Increasing statistical power in medical image analysis

Alexei M. C. Machado
Pontifical Catholic University of Minas Gerais
Av. Dom Jose Gaspar, 500, Belo Horizonte, MG, Brazil
alexei@pucminas.br

Abstract

In this paper, we present a novel method for estimating the effective number of independent variables in imaging applications that require multiple hypothesis testing. The method increases the statistical power of the results by refuting the assumption of independence among variables, while keeping the probability of false positives low. It is based on the spectral graph theory, in which the variables are seen as the vertices of a complete undirected graph and the correlation matrix as the adjacency matrix that weights its edges. By computing the eigenvalues of the correlation matrix, it is possible to obtain valuable information about the dependence levels among the variables of the problem. The method is compared to other available models and its effectiveness illustrated in a case study on the morphology of the human corpus callosum.

1 Introduction

An important problem related to medical image analysis is the significance evaluation of the results. The information obtained from a study must provide enough and significant evidence to support the hypothesis under investigation. For example, one may be interested in proving that the shape of a specific structure or organ is significantly different when comparing two populations and this is usually accomplished by hypothesis testing. A problem arises when the imaging modality provides large amounts of information, placing the problem in a high-dimensional variable space. Under these circumstances, the problem may require the simultaneous testing of multiple hypotheses, which requires special attention, so as to avoid the consideration of false positive results. While testing a false hypothesis on a set of 100 variables, at the significance level of 0.05, it is expected that 5 variables will show significant when, in fact, they are not. Therefore, an adjustment on the significance values is required, in order to reduce the probability of Type I error.

Milestones on the theory of multiple comparison correction are the works of Bonferroni [4] and Sidak [21], who proposed a classic and conservative method to adjust the significance values based on the number of hypotheses being jointly evaluated. The problem about correcting significance values is that, while reducing the Type I error, it *increases* the Type II error, i.e., it increases the probability that a true hypothesis will be considered false, reducing the statistical power of the analysis. Other less conservative methods followed the work of Bonferroni, but none of them presented a suitable solution for dealing with high-dimensional problems, leading many researches to question the adequacy of multiple comparison correction [2, 17, 19].

In this work, we present a novel method for multiple comparison correction based on spectral graph theory. It increases the statistical power of the results in medical imaging applications by refuting the assumption of independence among variables, which is usually a premise in other approaches. We first critically review the main available methodologies for significance adjustment, showing their inadequacy on typical problems related to medical image analysis. The proposed method is then presented and evaluated using a real study in image registration, followed by discussion and conclusions.

2 Multiple-hypothesis testing

The standard and perhaps most conservative method of adjustment was proposed by Bonferroni [4] (also known as the Sidak method [21]). If we have a set of n individual and independent hypotheses being tested at the significance level α, the probability of improperly rejecting k out of n hypotheses follows a Binomial distribution:

$$P(k) = \frac{n!}{(n-k)!k!}(1-\alpha)^{n-k}\alpha^k. \qquad (1)$$

In order not to make any false rejections, k must be equal to zero. The probability of making at least 1 false rejection can be computed with the aid of (1) as

$$P(k>0) = 1 - P(0) = 1 - (1-\alpha)^n. \qquad (2)$$

Since $(1-\alpha)^n \approx 1-n\alpha$, for small α, if we wish the whole experiment to have a false positive rate of π, each individual hypothesis will need to be tested at the level of

$$\alpha = 1-(1-\pi)^{1/n} \approx \pi/n. \quad (3)$$

The method of adjustment for individual significance level given by (3) is known as the Bonferroni correction. It is very conservative in the sense that it reduces the statistical power of the individual tests, i.e. the probability of correctly rejecting a false null hypothesis. If we have a set of $n = 100$ tests and wish the experiment to be tested at the level of $\pi = 0.05$, each test will have to be tested at a significance level $\alpha = 0.0005$, which is extremely low. Therefore, many potentially significant results may be uncovered, increasing the probability of false negatives. The correction is nevertheless necessary to avoid false positives that would be expected to appear once in every 20 tests.

