An Efficient Permutation Approach for Classical and Bioequivalence Hypothesis Testing of Biomedical Shape Study

Chunxiao Zhou\textsuperscript{1,2,5}, Yongmei Michelle Wang\textsuperscript{1,3,4,5}

Departments of \textsuperscript{1}Statistics, \textsuperscript{2}Electrical\&Computer Engineering, \textsuperscript{3}Psychology, \textsuperscript{4}Bioengineering, \textsuperscript{5}Beckman Institute,
University of Illinois at Urbana-Champaign, IL, USA
czhou4@uiuc.edu, ymw@uiuc.edu

Abstract

A new statistical permutation analysis method is presented in this paper to efficiently and accurately localize regionally specific shape differences between groups of 3D biomedical images. It can improve the system’s efficiency by approximating the permutation distribution of the test statistic with Pearson distribution series. This procedure involves the calculation of the first four moments of the permutation distribution, which are derived theoretically and analytically without any permutation. Furthermore, bioequivalence testing aims for practical significances between the two groups that are statistically significant with the shape differences larger than a desired threshold. Experimental results based on both classical and bioequivalence hypothesis tests using simulated data and real biomedical images are presented to demonstrate the advantages of the proposed approach.

1. Introduction

In biomedical image analysis, the distribution of the data is usually unknown and sample size is quite small. When hypothesis testing involved in the analysis, the parametric approaches may not be optimal despite their simplicity. Permutation tests are among the most powerful nonparametric tests that can be applied when parametric tests do not work. They obtain p-values from permutation distributions of a test statistic, rather than from parametric distributions. In addition, permutation tests require few assumptions concerning statistical distributions but exchangeability. They belong to the nonparametric “distribution-free” category of hypothesis testing and are thus flexible, and have been used successfully in biomedical MR image analysis [6]. There are three major approaches to construct the permutation distribution [5]. First, exact permutation enumerates all possible arrangements. The second approach is an approximate permutation distribution based on random sampling from all possible permutations. Third, permutation distribution approximation uses the analytical moments of the exact permutation distribution under the null hypothesis. The computational cost is the main disadvantage of the exact permutation, due to the factorial increase in the number of permutations with the increasing number of subjects. The second technique has the problem of replication and causes more type I errors. When a large number of repeated tests are needed, the random permutation strategy is also computationally expensive to achieve satisfied p-value accuracy. Sometimes, the moments of the exact permutation distribution do not actually exist. In addition, if they ever exist, it is difficult to obtain them. These two factors are two main limitations of the third approach.

In this paper, we propose to use a novel hybrid strategy to take advantage of nonparametric permutation tests and parametric Pearson distribution approximation for both efficiency and accuracy/flexibility. Specifically, we employ a general theoretical method to derive moments of permutation distribution for any linear test statistics. Here, the term “linear test statistic” refers to a linear function of test statistic coefficients, instead of that of data. The key idea is to separate the moments of permutation distribution into two parts, permutation of test statistic coefficients, instead of that of data. The key idea is to separate the moments of permutation distribution into two parts, permutation of test statistic coefficients, and function of the data. We can then obtain the moments without any permutation since the permutation of test statistic coefficients can be derived theoretically. Given the first four moments, the permutation distribution can be well fitted by Pearson distribution series. The p-values are then estimated without any real permutation. For multiple comparison of two-group difference, given the sample size \( n_1 = 21 \) and \( n_2 = 21 \), the number of tests is \( m = 2000 \). In this case, \( m \times (n_1 + n_2)!/ (n_1!) (n_2)! \approx 1.1 \times 10^{15} \) permutations are needed for an exact permutation test. Even for 20,000 random permutations per test, \( 4 \times 10^7 \) permutations are still required. Alternatively, our hybrid permutation method using Pearson distribution approximation only involves the calculation of analytically derived first four moments of exact permutation distributions while
achieve high accuracy (see section 4). Instead of calculating the test statistics in factorial scale with exact permutation, our hybrid permutation using mean difference test statistic only require $O(n)$ computation cost, where $n = n_1 + n_2$.

Accurate and efficient brain shape morphometry analysis is of great importance in detecting morphological changes in structures of interest for neuroscience research and medical diagnosis and treatment. There is increasing evidence that surface shape analysis of brain structures provides new information which is not available by conventional analysis [9]. A critical issue in surface morphometry is the shape description and representation. Various strategies have been investigated recently in the literature, such as [2], [9], [10]. The spherical harmonics (SPHARM) approach using spherical harmonics as basis functions for a parametric surface description was proposed in [2]. The correspondence across different surfaces is established by aligning the parameterizations via the first order ellipsoid. The present work employs the SPHARM-PDM shape description [8], which leads to corresponding location vectors across all surfaces for our subsequent statistical analysis of surface shape. At each corresponding position on the surfaces, we test whether there is a significant mean vector difference between location vectors of two groups. If a hypothesis test leads to a $p$-value smaller than the pre-chosen $\alpha$-level, we reject the null hypothesis and conclude that a significant shape difference exists at this surface location. In this paper, we focus on the surface shape analysis for two groups, though our method can be extended to the multi-group case.

