The 3D Moore-Rayleigh Test for the Quantitative Groupwise Comparison of MR Brain Images.

A.E.H. Scheenstra[1,2], M. Muskulus[5], M. Staring[1,2], A.M.J.V. van den Maagdenberg[3,4], S. Verduyn Lunel[5], J.H.C.Reiber[1,2], L. van der Weerd[1,4], and J. Dijkstra[1,2]

[1] Department of Radiology, [2] Division of Image Processing,
[3] Department of Neurology, [4] Department of Anatomy and Embryology,
Leiden University Medical Center, Postbus 9600, 2300 RC , Leiden, the Netherlands
[5] Leiden University, Mathematical Institute, Leiden, The Netherlands

Abstract. Non-rigid registration of MR images to a common reference image results in deformation fields, from which anatomical differences can be statistically assessed, within and between populations. Without further assumptions, nonparametric tests are required and currently the analysis of deformation fields is performed by permutation tests. For deformation fields, often the vector magnitude is chosen as test statistic, resulting in a loss of information. In this paper, we consider the three dimensional Moore-Rayleigh test as an alternative for permutation tests. This nonparametric test offers two novel features: first, it incorporates both the directions and magnitude of the deformation vectors. Second, as its distribution function is available in closed form, this test statistic can be used in a clinical setting. Using synthetic data that represents variations as commonly encountered in clinical data, we show that the Moore-Rayleigh test outperforms the classical permutation test.[1]

1 Introduction

Mice have been used in genetic research as models for a variety of diseases occurring in the human population. They allow researchers to study the development of genetic diseases, to improve early diagnoses and subsequent treatment. Non-invasive imaging techniques, e.g. MRI, allow localized investigation of 3D anatomical structures of interest [1]. This provides a useful tool for *in vivo* structural and functional phenotyping, especially in the brain [2]. Since the introduction of non-rigid registration of brain images, a variety of new applications for brain research have emerged. Non-rigid registration is used in clinical practice to register MR images taken from different biological populations to a common average. The resulting deformation fields indicate and localize differences between pairs of images. Their second order statistics are stored in atlases to characterize variability within a population (intra-group variability) [3, 4].

[1] J.L. Prince, D.L. Pham and K.J. Meyers (Eds.): IPMI 2009, LNCS 5636, pp 564 – 575, 2009. © Springer-Verlag Berlin Heidelberg 2009

A major challenge in this area of research is not only to highlight the intra-group variability but to also assess inter-group variability, especially with regard to structural anatomical differences and their possible causes. For instance, in genetic research with transgenic mice, mutants are compared with their wild-types, where the group difference is determined by only one gene. To test and localize possible anatomical differences, statistical testing is required. In human brain research, similar problems have been addressed in functional brain images. BOLD signals are compared between and within groups to characterize differences in brain activation. A logical first choice for the statistical analysis of such data are permutation tests with their minimal assumptions [5, 6].

In contrast to the statistical analysis of fMRI signals, the statistical analysis of deformation fields requires handling vector data instead of scalars. Furthermore, in genetic research, usually transgenic mouse with the same genetic background (apart from the gene of interest) are compared, resulting in populations with low intra-variability[3], but often also very subtle inter-group differences, thus requiring a highly sensitive test. Chen et al. presented a test statistic using the standard deviation of the lengths of deformation vectors, for which different mouse strains (129S1, SvImJ, C57/B16 and CD1) were subjected to permutation tests [7]. However, since they were analyzing the complete brain, they had to limit their tests to 500 permutations because of time considerations. In mouse strains for which there result large inter-group differences, this might be sufficient. The minimal p-value that can be resolved with 500 permutations is $\frac{1}{500+1}$, without correcting for multiple comparisons. Obviously, a larger number of permutations is required to resolve smaller significance probabilities. Furthermore, by using only the lengths of the deformation vectors, valuable information, in particular that encoded in the directionality, is lost. A first step to improve on this situation is the use of Hotelling's t^2-test with voxelwise estimated covariances of the vector field. This test statistic can also be used in the setting of a permutation test as in [6]. Unfortunately permutation tests are highly computationally expensive, even taking into account algorithmic improvements [8], so alternatives for the permutation tests have been considered: among others, the Brunner - Munzel test[9].