A less conservative variation of the Bonferroni correction considers that, after the first test is performed, the number of remaining tests reduces to $n-1$. The tests may therefore be taken sequentially, as proposed by Holm [12]. In this approach, the most significant test is adjusted first and, if it is able to reject the null hypothesis, the process moves to the next test, now considering $\alpha = \pi/(n-1)$. As soon as one test fails to reject the null hypothesis, all subsequent tests are considered not significant. A variation to Holm's method, presented by Simes [22] and Hochberg [11] starts with the test presenting the largest p-value, taking $n = 1$, and goes backwards until a test fails to be rejected. Hommel [13] also proposes a sequential method based on the evaluation of ordered tests. Although less conservative than Bonferroni, the sequential methods of Holm, Simes-Hochberg and Hommel are still inadequate to applications such as the ones in medical imaging analysis that usually involve large number of variables.

The methods derived from the Bonferroni correction aim to control the false positive rate (FPR) of multiple tests, i.e., to reduce the fraction of null hypotheses that are erroneously called significant. An alternative approach on multiple comparison proposed by Benjamini and Hochberg [3] considers the fraction of the false positives over the amount of tests declared significant. This metric is called the false discovery rate (FDR). While a FPR α=0.05 means that, on average, 5% of all tests will be declared significant when they are not, a FDR δ=0.05 expects that 5% of the tests declared significant (and only those) will be false positives. More formally, the FPR is the probability of a test being considered significant, given that the null hypothesis is null, whether the FDR is the probability of a null hypothesis being true, given that the test was considered significant. There is a flavor of the Bayes' Theorem on the relationship between both concepts.

The FDR was originally estimated by Benjamini and Hochberg, by sequentially taking the n tests, ordered from the smallest to the largest p-value. The k-th test is considered significant if its p-value, p_k is such that $p_k \leq \delta k/n$, where δ is the experiment-wise FDR. The problem with the FDR estimator is that it still considers the number of tests, n, to decide on the significance of an individual test. The first test in the sequence (the one with smallest p-value) will actually be computed in the same way as with the Bonferroni method. The approach is therefore too conservative for problems in high-dimensional variable spaces.

3 The assumption of independence

Multiple comparison adjustments on significance levels usually rely on the assumption of independence among the individual tests. The Bonferroni correction, for instance, assumes that n independent null hypothesis H_0^i, $i = 1 \ldots n$ are tested and the corresponding probability of false rejection evaluated. The joint probability of Type I error for $H_0^1 \wedge H_0^2 \wedge \ldots \wedge H_0^n$ is therefore the product of the individual probabilities. The significance level α that should be applied to each hypothesis, so as to keep the experiment-wise level of significance π, is $\alpha = \pi/n$. However, if some degree of dependence exists among the variables being tested, the significance threshold applied to each individual hypothesis will be smaller than it should be. In the context of dependence, we are actually not testing n independent hypotheses, but m independent groups of dependent hypotheses. Of course, dependence comes in different degrees, what makes the problem non-trivial.

Many problems in medical image analysis are defined over a variable space that embeds intricate nets of dependence. Image registration algorithms, for example, induce a high level of smoothness in the results. The tissues and organs being imaged generally present smooth shapes or intensity variation. Interpolation may be used in many cases, adding another level of smoothness. In summary, determining a realistic estimation for the number of independent hypotheses considered in those problems becomes a very important issue.

The development of Spectral Graph Theory brought a new perspective to the evaluation of dependence among the variables [6]. The set of variables can be seen as the vertices of a complete undirected graph and the correlation matrix as the adjacency matrix that weights their connections. It is known that the eigenvalues of the correlation matrix give valuable information about the dependence levels among the variables of a problem, although the quantification of this phenomenon is not yet completely understood [5, 24]. Eigendecomposition is used in Principal Component Analysis (PCA) as a tool to reduce the dimensionality of a problem. In PCA, the set of original variables is rotated in order to find the orthogonal axes along which the data is maxi-

mally spread out. Data reduction is achieved by changing the basis of the variable space, so that the new orthogonal axes represent most of the variance embedded in the dataset, and by ignoring the axes in which the data present small variance. Each new variable (principal component) associated with an eigenvector is a linear combination of the original variables and the corresponding eigenvalue represents its variance. Without loss of generality, we will consider that variables are standardized, i.e., their mean and variance are zero and one, respectively. The covariance and correlation matrices are, therefore, the same.

The two extreme cases of the spectral decomposition of a correlation matrix occur when the n variables are either completely correlated to each other or completely uncorrelated. In the first case, there will be a single non-null component accounting for the variance of all variables (eigenvalue equals to n). In the second case, there will be no change in the basis, since the variables are already the n principal components, with all eigenvalues equal to one. For the intermediary cases, there is no simple solution. Therefore, a measure of dependency, representing the number of groups of correlated variables, should be estimated.

