Statistics and Its Interface Volume 3 (2010) 59–67

Nonparametric evaluation of heterogeneity of brain regions in neuroreceptor mapping applications[*]

R. Todd Ogden[†] and Huiping Jiang

In many applications of modeling positron emission tomography (PET) data for neuroreceptor mapping studies, it is necessary to define one or more regions of interest (ROI) and analyze aggregate data from each region, often assumed to be homogenous. We propose a simple method for assessing the level of heterogeneity within any given ROI along with a procedure for testing a null hypothesis of regional homogeneity that uses a wild bootstrap algorithm. Estimation of outcome measures is accomplished using a mixture modeling approach. We provide results of a simulation study along with analysis of an imaging dataset, which indicates that most of the ROIs considered are quite heterogeneous.

Keywords and phrases: Wild bootstrap, Mixture modeling, Positron emission tomography, Compartment modeling.

1. INTRODUCTION

In a typical neuroreceptor mapping study, a radioactively labeled ligand is injected into a subject's bloodstream and allowed to move throughout the body. The flow of the ligand throughout the brain may be observed using imaging modalities such as positron emission tomography (PET), which measure the concentration of the ligand over time at each location in the brain. These measures are typically aggregated over time into several pre-specified intervals, termed *frames*, and the resulting data for a given location considered over all frames comprises a *time activity curve*. Each frame consists of a three-dimensional grid of concentration measurements; each element of the grid is termed a *voxel*. By modeling this time activity curve it is possible to estimate various measures of the density of the target receptor throughout the brain. Once these estimates have been computed for each subject at each brain location, and once appropriate registration of images has been done, it is possible to compare these estimates among subjects (e.g., between a patient group and a control group).

The modeling of a ligand's kinetic behavior over time is often done using compartmental modeling techniques (Jacquez [1985]; Gunn et al. [2001]), which necessitates applying an iterative nonlinear regression algorithm in order to estimate model parameters. A number of alternative modeling strategies have been proposed with a variety of motivations, including computational feasibility, modeling flexibility, and relaxation of model assumptions.

There are two general complementary approaches to the analysis of brain imaging data. One is to analyze the data separately for each voxel. The other requires specifying some anatomically defined regions of interest (ROIs). This may be accomplished in practice by identifying a template indicating which voxels correspond to each ROI. When the sequence of images for a subject are coregistered to each other and also to the template, the time activity curve for the ROI may be extracted by calculating some summary (typically, the mean) of the measured concentration across voxels for each frame. Then a compartmental model (or any of the various alternatives) can be fit to the aggregated ROI data to obtain parameter estimates specific to each ROI. Subsequently, we can compare estimates of any of several outcome measures of interest, which may then be compared among many subjects to determine possible differences due to diagnosis, treatment, sex, etc.

When neuroreceptor mapping was a newly emerging technology, ROI-level analysis was the standard for any kind of quantitative analysis, since the capacity to fit hundreds of thousands of non-linear models to voxel-level time series was not widely available. ROI analysis is still routinely performed decades later, although advances in computing power and algorithmic development have made fitting of each voxel time series computationally feasible. One reason to consider ROI-level analysis is that voxel-level data are typically very noisy, and model fitting at this level is often beset with lack of convergence and/or unrealistic parameter estimates, problems that arise much less frequently when fitting at the ROI level. Other than the stability of modeling and the lower computational requirements, however, there are other reasons to consider modeling at the ROI level in some situations. For one, defining an ROI based on a structural magnetic resonance image (MRI) that is specific for

[*] The authors are grateful to Chung Chang for helpful discussions about bootstrap algorithms. The first author was funded in part by National Institute of Mental Health grant MH-062185.

[†] Corresponding author.

each subject gives some assurance of "apples to apples" comparisons among many subjects. Also, in the second stage of analysis (comparing estimates of density across subjects), the statistical modeling is generally quite straightforward (e.g., linear mixed models) because the dimensionality of each subject's data is manageable.

On the other hand, determination of anatomically defined ROIs involves either very painstaking work on the part of a technician or the implementation of some automated utility. And while ROI analysis readily enables confirmatory analysis (e.g., about differences in binding measures at several ROIs between groups), a voxel-level analysis would be more appropriate for generating new hypotheses, as interesting areas of difference may not be contained within any predefined ROI. In many applications it is appropriate to perform both an ROI-level analysis and a complementary voxel-level analysis.

Compartmental models that are typically applied to data in neuroreceptor mapping studies include a "plasma" compartment which represents the ligand in the subject's bloodstream ([Gunn et al., 2001]). If measurements are made on this compartment over time, typically by drawing a sequence of blood samples from a subject's artery during scanning, then this can provide an "input function" which will allow estimation of all kinetic measures of common interest. This procedure is rather invasive, however, and the gathering, analysis, and subsequent modeling of the blood date requires significant effort and can be prone to error. It is possible to avoid this blood sampling altogether by identifying a "reference region," such as the cerebellum, that is assumed to be devoid of the target receptor. In the subsequent modeling of the imaging data, aggregate data from this region are gathered and analyzed, and used to constrain the modeling of the data from other regions or voxels (see, e.g., [Lammertsma and Hume, 1996]). Even when blood data are available, it is often useful to identify a reference region thought to be devoid of receptors, as this may enable constraints to be placed on the compartmental modeling that can avoid problems of identifiability.