The goal of this work is to show the applicability of the 3D Moore-Rayleigh test for the quantitative groupwise comparison of images, and to propose the Moore-Rayleigh test as an alternative to permutation testing. The 2D case of this test was first introduced by Moore [11] who numerically obtained critical values. We generalized his idea to k dimensions and determined the closed form of the density and the distribution function for three dimensional vector data. Because its test statistic is available in closed form, the test does not require much computational effort. We evaluate it empirically with simulated clinical data of known ground-truth and compare it to the performance of a permutation test and a variant of the Mann-Whitney test. Furthermore, we show that the Moore-Rayleigh test outperforms permutation testing on account of sensitivity. For the compactness of this paper, in Section 2 we introduce the Moore-Rayleigh test only for the three-dimensional case and describe how this method is applied in the two-sample problem. In Section 3 experimental results on simulated data are given and the method is compared to currently popular methods. Finally, we conclude the paper with a summary and discussion in Section 4.

2 3D Moore-Rayleigh Test

We consider a finite sample of N real-valued vectors $X = (X_1, ..., X_N)$. For the application to deformation fields we only consider vectors in $\mathbb{R}^3$, such that $X_n = (X_{n,1}, X_{n,2}, X_{n,3})$. For the general Moore-Rayleigh test in k dimensions, we refer to our publication [10].

If we assume that the X_n are independently drawn from a common continuous distribution, the null-hypothesis is that the probability density $f : \mathbb{R}^3 \to [0, \infty)$ is spherically symmetric. This implies that the density f factors into the product of a radial probability density $p_r : [0, \infty) \to [0, \infty)$ and the uniform distribution on each hypersphere $r\,S^2 = \{x \in \mathbb{R}^3 \mid \|x\| = r\}$, such that $f(x) = p_r(\|x\|)/\mathrm{vol}(rS^2)$. The random sum $\sum_{n=1}^{N} X_n$ represents a random flight with N steps whose lengths are distributed according to p_r. In the one dimensional case (not discussed here), this sum corresponds to a random walk.

To render the test nonparametric, the vectors are scaled by the rank of their lengths:

$$S_N = \sum_{n=1}^{N} \frac{nX_{(n)}}{\|X_{(n)}\|}, \tag{1}$$

where $X_{(n)}$ is the nth largest vector in the sample. The distribution of S_N is now independent of the actual p_r. Note that $\sum_{n=1}^{N} \frac{X_{(n)}}{\|X_{(n)}\|}$ is a Rayleigh random flight [12], and our S_N is a Rayleigh random flight with increasing steps. The addition of the vectors incorporates the directionality information and by weighting the vectors by their ranks, also the vector magnitude influences the test statistic.

The test statistic of interest R_N^* is then obtained by scaling S_N by $N\sqrt{N}$ for asymptotic simplicity:

$$R_N^* = \frac{S_N}{N^{3/2}} \tag{2}$$

Let $\alpha_N = N^{3/2}$. The distribution function of $R_N = \alpha_N R_N^*$ in 3 dimensions is given by:

$$\mathrm{pr}(R_N = r) = \frac{2r}{\pi} \int_0^{\infty} t \frac{\sin rt/\alpha_N}{r} \prod_{n=1}^{N} \frac{\sin nt}{nt} \mathrm{d}t \tag{3}$$

This function can be derived by way of characteristic functions [12].

Asymptotically the distribution of R_N^* approaches a χ_3^2 distribution. Of course, for small values of N, the exact form of $\mathrm{pr}(R_N = r)$ should be used whenever possible. As shown in [10] the oscillating integral in eq. (3) can be evaluated in the form of a finite series.