Cheverud [5], studying the linkage of traits across genomes, considered the variance of the eigenvalues, V_λ, as a measure of dependence. In the case of completely correlated variables, V_λ would be at its maximum, n; in the case of uncorrelated variables, V_λ would be at its minimum, zero. He proposed that an estimate for the intermediary cases could be given by the proportional reduction of the number of independent groups of variables, as the ratio of V_λ to its maximum, n. Rescaled to vary from one to n, the effective number of independent variables, m, was defined as

$$m = 1 + (n-1)(1 - \frac{V_\lambda}{n}). \tag{4}$$

The new value, m, replaced n, in (2), for computing the individual significance level threshold.

Although the method proposed by Cheverud was effective for computing the value of m in the extreme cases, it overestimated the value in situations of partial correlation. Li and Ji [14] showed that, when the n tests could be partitioned into c, $1 \leq c \leq n$, groups of n/c completely correlated tests, the spectral decomposition of the correlated matrix would yield a set of n/c identical eigenvalues c and $n - n/c$ null eigenvalues. In this scenario, the application of Cheverud's method defined in (4) would result on $m = n + 1 - c$, when in fact it should be n/c. They proposed a new method in which the eigenvalues were decomposed into their integral and fractional parts. Each integral part would represent a cluster of correlated variables and be counted as a single variable to the computation of m. The fractional parts were considered as partially correlated tests and added to m. Formally,

$$m = \sum_{j=1}^{n} [I(\lambda_j \geq 1) + (\lambda_j - \lfloor \lambda_j \rfloor)], \tag{5}$$

where I is the indicator function that returns 1 when the argument is true or 0 otherwise. The model is precise in the cases of spectral decompositions that result in integer eigenvalues, but still overestimates m in other cases, as we shall see in the next section.

A crucial issue on the evaluation of estimates for m in the case of partial correlations is to determine a ground truth for comparison. In fact, no procedure has been proved optimum for this purpose, although permutation tests are considered the best estimators for m. A permutation test [9, 15, 25] aims at determining the actual individual significance level that should be applied to each hypothesis so as to yield the desired experiment-wise FPR π. The samples being compared have their subjects permuted in order to generate all possible false configurations. The hypothesis tests are applied to the configurations, for each variable at the individual significance level α, and the number of joint hypotheses considered significant is determined. Under the assumption of independent variables, the procedure should result on a value close to π. If it is greater than π, there is evidence that α was overestimated and that the assumption of independence is false. Therefore, the value of m can be estimated based on the value obtained from the iterative process. When the number of permutations is intractable, the iteration may be repeated a reasonable number of times. At each iteration, a random configuration is generated. In this case, the method is known as *resampling* and its algorithm can be summarized as follows:

1. Determine the individual significance level threshold, α for testing each of the n hypothesis: Under the assumption of independence (that we want to refute), α can be computed from (2) as $\alpha = 1 - (1 - \pi)^{1/n}$, where π is the desired experiment-wise FPR threshold (e.g. 0.05).

2. Generate a configuration: for each pair of subjects in the sample, s_1 and s_2, belonging to class 1 and 2, respectively, generate a random number in the interval [0,1]. If this number is greater than 0.5, assign s_1 to class 2 and s_2 to class 1, otherwise keep them in their original classes.

3. Test the null hypothesis for each variable of the configuration and obtain a significance level for its rejection. If at least one of the variable presents a significance level less than or equal to α, the configuration is counted as a false positive.

4. Repeat steps 2 and 3, R times, in order to compute the number of false positives, F.

5. Compute the FPR for the experiment:$\pi_r = F/R$.

6. Compute the individual significance level threshold, α_r, that would have to be used in the resampling experiment in order to result in a experiment-wise FDR π_r: $\alpha_r = 1 - (1 - \pi_r)^{1/n}$.

7. Compute the effective number of independent hypotheses: If $\alpha_r > \alpha$, n is overestimated. Based on (2), the effective number of independent variables, m, that will result in a experiment-wise FPR π can be computed as:

$$m = \frac{\ln(1-\pi)}{\ln(1-\alpha_r)}. \quad (6)$$

The effectiveness of the resampling method is based on the fact that, while the subjects are permuted from one class to the other, the relationship among the variables in the dataset remains unchanged. The drawback of the method is its computational complexity that limits its application in the case of high-dimensional variable spaces. The number of iterations required to achieve a more precise estimate of m is proportional to n and inversely proportional to π. Since, at each iteration, n hypothesis tests should be performed, the computational complexity of the method is $O(n^2)$.

