A Statistical Analysis of Brain Morphology Using Wild Bootstrapping

Hongtu Zhu*, Joseph G. Ibrahim, Niansheng Tang, Daniel B. Rowe, Xuejun Hao, Ravi Bansal, and Bradley S. Peterson

***Abstract*—Methods for the analysis of brain morphology, including voxel-based morphology and surface-based morphometries, have been used to detect associations between brain structure and covariates of interest, such as diagnosis, severity of disease, age, IQ, and genotype. The statistical analysis of morphometric measures usually involves two statistical procedures: 1) invoking a statistical model at each voxel (or point) on the surface of the brain or brain subregion, followed by mapping test statistics (e.g., t test) or their associated p values at each of those voxels; 2) correction for the multiple statistical tests conducted across all voxels on the surface of the brain region under investigation. We propose the use of new statistical methods for each of these procedures. We first use a heteroscedastic linear model to test the associations between the morphological measures at each voxel on the surface of the specified subregion (e.g., cortical or subcortical surfaces) and the covariates of interest. Moreover, we develop a robust test procedure that is based on a resampling method, called wild bootstrapping. This procedure assesses the statistical significance of the associations between a measure of given brain structure and the covariates of interest. The value of this robust test procedure lies in its computationally simplicity and in its applicability to a wide range of imaging data, including data from both anatomical and functional magnetic resonance imaging (fMRI). Simulation studies demonstrate that this robust test procedure can accurately control the family-wise error rate. We demonstrate the application of this robust test procedure to the detection of statistically significant differences in the morphology of the hippocampus over time across gender groups in a large sample of healthy subjects.**

***Index Terms*—Heteroscedastic linear model, hippocampus, multiple hypothesis test, permutation test, robust test procedure.**

I. Introduction

VARIOUS methods for modeling the morphology of the brain, including voxel-based, surface-based, and tensor-based morphometries, provide invaluable tools for understanding neuroanatomical differences in brain structure across subjects [1]–[7]. Statistical analysis of these morphometric measures can subsequently be used to understand normal brain development, the neural bases of neuropsychiatric disorders, and how environmental and genetic factors interact to determine brain structure and function. For instance, a joint analysis of brain morphometry and genotype may reveal brain regions with strong heritability in healthy subjects [8], [9]. Moreover, some measures of brain structure may be used as an endophenotypic marker for a disease if statistical analyses show that they are associated with behavioral, cognitive, or clinical outcomes [6], [10]–[14]. Studies of brain morphology have been conducted widely to characterize differences in brain structure across differing populations, such as patients with schizophrenia and healthy subjects [7], [15]–[19].

The statistical analysis of morphometric measures usually involves two procedures executed in sequence. The first procedure entails fitting a general linear model (LM) to the morphometric data from all subjects at each voxel and generating a statistical parametric map that contains a statistic (or a p value) at each voxel [20]. The second procedure entails using various statistical methods (e.g., random field theory, false discovery rate, permutation method) to calculate adjusted p values that account for the multiple statistical tests that are conducted across the many voxels of the brain region [21], [22]. All these statistical methods are implemented in existing neuroimaging software platforms, such as SPM, FSL, and SnPM.

The existing methods for these two procedures, however, have at least three limitations. First, the general linear model used in the neuroimaging literature usually involves two key assumptions: that the variance of the imaging data are homogeneous across subjects and that the data conform to a Gaussian distribution at each voxel. These two assumptions are critically important for the valid calculation of parametric distributions (e.g., t, F, and T) in conventional tests (e.g., t test) that assess the statistical significance of parameter estimates in the general linear model [3], [23]. Diagnostic procedures have been proposed to test these assumptions of the general linear model [24], [25], yet few statistical methods have been developed to analyze imaging data when these two assumptions are not

Manuscript received November 14, 2006; revised February 2, 2007. This work was supported in part by the Suzanne Crosby Murphy Endowment at Columbia University Medical Center and in part by the Thomas D. Klingenstein and Nancy D. Perlman Family Fund. The work of H. Zhu was supported by the NSF under Grant SES-0643663. The work of J. G. Ibrahim was supported by the NIH under Grant GM 70335 and Grant CA 74015. The work of N. Tang was supported by in part by NSFC under Grant 10561008 and in part by NSFYN under Grant 2004A0002M. The work of D. B. Rowe was supported by the NIH under Grant EB000215. The work of B. S. Peterson was supported in part by the NIDA under Grant DA017820 and in part by NIMH under Grant MH068318 and Grant K02-74677. *Asterisk indicates corresponding author.*

*H. Zhu is with Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC 27599-7420 USA (e-mail: hzhu@bios.unc.edu).

J. G. Ibrahim is with Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420 USA (e-mail: ibrahim@bios.unc.edu).

N. Tang is with Department of Statistics, Yunnan University, Kunming, 650091, China (e-mail: nstang@ynu.edu.cn).

D. B. Rowe is with Department of Biophysics, Medical College of Wisconsin, Milwaukee, WI 53000 USA (e-mail: dbrowe@mcw.edu).

X. Hao, R. Bansal, and B. S. Peterson are with New York State Psychiatric Institute, New York, NY 10032 USA and with the Department of Psychiatry, Columbia University, New York, NY 10032 USA (e-mail: haox@childpsych.columbia.edu; bansalr@childpsych.columbia.edu; petersob@childpsych.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2007.897396

satisfied. Second, the methods of random field theory that account for multiple statistical comparisons depend strongly on these assumptions of the general linear model, as well as several additional assumptions (e.g., smoothness of autocorrelation function) [21]. Third, permutation methods require the so-called "complete exchangeability" [26]–[28]. Complete exchangeability, however, is in fact a very strong assumption. For instance, consider two diagnostic groups (healthy controls and a disease group) and suppose that the null hypothesis is that the morphometric measures in all voxels from the two groups have the same mean. A permutation null distribution actually enforces equal distributions in the two groups in all voxels, which is a much stronger assumption than that of equal means across groups [26], [28].

The aim of this paper is to use new statistical methods to address these three limitations of extant methods for morphometric analyses. Specifically, we propose to apply two statistical techniques to the analysis of brain morphology: a heteroscedastic linear model, which avoids the two key assumptions of the general linear model, and a robust test procedure to correct for multiple statistical tests.

First, we use a heteroscedastic linear model together with a Wald-type statistical test to test linear hypotheses of brain morphology. The heteroscedastic linear model does not assume the presence of homogeneous variance across subjects, and it allows for a large class of distributions in the imaging data. These extensions are desirable for the analysis of real-world imaging data (e.g., anatomical and functional magnetic resonance imaging (fMRI) data, positron emission tomography measures), because between-subject and between-voxel variability in the imaging measures can be substantial [29]–[31]. Moreover, the distribution of the imaging data often deviates from the Gaussian distribution (see example in Sections III and IV) [2], [6], [23]. Under the heteroscedastic linear model, we calculate the ordinary least squares (OLS) estimator (denoted by $\hat{\beta}$) to estimate the associations (denoted by β) between the measures of a brain region and the covariates of interest. We then use a Wald-type test statistic based on a consistent estimator of the covariance matrix (CECM) for $\hat{\beta}$ under the null hypothesis [32], [33]. Although the Wald-type test statistic does not have a simple parametric distribution, we can use a wild bootstrap method to improve the finite performance of the Wald-type test statistic. The wild bootstrap method has been shown to have good theoretical properties and excellent performance in practice [34], [35].

To test multiple hypotheses across all voxels of a brain region, we propose a robust test procedure to control the family-wise error rate. Specially, we perform the Wald-type test statistic using the wild bootstrap method simultaneously at all voxels of the brain region, while preserving the dependence structure among the test statistics. In addition, the wild bootstrap method does not involve repeated analyses of simulated datasets and therefore is not computationally intensive. Specifically, the wild bootstrap method requires neither complete exchangeability nor a Gaussian distribution for the imaging data. The robust test procedure is, thus, widely applicable to other imaging modalities, including fMRI and positron emission tomography (PET) data.

II. Methods

Here, we formally introduce the heteroscedastic linear model and use a Wald-type test statistic for testing linear hypotheses of β. We then present a robust test procedure based on the wild bootstrap as a method for correcting p values for multiple statistical comparisons.

A. Heteroscedastic Linear Model and a Wald-Type Test Statistic

In a particular voxel d on the brain structure, we consider the following heteroscedastic linear model:

$$y_t = X_t^T\beta + \varepsilon_t, \quad E(\varepsilon_t|X_t) = 0, \quad E\left(\varepsilon_t^2|X_t\right) = \sigma_t^2 \tag{1}$$

for $t = 1, \ldots, n$, where t represents the tth subject, y_t represents a measure of brain morphology (e.g., signed Euclidean distance, grey matter density), X_t is an exogenous $k \times 1$ vector (e.g., age, gender, and genotype), β is a $k \times 1$ vector of unknown parameters, and ε_t is a random error term. Let $Y = (y_1, \ldots, y_n)^T$, $X = (X_1, \ldots, X_n)^T$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, and $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, where the superscript T represents transpose. Then, (1) can be rewritten as

$$Y = X\beta + \varepsilon, \quad E(\varepsilon|X) = \mathbf{0}, \quad E(\varepsilon\varepsilon^T|X) = \Omega. \tag{2}$$

Here, without loss of generality, we assume that X is a column full rank matrix, i.e. $\mathrm{rank}(X) = k$. The ordinary least squares estimate of parameter β, given by $\hat{\beta} = (X^TX)^{-1}X^TY$, has been implemented in SPM[1] and widely used in many neuroimaging studies, because of its computational simplicity. In contrast, if Ω were known, we could use the generalized least squares estimate of β, which is more efficient than $\hat{\beta}$ [36]. However, except for a few special cases (e.g., fMRI), we rarely have prior information to consistently estimate Ω [32], [34].

Let I_n be an $n \times n$ identity matrix, $P_X = X(X^TX)^{-1}X^T$, and $h_t = X_t^T(X^TX)^{-1}X_t$ for $t = 1, \ldots, n$. The covariance matrix of $\hat{\beta}$ is given by

$$\mathrm{Cov}(\hat{\beta}) = (X^TX)^{-1}X^T\Omega X(X^TX)^{-1}$$

while a consistent estimator of the covariance matrix (CECM) of $\hat{\beta}$ in model (2) has the following form:

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = (X^TX)^{-1}X^T\widehat{\Omega}X(X^TX)^{-1} \tag{3}$$

where $\widehat{\Omega} = \mathrm{diag}(a_1^2\hat{\varepsilon}_1^2, \ldots, a_n^2\hat{\varepsilon}_n^2)$, $a_t = 1/(1 - h_t)$, and $\hat{\varepsilon}_t$ is the tth component of $\hat{\varepsilon} = Y - X\hat{\beta} = (I_n - P_X)Y$ [34]. It should be noted that $\widehat{\Omega}$ is not a consistent estimate of Ω, whereas because $n^{-1}X^T(\widehat{\Omega} - \Omega)X$ converges to zero, $\widehat{\mathrm{Cov}}(\hat{\beta})$ is a CECM [32].