2.1 The Two-Sample Problem

In the two sample problem, we are given two vector-valued random variables $X = (X_1, ..., X_N)$ and $Y = (Y_1, ..., Y_N)$. Under the null hypothesis that X_i and Y_j are identically and independently distributed according to a common probability density $f : \mathbb{R}^3 \to [0, \infty)$, the differences $X_i - Y_j$ are distributed according to the symmetrizing convolution $f * (-f)$, whose density is given by

$$\mathrm{pr}(X - Y = x) = \int \mathrm{pr}(X = u)\mathrm{pr}(Y = u + x)\,\mathrm{d}u \tag{4}$$

Under the null hypothesis that f is spherically symmetric around its mean, significance probabilities can be calculated from eq.(3). If it is assumed that X is distributed according to a multivariate normal distribution, the use of the Moore-Rayleigh test is justified. However the distribution of $g = f * (-f)$ is in general only symmetric, i.e. $g(x) = g(-x)$ for all $x \in \mathbb{R}^3$ and therefore the Moore-Rayleigh test is only approximately valid.

Of course, one would prefer to test conservatively for mere symmetry, but the available tests are either only asymptotically nonparametric or require further randomization of the underlying distribution [13–18].

Therefore, we suggest to use the Moore-Rayleigh test, but to bootstrap the empirical distributions of the two samples X and Y by random sampling without replacement (to avoid degeneracy issues when two or more vector differences are equal) and to compare the mean of R_N^* obtained under M such samples with eq. (3). In theory, the bootstrapping reduces the error made when the assumptions of the Moore-Rayleigh test are only approximately fulfilled by a factor of almost $1/\sqrt{M}$, although these considerations are beyond the scope of this paper. Here, the properties of the test so obtained are evaluated by computer experiments.

2.2 Clinical Interpretation

Consider two sets of 3D MR images taken from different populations of equal size. The first step in the analysis is the affine registration to an atlas A. This normalization step brings all images to the same coordinate system and removes all non-specific anatomical differences, like global orientation and the scale parameters. From now on, we consider only the normalized sets of images $I = (I_1, ..., I_N)$ and $J = (J_1, ..., J_N)$.

A non-rigid registration defines the relation between the average and an image I, which is found by the minimum of a similarity measure ρ:

$$\mathcal{T}_I = \min \rho(I, A) \tag{5}$$

Assuming that the similarity measure returns the best approximation of the unknown relation between I and A, $\mathcal{T}_I$ indicates the local anatomical differences between I and A, which are coded by vectors in $\mathbb{R}^3$.

Non-rigidly registering I and J to the atlas results in two sets of deformation fields $\mathcal{T}_I = (\mathcal{T}_{I1}, ..., \mathcal{T}_{IN})$ and $\mathcal{T}_J = (\mathcal{T}_{J1}, ..., \mathcal{T}_{JN})$. Each anatomically homologous point h in $\mathcal{T}_I$ and $\mathcal{T}_J$ can then be subjected to the two sample Moore-Rayleigh test, as described in Section 2.1.

The resulting p-value in a point h indicates the probability that an observed difference between groups I and J occurs by chance. A small p-value is an indication that there is a difference between I and J. It can occur that both groups I and J are significantly different from A, but as long as the difference of I with A is similar to the difference of J with A, the Moore-Rayleigh test will not return a significant difference between I and J.

3 Validation

3.1 Image Formation

For the validation of the Moore-Rayleigh test an average MR volume of the C57Bl6/Jico mouse brain was used, which was cropped to a volume of 50 x 50 x 80 voxels due to considerations of running time. This subvolume included the ventricles, thalamus, and several fiber tracts, as illustrated in Figure 1(a).