In the next section, we propose a more precise model for multiple comparison correction based on Spectral Graph Theory that better approximates the estimation of the effective number of independent hypotheses, at a much lower computational cost.

4 A new multiple comparison correction model

Although the method proposed by Li and Ji circumvented the limitations of Cheverud's method in the case of intermediary number of correlated variable groups, it still overestimates the number of independent variables, m, when the correlation is partial. If we analyze equation (5), we notice that m will always be an integer number, since all the fractional parts of the eigenvalues are summed. In fact, the integer number that exceeds 1 in each eigenvalue is subtracted from n to yield the effective number of independent variables, m. Partial (fractional) correlation, in this case, is actually not accounted for.

Another limitation of equation (5) is that it overestimates m when the rank of the correlation matrix is less than n. It is known that that the rank of a matrix is less than or equal to its minimum dimension [23]. Furthermore, when 2 matrices are multiplied, the rank of the resulting matrix cannot exceed the rank of each factor. The correlation matrix, $\mathbf{R}$, is given as $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}/N$, where $\mathbf{Z}$ is the $N \times n$ data matrix whose rows contain the N subjects. Therefore, as far as the number of subjects is less than or equal to the number of variables, the rank of $\mathbf{R}$ will be at most N, meaning that there will be at least $n-N$ linearly dependent variables. Actually, the eigendecomposition of $\mathbf{R}$ in this scenario yields at most $N-1$ non-null eigenvalues. Consequently, the number of independent variables, m, has an upper bound equal to $N-1$.

4.1 Description of the model

The model we propose for computing the number of independent variables is derived from the interpretation of eigendecomposition, that aims at obtaining a lower-dimensional variable space for the problem, in which the new variables (principal components) are uncorrelated. The eigenvalues associated to each component represent the variance of the new variable. When rotation is applied to $\mathbf{R}$, in order to find its axes of maximum variance, high-correlated groups of original variables result in new components with large variances. Each new component is an uncorrelated variable in the new basis. Original variables that cannot be grouped result in components that present smaller variances. Their contribution to data description is at most equal to one original variable and should be counted as proportional to the variance they explain, given by their eigenvalues. Based on this rationale, the number of independent variables, m, is given as

$$m = \sum_{j=1}^{n} [I(\lambda_j \geq 1) + I(\lambda_j < 1)\lambda_j], \quad (7)$$

where I is the indicator function that returns 1 when the argument is true or 0 otherwise.

4.2 Comparison with related work

The proposed model described in (7) satisfies all the extreme cases explored by Cheverud [5] and Li and Ji [14]. In the case of completely independent variables, all eigenvalues will be equal to 1 and $m = n$. For completely correlated variables, all eigenvalues will be 0, except one that equals to n. In this case, $m = 1$. For intermediary number of completely correlated variables, the method behaves as the model of Li and Ji and better than Cheverud's. If, for instance, the n variables can be partitioned into c, $1 \leq c \leq n$, mutually independent groups of n/c completely correlated variables, there will be n/c identical eigenvalues c, $c > 1$, and m will be equal to n/c, as expected.

The advantages of the new method are evident when the number of subjects, N, is less than or equal to the number of variables, n, as generally happens in medical imaging applications. In this case, the maximum number of non-zero eigenvalues computed for the correlation matrix

$\mathbf{R}$ is $N-1$, which is the rank of $\mathbf{R}$. The value for m computed by the proposed method has an upper bound of $N-1$, whereas the methods proposed by Cheverud and Li and Ji may exceed this limit when partial correlation is observed. For example, consider a study composed of n variables that can be partitioned into $N-1$ independent groups of $c=(n-1)/(N-1)$, $c>1$, completely correlated variables and a single variable that is only partially correlated to the groups. Let $n > N+c$, so that the rank of $\mathbf{R}$ is $N-1$. In this case, there will be $N-1$ eigenvalues greater than 1 and $n-N+1$ null eigenvalues. Furthermore, at least 2 eigenvalues will have fractional parts greater than 0, since there is a variable that is only partially correlated to the others. The $N-1$ non-zero eigenvalues can be denoted as $\lambda_j = c+d_j$, such that $0 \leq d_j < 1$ and $\sum_{j=1}^{N-1} d_j = 1$. In this case, the model of Li and Ji will result on $m=N$, exceeding the rank of $\mathbf{R}$. For the method proposed by Cheverud in equation (4), V_λ will be computed as