Thus, the need to identify at least one region of interest, aggregate the data within it for each time point, and model the result can arise in each several analysis strategies in neuroreceptor mapping studies with PET data. It is generally understood that an ROI-level analysis is intended to make inference about aggregate measures across each brain region, but it is reasonable to question the interpretation of such an analysis when the regions exhibit a high degree of internal heterogeneity. Cselényi et al. (2002) also raise this issue and note further that differences in receptor density in subsets may be lost in an ROI-level analysis. It is therefore of interest to examine the level of heterogeneity within ROIs. Here we propose a simple measure (the sample variance of the estimated outcome measures across all voxels in an ROI) of regional heterogeneity along with a bootstrap-based nonparametric testing procedure to test a null hypothesis of homogeneity. The methodology is laid out in Section 2 and results from a simulation study are described in Section 3. Application to an existing imaging dataset is reported in Section 4 and some discussion is given in Section 5.

2. METHODOLOGY

In neuroreceptor mapping, either some compartmental model or one of its alternatives is typically fit in turn to the data for each voxel or ROI. For the jth voxel (or ROI) the "true" concentration at time t will be denoted $C_j(t)$ and independent observations $Y_{1j}, \ldots, Y_{nj}$ are taken at time points $t_1 < \cdots < t_n$, respectively:

$$Y_{ij} = C_j(t_i) + \frac{1}{w_i}\epsilon_{ij}, \tag{1}$$

where $\epsilon_{1j}, \ldots, \epsilon_{nj}$ are assumed to be iid $\mathrm{N}(0, \sigma^2)$ random variables and the factors $w_1, \ldots, w_n$ are known.

2.1 Standard compartmental modeling

The most common way to model the concentration curve in (1) is to use compartmental modeling ([Jacquez, 1985; Gunn et al., 2001]). With such an approach, a particular compartmental structure is assumed and then the "true" concentration curve for voxel j can be expressed generally as

$$C_j(t) = \sum_{k=1}^{K} \beta_{jk} \left(e^{-\gamma_{jk}\cdot} \otimes C_p\right)(t),$$

where $C_p(t)$ represents the concentration of the ligand in the plasma and $\otimes$ represents the convolution operator. The parameters $\beta_{j1}, \ldots, \beta_{jK}$ and $\gamma_{j1}, \ldots, \gamma_{jK}$ are functions of the various rate parameters of the assumed compartmental structure. As mentioned in Section 1, the C_p function may be estimated from measurements made on blood samples drawn during the scanning. We will assume henceforth that such measurements are available and thus that C_p may be estimated by $\hat{C}_p$. If such data are not available, some straightforward modifications to the methodology developed here would be required. With this estimate of C_p, fitting of the PET data may be accomplished separately for the jth voxel or region by minimizing over all choices of parameters $\beta_{j1}, \ldots, \beta_{jK}$ and $\gamma_{j1}, \ldots, \gamma_{jK}$:

$$\sum_{i=1}^{n} w_i^2 \left(Y_{ij} - \sum_{k=1}^{K} \beta_k \left(e^{-\gamma_k\cdot} \otimes \hat{C}_p\right)(t_i) \right)^2 .$$

One outcome measure of interest is the "total volume of distribution" defined as the integral of the *impulse response function*:

$$V_j = \int_0^\infty \sum_{k=1}^{K} \beta_{jk} e^{-\gamma_{jk} t}\, dt = \sum_{k=1}^{K} \frac{\beta_{jk}}{\gamma_k}. \tag{2}$$

After modeling the data for any voxel/region, the total volume of distribution may be estimated simply by replacing the parameters in (2) by their estimates.

2.2 Evaluating heterogeneity of regions

If a region is homogeneous, then the "true" V_j will be the same for every voxel in the region. If R represents a set of indices corresponding to a given region of interest then the null hypothesis of interest may be written

$$H_0 : V_j = V_{j'} \text{ for every } j, j' \in R.$$

Thus, a reasonable strategy for testing this hypothesis would be to compute estimates of V_j for every voxel, then compute some statistic that measures the variability of the estimates. One possibility would be simply to calculate the variance of the estimates across the region. For R representing a set of voxel indices corresponding to a given region of interest, the test statistic may be

$$(3) \qquad \frac{1}{J}\sum_{j\in R}\left(\hat{V}_j - \bar{\hat{V}}\right)^2,$$

where J is the number of voxels in R and $\bar{\hat{V}}$ is the mean of the estimated total volume of distribution for the region: $\bar{\hat{V}} = \frac{1}{J}\sum_{j\in R}\hat{V}_j$.

To use (3) as a test statistic, it would be necessary to determine its null distribution. This could be approached in a parametric way by utilizing likelihood methods and applying large-sample results, but the validity of the procedure would then depend on some rather strong parametric assumptions. Instead, we propose a nonparametric approach to determining the null distribution based on a wild bootstrap algorithm. Our approach to this will make use of estimates of V computed using mixed modeling, a flexible alternative to kinetic modeling. This is in part to avoid specifying a particular compartmental structure, in part for computational tractability, and in part because it readily extends to testing homogeneity of each of the mixture components separately. We note, however, that a similar approach to the testing for homogeneity could be taken for a variety of methods for estimating V.

2.3 Estimation of V using mixture modeling

When estimating each voxel (or region) separately from the others, only the models with the simplest compartmental structure (typically at most $K = 2$) can be fit identifiably ([Slifstein and Laruelle, 2001]). By placing some constraints on the functions (namely, that for each k, the γ_{jk} values are the same for all voxels) and fitting the voxels/regions simultaneously, more complex models may be fit without greatly increasing the number of parameters to be estimated ([O'Sullivan, 2006; Jiang and Ogden, 2008]). This is an instance of mixture modeling in which the components are the same for all voxels but the coefficients are allowed to differ among the voxels/regions.