Since the distribution of the location vectors is unknown, only a limited number of subject samples are available, and the same tests are repeated on thousands of locations, we propose to use our hybrid permutation approach to the brain shape analysis. Without any permutation, we can accurately estimate the $p$-value of the permutation distribution of the test statistic at each surface location. The analysis of the location vectors involves testing thousands of hypotheses, i.e. multiple testing. False positives must be controlled over all tests. In this work, we apply a Region of Interest (ROI) constrained adaptive FDR to enhance the power of finding true discoveries [11].

2. Hypothesis

2.1. Classical Hypothesis

Given registered location vectors across all subjects, surface shape morphometry analysis becomes a two-sample test for equality of means at each surface location. The hypothesis is typically constructed as:

$$H_0 : \mu_A = \mu_B \hspace{1cm} \text{vs.} \hspace{1cm} H_a : \mu_A \neq \mu_B$$

where $\mu_A = [\mu_A^{(x)}, \mu_A^{(y)}, \mu_A^{(z)}]^{\top}$ and $\mu_B = [\mu_B^{(x)}, \mu_B^{(y)}, \mu_B^{(z)}]^{\top}$ are three dimensional mean vectors of group A and B.

2.2. Bioequivalence Hypothesis

In many applications, statistical significance is not equivalent to practical significance since smaller differences of two group location vectors can be more statistically significant than the larger ones. Statistical significance means that the observed difference is not a consequence of sampling error. Practical significance indicates whether the difference is large enough to be of value in a practical sense. Statistical significance does not necessarily indicate practical significance because extremely small and non-notable differences can be statistically significant. For example, there are two pairs of observed mean location vectors $(\mu_{A1}, \mu_{B1})$ at location 1 and $(\mu_{A2}, \mu_{B2})$ at location 2, with $\mu_{A1} = [1,1,1]^{\top}$, $\mu_{B1} = [0.999,0.999,0.999]^{\top}$, $\mu_{A2} = [1,1,1]^{\top}$, and $\mu_{B2} = [0.7,0.7,0.7]^{\top}$. We assume that the variance of location vectors at location 2 is much larger than that at location 1, and their $p$-values of the observed mean differences are $p_1 = 0.001$ and $p_2 = 0.01$ respectively. The mean difference at location 1 is physically very small, although it is more statistical significant than the one at location 2. In this case, it is more reasonable to identify practical or physical shape difference at location 2 rather than at location 1. In order to achieve this, we propose to use the multivariate bioequivalence hypothesis test for our surface morphometry analysis:

$$H_0 : \max_{s \in \{x, y, z\}} |\mu_A^{(s)} - \mu_B^{(s)}| \leq \Delta$$

$$H_a : \max_{s \in \{x, y, z\}} |\mu_A^{(s)} - \mu_B^{(s)}| > \Delta$$

where $\Delta$ is the desired threshold. That is, the shape difference is detected as significant if the mean vector difference is large enough in either $x$, $y$ or $z$ direction. Bioequivalence tests were originally introduced in the pharmaceutical industry to determine the bioequivalence [3]. In this work, we employ bioequivalence concept though for detecting bioinequivalence as in Eq. (2) we constructed, instead of bioequivalence as in the standard pharmaceutical studies.

A permutation test is valid if the observations are exchangeable under the null hypothesis. However, the condition of exchangeability under null hypothesis is...
not satisfied in hypothesis Eq. (2). We thus propose to utilize a two-step permutation test.

**Step 1:**
\[ H_0^{(1)}: \mu_i^{(1)} = \mu^{(1)} + \Delta, \quad s \in [x, y, z] \]
\[ H_1^{(1)}: \mu_i^{(1)} > \mu^{(1)} + \Delta \text{ or } \mu_i^{(1)} < \mu^{(1)} + \Delta \]

**Step 2:**
\[ H_0^{(2)}: \mu_i^{(2)} = \mu^{(2)} - \Delta, \quad s \in [x, y, z] \]
\[ H_1^{(2)}: \mu_i^{(2)} < \mu^{(2)} - \Delta \text{ or } \mu_i^{(2)} > \mu^{(2)} - \Delta \]

If a test of significance in hypothesis Eq. (3) or Eq. (4) gives a p-value lower than the \(\alpha/2\)-level, we reject the null hypothesis and significant shape difference exists. The total significance level in this case is still \(\alpha\) due to the involved two steps in Eq. (3) and Eq. (4). Note that the classical hypothesis is a special case of the bioequivalence hypothesis when \(\Delta = 0\). Classical hypothesis is used in applications where statistical and practical significances are consistent. Otherwise, bioequivalence test is preferred if there is any non-negligible difference between practical significance and statistical significance.