$$\begin{aligned} V_\lambda &= \tfrac{1}{n-1}[\textstyle\sum_{j=1}^{N-1}(c+d_j-1)^2 + \sum_{j=N}^{n}(0-1)^2] \\ &= \tfrac{1}{n-1}[(N-1)(c-1)^2 + 2(c-1) + x + n - N + 1] \\ &= \tfrac{1}{n-1}[cn + c - n + x], \end{aligned} \tag{8}$$

where $x = \sum_{j=1}^{N-1} d_j^2$, $0 < x < 1$. From (4) and (8), m is given as

$$m = n - \frac{1}{n}(cn + c - n + x).$$

In order for the method to yield a value of m that is less than or equal to the rank of $\mathbf{R}$, the following inequality must be true:

$$m = n - c + \frac{c}{n} - 1 + \frac{x}{n} \leq N - 1. \tag{9}$$

Since both c/n and x/n are positive but smaller than 1, the inequality in (9) is equivalent to $n < N + c$ which contradicts the condition of n being greater than $N + c$. Therefore, we conclude that neither the model of Cheverud nor the method of Li and Ji guarantee an upper bound equal to the rank of $\mathbf{R}$ when computing the effective number of independent variables for multiple comparison correction.

4.3 Complexity analysis

The critical step on the computation of the effective number of independent variables, following the approach of spectral graph theory, is the eigendecomposition of the correlation matrix $\mathbf{R}$. This procedure is $O(n^3)$ with respect to computational cost, where n is the number of variables. In the case of high-dimensional variable spaces, the computation of $\mathbf{R}$ itself is already intractable. It is known, however, that the eigenvalues of $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}/N$ and $\mathbf{Z}\mathbf{Z}^T/N$ are the same [7], so that the eigenvalues can be obtained in time proportional to $O(N^3)$, where N, the number of subjects, is usually small. The critical step becomes the computation of $\mathbf{Z}\mathbf{Z}^T/N$, which is $O(N^2 n)$.

5 A case study on the morphology of the corpus callosum

In this section, we illustrate the effectiveness of the proposed method in a study of gender-related shape differences on the morphology of the human corpus callosum.

5.1 Materials

The MRI images used in the experiments, gently shared by the Mental Health Clinical Research Center of the University of Pennsylvania, are 56 normal controls recruited for a larger study on schizophrenia. The subjects are right-handed with average age and standard deviation of 29.9 ($\pm$13.4) years for the male group (N=28) and 27.5 ($\pm$11.1) years for the female group (N=28). The images were acquired on a GE 1.5 Tesla instrument, using a spoiled GRASS pulse sequence optimized for high resolution, near isotropic volumes (flip angle = 35^o, TR = 35 ms, TE = 6 ms, field of view = 24 cm, 0.9375 $\times$ 0.9375 mm^2 in-plane resolution, 1.0 mm slice thickness, no gap). The images were obtained in the axial plane and the midsagittal slice extracted and reformatted into 256 $\times$ 256 8-bit images. Details about the sample recruitment and acquisition procedures are described in Shtasel *et al.* [20] and Gur *et al.* [10]

Although the data used in the experiments are real MRI images, the subjects were deliberately selected based on the shape of the callosal splenium, so that the female group would present more "bulbous" splenium (posterior-most part of the corpus callosum). The purpose of working with a biased sample was to ensure that hypothesis testing would show significant differences in this particular region of interest.

5.2 Experimental procedure

The images were first partitioned into white matter, gray matter and cerebro-spinal fluid components using the adaptive K-means clustering algorithm of Pappas [16]. The corpus callosum structure was extracted by manual delineation and the segmented callosa rigidly registered to the template by computing the center of area and orientation of the structures. Translation and rotation were performed to bring them into global registration. The boundaries of the callosa were automatically determined using the Rosenfeld algorithm for 8-connected contours [18]. Local registration was performed based on a parametric curve matching algorithm [1] in which the boundary of the template is registered to each subject's boundary, maximizing their geometric correspondence. The resulting displacement field was then extrapolated to the whole structure [1].