Constraining the γ_{jk} values to be the same across voxels, i.e., $\gamma_{jk} = \gamma_k$ for all j and for $k = 1, \ldots, K$, this mixture model may be fit simultaneously to all voxels by minimizing

$$(4) \qquad \sum_{j=1}^{N}\sum_{i=1}^{n} w_i^2 \left(Y_{ij} - \sum_{k=1}^{K}\beta_{jk}\left(e^{-\gamma_k t}\otimes C_a\right)(t)\right)^2$$

over $\gamma_1, \ldots, \gamma_k, \beta_{11}, \ldots, \beta_{NK}$, with the further constraint that all β_{jk}'s be nonnegative. Jiang and Ogden (2008) describe an algorithm for fitting this model to data while accounting for the spatial dependency of the PET noise. Applying a conditional autoregressive (CAR) model to account for the spatial correlation structure, model fitting involves, for a given value of the autoregressive parameter, a nonlinear least squares algorithm with a nonnegative least squares algorithm ([Lawson and Hanson, 1974]) nested (at each iteration). The autoregressive parameter is estimated at the top level using a grid search algorithm. The number of components K may be chosen using standard model-selection criteria such as AIC ([Akaike, 1973, 1974]).

2.4 General bootstrap algorithm for testing homogeneity of V_j within a region

For arbitrary values of parameters γ and β define the function

$$f(t;\gamma,\beta) = \beta\left(e^{-\gamma\cdot}\otimes C_p\right)(t).$$

According to the mixture model, the "true" concentration at time t in voxel j is given by $\sum_{k=1}^{K} f(t;\gamma_k,\beta_{jk})$.

In the bootstrap algorithm given here it is understood that the plasma concentration function C_p is to be replaced by its estimate $\hat{C}_p$ when doing model fitting. Let R be the set of indices of the region in question and let J be its cardinality.

1. Fit the mixture model to all data, choosing K by AIC and, computing $\hat{\gamma}_1, \ldots, \hat{\gamma}_K$ and $\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK}$ for all j.
2. Compute $\hat{V}_j = \sum_{k=1}^{K}\hat{\beta}_{jk}/\hat{\gamma}_k$ for each j in R.
3. Compute the test statistic $T = \frac{1}{J}\sum_{j\in R}\left(\hat{V}_j - \bar{\hat{V}}\right)^2$, where $\bar{\hat{V}} = \frac{1}{J}\sum_{j\in R}\hat{V}_j$.
4. Compute the residuals $e_{ij} = Y_{ij} - \sum_{k=1}^{K} f(t_i;\hat{\gamma}_k,\hat{\beta}_{jk})$ for $i = 1, \ldots, n$ and for all $j \in R$.
5. Compute aggregate ROI data: $\bar{Y}_i^R = \frac{1}{J}\sum_{j\in R} Y_{ij}$ for $i = 1, \ldots, n$.
6. Fit aggregate ROI data: choose $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$ to minimize

$$\sum_{i=1}^{n} w_i^2\left(\bar{Y}_i^R - \sum_{k=1}^{K} f(t_i;\hat{\gamma}_k,\beta_k)\right)^2$$

over $\beta_1, \ldots, \beta_K$, subject to the constraint that each $\beta_k \geq 0$, using a nonnegative least squares algorithm.

7. The bootstrap loop: For $b = 1 \ldots B$:
   (a) "Coin tosses": Generate $U_1, \ldots, U_n$, independently, with $P(U_i = 1) = P(U_i = -1) = 1/2$.
   (b) Compute the bootstrapped data: $Y_{ij}^{(b)} = \sum_{k=1}^{K} f(t_i; \hat{\gamma}_k, \tilde{\beta}_{jk}) + U_i e_{ij}$ for $j \in R$ and for each $i = 1, \ldots, n$.
   (c) Fit the bootstrapped data: For each $j \in R$, compute $\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK}$ that minimize $\sum_{i=1}^{n} w_i^2 \left(Y_{ij}^{(b)} - \sum_{k=1}^{K} f(t_i; \hat{\gamma}_k, \beta_k) \right)^2$ over $\beta_1, \ldots, \beta_K$, again subject to the constraint that each $\beta_k \geq 0$, using a nonnegative least squares algorithm.
   (d) Compute the estimates of the outcome measure for the bootstrapped data: $\hat{V}_j^{(b)} = \sum \hat{\beta}_{jk}^{(b)} / \hat{\gamma}_k$ for each $j \in R$.
   (e) Compute the bootstrapped value of the test statistic: $T^{(b)} = \frac{1}{J} \sum_{j \in R} \left(\hat{V}_j^{(b)} - \bar{\hat{V}}^{(b)} \right)^2$, where $\bar{\hat{V}}^{(b)} = \frac{1}{J} \sum_{j \in R} \hat{V}_j^{(b)}$.
8. The bootstrap p-value is the proportion of times the actual test statistic T exceeds the bootstrapped test statistic values: $p = \frac{1}{B} \sum_{b=1}^{B} I(T > T^{(b)})$

Bootstrap replication of the data is done according to a wild bootstrap algorithm described by Liu (1988) and studied by Flachaire (2005) and many others. Such an algorithm has been applied to brain imaging data by Zhu et al. (2007), Whitcher et al. (2008), and others. Theoretical aspects of some applications of such an algorithm to general functional data, of which voxel-level brain imaging applications are one instance, were studied by Chang and Ogden (2009). In the wild bootstrap algorithm, each residual is thus multiplied either by 1 or −1, keeping this factor the same for all voxels for each time point, in order to maintain the spatial correlation structure that exists among the voxels within the region. Another appealing aspect of this type of bootstrap procedure is that it will allow for different variances at each time point. In our implementation, each bootstrapped data set is created so that the voxels in R have the same values of the parameters as those estimated for the aggregate ROI data — i.e., so that, for each bootstrap replication, the null hypothesis of homogeneity will be true.