Ignoring heteroscedasticity in model (1) leads to using $\hat{\sigma}^2(X^TX)^{-1}$ as an estimate of the covariance matrix for the

[1] Available: http://www.fil.ion.ucl.ac.uk/spm/

OLS estimator $\hat{\beta}$, where $\hat{\sigma}^2 = Y^T(I - P_X)Y/(n-k)$. However, failure to account for interindividual variance can lead to the following consequences: 1) $\hat{\sigma}^2(X^TX)^{-1}$ may be inconsistent; 2) conventional statistics for testing linear hypotheses of β do not follow t and F distributions; 3) invalid inferences based on $\hat{\sigma}^2(X^TX)^{-1}$ lead to large Type I and/or Type II error rates for testing linear hypotheses of β [32], [34], [37].

We consider testing the linear hypotheses

$$H_0 : R\beta = b_0 \quad \text{versus} \quad H_1 : R\beta \neq b_0 \tag{4}$$

where R is an $r \times k$ matrix of full row rank and b_0 is an $r \times 1$ specified vector. We test the null hypothesis $H_0 : R\beta = b_0$ using a Wald-type test statistic

$$W_n = (R\hat{\beta} - b_0)^T \Sigma_{\tilde{\Omega}}^{-1} (R\hat{\beta} - b_0) \tag{5}$$

where $\Sigma_{\tilde{\Omega}}$ is a consistent estimate of the covariance matrix of $R\hat{\beta} - b_0$ under H_0. Explicitly, $\Sigma_{\tilde{\Omega}}$ is given by

$$\Sigma_{\tilde{\Omega}} = R(X^TX)^{-1}X^T\tilde{\Omega}X(X^TX)^{-1}R^T \tag{6}$$

where $\tilde{\Omega} = \mathrm{diag}(a_1^2\tilde{\varepsilon}_1^2, \cdots, a_n^2\tilde{\varepsilon}_n^2)$ and $\tilde{\varepsilon}_t$ is the tth component of $\tilde{\varepsilon}$ as given in (8). Various simulation studies have shown that the use of $\tilde{\varepsilon}$ leads to a better control of Type I error rates [34], [38].

Under H_0, a restricted least squares (RLS) estimate of β, denoted by $\tilde{\beta}$, is given by (Appendix I)

$$\tilde{\beta} = \hat{\beta} - (X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1}(R\hat{\beta} - b_0) \tag{7}$$

and a restricted residual vector $\tilde{\varepsilon} = Y - X\tilde{\beta}$ is calculated to be

$$\tilde{\varepsilon} = \hat{\varepsilon} + X(X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1}(R\hat{\beta} - b_0). \tag{8}$$

Because W_n is asymptotically distributed as $\chi^2(r)$, a chi-square distribution with r degrees-of-freedom, under the null hypothesis H_0, an asymptotically valid test can be obtained by comparing sample values of test statistic with the critical value of the right-hand tail of $\chi^2(r)$ distribution at a prespecified significance level α [32]. That is, we reject H_0 if $W_n = w_n \geq \chi_\alpha^2(r)$, and do not reject H_0 otherwise, where $\chi_\alpha^2(r)$ is the upper α-percentile of the $\chi^2(r)$ distribution. However, for small n, numerical results have shown that W_n may yield misleading results (large Type I and/or Type II error rates) [34], [35], [37], [39].

B. Wild Bootstrap

We present a wild bootstrap method to improve the finite performance of W_n in testing the null hypothesis H_0. This wild bootstrap method has been extensively studied in the literature [34], [35]. To use wild bootstrapping to test $H_0 : R\beta = b_0$, we generates bootstrap samples that conform to the null hypothesis. Thus, we estimate the unknown parameters of β under the constraint $R\beta = b_0$, which is exactly the RLS estimator of β, $\tilde{\beta}$. Then, a p value can be calculated based on the generated bootstrap samples.

To produce a bootstrap sample $\{(y_t^*, X_t) : t = 1, \cdots, n\}$, we use the following data-generating process (DGP):

$$y_t^* = X_t^T\tilde{\beta} + a_t\tilde{\varepsilon}_t\varepsilon_t^* \tag{9}$$

where $\tilde{\beta}$ and $\tilde{\varepsilon}$ are, respectively, defined in (7) and (8), and ε_t^* are independently and identically distributed as a distribution F. Following Flachaire [34], F is chosen as

$$\varepsilon_t^* = \begin{cases} 1, & \text{with probability } 0.5 \\ -1, & \text{with probability } 0.5 \end{cases}. \tag{10}$$

Thus, a bootstrap sample $\{(y_t^*, X_t) : t = 1, \cdots, n\}$ can be obtained using the data-generating process (9). Let $Y^* = (y_1^*, \cdots, y_n^*)^T$, and $\varepsilon^* = (\varepsilon_1^*, \cdots, \varepsilon_n^*)^T$. Equation (9) can be rewritten as

$$Y^* = X\tilde{\beta} + \tilde{\Omega}^{1/2}\varepsilon^*. \tag{11}$$

We now calculate the Wald-type test statistic for the bootstrap sample. It follows from (11) that the ordinary and restricted least squares estimates of β are, respectively, given by

$$\begin{aligned} \hat{\beta}^* &= \tilde{\beta} + (X^TX)^{-1}X^T\tilde{\Omega}^{1/2}\varepsilon^*, \text{ and} \\ \tilde{\beta}^* &= \hat{\beta}^* - (X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1} \\ &\quad \times (R\hat{\beta}^* - b_0). \end{aligned} \tag{12}$$

Thus, the ordinary residual vector of model (11) is given by

$$\hat{\varepsilon}^* = Y^* - X\tilde{\beta} - X(X^TX)^{-1}X^T\tilde{\Omega}^{1/2}\varepsilon^*.$$

Furthermore, the restricted residual vector of model (11), denoted by $\tilde{\varepsilon}^*$, is given by

$$\tilde{\varepsilon}^* = \hat{\varepsilon}^* + R_X\left[R(X^TX)^{-1}R^T\right]^{-1}R_X^T\tilde{\Omega}^{1/2}\varepsilon^* \tag{13}$$

where $R_X = X(X^TX)^{-1}R^T$. Let $\tilde{\Omega}^* = \mathrm{diag}(a_1^2\tilde{\varepsilon}_1^{*2}, \cdots, a_n^2\tilde{\varepsilon}_n^{*2})$, where $\tilde{\varepsilon}_t^*$ is the tth element of $\tilde{\varepsilon}^*$. Since $R\hat{\beta}^* - b_0 = R(X^TX)^{-1}X^T\tilde{\Omega}^{1/2}\varepsilon^*$, the Wald-type test statistic W_n^* based on the bootstrap sample is given by

$$W_n^* = \left[R_X^T\tilde{\Omega}^{1/2}\varepsilon^*\right]^T \Sigma_{\tilde{\Omega}^*}^{-1}\left[R_X^T\tilde{\Omega}^{1/2}\varepsilon^*\right] \tag{14}$$

where $\Sigma_{\tilde{\Omega}^*} = R(X^TX)^{-1}X^T\tilde{\Omega}^*X(X^TX)^{-1}R^T$.

We can approximate the p value of W_n as follows:

Step 1) Independently generate S bootstrap samples $\{(y_t^{*,(s)}, X_t) : t = 1, \ldots, n\}$ for $s = 1, \ldots, S$ using the bootstrap DGP (11).

Step 2) Calculate $W_n^{*,(s)}$ for each bootstrap sample.

Step 3) Approximate the p value of W_n by

$$\frac{1}{S}\sum_{s=1}^{S} I\left(W_n^{*,(s)} \geq W_n\right)$$

where $I(\cdot)$ is an indicator function.

We may consider other bootstrap methods, such as the pairs bootstrap, and other distributions F, such as the two-point distribution of Mammen [34], [40], [41]. For instance, we may use the pairs bootstrap method, but bootstrap samples generated by the pairs bootstrap method may not come from model (2) conforming to the null hypothesis $R\beta = b_0$. Thus, some appropriate modifications of the pair bootstrap are needed and these modifications lead to the use of the wild boostrap [34]. In addition, the use of the pair bootstrap can lead to a loss of power [34]. Various simulation studies have clearly shown that the wild bootstrap outperforms the pairs bootstrap in the literature [34], [42], [43]. The noise distribution (10) is justified by theoretical underpinnings and numerical simulations [34], [35].

The above heteroscedastic linear model and the wild bootstrap method can be used to analyze $\{(y_t, X_t) : t = 1, \ldots, n\}$ in each voxel d of the brain region. Henceforth, we use d in our notation if necessary, such as $\{W_n(d), W_n^{*,(s)}(d)\}$.

C. Robust Test Procedure

To test whether $H_0 : R\beta = b_0$ holds in all voxels of the brain region under investigation, we consider a maximum statistic, the maximum of the Wald-type test statistics, as

$$W_D = \max_d W_n(d). \tag{15}$$

The maximum statistic W_D plays a crucial role in controlling the family-wise error rate. In order to use W_D as a test statistic, we need to approximate the distribution of W_D under the null hypotheses in all voxels of the brain structure. We may apply random field theory for χ^2 processes to approximate the upper tail of W_D, because $W_n(d)$ converges to a $\chi^2(r)$ distribution under certain conditions as the number of subject n is sufficiently large [44]–[46]. However, the random field theory for χ^2 processes may be conservative because the asymptotic test of $W_n(d)$ leads to large Type I (and/or Type II) error rates in a single voxel d [37].

We propose a robust test procedure based on the wild bootstrap method to approximate the distribution of W_D. This procedure is implemented as follows.

Step 1) In each voxel d of the brain structure, calculate the Wald-type test statistic $W_n(d)$ given in (5) based on the observed data $\{(y_t(d), X_t) : t = 1, \ldots, n\}$. Compute $W_D = \max_d W_n(d)$.

Step 2) Generate a random sample $\{\varepsilon_t^* : t = 1, \ldots, n\}$ from the distribution F. In all voxels d, generate observations $\{y_t^*(d) : t = 1, \ldots, n\}$ from model (11) using the same sample $\{\varepsilon_t^* : t = 1, \ldots, n\}$.

Step 3) Calculate the Wald-type test statistic $W_n^*(d)$ based on the bootstrap sample $\{(y_t^*(d), X_t) : t = 1, \ldots, n\}$ and $W_D^* = \max_d W_n^*(d)$.

Step 4) Repeated Steps 2–3 S times and calculate $\{W_D^{*,(s)} : s = 1, \cdots, S\}$. Finally, the p value is approximated by

$$p_D \approx \frac{1}{S} \sum_{s=1}^{S} I\left(W_D^{*,(s)} \geq W_D\right). \tag{16}$$

We reject that the null hypothesis $H_0 : R\beta = b_0$ is true in all voxels of the brain structure if p_D is smaller than a prespecified value α.

Step 5) Calculate adjusted p value in each voxel d according to

$$p_D(d) \approx \frac{1}{S} \sum_{s=1}^{S} I\left(W_D^{*,(s)} \geq W_n(d)\right). \tag{17}$$

We note here at least four important advantages of this test procedure compared with existing procedures:

i) the wild bootstrap method performs well at each point d even for relatively small n (e.g., $n \leq 40$);

ii) the above test procedure asymptotically preserves the dependence structure among the $W_n(d)$;

iii) the above test procedure is not computationally intensive, because it does not involve the repeated analysis of simulated datasets;

iv) the above test procedure does not require complete exchangeability.