When the template image is warped to match a subject image, some regions may get enlarged and some may be re-

(a) (b)

Figure 1. Average callosal shape for the male (a) and female (b) groups.

duced. It is possible to determine the amount of scaling applied to an infinitesimal area around each point of the template, by computing the Jacobian determinant of the mapping function [8]. The pointwise Jacobian determinants are the variables that will be tested for size differences. Since the result of image registration is a smooth displacement field, it is expected that the Jacobian determinants of neighboring points be correlated.

In addition, volumetric variation was computed for equal-spaced segments taken perpendicularly to the callosum medial axis. The medial axis is a curve that splits the corpus callosum into dorsal and ventral regions, such that, at any point along the axis, the two perpendicular line segments emanating from the point connecting the axis to dorsal and ventral points of the boundary have the same length. Medial axis extraction was performed using a variation of the thinning algorithm described in Rosenfeld and Kak [18], with subsequent pruning of spurious branches. The curve representing the medial axis was extended to terminate at the tips of the rostrum and splenium, and a sample of 67 equally spaced interpolated points were taken, so as to yield an isotropic rotation-invariant representation for the template axis. Volumetric variation was computed after the registration of the curves representing the medial axes of the template and the subjects. For each segment of the template, with volume v_T, registered to a segment of the subject, with volume v_S, the amount of scaling was defined as v_S/v_T. The analysis of callosal volumetric differences between the samples was performed based on multiple analyses of variance (ANOVA).

5.3 Experimental results

The average shapes of the male and female callosa are shown in Fig. 1 and were generated based on the displacement fields resulting from image registration. A grid was superimposed to the images so as to facilitate the observation of the significant differences in the size of the splenium (larger in females) and in the isthmus (larger in males). The genu (anterior-most part of the callosum) also appears to be larger in the males. The template image used in image registration was composed of 2830 pixels. The results of individual F-tests applied to each of the 2830 Jacobian determinants, showing the regions in which there is enough evidence to reject the null hypothesis of equal means, are displayed in Fig. 2(a). The figure shows the significance (p-value) of each test without multiple comparison correction. The tests are 2-tailed, with α=0.05 (0.025 in each tail).

The effective number of independent variables, m, was computed based on the spectral decomposition of the correlation matrix $\mathbf{R}$. As expected, since the number of variables, 2830, was much larger than the number of subjects, 56, the eigendecomposition of $\mathbf{R}$ yielded only 55 non-zero eigenvalues ranging from 517.44 to 2.44. The resulting values for m computed by the methods of Cheverud , Li and Ji, and the model proposed in this paper were 2580.64, 81 and 55, respectively. While our method achieved the upper limit for m (the rank of $\mathbf{R}$), the other methods exceeded this limit. The resampling procedure could not be used to estimate the true value of m, in this case, since such large number of variables would require the evaluation of an intractable number of permutations.

The adjusted p-values were computed from (3), using the computed values of m in place of n. Fig. 2(b-e) show the results based on Bonferroni, Cheverud, Li and Ji, and the method proposed in this paper. The adjusted significance threshold, α, were respectively 0.000018, 0.000019, 0.000617 and 0.000909. It can be seen that the Bonferroni correction and the method of Cheverud were unable to detect any region in which the average male callosum was significantly larger than the female. Only the splenium was considered significantly larger in the females, at the experiment-wise false positive rate $\pi = 0.05$. The model of Li and Ji was able to additionally detect a difference in the isthmus, in which the males had larger Jacobian determinants. However, only the method we proposed was sufficiently powerful to detect some difference in the genu.

The computation of the eigenvalues of $\mathbf{R}$ took 0.35 second. All methods were implemented in IDL language (Research Systems) and run in a 1.1 GHz Intel Celeron processor computer with 256 MB of RAM, under Windows XP operating system.

The second experiment was based on the volumetric variation taken at each of the 67 segments of the template. In this case, it was possible to estimate a reference ground truth for the number of independent variables, based on resampling. The results of individual F-tests applied to the volume of each segment, showing the regions in which there is enough evidence to reject the null hypothesis of equal means, are displayed in Fig. 3(a). The figure shows the significance (p-value) of each test without multiple comparison correction, at the significance level α=0.05.