Individual subjects ($n = 30$) were simulated by the introduction of artificial, spherically shaped deformations. Each such local deformation is completely characterized by a center point c and a radius r. The length of each deformation vector in the sphere takes a maximum of $\frac{1}{2}r$ at the center and drops radially and linearly until the edge of the sphere, where the deformation is zero to ensure continuity (in fact, smoothness) with the surrounding field. By varying the parameters c and r two groups (G_1 and G_2) were generated with 15 individuals each.

The goal of this numerical experiment was to test whether and under which conditions the Moore-Rayleigh test picks up group differences (the inter-group variation) only or whether also inter-group variation is detected. For this reason two spherical deformations were used in each subject, which are referred to as *sphere 1* (S_1) and *sphere 2* (S_2) and which are shown in Figure 1(b). The average radius and center of S_1 was taken to be identical for both groups, but in S_2 a systematic difference of 5 voxels in the average radius ($\Delta r = 5$) between G_1 and G_2 was introduced, the average center was kept constant. Intra-group variation was simulated in both groups by randomly adding small variations in the center point and uniform radius of the spheres ($r \pm 5$ and $c \pm 2.5$) from an uniform distribution. The values for the inter-group and intra-group variability, are corresponding to the ones described in the literature[3, 7].

After the creation of the individual subjects, Gaussian noise (μ=0, sd=300) was superimposed on each image. The result is shown in Figure 1(c) for one subject. A spherical mask was created (Figure 1(d)) that indicates the average locations of the spherical deformations S_1 and S_2, and which is later used for validation purposes. Figure 1(d) shows the average locations of S_1 (lower sphere) and S_2 (upper sphere), where in S_2 also the differences between the two groups (G_1 and G_2) is shown.

The non-rigid registration of the 30 subjects to the average was performed using the symmetric demons algorithm [19], as implemented in ITK [20]. This registration algorithm uses the mean squared difference with a smoothing factor of 1.0 and was performed with 60 iterations. The output of the non-rigid registration is a deformation field, i.e. vectors in $\mathbb{R}^3$ that represent the geometric translations of voxels to their corresponding points in the atlas image.

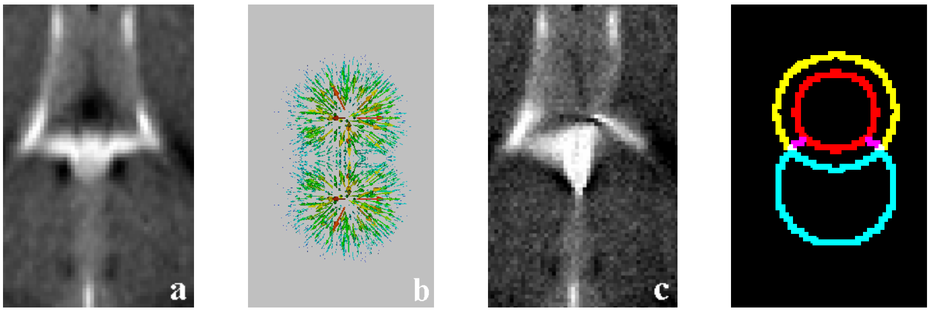

Fig. 1. The various stages during the creation of one of the synthetic images: The midsection of the average image (a), which is deformed by two spheres (b), and the final image after adding gaussian noise (c). For each dataset a sphere mask is created (d) which indicates the average locations of the spherical deformations S_1 and S_2 for G_1 (outer sphere) and G_2 (inner sphere).

3.2 Validation Method

The output of the two-sample Moore-Rayleigh test is a probability image with a p-value per voxel that indicates how likely it is that the null hypothesis holds at that particular point, thus indicating how likely it is that an observed difference between groups I and J occurs by chance in that voxel. Thresholding the probability image with a critical value α results in a binary image that shows the regions where a significant group difference has been detected. For the synthetically generated data the locations are known where there exists a simulated structural difference between the two groups; these are all the voxels lying in S_2 (Figure 1(d)) and this knowledge is taken as ground-truth for the assessment of the statistical tests. We use this information to quantify the performance of the Moore-Rayleigh test by calculating the following:

True positives (TP) Amount of voxels found significant inside S_2
False positives (FP) Amount of voxels found significant outside S_2
false negatives (FN) Amount of voxels found not significant inside S_2
true negatives (TN) Amount of voxels found not significant outside S_2

For the various tests, we report the sensitivity and specificity for α=0.05. The sensitivity is the ratio of significant voxels which are detected correctly and specificity is the ratio of not significant voxels which are detected correctly.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

From these two measures we also calculated Receiver Operator Curves (ROCs) that show the dependence of sensitivity and specificity on the critical values $0.0 \leq \alpha \leq 1.0$.

3.3 Comparison of Test Statistic

To compare the performance of the Moore-Rayleigh test to other nonparametric methods, we implemented a permutation test (m=10.000 labellings) with Hotelling's t^2 as test statistic as in [6]. Furthermore, we also implemented the Mann–Whitney test [21], which is the nonparametric equivalent of the t-test, on a rank one approximation (see Appendix A for details).

A visualization of the significant voxels for the three test statistics are given in Figure 2. By visual inspection, better classification results were obtained using the Moore-Rayleigh test. The performance of the three test statistics are given in Figure 3. The sensitivity and the specificity of the cut-off value of $\alpha = 0.05$ is given for the three test statistics in Table 1. The relatively low sensitivity of the permutation tests, might be increased if more permutations were used. Furthermore, the sensitivity of the Mann-Whitney test is comparable with the Moore-Rayleigh test, where the Moore-Rayleigh test outperforms the Mann-Whitney on specificity. Furthermore, the presented p-values are not corrected for multi testing[23]. If multitest correction would have been applied, the Mann-Whitney test would have shown no significance, since the Mann-Whitney test classified voxels significantly different with a probability between 0.01 and 0.05, whereas the Moore-Rayleigh test indicates significance with a probability of between 10^{-2} and 10^{-10}.

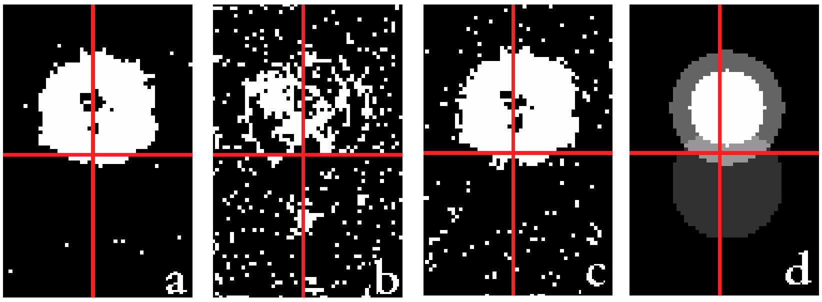

Fig. 2. The classification result with cut-off value $\alpha = 0.05$ for the Moore-Rayleigh test (a), the permutation test (b), and the Mann-Whitney test (c) as compared to the ground truth(d)

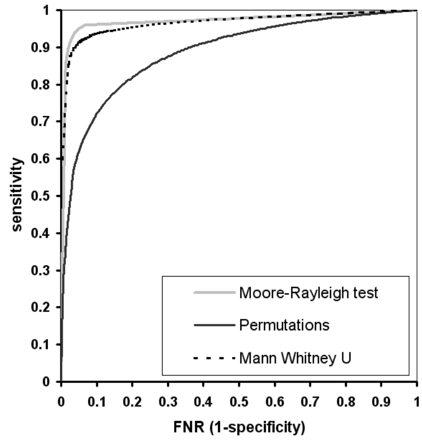

Fig. 3. The ROC of the different nonparametric test statistics for the quantitative analysis of deformation fields.