We can show that the robust test procedure asymptotically preserves the dependence structure among the $W_n(d)$. According to $W_n^*(d)$ in (14), the correlation between $W_n^*(d)$ and $W_n^*(d')$ is primarily determined by the correlation between $R_X^T \tilde{\Omega}^{1/2}(d)\varepsilon^*$ and $R_X^T \tilde{\Omega}^{1/2}(d')\varepsilon^*$. We can show that

$$\begin{aligned}\mathrm{Cov}&\left[R_X^T \tilde{\Omega}^{1/2}(d)\varepsilon^*, R_X^T \tilde{\Omega}^{1/2}(d')\varepsilon^* |\mathrm{data}\right]\\ &= R(X^T X)^{-1} \sum_{t=1}^{n} X_t X_t^T a_t^2 \tilde{\varepsilon}_t(d) \tilde{\varepsilon}_t(d') (X^T X)^{-1} R^T\end{aligned}$$

holds for any two points d and d'. Similarly, the correlation between $W_n(d)$ and $W_n(d')$ is primarily determined by the correlation between $R_X^T \varepsilon(d)$ and $R_X^T \varepsilon(d')$, which is given by

$$\begin{aligned}\mathrm{Cov}\left[R_X^T \varepsilon(d), R_X^T \varepsilon(d')\right] &= R(X^T X)^{-1}\\ &\times \sum_{t=1}^{n} X_t X_t^T \mathrm{E}\left[\varepsilon_t(d)\varepsilon_t(d')\right] (X^T X)^{-1} R^T.\end{aligned}$$

Thus, under some conditions [32]

$$n^{-1} \sum_{t=1}^{n} X_t X_t^T \left\{a_t^2 \tilde{\varepsilon}_t(d) \tilde{\varepsilon}_t(d') - \mathrm{E}\left[\varepsilon_t(d)\varepsilon_t(d')\right]\right\}$$

converges to zero in probability, and thus we have proved the advantage (ii).

III. Simulation Studies and Real-World Studies

We conducted two sets of Monte Carlo simulations. The first examined the finite performance of the wild bootstrap method for W_n at the single-voxel level. In particular, we compared its performance to the F test, the asymptotic test for W_n, and the permutation method based on the t test statistic. The second set of Monte Carlo simulations was to evaluate the family-wise error rate and power of the robust test procedure at the level of the whole surface (or brain). Then, we compared its performance to the permutation method based on the t test statistic and random field theory for F and χ^2 fields.

A. Monte Carlo Simulations: Set I

1) Design: For the first set of Monte Carlo simulations, we simulated data from the heteroscedastic linear model

$$y_t = X_t^T \beta + \epsilon_t \tag{18}$$

for $t = 1, \cdots, n$, where ϵ_t is a random error with zero mean, β is a $k \times 1$ vector of unknown parameters, and X_t is a $k \times 1$ vector of covariates of interest. Because of prior extensive simulations reported in the literature [34], [35], [37], [39], we chose a simple X_t as follows: $X_t = (1, 0)^T$ for $t = 1, \cdots, [n/2]$ and $X_t = (1, 1)^T$ for $t = [n/2]+1, \cdots, n$, where $[n/2]$ denotes the largest integer smaller than $n/2$. We set $n =$ 10, 20, and 40.

2) Random Errors: We considered the effects of three differing distributions of ϵ_t to examine the effects of these distributions on the finite performance of the four test statistics at the single-voxel level, including the wild bootstrap method for W_n, the F test, the asymptotic test for W_n, and the permutation method based on the t test statistic, at the single-voxel level. First, ϵ_t is a Gaussian error from $N(0, 1)$, where $N(\mu, \sigma^2)$ denotes a Gaussian distribution having a mean μ and standard deviation σ. The Gaussian errors with unit variance were generated to meet the assumptions of the general linear model. Second, we assumed $\epsilon_t = \chi^2(2) - 2$, in which $\chi^2(2)$ represents a chi-squared random variable with 2 degrees-of-freedom. The skewed distribution $\chi^2(2) - 2$ differs substantially from any Gaussian distribution. Third, we assumed that $\epsilon_t = \sigma_t z$ and z were independently generated from a $N(0, 1)$ distribution. Moreover, $\sigma(t) = \exp(u)$ when $X_{t,2} = 0$ and $\sigma(t) = \exp(u+1)$ when $X_{t,2} = 1$, where u were independently generated from a $N(0, 1)$ distribution. Conditional on u, the variances of ϵ_t were highly heterogeneous.

3) Hypothesis: We assumed $\beta = (\beta_0, \beta_1)^T = (1, 0)^T$ and set the null hypothesis $H_0 : \beta_1 = 0$ to assess the Type I error rates for the four test statistics. Furthermore, we assumed $\beta = (\beta_0, \beta_1)^T = (1, 2)^T$ and test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Then, we examined the Type II errors for the four test statistics (e.g., F test). In both cases, $R = (0, 1)$ and $b_0 = (0)$.

For each simulation, the significance level was set at $\alpha = 5\%$, and 20 000 replications were used to estimate the rejection rates. For a fixed α, if the Type I rejection rate is smaller than α, then the test is conservative, whereas if the Type I rejection rate is greater than α, then the test is anticonservative, or liberal [57].

B. Monte Carlo Simulations: Set II

1) Basic Design: We used a heteroscedastic linear model to generate data in all $m = 2064$ points on the surface of a reference sphere for all n subjects (or objects) (Fig. 1). For the tth subject, $\mathbf{Y}(t)$ denotes an $m \times 1$ vector that contains all morphometric measures (e.g., signed Euclidean distance, grey matter density) in all m points, B denotes an $m \times k$ matrix of unknown parameters, and $\mathbf{x}(t)$ is a $k \times 1$ vector of covariates of interest. The heteroscedastic linear model can be written as

$$\mathbf{Y}(t) = \mathbf{B}\mathbf{x}(t) + \sigma(t)\mathbf{C}^{1/2}\mathbf{e}(t), \quad E[\mathbf{e}(t)|\mathbf{x}(t)] = 0 \tag{19}$$

for $t = 1, \cdots, n$, where $\mathbf{e}(t)$ is an $m \times 1$ vector of independent Gaussian errors with zero mean and unit variance and $\mathbf{C}$ is an $m \times m$ correlation matrix. In addition, $\mathbf{x}(t)$, $\mathbf{C}$, and $\sigma^2(t)$ are specified below.

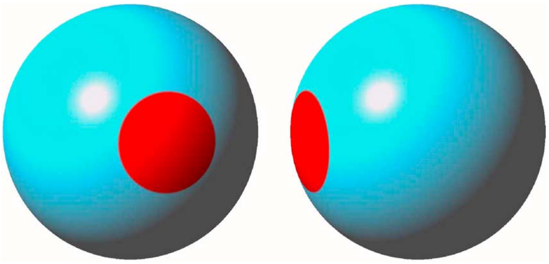

Fig. 1. Simulation study ROI. ROI is highlighted in red on the surface of a reference sphere: (a) anterior and (b) right lateral views.

2) Covariates of Interest: Our choice of statistical covariates was motivated by two scientific aims. The first was to compare brain structure across diagnostic groups (e.g., healthy controls (HC) and persons with schizophrenia) [8], [50]. To compare the performance of our robust test procedure with that of the permutation test, we choose a simple $\mathbf{x}(t)$ given by

$$\mathbf{x}(t) = (1, I(\text{the } t\text{th subject is HC}))^T \tag{20}$$

where HC denotes the healthy control group. Moreover, the first $[n/2]$ subjects were assumed to be healthy controls, and the rest were assumed to be patients.

The second scientific aim was to understand differences in brain structure across genders in the sample of healthy controls [19]. We choose $\mathbf{x}(t)$ as

$$\mathbf{x}(t) = (1, \text{Age}, \text{Gender})^T \tag{21}$$

where Gender equals 0 for males and 1 for females. We assumed that the first $[n/2]$ subjects were males, and the rest were females. The Age variable was uniformly generated from the interval $[1, n]$.

3) Variance Structure: We considered two types of variance structures: a homogeneous case and a heterogeneous case. For the homogeneous case, we assumed that $\sigma(t) \equiv 1$ for all $t = 1, \cdots, n$. However, for the heterogeneous case, we assumed that we observed larger variability from the male group as compared with the female group. Thus, for the covariates of interest in (20) and (21), we set $\sigma(t) = \exp(z)$ and generate z from a $N(0, 1)$ distribution for each man, and from a $N(1, 1)$ distribution for woman.

4) Correlation Structure: We considered a stationary and exponential correlation matrix $\mathbf{C}$, in which the correlation between any two points d and d' on the surface was given by $\rho^{\|d-d'\|}$, where $\rho \in [0, 1]$ and $\|d - d'\|$ represents Euclidean distance between d and d' [51]. We denote such an exponential correlation matrix by $\mathbf{C}(\rho)$. We simulated images based on $\mathbf{C}(\rho)$ using $\rho =$ 0, 0.25, 0.5, and 0.75 in order to mimic the differing degrees of smoothness in the simulated images [52].

5) Hypotheses: For $\mathbf{x}(t)$ in (20), we first assumed $\beta = (\beta_0, \beta_1)^T = (1, 0)^T$ in all points on the reference sphere to assess the family-wise error rate. In addition, to assess both the power and family-wise error rate, we selected a region-of-interest (ROI) with 64 points on the reference sphere and set $\beta_1 = 5$ for any point d in ROI [Fig. 1(a) and (b)]. In this case, k, the dimension of β, was 2. We were interested in testing the null hypothesis $H_0 : \beta_1 = 0$ at all points on the surface of the reference sphere. In this case, $R = (0, 1)$ and $b_0 = (0)$.

For $\mathbf{x}(t)$ in (21), we first assumed $\beta = (\beta_0, \beta_1, \beta_2)^T = (1, 0, 0)^T$ in all points on the surface of the reference sphere to assess the family-wise error rate. Furthermore, we used the same ROI on the reference sphere and set $\beta_2 = 5$ for any point d in ROI [Fig. 1(a) and (b)]. In this case, the dimension of β was 3. We were interested in testing whether any differences of morphology changes existed across gender groups, that is, $H_0 : \beta_3 = 0$. In this case, $R = (0, 0, 1)$ and $b_0 = (0)$.

6) Type I Error Rate and Average Power: For each simulation study, we calculated the family-wise error rate $(\mathrm{FWER} = P(V \geq 1))$ for the Type I error rates [26], [55]. The significance level was set at $\alpha = 5\%$, and 1000 replications were used to estimate the FWER. For a fixed α, if the FWER is smaller than α, then the test is conservative, whereas if the FWER is greater than α, then the test is anticonservative, or liberal [57]. We also calculated an average power, that is, the average of the probabilities of rejecting each of the 64 vertices in ROI.