### 3. Proposed New Permutation Approach

#### 3.1. Pearson Distribution Series

The Pearson distribution series (Pearson I ~ VII) are a family of probability distributions that are more general than the normal distribution [5]. It covers all distributions in the \((\beta_1, \beta_2)\) plane including normal, beta, gamma, log-normal and etc., where distribution shape parameters \(\beta_1, \beta_2\) are the square of the standardized skewness and kurtosis measurements, respectively. According to some permutation of the numbers 1 to \(n\), it is natural to partition the index space \(U = \{1 \ldots n\}^T\) into
\[ U^{(r_1, r_2, \ldots, r_q)} \]
where \(U^{(r_1, r_2, \ldots, r_q)}\) means that all \(r\) indices are permuted into \(q\) different numbers. Each number corresponds to \(\lambda\) indices. When \(r = 3\), \(U = U^{(1,1)} \cup U^{(1,2)} \cup U^{(3)}\), where \(U^{(1,1)}\) is the set of \(\{i_1 \neq i_2 \text{ and } i_2 \neq i_3\}\) with \(q = 3\) and \(\lambda_1 = \lambda_2 = \lambda_3 = 1\), \(U^{(1,2)}\) is the set of \(\{i_1 = i_2 \neq i_3\}\) with \(q = 3\) and \(\lambda_1 = \lambda_2 = 2\), \(U^{(3)}\) is the set of \(\{i_1 = i_2 = i_3\}\) with \(q = 3\). Since permutation is equally related to all \(r\) indices, \(\frac{1}{n!} \sum_{x \in \mathbb{Z}} \prod_{k=1}^{r} c_{\pi(x)}\) is invariant in each category, we define \(\frac{1}{n!} \sum_{x \in \mathbb{Z}} \prod_{k=1}^{r} c_{\pi(x)}\) as moment coefficient, \(a_{\lambda_1, \lambda_2, \ldots, \lambda_q}\), if \(\{i_1 t_2 \ldots i_r\} \in U^{(r_1, r_2, \ldots, r_q)}\).

Eventually, the \(r\)-th conditional moment is:
\[ E(T^r(X, \pi)|X) = \sum_{\lambda_1, \lambda_2, \ldots, \lambda_q} a_{\lambda_1, \lambda_2, \ldots, \lambda_q} \sum_{(i_1, i_2, \ldots, i_r) \in U^{(r_1, r_2, \ldots, r_q)}} \prod_{k=1}^{r} x_i \]

Eq. (6) separates the permutation from the data. To get the moments, we only need to derive the permutation of the coefficients of pre-chosen test statistics and calculate the summation terms of data. Due to the simple pattern of the coefficients of test statistics which is the same for repeated tests, we can derive the moments of permutation distribution without permutation of the data [12]. Alternatively, all \(a_s\) can also be calculated by computer simulation without analytical derivation. In addition, the discussed
approach can be easily extended to the multivariate case [12].

3.3. Mean Difference Test Statistic

In surface shape analysis, we use mean vector difference as test statistic, and $T = [TX, TY, TZ] = [CXY, CYY, CZY]$, where the mean difference vector $C = \frac{1}{n_1}1_{n_1} - \frac{1}{n_2}1_{n_2}$, $X' = X + [0_{n_2}, 1_{n_2}]^T \Delta$, $Y' = Y + [0_{n_2}, 1_{n_2}]^T \Delta$, and $Z' = Z + [0_{n_2}, 1_{n_2}]^T \Delta$. is the desired threshold for bioequivalence test, and is equal to zero in classical hypothesis test case. The detailed and complete formulas of corresponding $a$s is derived and listed in [12]. For the mean difference test statistic, the computation cost of data summation terms for the $r$-th moment in each index subspace can be reduced to $O(n)$ from $O(n^2)$.

4. Experiments and Results

4.1. Simulated Data

In this experiment, we generated six different simulated data sets to evaluate our hybrid permutation tests. In case #1 and case #2, two group data are normal distributed with different mean and variance(Normal(0,1) vs Normal(1,0.5)) in balanced design ($n_1 = n_2 = 10$) and unbalanced design($n_1 = 6$, $n_2 = 18$), respectively. Each group has gamma distribution in case #3 (gamma(3,3) vs gamma(3,2), $n_1 = n_2 = 10$) and case #4 (gamma(3,3) vs gamma(3,2), $n_1 = 6$, $n_2 = 18$). In case #5 and case #6, two group data has beta distribution with different parameters (Bets(0.8,0.8) vs Bets(0.1,0.1)) in balanced design ($n_1 = n_2 = 10$) and unbalanced design($n_1 = 6$, $n_2 = 18$).