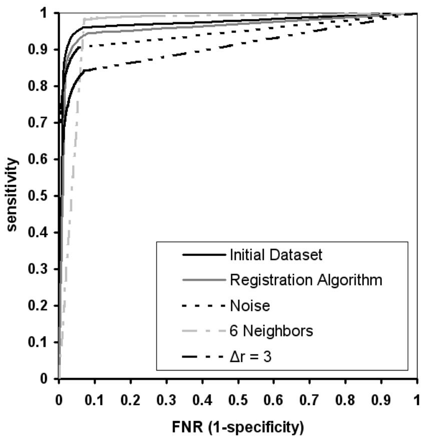

Fig. 4. The ROC of the different parameter settings for the robustness testing of the Moore-Rayleigh test

3.4 Robustness Testing

Several parameters have been varied to investigate the robustness of the Moore-Rayleigh test under various conditions. The dataset as described in 3.1 was used as input and for each experiment only one parameter was changed to measure its effect. The several tests are described below:

Registration Algorithm Instead of using the Demons non-rigid registration method, a B-Spline transform[24] was applied, as implemented in elastix[25]. Since the Moore-Rayleigh test is a voxel based test, the gridspacing was set at 2.5 voxels with a 3-level pyramid registration. Furthermore, the mutual information metric was used.

Noise Noise is known for having a major influence on the quality of the images and their postprocessing. The influence of noise on the hypothesis testing is quantified by increasing the standard deviation from 300 (used in Section 3.1) to a value of 700.

6 Neighbors With the hope of increasing the sensitivity of the test and decreasing the influence of noise, the deformation vectors of the six closest neighbors (two in each coordinate dimension) for each voxel z are pooled.

Δr In this test, 15 subjects of one group are generated with a systematic difference in radius of 3 voxels between G_1 and G_2 (for S_2), instead of 5 voxels as in Section 3.1. The intra-group variation is kept constant at ±5 voxels as before.

The ROCs of the Moore-Rayleigh test for these various settings are shown in Figure 4. The sensitivity and the specificity for the cut-off value $\alpha = 0.05$ are additionally given in Table 2. The performance of the Moore-Rayleigh test is seen to be least influenced by the change in the nonlinear registration method.

	Moore-Rayleigh test	Permutation test	Mann-Whitney test
sensitivity	0.91	0.73	0.41
specificity	0.97	0.89	0.99

Table 1. Sensitivity and specificity for the Moore-Rayleigh test, the permutation test and the Mann-Whitney test ($\alpha = 0.05$ each).

	Moore-Rayleigh test	Registration Algorithm	Noise	Pooling	Δr
sensitivity	0.91	0.88	0.77	0.99	0.70
specificity	0.97	0.97	0.99	0.90	0.98

Table 2. The sensitivity and specificity ($\alpha = 0.05$) for the different parameter settings for the robustness testing of the Moore-Rayleigh test.

However, the Moore-Rayleigh test is quite sensitive on the influence of noise, loosing sensitivity when the noise level is increased. Pooling neighboring voxels, however, increases the sensitivity, with an accompanying loss in specificity due to the smoothing effect this introduces. As can also be seen in Figure 4, a decrease in inter-group variation (scenario Δr) decreases the sensitivity. This is well explained, as the intra-group variation was kept constant, while the inter-group variation was decreased to only 3 voxels.

4 Discussion

In this paper we presented a novel nonparametric statistical method to detect and quantify anatomical differences between groups of MR images. Our method is based on (a generalization of)the nonparametrical Moore-Rayleigh test, which tests for spherical symmetry in vector data. This method uses as input the deformation fields which are obtained by the non-rigid registration of all subjects to a common reference image. Under the assumption that no registration errors were made in the creation of the deformation fields, significant anatomical differences between the two groups can be assessed.

Permutation tests are not routinely applicable in a clinical setting because of the large number of permutations required. In clinical practice the number of labellings is often reduced to speed up the analysis, which reduces the power of the permutation test severely and makes it almost impossible to correct for multiple comparisons. The method presented here, on the other hand, is computationally fast and offers an interesting alternative to permutation tests. Although the null hypothesis of mere symmetry is not tested, i.e. the Moore-Rayleigh test is only approximately valid, the results are quite convincing, as shown in Section 3.3.