7) Test Procedures: We evaluated the family-wise error rate of the four test procedures as follows. First, we considered the robust test procedure based on the maximum statistic W_D (Section II-C), in which $S = 699$ boostrap samples were generated to calculate the adjusted p value. Second, we also used the Wald-type test statistics $W_n(d)$ and W_D, but we calculated the p value of W_D and the corrected p values of $W_n(d)$ using the theoretical results of $\chi^2(1)$ field [21], [22], [56]. Third, we calculated the F statistic for the general linear model at each voxel and the maximum of the F statistics across those voxels. Then, we approximated the adjusted p values of all F statistics using the results of the F field [22], [56]. Finally, we only applied the permutation test based on the maximum of absolute values of the t statistics with 699 permutations to model (19) with $\mathbf{x}(t)$ given in (20), because the permutation method based on the t statistic may not be applicable when $\mathbf{x}(t)$ given in (21) has multiple covariates.

8) Random Field Theory: We applied the results for the χ^2 and F fields to the calculation of the corrected p value of the local maxima of the F statistics (or $W_n(d)$) and the adjusted p value in each point of the reference sphere. Explicitly, for the second test procedure, the corrected p value of W_D in a 2-D search region D is well approximated by

$$p(W_D > w) \approx \sum_{c=0}^{2} \mathrm{Resels}_c(D)\mathrm{EC}_c(w) \tag{22}$$

where Resels_c and EC_c, respectively, represent the resels of the search region and the Euler characteristic density of the $\chi^2(1)$ field in c dimension. Equation (22) can be used to calculated the adjusted p value for large $W_n(d)$ in each vertex of the reference sphere. For the triangular mesh on the reference sphere (2-D), we have $\mathrm{Resels}_0 = 2$, $\mathrm{Resels}_1 = 0$, and

$$\mathrm{Resels}_2 = 0.5 \sum_{\text{all triangulars}} |\Delta\mathbf{u}^T\Delta\mathbf{u}|^{1/2}\,[4\log(2)]^{-1}. \tag{23}$$

Moreover, let $\mathbf{u}_0$, $\mathbf{u}_1$, and $\mathbf{u}_2$ be three $n \times 1$ vectors of the normalized residuals at each vertex of a triangular of the reference sphere, we define $\Delta\mathbf{u} = (\mathbf{u}_1 - \mathbf{u}_0, \mathbf{u}_2 - \mathbf{u}_0)$ [56]. Expressions of the Euler characteristic densities for the χ^2 field and $d \leq 2$ can be found in [22]. Similarly, we used (23) to calculate the resels of the search region D and then applied the expected Euler characteristic for the F field to calculate the corrected p value of the local maxima of the F statistics [22], [56].

C. Real-World Example

The robust test procedure was used to model morphological changes in the hippocampus over time across gender groups.

1) Subjects: All 123 healthy subjects were recruited from a telemarketing list of families in southern Connecticut. The ages of all subjects range from 7 to 62 years (mean 20.14, SD: 13.2 years). The sample was similarly distributed across gender (males: 67; female: 56). Subjects were predominantly right handed (93.5%).

2) Image Acquisition Protocol: Head positioning in the head coil of the magnetic resonance imaging (MRI) scanner was standardized using cantho-meatal landmarks. We acquired high-resolution T1-weighted MRIs on a single 1.5-T scanner (GE Signa; General Electric, Milwaukee, WI) using a sagittal 3-D volume spoiled gradient echo sequence. Parameters included repetition time $=$ 24 msec, echo time $=$ 5 msec, 45° flip angle, frequency encoding superior/inferior, no wrap, 256×192 matrix, FOV $=$ 30 cm, 2 excitations, slice thickness $=$ 1.2 mm, and 124 contiguous slices encoded for sagittal slice reconstruction, voxel dimensions $1.17 \times 1.17 \times 1.2$ mm^3.

3) Selection of the Reference Structure: We first selected a preliminary brain of one subject (a 32.5 year-old right-handed, Caucasian male). Then, we registered the brains of other subjects in this study to this preliminary reference brain. We determined the point correspondences across their surfaces according to the methods described below and calculated the distances of those points from the corresponding points on the preliminary reference. Finally, we selected the brain for which all points across its surface were closest (in the least squares sense) to the average of the distances across those points for the entire sample as the final reference.

4) Morphological Descriptions of the Surface of the Hippocampus Surface: A four-step procedure described below was developed to obtain the morphological descriptions of the hippocampus surface. First, we registered the brains of all subjects to the cerebrum of the selected reference subject by using a rigid-body similarity transformation. The method of mutual information [58] was employed to calculate seven parameters (three translations, three rotations, and a global scale). Second, we rigidly coregistered to one another the hippocampus within the coregistered brains using a rigid-body transformation. Third, we identified correspondences between the points on the surfaces of the hippocampus by deforming these structures into the hippocampus of the reference brain using an algorithm based on fluid dynamics [59], [60]. Fourth, we calculated signed Euclidean distances of each point in the hippocampus of each subject from the corresponding point in the reference hippocampus. Distances of the points on the undeformed surface of the hippocampus of each subject that were

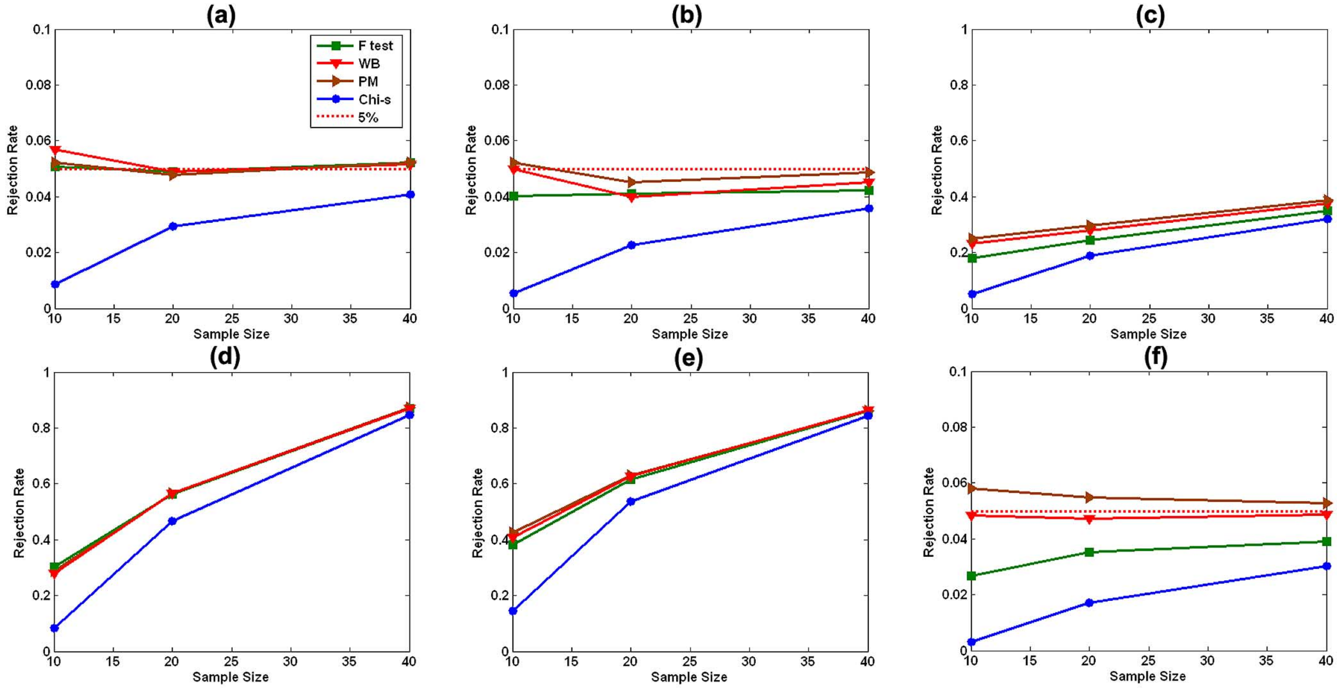

Fig. 2. Simulation study: Type I and Type II error rates. Rejection rates of the wild bootstrap method (WB), the permutation method (PM), the F test, and the χ^2 test for W_n are calculated for sample sizes of 10, 20, 40 subjects and for differing error distributions at the 5% significance level. (a)–(c) The estimated Type I error rates under the null hypothesis. (d)–(f) The estimated Type II error rates under the alternative hypothesis. Three distributions of error terms are Gaussian $N(0,1)$ [(a) and (d)], $\chi^2(2)-2$ [(b) and (e)], and Gaussian with heterogeneous variances [(c) and (f)].

positioned inside the boundary of the reference structure were labeled as negative, whereas distances for points positioned outside of the reference structure were labeled as positive.

5) Heteroscedastic Linear Model: To control the effects of covariates (age and gender) on our models of surface morphology, we considered a heteroscedastic linear model in each point on the reference surface

$$y_t(d) = \beta_0 + x_t\beta_1 + x_t^2\beta_2 + g_t\beta_3 + g_tx_t\beta_4 + g_tx_t^2\beta_5 + \varepsilon_t(d) \tag{24}$$

where x_t and g_t denote the $\log(\text{age})$ and gender of the tth subject, respectively, and $y_t(d)$ is the signed Euclidean distance for the tth subject in the dth point. In model (24), we do not include an adjustment term for overall intracranial volume, because the effects of brain size already have been taken into account by first coregistering the cerebrums of different subjects to the cerebrum of a reference subject (see Section III-C-4).

We are primarily interested in testing the morphological changes of the hippocampus over time across gender groups, i.e., we are testing the null hypotheses $H_0 : \beta_4 = \beta_5 = 0$ at all points on the surface of the hippocampus. Thus, we have

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad b_0 = (0,0)^T.$$

6) Smoothing the Surface of Hippocampus: We smoothed the signed Euclidean distance measures of all 123 subjects using the heat kernel smoothing with parameters $\sigma = 1$ and 16 iterations, which gave an effective smoothness of about 4 mm [53].

IV. Results

A. Simulation Studies: Set I

1) Type I Error Rates: Overall, the rejection rates for the permutation test and wild bootstrap method were accurate for all sample sizes ($n =$ 10, 20, or 40) and for the three differing distributions of error terms [Fig. 2(a)–(c)]. In particular, the wild bootstrap performed well even when the data were Gaussian distributed with heterogeneous variances, because W_n in the wild bootstrap accounted for inhomogeneity of variance across subjects. Although the F test was accurate in the presence of Gaussian N(0, 1) errors Fig. 2(a), Type I error rates associated with application of the F test declined for error terms that followed either the skewed distribution $\chi^2(2) - 2$ Fig. 2(b) or the Gaussian distribution with heterogeneous variances Fig. 2(c). This decline in Type I error was caused by applying the upper percentile of the F distribution to the F test, whereas when not assuming that the data were Gaussian distributed with homogeneous variance, the distribution of the F test was not in fact F distributed. Moreover, in all cases, the asymptotic χ^2 test for W_n was highly conservative because it was applied in the context of a small sample size.