Note that in the algorithm given above, the γ_k values are only estimated once for the entire brain. These estimates are used to generate the bootstrapped data and when the mixture model is fitted to the bootstrapped data, estimation of γ_k is not repeated, but previous estimates are used when estimating the $\beta_{j,k}^{(b)}$ values. Our experience with such data has shown that all voxels/regions are not needed to estimate the γ_k values. These estimates are quite stable when selecting as few as 5,000 voxels at random and thus we are spared the additional computational expense that would be entailed by estimating the γ_k values for each bootstrap sample.

2.5 Bootstrap algorithm for testing for homogeneity of each component

One of the advantages of taking the mixture modeling approach in such an analysis is that each of the mixture components may be examined separately. The test in Section 2.4 can test whether the total distribution volume is uniform across a region, and if it is decided that it is not, it may be of further interest to determine the level of heterogeneity of each of the components separately. This can be accomplished by a straightforward modification of the bootstrap algorithm in Section 2.4.

Here, the null hypothesis for component k is

$$(5) \qquad H_0 : \frac{\beta_{j,k}}{\gamma_k} = \frac{\beta_{j',k}}{\gamma_k} \text{ for } \quad \text{every } j, j' \in R.$$

We note that the hypothesis in (5) could be stated more succinctly as specifying that $\beta_{j,k} = \beta_{j',k}$ but we prefer to keep the γ_k in the denominator so that this test procedure will more closely parallel that of Section 2.4. The testing procedure would be equivalent in either case.

The algorithm for testing (5) is the same as that in Section 2.4 except for three differences. First, Step 3 should be replaced by

3′. Compute the test statistic $T_k = \frac{1}{J\hat{\gamma}_k^2} \sum_{j \in R} \left(\hat{\beta}_{jk} - \bar{\hat{\beta}}_{\cdot k} \right)^2$, where $\bar{\hat{\beta}}_{\cdot k} = \frac{1}{J} \sum_{j \in R} \hat{\beta}_{jk}$.

Also, the bootstrapped data in Step 7b should be generated according to the sub-hypothesis of homogeneity of the kth component. Instead of adding the bootstrapped noise to the entire region-wise fitted concentration curve, the components for each component except the one being tested should be retained for each voxel, but the component being tested should be replaced by that component for the region-level data. To make this more clear, Step 7(b) should be replaced by:

7(b′) Compute $Y_{ij}^{(b)} = \sum_{k'=1}^{K} f(t_i; \hat{\gamma}_{k'}, \hat{\beta}_{jk'}) - f(t_i; \hat{\gamma}_k, \hat{\beta}_{jk}) + f(t_i; \hat{\gamma}_k, \tilde{\beta}_k) + U_i e_{ij}$ for $j \in R$ and for each i.

Thus, the "fitted value" for each voxel will consist of an ROI-level fit for the component under consideration and voxel-level fits for the other components. Accordingly, step 7e should be replaced by

7(e′) Compute $T_k^{(b)} = \frac{1}{J\hat{\gamma}_k^2} \sum_{j \in R} \left(\hat{\beta}_{jk}^{(b)} - \bar{\hat{\beta}}_{\cdot k}^{(b)} \right)^2$, where $\bar{\hat{\beta}}_{\cdot k}^{(b)} = \frac{1}{J} \sum_{j \in R} \hat{\beta}_{jk}^{(b)}$.

Then in step 8, the bootstrap p value is determined the same way, comparing the computed value of T_k with the distribution of $T_k^{(b)}$ values.

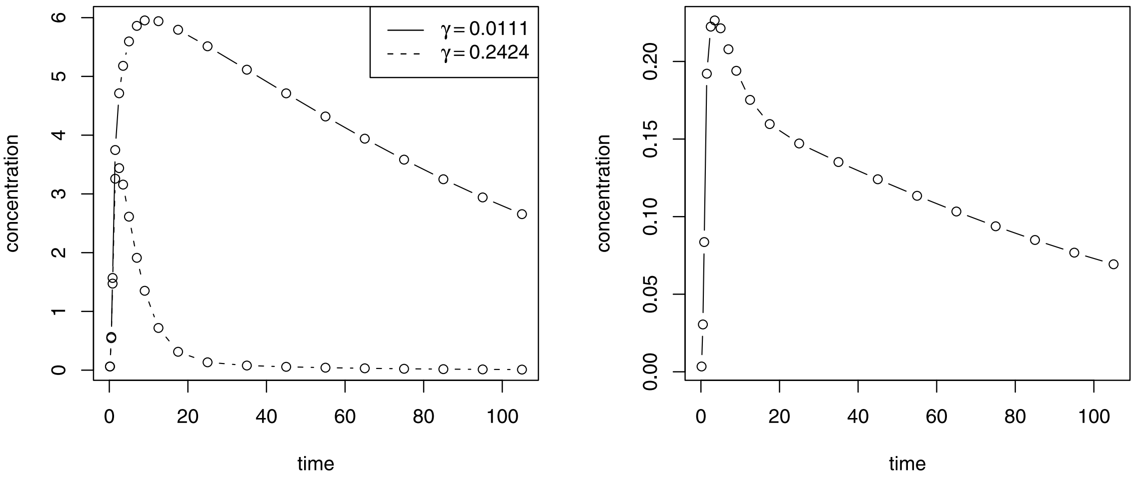

Figure 1. Settings for the simulation. The panel on the left displays the two mixture components; that on the right shows the "true" concentration curve for each voxel in the "region".