The eigendecomposition of $\mathbf{R}$ yielded 55 non-zero eigenvalues: 11 greater than 1, ranging from 23.82 to 1.18, and 44 smaller than 1. The resulting values for m computed by the methods of Cheverud, Li and Ji and the proposed

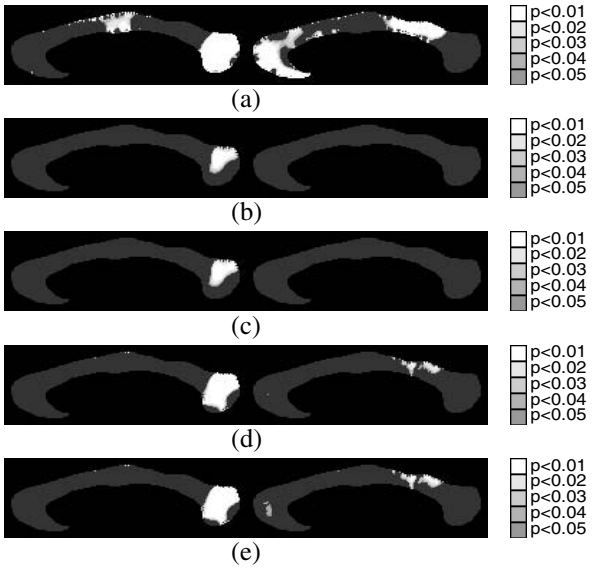

Figure 2. Results of Jacobian analysis. Regions in which H_0 is rejected based on unadjusted tests (a) Bonferroni (b), Cheverud (c), Li (d) and the proposed method (e).

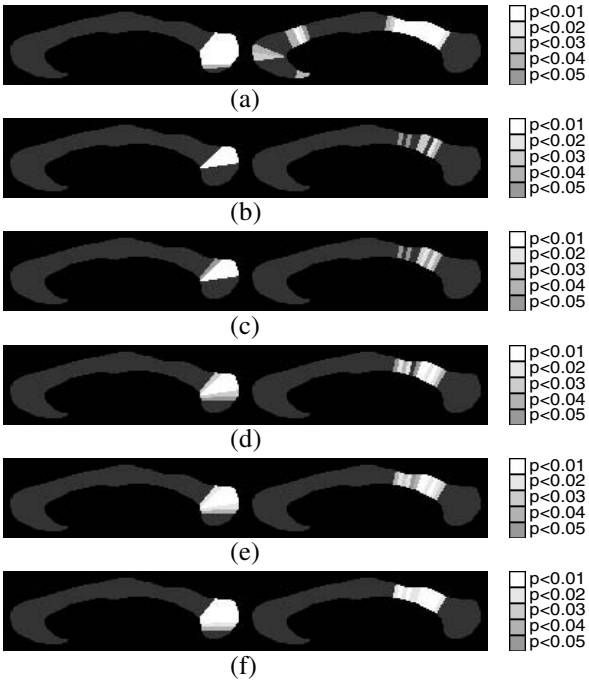

Figure 3. Results of volumetric analysis. Regions in which H_0 is rejected based on unadjusted tests (a) Bonferroni (b), Cheverud (c), Li and Ji (d), the proposed method (e) and by resampling (f).

model were 56.78, 24 and 17.86, respectively. The computation of the eigenvalues took only 0.01 second. The true value of m was estimated as 8.28 by the resampling procedure using 200,000 permutations. The iterative algorithm took 6227.27 seconds to compute.

The adjusted p-values computed based on Bonferroni, Cheverud, Li and Ji, the method proposed in this paper and from resampling are shown in Fig. 3(b-f). The adjusted significance threshold, α, were 0.000746, 0.000880, 0.002083, 0.002799 and 0.006175, respectively. In this experiment, all methods failed to detect differences in the genu. Since a segment encompassed a larger region of the structure, the computation of volumetric variation actually considered an average of significant and non-significant values. Therefore, the differences at the anterior half of the corpus callosum were much less significant than at the splenium and isthmus. The model of Li and Ji and our method provided the results that best approximated the values computed by resampling. The value for the effective number of independent tests computed by our model was 2.16 times larger than the one obtained by resampling, whereas the model of Li and Ji resulted in a value approximately 3 times larger.