2) Type II Error Rates: We observed that Type II error rates for the F test, the permutation test, and the wild bootstrap method were similar under N(0, 1) errors and for all sample sizes [Fig. 2(d)–(f)]. Compared with the rates of Type II error during application of the permutation test and the wild bootstrap method, however, the power of the F test to reject the null hypothesis declined modestly when the distributions of errors either were skewed Fig. 2(e) or were Gaussian with heterogeneous variance Fig. 2(f); this increase in Type II

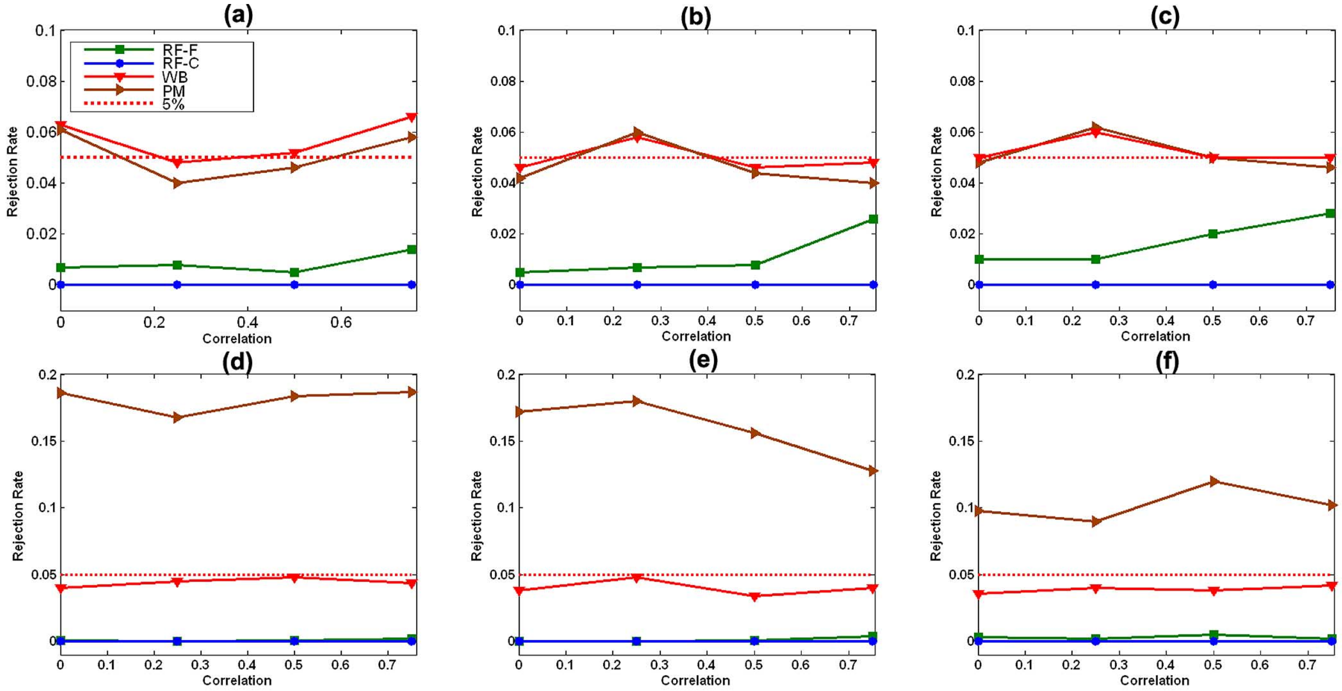

Fig. 3. Family-wise error rates with two covariates: family-wise error rates of the robust test procedure (WB), the permutation method (PM), random field F tests (RF-F), and the χ^2 field based on $W_n(d)$ (RF-C) under the linear model (19) with two covariates (Covs) (20). We consider sample sizes of 10, 20, and 40 subjects, four differing correlations $\rho = 0, 0.25, 0.5$, and 0.75, and two differing distributions, including homogeneous variance (HMV) and heterogeneous variance (HTV), at the 5% significance level. (a) $n = 10, 2$ Covs, and HMV; (b) $n = 20, 2$ Covs, and HMV; (c) $n = 40, 2$ Covs, and HMV; (d) $n = 10, 2$ Covs, and HTV; (e) $n = 20, 2$ Covs, and HTV; (f) $n = 40, 2$ Covs, and HTV.

error when noise was not Gaussian distributed reflected the consequence of applying the F test when the F test was not F distributed. Under all sample sizes and distributions of errors, the asymptotic test for W_n produced the highest rates of Type II error, because the upper 95th percentile of the χ^2-distribution was much higher than the upper 95th percentile of the sample distribution of W_n when the sample size was small. Consistent with our expectations, the statistical power for rejecting the null hypothesis increased with the sample size n.

B. Simulation Studies: Set II

1) Family-Wise Error Rates: In the presence of random errors with homogeneous variance, the permutation test based on the t statistic performed very well for all sample sizes [Fig. 3(a)–(c)]. In the presence of random errors with heterogeneous variance, in contrast, the permutation test was excessively liberal under all sample sizes [Fig. 3(d)–(f)], though less so as n increased [Fig. 3(d)–(f)]. In the presence of random errors with heterogeneous variance, the distributions of data in the two groups differed substantially from one another, invalidating the assumption of complete exchangeability, and causing the inflation of family-wise error rates during application of the permutation test.

For model (19) with two covariates, our robust test procedure worked well for relatively small sample sizes ($n = 10$, 20, and 40) and in the presence of random errors with either homogeneous or heterogeneous variance [Fig. 3(a)–(f)]. Under model (19) with three covariates, however, the family-wise error rates for our robust test procedure were not particularly accurate in the presence of random errors with either homogeneous or heterogeneous variance for the smallest sample size, $n = 10$ [Fig. 4(a) and (d)]; in contrast, they approximated the 5% significance level at the larger sample sizes of $n = 20$ and 40. Thus, sample size and the number of covariates can influence somewhat the finite performance of our robust test procedure.

The F field for the F statistic and the χ^2 field for W_n were highly conservative for relatively small sample sizes ($n = 10$, 20, and 40), when including two or three covariates, and in the presence of random errors with either homogeneous or heterogeneous variance (Figs. 3 and 4). Differing distributions of random errors significantly influenced the family-wise error rates of the F field for the F statistic and the χ^2 field for W_n. Larger correlations (or heavier smoothing) improved the performance of the F field for the F statistic when we compared its family-wise error rates with the 5% significance level. Overall, the F field for the F statistic yielded a highly conservative test, even with sample sizes as high as 40 and when the correlation of errors across two neighboring voxels on the reference sphere was as high as 0.75.

2) Average Power: For model (19) with two covariates, compared with our robust test procedure, the permutation test based on the t statistic had slightly larger average power in detecting statistically significant vertices in an ROI [Fig. 5(a) and (d)–(f)], and both the permutation test and our test procedure had much larger average power than did the F field for the F statistic and the χ^2 field for W_n [Fig. 5(a), (b), and (d)–(f)].

For model (19) with three covariates, the robust test procedure had larger average power than did the F field for the F statistic and the χ^2 field for W_n under heterogeneous variances [Fig. 6(d)–(f)]. However, under homogeneous variance, large

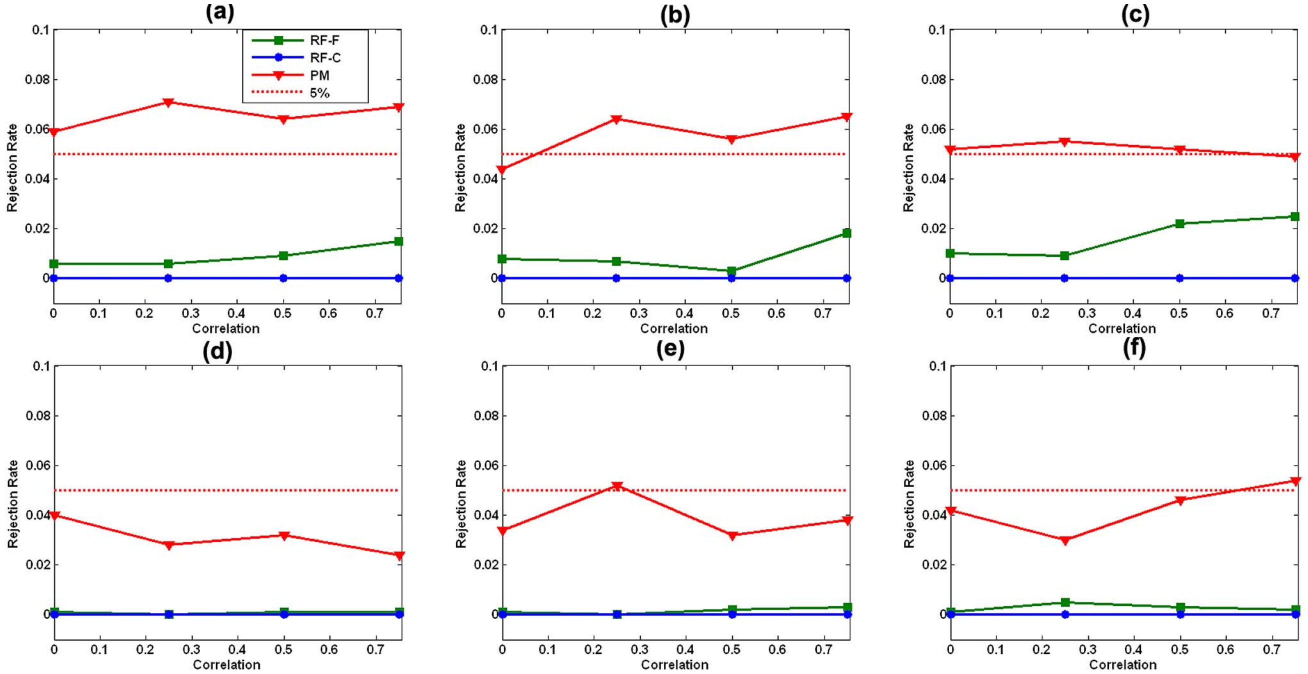

Fig. 4. Family-wise error rates with three covariates: family-wise error rates of the robust test procedure (WB), random field F test (RF-F), and the χ^2 field for the Wald-type test statistics (RF-C) under the linear model (19) with three covariates (Covs) in (21). We consider sample sizes of 10, 20, and 40 subjects, four differing correlations $\rho = 0$, 0.25, 0.5, and 0.75, and two differing distributions, including homogeneous variance (HMV) and heterogeneous variance (HTV), at the 5% significance level. (a) $n = 10$, 3 Covs, and HMV; (b) $n = 20$, 3 Covs, and HMV; (c) $n = 40$, 3 Covs, and HMV; (d) $n = 10$, 3 Covs, and HTV; (e) $n = 20$, 3 Covs, and HTV; (f) $n = 40$, 3 Covs, and HTV.