3. SIMULATION STUDY

To investigate the size of the test determined by the bootstrap procedure as laid out in Section 2.4, a simulation study was conducted. Parameters for the simulation were set to estimates obtained by fitting a real data set so as to be as realistic as possible. Time points were set to 1/6, 1/2, 5/6, 1.5, 2.5, 3.5, 5, 7, 9, 12.5, 17.5, 25, 35, 45, 55, 65, 75, 85, 95, 105 and the w_i for the weights were set to the reciprocals of the frame durations. We set K, the number of components to 2 and the "true" values of parameters $\gamma_1 = 0.0111$ and $\gamma_2 = 0.242$ with the (common) coefficient values for the region set as $\beta_1 = 0.0260$ and $\beta_2 = 0.0291$. The two components for the simulated data are displayed in the left-hand panel of Figure 1. The first component ($\gamma_1 = 0.0111$) exhibits a relatively slow decline that would correspond to having available receptors to which the ligand can bind. The time course of the other component rises quickly with blood flow but then drops quickly as well; such a time course would be expected in a region which has no receptors available for binding with the ligand. Most areas of the brain will have some contribution from each of these two components. The resulting "true" concentration curve for each voxel in the region is shown in the right-hand panel of Figure 1. The noise level σ was set to 0.1469 and the number of bootstrap samples was $B = 500$.

Time activity curves every voxel within the "region" were generated using Gaussian noise, scaled by the w_i factors as in (1), independently over time, but correlated across the region by an autoregressive process, added to the "true" concentration curve. The bootstrap algorithm in Section 2.4 was applied to each dataset in turn and the p-value determined. This was done for a range of region sizes from 10 to 1,000 voxels, and for three values of the autoregressive parameter ϕ (0.5, 0.7, and 0.9). For each region size and autoregressive parameter choice, 10,000 datasets were generated, and the bootstrap algorithm with $B = 500$ was applied to each.

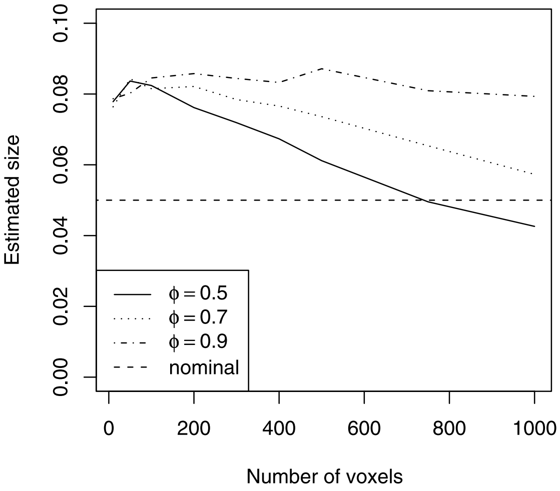

Figure 2. Estimated size of the bootstrap test with nominal Type I error rate $\alpha = 0.05$, using simulated data for various region sizes and for varying amounts of spatial correlation. Each estimated size is based on 10,000 replications with $B = 500$ bootstrap samples computed for each generated dataset.

This entire simulation procedure was repeated two additional times with different parameter values chosen to match those estimated from real data. Conclusions are similar for the three settings, and thus we present results that are aggregated over all settings. Results for the actual size of the test are given in Figure 2, for which the nominal size is $\alpha = 0.05$. For regions with relatively few voxels the actual size of the test is somewhat higher than the nominal size for all values of the autoregressive parameter. As the number

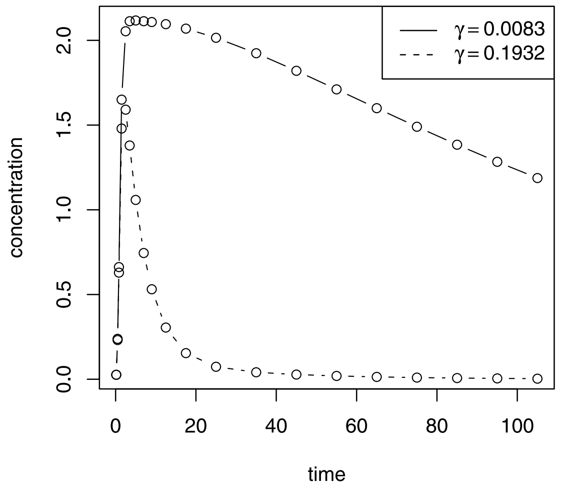

Figure 3. The two mixture components estimated for the real dataset.

of voxels in the region increases, the actual size tends to gradually decrease. For smaller regions there is little difference in the size of the test for the three values of ϕ, but for larger regions this difference is more pronounced. We note that the effect of region size on the size of the test is greater for the situations with lower autocorrelation — possibly because the equivalent number of uncorrelated observations increases more quickly as region sizes increases when there is less autocorrelation.

4. APPLICATION TO BRAIN DATA

The procedures outlined in Section 2 were applied to existing imaging data. The ligand $[^{11}C]$WAY-100635 binds to the serotonin 1A receptor, which has been implicated in depression studies ([Drevets et al., 1999; Sargent et al., 2000; Parsey et al., 2006]). The imaging data was obtained as described in Parsey et al. (2006) with full arterial sampling. Forty regions of interest were traced on an individual structural MRI scan based on brain atlases and published reports and subsequently verified by a neuroanatomist. For each region, p-values were computed using the bootstrap algorithm from Section 2.4 generating $B = 1,000$ replications for each.

Fitting the mixture model to this dataset as described in Section 2.3 indicates that two components are sufficient. The two components are plotted in Figure 3 and are similar to those used in the simulations (cf. Figure 1).

