<th>$s_v^2$</th>
<th>$\frac{\hat{\sigma}_v^2}{s_v^2}$ (%)</th>
<th>$\frac{\sigma_v^2}{s_v^2}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>9.95e-12</td>
<td>9.57e-12</td>
<td>1.04e-11</td>
<td>96.05</td>
<td>92.38</td>
</tr>
<tr>
<td>II</td>
<td>1.94e-08</td>
<td>1.77e-08</td>
<td>1.44e-08</td>
<td>134.64</td>
<td>122.85</td>
</tr>
<tr>
<td>III</td>
<td>5.96e-12</td>
<td>5.57e-12</td>
<td>6.50e-12</td>
<td>91.58</td>
<td>85.67</td>
</tr>
</tbody>
</table>

Table 6: Mixtures of normals residual variance estimation comparison on MSE ratios. MSE ratio(%) = $\frac{\text{MSE for Shrunken}}{\text{MSE for classical}}$ * 100%.
<table>
<thead>
<tr>
<th>AUC</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{\mu(N)}$</td>
<td>$\sigma^2_{\mu(M)}$</td>
<td>$s^2_{\mu}$</td>
</tr>
<tr>
<td>I</td>
<td>0.9999481, 0.9999479, 0.9999478</td>
<td>0.99988, 0.99988, 0.99987</td>
</tr>
<tr>
<td>II</td>
<td>0.9999350, 0.9999350, 0.9999349</td>
<td>0.99982, 0.99982, 0.99982</td>
</tr>
<tr>
<td>III</td>
<td>0.9999745, 0.9999743, 0.9999743</td>
<td>0.99991, 0.99991, 0.99991</td>
</tr>
</tbody>
</table>

Table 7: Mixtures of normals modified and classical t-statistics comparison.
References