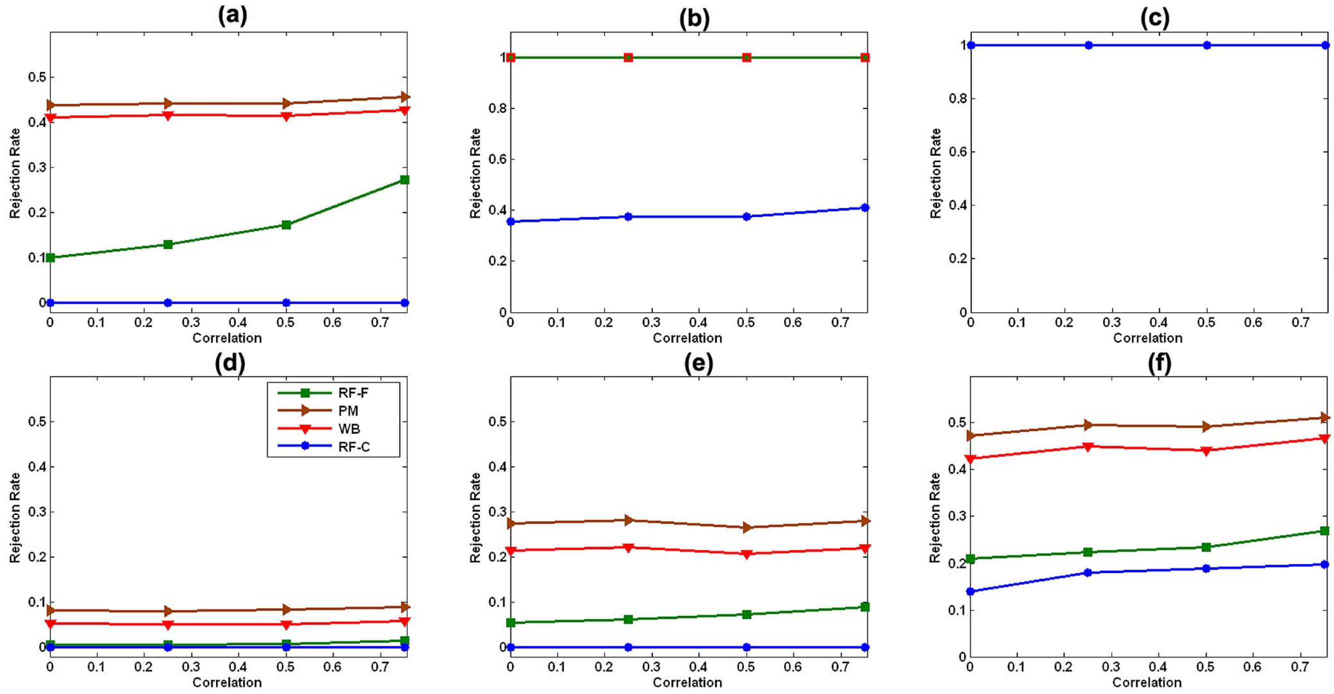

Fig. 5. Statistical power with two covariates: average powers of the robust test procedure (WB), the permutation method (PM), the F field F test (RF-F), and the χ^2 field for $W_n(d)$ (RF-C) under the linear model (19) with two covariates (Covs) (20). We consider sample sizes of 10, 20, and 40 subjects, four differing correlations $\rho = 0$, 0.25, 0.5, and 0.75, and two differing distributions, including homogeneous variance (HMV) and heterogeneous variance (HTV), at the 5% significance level. In (b), the lines for PM, WB, and RF-F overlay one other, while in (c), all four lines overlay one other. (a) $n = 10$, 2 Covs, and HMV; (b) $n = 20$, 2 Covs, and HMV; (c) $n = 40$, 2 Covs, and HMV; (d) $n = 10$, 2 Covs, and HTV; (e) $n = 20$, 2 Covs, and HTV; (f) $n = 40$, 2 Covs, and HTV.

correlations (e.g., $\rho \geq 0.5$), and $n = 10$, the F field for the F statistics was more sensitive than was the robust test procedure Fig. 6(a).

C. Real-World Example

1) Assessing Assumptions of the Model: We investigated whether the general linear model was appropriate for this

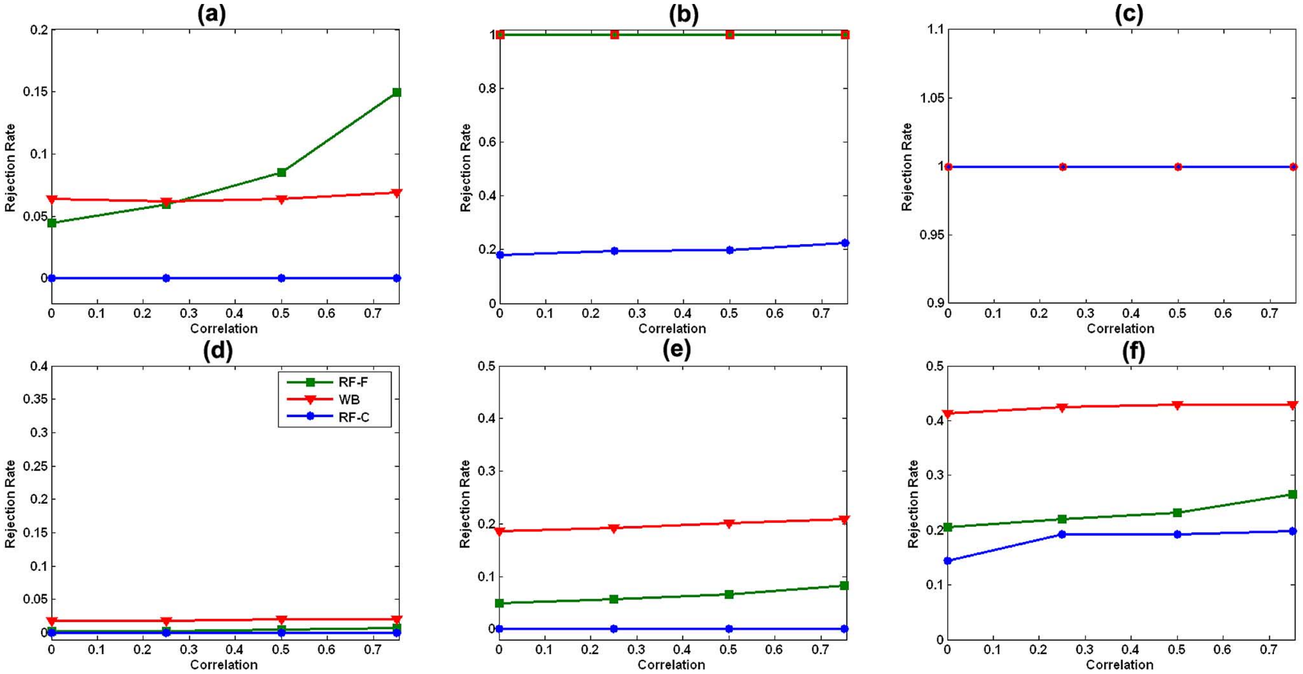

Fig. 6. Statistical power with three covariates: average powers of the robust test procedure, the F field F test (RF-F), and the χ^2 field for $W_n(d)$ (RF-C) under the linear model (19) with three covariates (Covs) (21). We consider sample sizes of 10, 20, and 40 subjects, four different correlations $\rho = 0, 0.25, 0.5$, and 0.75, and two differing distributions, including homogeneous variance (HMV) and heterogeneous variance (HTV), at the 5% significance level. In (b), the lines for RF-F and WB overlay one other, while in (c), all lines overlay one other. (a) $n = 10$, 3 Covs, and HMV; (b) $n = 20$, 3 Covs, and HMV; (c) $n = 40$, 3 Covs, and HMV; (d) $n = 10$, 3 Covs, and HTV; (e) $n = 20$, 3 Covs, and HTV; (f) $n = 40$, 3 Covs, and HTV.

study. We calculated test statistics for assessing the validity of the two assumptions of the general linear model: normality and homogeneous constant variance of the data. Based on the residuals after fitting the general linear model, we calculated the Shapiro–Wilk and Cook–Weisberg statistics to test the assumptions of a Gaussian distribution and the homogeneous variance for the error terms [61], [62]. These statistics rejected the assumptions of normality (Fig. 7) and homogeneous variance (not presented here) at many points on the surfaces of the both left and right hippocampus for both the original distance measures [Fig. 7(a)–(d)] as well as the smoothed distance measures [Fig. 7(e)–(h)]. The application of smoothing techniques, however, improved the normality of the random errors [Fig. 7(e)–(h)].

Because the assumptions of the general linear model are invalid, the use of random field theory to analyze these imaging data is inappropriate, at least without prior spatial smoothing of the data [2], [6], [23]. Moreover, the permutation method based on the t statistic cannot be applied directly to the model (24), which contains multiple covariates.

2) Analysis of Hippocampal Surface: We used the signed Euclidean distances to detect and localize statistically significant differences in the morphology of the hippocampus over time across gender groups. We tested these differences using gender-by-$\log(\text{age})$ and gender-by-$[\log(\text{age})]^2$ interactions in model (24) at each point of the surface of the hippocampus. The p-values based on the asymptotic χ^2 test were color-coded in each point of the reference hippocampus [Fig. 8(a), (b), (f) and (g)]. To correct for multiple comparisons, we applied our robust test procedure to calculate the adjusted p value at each point on the surface of the reference hippocampus [Fig. 8(c), (d), (h), and (i)]. Color-coded maps of p value maps using either the uncorrected χ^2 test alone or the corrected resampling method indicated large-scale differences

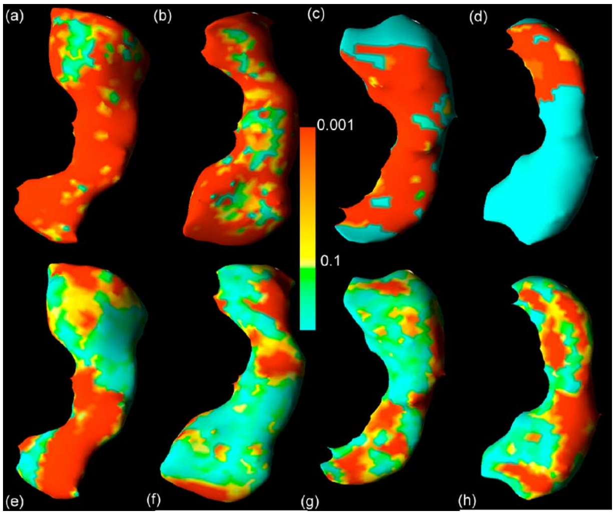

Fig. 7. Assessing normality at the surface of the hippocampus. Color-coded maps display the uncorrected p values for the Shapiro–Wilk test of normality at the surface of the hippocampus in data from 123 healthy children and adults registered to a template surface. Top row shows the unsmoothed images. Bottom row shows the images after smoothing with a heat kernel. Panels (a and e) right hippocampus dorsal view, with the anterior portion of the hippocampus located at the top of the figure; (b and f) right hippocampus ventral view, with the anterior portion of the hippocampus located at the bottom of the figure; (c and g) left hippocampus dorsal view, with the anterior hippocampus located at the top; (d and h) left hippocampus ventral view, with the posterior hippocampus located at the bottom. Smoothing reduces substantially the number of voxels at the surface of the hippocampus that violate assumptions of the normality of distribution of signed Euclidean distances of points on the surface of the hippocampus of each subject in the dataset from corresponding points on the surface of the reference hippocampus.