The test statistic values for the regions are plotted against region size in Figure 4. Evidence for heterogeneity was very strong for almost all regions, as all but two regions resulted in p-values less than 0.05. The two exceptions were the left dorsal caudate and the right dorsal caudate, indicated on the plot by the two solid circles. The largest region, the cerebellum (commonly used as a reference region for this ligand) with 10,773 voxels, resulted in a variance of 1.57 and a p-value less than 0.001 (not plotted).

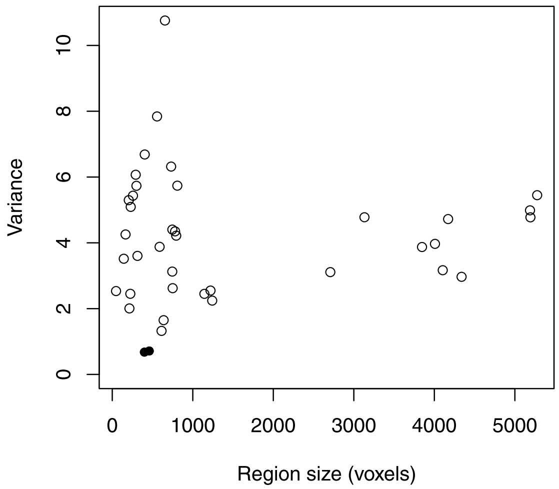

Figure 4. Region size vs. variance of several regions for a WAY study. The solid circles indicate regions for which the p-value for heterogeneity were greater than 0.05.

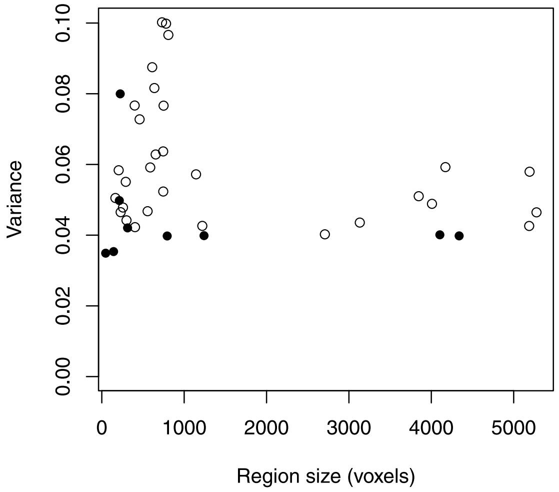

Figure 5. Region size vs. variance of the second component only of several regions for a WAY study. The solid circles indicate regions for which the p-value for heterogeneity were greater than 0.05.

Further analysis shows that for each of the regions considered, almost all of the variance of the voxel-level estimates of V is due to variance in the first component, which roughly corresponds to specific binding. Measuring the amount of variability in the first component and testing for its significance yields results very similar to that of the overall variance: only two regions (again, the left and right dorsal caudate) result in a p-value of greater than 0.05. The variance of the contribution of the second component is more varied, however. Figure 5 shows the variance of the regions plotted against their size. Again, the regions for which the bootstrap p-value is greater than 0.05 are plotted with solid black circles; the open circles correspond to p-values less than 0.05.

The results of this analysis indicate that assuming regional homogeneity for this dataset would not be appropriate for any of the regions considered (except for the two noted earlier). It is interesting to note the wide range of variances across the regions, and that there is no clear relationship between region size and variance of the volume of distribution. As mentioned earlier, the variances in total volume of distribution largely consist of variance in the first component, roughly corresponding to variation in the amount of specific binding. The other component, primarily representing "free" (and non-specifically bound) tracer, typically assumed to be constant across the entire brain, is still quite variable even within regions. Analysis on other imaging datasets not displayed here, reveal similar patterns as those shown here.

We acknowledge that one potential contributor to measures of heterogeneity of a region is the well-known partial volume effect, in which a truly homogenous region may be "contaminated" by surrounding areas of the brain, and thus the estimated variance might tend to overstate the truth. Relatively small regions are particularly susceptible to this effect. This can be avoided to some extent when assessing homogeneity by performing the procedure only on voxels in the interior section of a region, although this is not a completely satisfactory solution. Alternatively, testing for homogeneity of regions could be performed on data that has undergone partial volume "correction" (PVC), for which a number of algorithms have been proposed (see, e.g., Aston et al. [2002]).

To investigate the possible effect of such a correction on the procedures described here, we have repeated the analysis on the same data after applying the PVC procedure of Meltzer et al. (1990). As would be reasonably expected, the estimated variances for the regions of the PVC data tend to be larger than the corresponding variances of the original data (Figure 6, upper left-hand panel). PVC appears to have less of an effect on the estimated significance of the tests for heterogeneity, however; for the PVC data, only five regions had calculated p-values greater than 0.

As with the original data, the variance of the regions is dominated by the first component, and the conclusions made on the first component for the PVC data are similar to those made on the first component of the original data. The second component allows for somewhat more interesting comparison, however. Test statistics for the second component are plotted against each other in the upper right-hand panel of Figure 6 with an identity line added for reference. Estimated variance tends to be higher for the PVC data than for the original data. Corresponding p-values for the second component are plotted against each other in the lower left-hand panel of Figure 6 (both axes on the log scale). A reference line at 0.05 was included in each direction. (Note that since many of the bootstrap-based p-values were zero and so in order to make all points appear on the plot, 0.0001 was added to all p-values.) The variance of each region for the second component was plotted against region size for the PVC data (cf. Figure 5), again with solid circles corresponding to regions with $p > 0.05$, demonstrating a similar pattern.