6 Conclusion

In this paper we proposed a novel method for computing the effective number of independent variables used to adjust significance values in multiple comparison testing. The approach is based on the spectral decomposition of the correlation matrix and is both statistically powerful and computationally efficient. The method has two major advantages over other available methods: it is less conservative under partial correlation and has an upper bound equal to the rank of the correlation matrix. The effectiveness of the method was illustrated in a case study on the morphology of the human corpus callosum. A set of significantly different shapes drawn from male and female samples was used to evaluate the statistical power of the model compared to related works. The set of null hypotheses tested for significance was composed of 2830 Jacobian determinants, describing regional volumetric variability. Typical problems in medical imaging analysis are usually represented in such high-dimensional spaces, making it impossible to estimate the true number of independent tests by permutation or resampling procedures. Therefore, volumetric variation was additionally investigated based on the volume of segments taken perpendicular to the medial axis of the structure, reducing the number of tests to 67. In this scenario, it was possible to compare the results of the method with the estimated ground truth computed from resampling. The num-

ber of independent variables, in both experiments, was significantly lower than the ones computed by other methods based on spectral decomposition. In the analysis of Jacobian determinants, the method was the only one to satisfy the upper bound of the problem, defined by the rank of the correlation matrix.

The difference between the results obtained with the method and the ones obtained from resampling suggests that further improvement is possible. The way in which dependence is represented by the eigenvalues of the correlation matrix is still unclear. Principal component analysis aims to obtain new variables that maximally represent the variance of the sample. In this process, the covariance between original variables is taken into account, but the relationship between the original values and the variance of the new components are not simple. Therefore, any improvement in the computation of the number of independent variables is justified, as far as it contributes to increase statistical power, while keeping the probability of false positives low.

Acknowledgments — This work was partially supported by CNPq-Brazil. The author is grateful to the University of Pennsylvania for sharing the corpus callosum data.

References

[1] B. Avants and J. C. Gee. Soft parametric curve matching in scale-space. In *Proceedings of the SPIE Medical Imaging 2002: Image Processing*, pages 1139–1151, San Diego, 2002.

[2] R. Bender and S. Lange. Adjusting for multiple testing: When and how? *J Clin Epidemiol*, 54:343–349, 2001.

[3] Y. Benjamini and T. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 85:289–300, 1995.

[4] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[5] J. M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87:52–58, 2001.

[6] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Fresno, 1997.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models: Their training and applications. *Comput Vis Image Und*, 61(1):38–59, 1995.

[8] C. Davatzikos, M. Vaillant, S. Resnick, J. Prince, S. Letovsky, and R. Bryan. A computerized approach for morphological analysis of the corpus callosum. *J Comput Assist Tomo*, 20(1):88–97, 1996.

[9] P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image-based studies. *Lect Notes Comput Sci*, 2732:330–341, 2003.

[10] R. E. Gur, B. I. Turetsky, W. D. Bilker, and R. C. Gur. Reduced gray matter volume in schizophrenia. *Arch Gen Psychiat*, 56:905–911, 1999.

[11] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803, 1988.

[12] S. Holm. A simple sequential rejective multiple test procedure. *Scand J Stat*, 6:65–70, 1979.

[13] G. Hommel. A stagewise rejective multiple test procedure on a modified bonferroni test. *Biometrika*, 75:383–386, 1988.

[14] J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227, 2005.

[15] D. R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*, 74:765–769, 2004.

[16] T. N. Pappas. An adaptive clustering algorithm for image segmentation. *IEEE T Signal Proces*, 40:901–914, 1992.

[17] T. V. Perneger. What's wrong with bonferroni adjustments. *BMJ*, 316:1236–1238, 1998.

[18] A. Rosenfeld and A. Kak. *Digital Picture Processing*. Academic Press, Orlando, 1982.

[19] K. J. Rothman. No adjustments are needed for multiple comparisons. *Epidemiol*, 1:43–46, 1990.

[20] D. L. Shtasel, R. E. Gur, P. D. Mozley, J. Richards, M. M. Taleff, C. Heimberg, F. Gallacher, and R. C. Gur. Voluntiers for biomedical research: recruitment and screening of normal controls. *Arch Gen Psychiat*, 48:1022–1025, 1991.

[21] Z. Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 62:626–633, 1967.

[22] J. R. Simes. An improved bonferroni procedure for multiple test of significance. *Biometrika*, 73(3):751–754, 1986.

[23] G. Strang. *Linear Algebra and Its Applications*. Academic Press, San Diego, 1980.

[24] G. Wagner. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: Evidence for a nonrandom organization of quantitative character variation. *J Math Biol*, 21:77–95, 1984.

[25] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. John Wiley & Sons, New York, 1993.