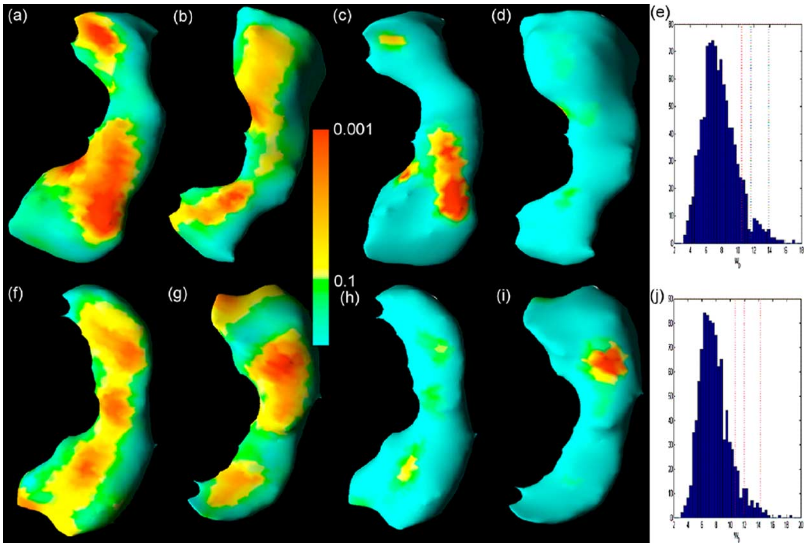

Fig. 8. Significance testing at the surface of the hippocampus: color-coded maps of p values and adjusted p values for the Wald-type test statistics. Row 1: right hippocampus. Row 2: left hippocampus. Columns 1 and 2: raw p values of the Wald-type test statistics based on a χ^2 distribution. Columns 3 and 4: adjusted p values of the Wald-type test statistics based on our robust test procedure for the correction of multiple comparisons. Panels (e and j): histograms of W_D^* based on the bootstrap samples. Spatial orientations of the hippocampus are the same as the corresponding views in Fig. 7. After correction for multiple comparisons, statistically significant interactions of gender-by-$\log(\mathrm{age})$ and gender-by-$[\log(\mathrm{age})]^2$ remain in the head of both the right (c) and left (i) hippocampus.

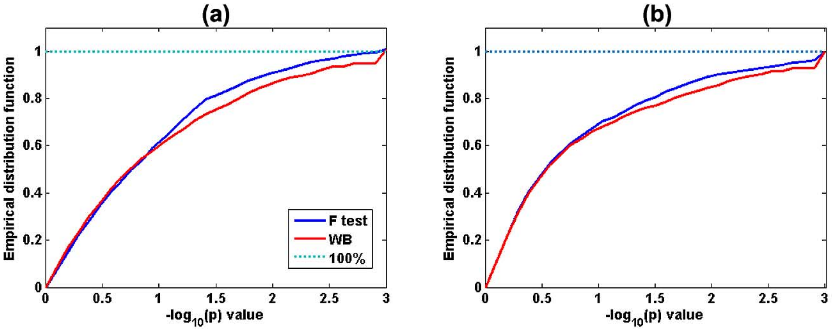

Fig. 9. Distribution functions of $-\log_{10}(p)$ values of test statistics at the surface of the Hippocampus. Empirical distribution functions of $-\log_{10}(p)$ values of the F statistic and $-\log_{10}(p)$ values of W_n based on the wild bootstrap method are shown. Combining the heteroscedastic linear model with the wild bootstrap increases the number of significant points $(-\log_{10}(p) > 1.3)$ on the surface of the hippocampus in each hemisphere: (a) left hippocampus; (b) right hippocampus.

in surface morphology across gender groups. The resampling method, however, captured far fewer points of differences in morphology of the hippocampus mainly in the head portion of the hippocampus (Fig. 8) [63].

We compared findings using the linear model and our heteroscedastic linear model for these data. When testing gender-by-$\log(\mathrm{age})$ and gender-by-$[\log(\mathrm{age})]^2$ interactions, we calculated the $-\log_{10}(p)$ value of the F statistic and the $-\log_{10}(p)$ value of W_n based on the wild bootstrap method at each point of the surface of the reference hippocampus. We observed that combining the heteroscedastic linear model with the wild bootstrap method increased the number of statistically significant points $(-\log_{10}(p) > 1.3)$ on the surface of the hippocampus in each hemisphere [Fig. 9(a) and (b)].

V. Conclusion and Discussion

We have developed a method for the analysis of anatomical imaging data based on a heteroscedastic linear model and a wild bootstrap method. The use of the heteroscedastic linear model avoids the assumptions of homogeneous variance across subjects and the Gaussian distribution of imaging data, that we have shown to be invalid in one real-world imaging dataset. The robust test procedure not only accounts for multiple comparisons across all voxels of the brain region under investigation, but it also asymptotically preserves the dependence structure among the Wald-type test statistics. We have used simulation studies to show that the robust test procedure provides accurate control of the family-wise error rate for relatively small to moderate sample sizes. Our analysis of a real-world dataset demonstrates the applicability of our test procedure to anatomical imaging data, as well as fMRI and PET data.

Our robust test procedure differs from other multiple comparison procedures for controlling the Type I error, including random field theory methods, permutation methods, and the false discovery rate. Computationally simple methods that employ random field theory depend on the validity of several stringent assumptions, including a Gaussian distribution for the imaging data and the smoothness of the spatial autocorrelation function [21]. Without formally assessing the validity of these assumptions, the application of random field theory can yield a very conservative statistical test (Figs. 3–6) [21]. Permutation methods outperform those of random field theory methods in various settings, even when the small sample size is small, although when not accounting for the presence of heterogeneous variances across subjects, permutation methods can be anticonservative (see the simulation results in Section IV). Moreover, the permutation methods may not be widely applicable to neuroimaging studies that require the statistical control of multiple covariates (e.g., age, gender, diagnoses, or genotype) without invoking stringent assumptions about the data, such as the presence of identical and independently distributed random errors. However, when the assumption of complete exchangeability is valid, the permutation test is almost the best. Methods of statistical analysis based on the false discovery rate can accurately control the false discovery rate, whereas the robust test procedure can accurately control the family-wise error rate. Moreover, the false discovery rate requires an accurate estimation of the p value for each hypothesis at each voxel, and under the heteroscedastic linear model, calculating the p value accurately for each hypothesis requires a large number of bootstrap samples in the wild bootstrap method.

We also note several advantages and limitations of our robust test procedure for controlling Type I error. Type I error rates when using the wild boostrap method are reasonably small in the presence of either heterogeneous variances across subjects or skewed distributions of error terms (Fig. 2). The robust test procedure can accurately control the family-wise error rate under various scenarios examined (Section IV), and it can increase the sensitivity of detecting statistically significant differences in brain structure when the variances across subjects vary significantly across voxels. However, when the homogeneous variance and Gaussian assumptions underlying the general linear model are truly valid, the use of the wild bootstrap method yield slightly reduced statistical power (Section IV) [37]. Moreover, for small significance level, say $\alpha = 1\%$, the number of replications S in the robust test procedure must be increased in order to accurately estimate p_D and $p_D(d)$. Running the robust test procedure for large S can be computationally intensive [42].

Many aspects of this work merit further research. One is to examine the performance of our robust test procedure in the analysis of data from other imaging modalities, including PET and fMRI. Another is to extend our robust test procedure to the inclusion of cluster size inference in controlling the rate of Type I errors [22], [57], [64]–[66]. Our robust test procedure may lead to a simple test of cluster size in assessing the significance of all numbers of interconnected voxels greater than a given threshold (e.g., $\chi^2_{0.05}(r)$). We will formally study the cluster size test elsewhere. Finally, we may use the generalized least squares estimator of β instead of $\hat{\beta}$, when prior information concerning Ω is available.

Appendix I
Proof of the Restricted Least Squares Estimate

The restricted least squares estimate of β under $H_0 : R\beta = b_0$ can be obtained by minimizing the following objective function:

$$l(\lambda, \beta) = \sum_{i=1}^{n} \left(y_i - X_i^T\beta\right)^2 + 2\lambda^T(R\beta - b_0) \tag{A.1}$$

where $\lambda = (\lambda_1, \cdots, \lambda_r)^T$. Taking the first derivative of $l(\lambda, \beta)$ with respect to β and λ, respectively, yields

$$\partial_\beta l = -2\sum_{i=1}^{n} \left(y_i - X_i^T\beta\right) X_i + 2R^T\lambda$$
$$\text{and} \quad \partial_\lambda l = 2(R\beta - b_0).$$

Then, $\tilde{\beta}$ and $\tilde{\lambda}$ satisfy

$$\sum_{i=1}^{n} X_i X_i^T \tilde{\beta} + R^T\tilde{\lambda} = \sum_{i=1}^{n} X_i y_i \tag{A.2}$$
$$R\tilde{\beta} = b_0. \tag{A.3}$$

It follows from (A.2) that

$$\tilde{\beta} = \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \left(\sum_{i=1}^{n} X_i y_i - R^T\tilde{\lambda}\right). \tag{A.4}$$

Substituting the above $\tilde{\beta}$ into (A.3) yields

$$\tilde{\lambda} = \left[R(X^TX)^{-1}R^T\right]^{-1} \left[R(X^TX)^{-1}X^TY - b_0\right]. \tag{A.5}$$

Substituting $\tilde{\lambda}$ into (A.4), we have

$$\tilde{\beta} = \hat{\beta} - (X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} (R\hat{\beta} - b_0).$$

Appendix II
Proof of Equation (13)

Following the arguments of Appendix I, the restricted least squares estimate of β in model (11) can be expressed by

$$\tilde{\beta}^* = \hat{\beta}^* - (X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} (R\hat{\beta}^* - b_0).$$

Because $\hat{\beta}^* = (X^TX)^{-1}X^TY^*$, we have

$$\begin{aligned}\tilde{\beta}^* &= \hat{\beta}^* - (X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} \left[R_XY^* - b_0\right]\\ &= \hat{\beta}^* - (X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} R_X\\ &\quad \times (X\tilde{\beta} + \tilde{\Omega}^{1/2}\varepsilon^*) + (X^TX)^{-1}R^T\\ &\quad \times \left[R(X^TX)^{-1}R^T\right]^{-1} b_0\\ &= \hat{\beta}^* - (X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} R_X\tilde{\Omega}^{1/2}\varepsilon^*\end{aligned}$$

where $R_X = R(X^TX)^{-1}X^T$. From the above, we can easily prove (13).

Acknowledgment

The authors would like to thank the Editor, the Associate Editor, and two anonymous referees for valuable suggestions, which greatly helped to improve our presentation. They would also like to thank Dr. J. Royal for his invaluable editorial assistance and Dr. M. Chung for his helpful comments.

References

[1] J. C. Mazziotta, A. W. Toga, A. C. Evans, P. Fox, and J. Lancaster, "A probabilitic atlas of the human brain: Theory and rationale for its development," *NeuroImage*, vol. 2, pp. 89–101, 1995.

[2] J. Ashburner, "Computational neuroanatomy," Ph.D. dissertation, Univ. London, London, U.K., 2001.

[3] J. Ashburner and K. J. Friston, "Voxel-based morphometry: The methods," *NeuroImage*, vol. 11, pp. 805–821, 2000.

[4] M. K. Chung, "Statistical morphometry in neuroanatomy," Ph.D. dissertation, McGill Univ., Montreal, QC, Canada, 2001.

[5] P. M. Thompson and A. W. Toga, "A framework for computational anatomy," *Comput. Visual*, vol. 5, pp. 13–34, 2002.

[6] P. M. Thompson, J. L. Rapoport, T. D. Cannon, and A. W. Toga, "Automated analysis of structural MRI data," in *Brain Imaging in Schizophrenia*, S. Lawrie, E. C. Johnstone, and D. Weinberger, Eds. Oxford, U.K.: Oxford Univ. Press, 2003.