Table 1 indicates the high accuracy of our hybrid permutation technique, especially for the tail area (Note: the exact permutation results are considered as ground truth.) Furthermore, comparing with exact permutation or random 20,000 permutations, the hybrid permutation tests reduce more than 99% computation cost and can save much more computation time as the sample size increases.

We also generated a synthetic dataset to demonstrate that bioequivalence test plays an important role in identifying practical significance. There are 12 surfaces in group A and 9 in group B, which were generated by adding two types of Gaussian noises to the two flat patches, a $(5 \times 5)$ top patch and a $(21 \times 21 - 5 \times 5)$ bottom patch. For group A, Gaussian noise with mean zero and standard deviation $\sigma_a = 0.01$ was added to the bottom patch with $z = 0$; Gaussian noise with mean zero and standard deviation $\sigma_a = 0.09$ is added to the top patch with $z = 1$. The 9 surfaces in group B were generated with the same noise patterns as in group A but to different bottom patch $z = 0.01$ and top patch $z = 0.9$. Since the differences between the bottom patches of the two groups are very small ($z = 0$ vs. $z = 0.01$), the practical group differences should only occur on the top patch ($z = 1$ vs. $z = 0.9$). Fig. 1(c) shows that many significance locations on both the top and bottom patches are detected by the classical hypothesis tests. Also, the non-practical significances can not be revealed with a more conservative significance level (i.e., lower $\alpha$) because not all $p$-values on the top patch are lower than those of the bottom patch (see Fig. 1(d)). Using bioequivalence tests, we are able to precisely identify the practical significances that occur at the top patch (Fig. 1(e)).

<table>
<thead>
<tr>
<th>Case</th>
<th>$p_{HP}$</th>
<th>$p_{RP}$</th>
<th>$p_{EP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case #1</td>
<td>0.0113</td>
<td>0.0495</td>
<td>0.0498</td>
</tr>
<tr>
<td>Case #2</td>
<td>1.1584</td>
<td>0.1314</td>
<td>0.1301</td>
</tr>
<tr>
<td>Case #3</td>
<td>4.4389</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>Case #4</td>
<td>4.2795</td>
<td>0.0249</td>
<td>0.0250</td>
</tr>
<tr>
<td>Case #5</td>
<td>4.2983</td>
<td>0.0908</td>
<td>0.0925</td>
</tr>
<tr>
<td>Case #6</td>
<td>4.1320</td>
<td>0.0805</td>
<td>0.0803</td>
</tr>
</tbody>
</table>

Figure 1. (a): Mean shape of group A. (b): Mean shape of group B. (c) and (d): Results using conventional hypothesis tests with $\alpha = 0.05$ (c) and with $\alpha = 0.001$ (d). (e) Results using bioequivalence tests with $\alpha = 0.05$ and $\alpha = 0.025$. 4.2. Real Data of the Human Brain

The MRI hippocampi used in this experiment were semi-automatically segmented by human expert raters and manually grouped into 2 groups with 21 subjects in group A and 15 in group B.
A new statistical surface morphometry analysis method is presented and developed by using our hybrid permutation tests where the permutation distributions are accurately approximated through Pearson distributions for considerably reduced computation cost. The proposed hybrid strategy takes advantage of nonparametric permutation tests and parametric Pearson distribution approximation to achieve both accuracy/flexibility and efficiency. Note that the described theoretical derivations are general and can be applied to any linear test statistics on multivariate data, but not limited on the test statistic demonstrated in this paper. In addition, hybrid permutation schemes for both the conventional and bioequivalence tests are provided. Compared with the classical hypothesis tests, bioequivalence tests can screen out the non-notable differences and accurately locate the practical or physical significances. In real biomedical applications, either the standard or the bioequivalence hypothesis tests can be chosen, depending on the specific problems, i.e. whether the statistical and practical significance differences are negligible or not.

5. Conclusion

A new statistical surface morphometry analysis method is presented and developed by using our hybrid permutation tests where the permutation distributions are accurately approximated through Pearson distributions for considerably reduced computation cost. The proposed hybrid strategy takes advantage of nonparametric permutation tests and parametric Pearson distribution approximation to achieve both accuracy/flexibility and efficiency. Note that the described theoretical derivations are general and can be applied to any linear test statistics on multivariate data, but not limited on the test statistic demonstrated in this paper. In addition, hybrid permutation schemes for both the conventional and bioequivalence tests are provided. Compared with the classical hypothesis tests, bioequivalence tests can screen out the non-notable differences and accurately locate the practical or physical significances. In real biomedical applications, either the standard or the bioequivalence hypothesis tests can be chosen, depending on the specific problems, i.e. whether the statistical and practical significance differences are negligible or not.

References