5. DISCUSSION

A more parametric alternative to the test statistic put forth in Section 2 would be through standard likelihood ratio testing. This could entail assuming a Gaussian distribution for all data with the specified correlation structure and fitting the unconstrained mixture model as in Section 2. Then a restricted mixture model would be fit to all voxels, constraining the $\beta_{j,k}$ parameters to be the same for all voxels in the region. Such a procedure would rest strongly on the parametric assumptions, and the large number of degrees of freedom may decrease the power, so a less parametric alternative, such as that presented here, may be preferred in many situations.

The algorithm laid out in Section 2.4 could be applied with the estimates of V_j computed according to compartmental modeling or any of the other various alternative methods. Taking the mixture modeling approach as we've done here requires considerably less computational expense than standard compartmental modeling since the γ_k values do not have to be estimated for each bootstrap sample and because the non-negative least squares algorithm is relatively fast. Another reason to prefer estimating V_j using mixture modeling techniques is because of the easy extension to component-wise testing in Section 2.5.

That said, while the mixture model is less dependent on some of the particular parametric assumptions required by the usual compartmental model, it is still relatively strongly parametric, requiring that the impulse response function be expressed as a sum of exponentials. In some situations it may be preferable to relax this assumption as well, for instance the nonparametric approach of O'Sullivan et al. (2009). Though application of the bootstrap testing procedure outlined here to such a nonparametric approach would likely involve intensive computational demands, it would serve to provide analysis less affected by potential lack of model fit.

The other natural bootstrap technique to apply would be to compute residuals for each voxel and then resample them ([Davison and Hinkley, 1997]). If the permutation vectors are applied to all voxels and if the resampled residuals are assigned to the same voxel from which they were derived, then the covariance structure should be maintained. A separate simulation study following the settings outlined in Section 3 but resampling from residuals was conducted but the results indicated that even for relatively small regions the distribution of the variances of the bootstrapped estimates was always more variable than that for the original data and thus p-values in the null hypothesis case were almost all between 0.2 and 0.8. This may be remedied by assuming stationarity of the noise and then using a time-series

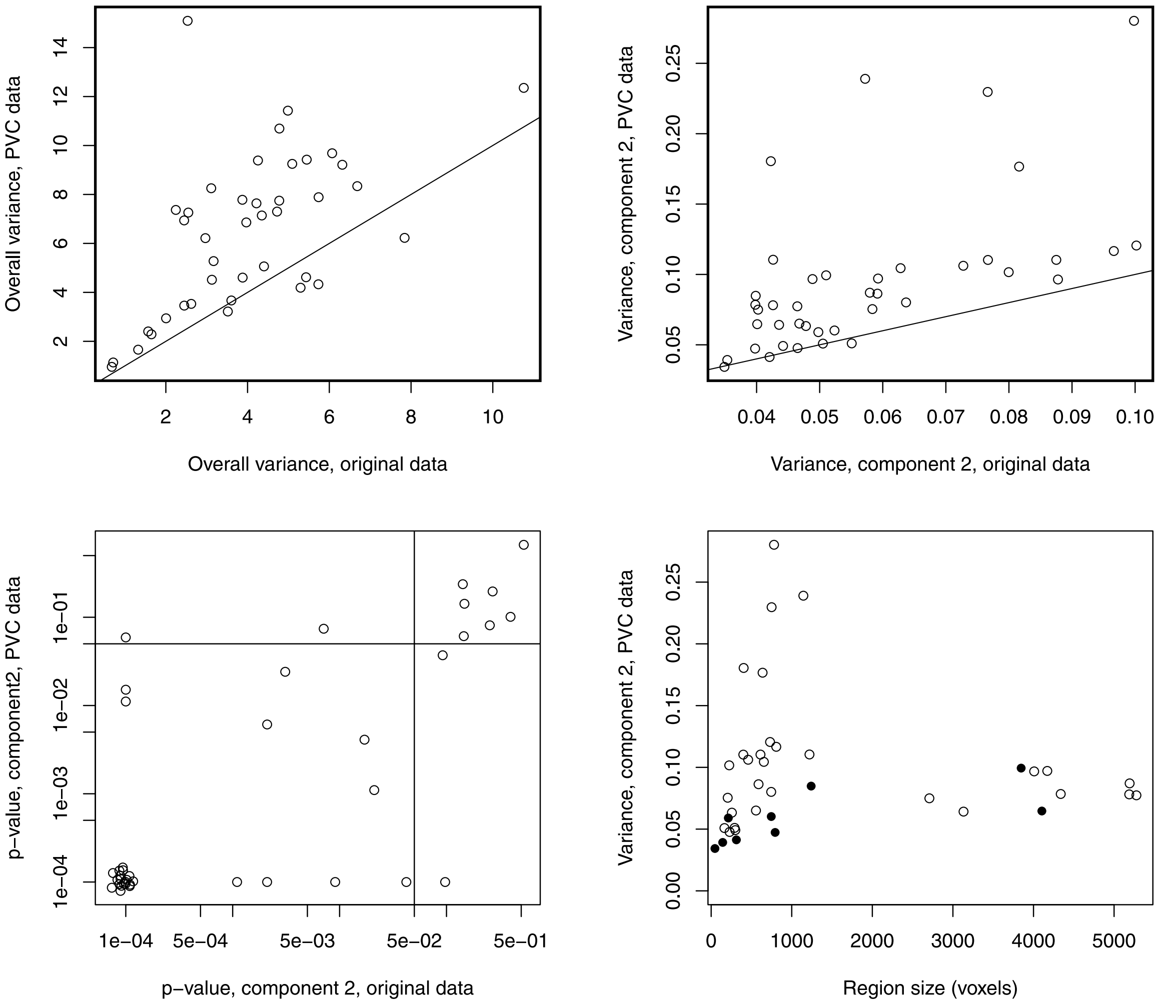

Figure 6. Top left: overall variance, original data vs. PVC data. Top right: Variance of second component, original data vs. PVC data. Bottom left: p-values of component 2, original data vs. PVC data, plotted on the log scale. The quantity 0.0001 was added to each p-value so that all p-value pairs will appear on the plot, and some jittering was performed for the regions for which both p-values were zero. Reference lines at 0.05 are given in both directions. Bottom right: Region size vs. variance of the second component. The solid circles indicate regions for which the p-value for heterogeneity were greater than 0.05.