[7] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: Methods and applications," *Curr. Med. Imag. Rev.*, vol. 1, pp. 105–113, 2005.

[8] P. M. Thompson, T. D. Cannon, K. L. Narr, T. van Erp, V. Poutanen, M. Huttunen, J. Lonnqvist, C. G. Standertskjold-Nordenstam, J. Kaprio, M. Khaledy, R. Dail, C. I. Zoumalan, and A. Toga, "Genetic influences on brain structure," *Nature Neurosci.*, vol. 4, pp. 1253–1358, 2001.

[9] P. M. Thompson, T. D. Cannon, and A. W. Toga, "Mapping genetic influences on human brain structure," *Ann. Med.*, vol. 24, pp. 523–536, 2002.

[10] J. G. Csernansky, S. Joshi, L. Wang, J. W. Haller, M. Gado, J. P. Miller, U. Grenander, and M. I. Miller, "Hippocampal morphometry in schizophrenia by high dimensional brain maping," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 11406–11411, 1998.

[11] R. Plomin and S. M. Kosslyn, "Genes, brain and cognition," *Nature Neurosci.*, vol. 4, pp. 1153–1154, 2001.

[12] K. L. Narr, T. D. Cannon, R. P. Woods, P. M. Thompson, S. Kim, D. Asunction, T. G. van Erp, V. P. Poutanen, M. Huttunen, J. Lonnqvist, C. G. Standerksjold-Nordenstam, J. Kaprio, J. C. Mazziotta, and A. W. Toga, "Genetic contribution to altered callosal morphology in schizophrenia," *J. Neurosci.*, vol. 22, pp. 3720–3729, 2002.

[13] I. C. Wright, P. Sham, R. M. Murray, D. R. Weinberger, and E. T. Bullmore, "Genetic contributions to regional variability in human brain structure: Methods and preliminary results," *NeuroImage*, vol. 17, pp. 256–271, 2002.

[14] M. Styner, J. A. Lieberman, R. K. McClure, D. R. Weinberger, D. W. Jones, and G. Gerig, "Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors," *Proc. Nat. Acad. Sci. USA*, vol. 102, pp. 4677–4872, 2005.

[15] I. Dryden and K. Mardia, *Statistical Shape Analysis*. New York: Wiley, 1998.

[16] P. Fletcher, S. Joshi, C. Lu, and S. M. Pizer, "Gaussian distribution on lie groups and their applications to statistical shape analysis," *Inf. Process. Med. Imag.*, pp. 450–462, 2003.

[17] P. M. Thompson, R. P. Woods, M. S. Mega, and A. W. Toga, "Mathematical/computational challenges in creating population-based brain atlases," *Hum. Brain Mapp.*, vol. 9, pp. 81–92, 2000.

[18] A. C. Evans and B. D. C. Group, "The NIH MRI study of normal brain development," *NeuroImage*, vol. 30, pp. 184–202, 2006.

[19] E. R. Sowell, B. S. Peterson, P. M. Thompson, S. E. Welcome, A. L. Henkenius, and A. W. Toga, "Mapping cortical change across the human life span," *Nature Neurosci.*, vol. 6, pp. 309–315, Jan. 2003.

[20] K. Friston, A. P. Holmes, K. J. Worsley, J. B. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Hum. Brain Mapp.*, vol. 2, pp. 189–210, 1995.

[21] T. Nichols and S. Hayasaka, "Controlling the family-wise error rate in functional neuroimaging: A comparative review," *Statist. Meth. Med. Res.*, vol. 12, pp. 419–446, 2003.

[22] J. Cao and K. J. Worsley, "Applications of random fields in human brain mapping," in *Spatial Statistics: Methodological Aspects and Applications*, M. Moore, Ed. New York: Springer, 2001, vol. 159.

[23] C. H. Salmond, J. Ashburner, F. Vargha-Khadem, A. Connelly, D. G. Gadian, and K. J. Friston, "Distributional assumptions in voxel-based morphometry," *NeuroImage*, vol. 17, pp. 1027–1030, 2002.

[24] D. R. Cook and S. Weisberg, *Residuals and Influence in Regression*. London, U.K.: Chapman Hall, 1982.

[25] W. Luo and T. Nichols, "Diagnosis and exploration of massively univariate fMRI models," *NeuroImage*, vol. 19, pp. 1014–1032, 2003.

[26] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statist. Sci.*, vol. 18, pp. 71–103, 2003.

[27] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences," *NeuroImage*, vol. 23, pp. 17–20, 2004.

[28] D. Y. Lin, "An efficient Monte Carlo approach to assessing statistical significance in genomic studies," *Bioinformatics*, vol. 6, pp. 781–787, 2005.

[29] M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith, "Multilevel linear modelling for FMRI group analysis using Bayesian inference," *NeuroImage*, vol. 21, pp. 1732–1747, 2004.

[30] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multilevel linear modeling for group analysis in FMRI," *NeuroImage*, vol. 20, pp. 1052–1063, 2003.

[31] K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom, and S. Kiebel, "Mixed-effects and fMRI studies," *NeuroImage*, vol. 24, pp. 244–252, 2005.

[32] H. L. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, pp. 817–838, 1980.

[33] F. Eicker, "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Ann. Math. Statist.*, vol. 34, pp. 447–456, 1963.

[34] E. Flachaire, "Bootstrapping heteroscedastic regression models: Wild bootstrap vs. pairs bootstrap," *Computational Statist. Data Anal.*, vol. 49, pp. 361–376, 2005.

[35] L. G. Godfrey and A. R. Tremayne, "The wild bootstrap and heteroscedasticity-robust tests for serial correlation in dynamic regression models," *Comput. Statist. Data Anal.*, vol. 49, pp. 377–395, 2005.

[36] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*. New York: Wiley, 1992.

[37] J. S. Long and L. H. Ervin, "Using heteroscedasticity consistent standard errors in the linear regression model," *Am. Statist.*, vol. 54, pp. 217–224, 2000.

[38] R. Davidson and J. G. MacKinnon, "Heteroskedasticity-robust tests in regression directions," *Ann. de l'INSEE*, vol. 59/60, pp. 183–218, 1985.

[39] J. G. MacKinnon and H. L. White, "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *J. Econometrics*, vol. 21, pp. 53–70, 1985.

[40] R. Y. Liu, "Bootstrap procedure under some non-i.i.d. models," *Ann. Statist.*, vol. 16, pp. 1696–1708, 1988.

[41] E. Mammen, "Bootstrap and wild bootstrap for high dimensional linear models," *Ann. Statist.*, vol. 21, pp. 255–285, 1993.

[42] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. London, U.K.: Chapman Hall, 1993.

[43] S. S. Kannurpatti and B. B. Biswal, "Bootstrap resampling method to estimate confidence intervals of activation-induced CBF changes using laser Doppler imaging," *J. Neurosci. Meth.*, vol. 146, pp. 61–68, 2005.

[44] A. Aronowich and R. J. Adler, "The behaviour of chi squared processes at critical points," *Adv. Appl. Probability*, vol. 17, pp. 280–297, 1985.

[45] K. J. Worsley, "Local maxima and the expected Euler characteristic of excursion sets of chi^2, F and t fields," *Adv. Appl. Probability*, vol. 26, pp. 13–42, 1994.

[46] A. W. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer, 1996.

[47] T. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Hum. Brain Mapp.*, vol. 15, pp. 1–25, 2002.

[48] S. Hayasaka, L. K. Phan, I. Liberzon, K. J. Worsley, and T. E. Nichols, "Nonstationary cluster-size inference with random field and permutation methods," *NeuroImage*, vol. 22, pp. 676–687, 2004.

[49] J. Mumford and T. E. Nichols, "The problem of inflated type I errors with simple group model," *NeuroImage*, vol. 26, p. 1258.

[50] J. P. Kerstin, R. Bansal, H. T. Zhu, R. Whiteman, J. Amat, G. Quackenbusch, L. Martin, K. Durkin, C. Blair, J. Royal, K. Hugdahl, and B. Peterson, "Hippocampus and Amydala morphology in attention-deficit/hyperactivity disorder," *Arch. Gen. Psychiatry*, vol. 63, pp. 795–807, 2006.

[51] A. C. Cressie, *Statistics for Spatial Data*, 2nd ed. New York: Wiley, 1993.

[52] B. R. Logan and D. B. Rowe, "An evalution of thresholding techniques in fMRI analysis," *NeuroImage*, vol. 22, pp. 95–108, May 2004.

[53] M. K. Chung, S. Robbins, K. M. Dalton, R. J. Davidson, A. L. Alexander, and A. C. Evans, "Cortical thickness analysis in autism via heat kernel smoothing," *NeuroImage*, vol. 25, pp. 1256–1265, 2005.

[54] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Statist. Soc., Ser. B-Statist. Methodol.*, vol. 57, pp. 289–300, 1995.

[55] J. P. Shaffer, "Multiple hypothesis testing: A review," *Ann. Rev. Psychol.*, vol. 46, pp. 561–584, 1995.

[56] K. J. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. C. Evans, "Detecting changes in nonisotropic images," *Hum. Brain Mapp.*, vol. 8, pp. 98–101.

[57] S. Hayasaka and T. Nichols, "Validating cluster size inference: Random field and permutation methods," *NeuroImage*, vol. 20, pp. 2343–2356, 2003.

[58] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vision*, vol. 24, pp. 137–154, 1997.

[59] G. Christensen, R. D. Rabbitt, and M. I. Miller, "3D brain mapping using a deformable neuroanatomy," *Phys. Med. Biol.*, vol. 39, pp. 609–618, 1994.

[60] R. Bansal, L. Staib, Y. Wang, and B. Peterson, "Roc-based assessments of 3D cortical surface-matching algorithms," *NeuroImage*, vol. 24, pp. 150–162, 2005.

[61] D. R. Cook and S. Weisberg, "Diagnostics for heteroscedasticity in regression," *Biometrika*, vol. 70, pp. 1–10, 1983.

[62] H. T. Zhu and H. P. Zhang, "A diagnostic procedure based on local influence measure," *Biometrika*, vol. 91, pp. 579–589, 2004.

[63] J. C. Pruessner, D. L. Collins, M. Pruessner, and A. C. Evans, "Age and gender predict volume decline in the anterior and posterior hippocampus in early adulthood," *J. Neurosci.*, vol. 21, pp. 194–200, 2001.

[64] J. B. Poline and B. Mazoyer, "Cluster analysis in individual functional brain images: Some new techniques to enhance the sensitivity of activation detection methods," *Hum. Brain Mapp.*, vol. 2, pp. 103–111, 1994.

[65] P. E. Roland, B. Levin, R. Kawashima, and S. Kerman, "Three-dimensional analysis of clustered voxels in 15O-butanol brain activation images," *Hum. Brain Mapp.*, vol. 1, pp. 3–19, 1993.

[66] K. J. Friston, K. J. Worsley, R. S. J. Frackowiak, J. C. Mazziotta, and A. C. Evans, "Assessing the significance of focal activations using their spatial extent," *Hum. Brain Mapp.*, vol. 1, pp. 214–220, 1994.