A further advantage of the Moore-Rayleigh test is that it is completely nonparametric and needs no assumptions on the underlying dataset. As the test statistic is continuous, the significance probabilities can be very low (up to 10^{-18} or less is numerically possible), so the Moore-Rayleigh test also results in significant voxels under correction for multiple comparisons (not shown). It was also found to be relatively unaffected by the (non-rigid) registration method used.

The sensitivity was most influenced by the decrease of the inter-variation to 3 voxels, but considering that the intra-variation was 5 voxels and that still more than 70 percent of all voxels assessed as significant were true positives, it can be concluded that the Moore-Rayleigh test is able to detect small differences between groups. Furthermore, it was expected and indeed observed that noise, which affects the registration algorithm, results in an increase of false positives. As the results in Section 3.4 indicate, this problem could be addressed by pooling the deformation vectors from neighboring voxels. Although this would result in a loss of specificity, it is plausible that the Moore-Rayleigh test would then be more robust to the effect of noise.

One important topic of our future work concerns the evaluation of this algorithm on real clinical data. For now, we have only assessed the Moore-Rayleigh test on simulated images, with spherically deformations. Although, it can be argued that these deformations are representative for structural deformations, as brain atrophy. Furthermore, the main goal of this paper is to show the performance of the Moore-Rayleigh test on the quantification of inter-group variability. This performance can only be validated on a dataset with known inter-group and intra-group variability. Therefore, to stay close to real data, the variabilities from the synthetic dataset are simulated according to the descriptions of variations of real mouse brain MRI data[3, 7].

Finally, we would like to encourage the reader to apply this method on their own data. Therefore, the code is made publicly available and can be obtained by sending an e-mail to the corresponding author.

Acknowledgments. This work was supported in part by funds from CYTTRON within the BSIK program (Besluit subsidies investeringen kennisinfrastructuur) and has been performed within the Molecular Imaging Laboratories Leiden (MILL). Furthermore, the authors would like to thank Boudewijn Lelieveldt and Luca Ferrarini for their useful comments on the manuscript and method development.

References

1. Driehuys. B., Nouls, J., Badea, A., Bucholz, E., Ghaghada, K., Petiet, A., Hedlund, L.W.: Small animal imaging with magnetic resonance microscopy. ILAR Journal **49** 35–53 (2008)
2. Benveniste, H., Blackband, S.: MR microscopy and high resolution small animal MRI: applications in neuroscience research. Progress in Neurobiology **67** 393–420 (2002)
3. Kovacević, N., Henderson, J.T., Chan, E., Lifshitz, N., Bishop, J., Evans, A.C., Henkelman, R.M., Chen, X.J.: A Three-dimensional MRI Atlas of the Mouse Brain with Estimates of the Average and Variability. Cerebral Cortex **15** 639–645 (2005)
4. Pohl, K.M., Fisher, J., Bouix, S., Shenton, M., McCarley, R.W., Grimson, W.E.L., Ron Kikinis, R., Wells W.M.: Using the logarithm of odds to define a vector space on probabilistic atlases. Medical Image Analysis **11** 465-477 (2007)