bootstrap algorithm such as that proposed by Romano and Thombs (1996). An extension of their method to an irregular shape in three-dimensional space may be challenging but we note that with a stationarity assumption residuals may be drawn from anywhere within the brain.

Received 1 September 2009

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki, eds., pp. 267–281. Akademiai Kiado, Budapest. MR0483125

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**:716–733. MR0423716

Aston, J. A. D., Cunningham, V. J., Asselin, M.-C., Hammers, A., Evans, A. C., and Gunn, R. N. (2002). Positron emission tomography partial volume correction: Estimation and algorithms. *Journal of Cerebral Blood Flow and Metabolism* **22**:1019–1034.

Chang, C. and Ogden, R. T. (2009). Bootstrapping sums of independent but not identically distributed continuous processes with applications to functional data. *Journal of Multivariate Analysis* **100**:1291–1303. MR2508388

Cselényi, Z., Olsson, H., Farde, L., and Gulyás, B. (2002). Wavelet-aided parametric mapping of cerebral dopamine D2 receptors using the high affinity PET radioligand [11C]FLB 457. *NeuroImage* **17**:47–60.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge. MR1478673

Drevets, W. C., Frank, E., Price, J. C., Kupfer, D. J., Holt, D., Greer, P. J., Huang, Y., Gautier, C., and Mathis, C. (1999). PET imaging of serotonin 1A receptor binding in depression. *Biological Psychiatry* **46**:1375–87.

Flachaire, E. (2005). Bootstrapping heteroscedastic regression models: Wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis* **49**:361–376. MR2137681

Gunn, R. N., Gunn, S. R., and Cunningham, V. J. (2001). Positron emission tomography compartmental models. *Journal of Cerebral Blood Flow and Metabolism* **21**:635–652.

Jacquez, J. A. (1985). *Compartmental Analysis in Biology and Medicine, Second Edition*. University of Michigan Press, Ann Arbor, Michigan.

Jiang, H. and Ogden, R. T. (2008). Mixture modeling for dynamic PET data. *Statistica Sinica* **18**:1341–1356. MR2468271

Lammertsma, A. A. and Hume, S. P. (1996). Simplified reference tissue model for PET receptor studies. *NeuroImage* **4**:153–158.

Lawson, C. L. and Hanson, R. J. (1974). *Solving least squares problems*. Prentice-Hall. MR0366019

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics* **16**:1696–1708. MR0964947

Meltzer, C. C., Leal, J. P., Mayberg, H. S., Wagner, H. J., and Frost, J. J. (1990). Correction of PET data for partial-volume effects in human cerebral cortex by MR imaging. *Journal of Computer Assisted Tomography* **14**:561–570.

O'Sullivan, F. (2006). Locally constrained mixture representation of dynamic imaging data from PET and MR studies. *Biostatistics* **7**:318–338.

O'Sullivan, F., Muzi, M., Spence, A. M., Mankoff, D. M., O'Sullivan, J. N., Fitzgerald, N., Newman, G. C., and Krohn, K. A. (2009). Nonparametric residue analysis of dynamic PET data with application to cerebral FDG studies in normals. *Journal of the American Statistical Association* **104**:556–571.

Parsey, R. V., Oquendo, M. A., Ogden, R. T., Olvet, D. M., Simpson, N., Huang, Y.-Y., Van Heertum, R. L., Arango, V., and Mann, J. J. (2006). Altered serotonin 1A binding in major depression: A [carbonyl-C-11]WAY100635 positron emission tomography study. *Biological Psychiatry* **59**:106–113.

Romano, J. and Thombs, L. (1996). Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association* **91**:590–600. MR1395728

Sargent, P. A., Kjaer, K. H., Bench, C. J., Rabiner, E. A., Messa, C., Meyer, J., Gunn, R. N., Grasby, P. M., and Cowen, P. J. (2000). Brain serotonin$_{1A}$ receptor binding measured by positron emission tomography with [11C]WAY-100635: Effects of depression and antidepressant treatment. *Archives of General Psychiatry* **57**:174–180.

Slifstein, M. and Laruelle, M. (2001). Models and methods for derivation of *in vivo* neuroreceptor parameters with PET and SPECT reversible radiotracers. *Nuclear Medicine and Biology* **28**:595–608.

Whitcher, B., Tuch, D. S., Wisco, J. J., Sorensen, A. G., and Wang, L. (2008). Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. *Human Brain Mapping* **29**:346–362.

Zhu, H. T., Ibrahim, J. G., Tang, N. S., Hao, X. J., Bansal, R., and Peterson, B. (2007). A statistical analysis of brain morphology using wild bootstrapping. *IEEE Transactions on Medical Imaging* **26**:954–966.

R. Todd Ogden
Department of Biostatistics
Columbia University
New York, NY 10032
E-mail address: todd.ogden@columbia.edu

Huiping Jiang
Biostatistics Division
New York State Psychiatric Institute
New York, NY 10032
E-mail address: huiping2004@gmail.com