5. Nichols, T.E. and Holmes, A.P.: Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples. Human Brain Mapping **15** 1-25(2001)
6. Ferrarini, L., Palm, W.M., Olofson, H., van Buchem, M.A., Reiber,J.H.C., Admiraal-Behloul, F.: Shape differences of the brain ventricles in Alzheimer's disease. NeuroImage **32** 1060–1069 (2006)
7. Chen, X.J., Kovacevic, N., Lobaugh, N.J., Sled, J.G., Henkelman,R.M., Henderson JT.: Neuroanatomical differences between mouse strains as shown by high-resolution 3D MRI. Neuroimage. **29** 99–105 (2005)
8. Heckel, D., Arndt, S., Cizadlo, T., Andreasen, N.C.: An efficient procedure for permutation tests in imaging research. Computers and biomedical research **31** 164–171 (1998)
9. Rorden, C., Bonilha, L., Nichols, T.E.: Rank-Order versus mean based statistics for neuroimaging. NeuroImage **35** 1531–1537 (2007)
10. *Removed for anonymous reviewing.*
11. Moore, B.R.: A modification of the Rayleigh test for vector data. Biometrika **67** 175–180 (1980)
12. Dutka, J.: On the problem of random flights. Archive for History of Exact Sciences **32** 351–375 (1985)
13. Aki, S.: On nonparametric tests for symmetry in $\mathcal{R}^m$. Annals of the Institute of Statistical Mathematics **45** 787–800 (1993)
14. Ngatchou-Wandji, J.: Testing for symmetry in multivariate distributions. Statistical Methodology *in press*
15. Henze, N. Klar, B., Meintanis, S.G.: Invariant tests for symmetry about an unspecified point based on the empirical characteristic function. Journal of Multivariate Analysis **87** 275–297 (2003)
16. Fernández, V.A., Gamero, M.D.J., García, J.M.: A test for the two-sample problem based on empirical characteristic functions. Computational Statistics & Data Analysis **52** 3730–3748 (2008)
17. Diks, C., Tong, H.: A test for symmetries of multivariate probability distributions. Biometrika **86** 605–614(1999)
18. Jupp, P.E.: A nonparametric correlation coefficient and a two-sample test for random vectors or directions. Biometrika **74** 887–890 (1987).
19. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Medical Image Analysis **2** 243–260 (1998)
20. Yoo, T.S., Ackerman, M.J., Lorensen,W.E., Schroeder,W., Chalana, V., Aylward, S., Metaxes, D., Whitaker, R.: Engineering and Algorithm Design for an Image Processing API: A Technical Report on ITK - The Insight Toolkit. In: Proc. of Medicine Meets Virtual Reality 586–592 (2002)
21. Mann, H.B. Whitney, D.R.: On a test of whether one of 2 random variables is stochastically larger than the other. The Annals of Mathematical Statistics. **18** 50-60 (1947)
22. Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioural Sciences (2nd edition). New York: McGraw-Hill ISBN: 978-00-7057-357-4 (1988)
23. Shaffer, J.P.: Multiple Hypothesis Testing. Annual Review of Psychology **46**, 561–584 (1995)
24. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: Application to breast MR images. IEEE Transactions on Medical Imaging **18** 712-721 (1999)
25. Staring, M.: Intrasubject Registration for Change Analysis in Medical Imaging. PhD thesis, Utrecht University, The Netherlands ISBN 978-90-8891-063-0 (2008) `http://elastix.isi.uu.nl`

A Mann-Whitney test

The Mann Whitney test is the nonparametric equivalent to the t–test. It makes the following assumptions:

1. The two samples are randomly and independently drawn from the same underlying distribution.
2. The dependent variable is continuous.
3. The values of the dependent variable are at least ordinal.

As item 3 is stating, the vectors for a point need to be ordered, based on a measure of a continuous scale (item 2). To order vectors in R^3, a rank–1 approximation has been performed: for each point of the average image, a covariance matrix Σ is calculated based on all subjects in the two groups under consideration. Using the singular value decomposition of this covariance matrix (principal component analysis), the eigenvectors V and eigenvalues Λ are obtained and represent the principal modes of variation. $\Sigma = U \Lambda V$,

The first mode of variation V_1, corresponding to the largest eigenvalue $\Lambda_{1,1}$, represents the direction of largest variance between the vectors considered for that particular point in the average. Projecting the vectors on this direction results in vectors all pointing in the same direction, and their lengths are then used in the usual Mann-Whitney test. A disadvantage of this test is that only the first principal mode of the covariance matrix is used, and therefore only partial information on the orientation of the vectors is used, decreasing the power of the test.