MULTIVARIATE VARIANCE-COMPONENTS ANALYSIS IN DTI

Agatha D. Lee[1], Natasha Leporé[1,5], Jan de Leeuw[2], Caroline C. Brun[1], Marina Barysheva[1], Katie L. McMahon[2], Greig I. de Zubicaray[3], Nicholas G. Martin[4], Margaret J. Wright[4], Paul M. Thompson[1]

[1]Laboratory of Neuro Imaging, Department of Neurology, UCLA School of Medicine, Los Angeles, CA, USA
[2]Department of Statistics, UCLA, Los Angeles, CA, USA
[3]Functional MRI Laboratory, Centre for Magnetic Resonance, University of Queensland, Brisbane, Australia
[4]Queensland Institute of Medical Research, Brisbane, Australia
[5]Department of Radiology, Children's Hospital Los Angeles, University of Southern California, Los Angeles, CA, USA

ABSTRACT

Twin studies are a major research direction in imaging genetics, a new field, which combines algorithms from quantitative genetics and neuroimaging to assess genetic effects on the brain. In twin imaging studies, it is common to estimate the intraclass correlation (ICC), which measures the resemblance between twin pairs for a given phenotype. In this paper, we extend the commonly used Pearson correlation to a more appropriate definition, which uses restricted maximum likelihood methods (REML). We computed proportion of phenotypic variance due to additive (A) genetic factors, common (C) and unique (E) environmental factors using a new definition of the variance components in the diffusion tensor-valued signals. We applied our analysis to a dataset of Diffusion Tensor Images (DTI) from 25 identical and 25 fraternal twin pairs. Differences between the REML and Pearson estimators were plotted for different sample sizes, showing that the REML approach avoids severe biases when samples are smaller. Measures of genetic effects were computed for scalar and multivariate diffusion tensor derived measures including the geodesic anisotropy (tGA) and the full diffusion tensors (DT), revealing voxel-wise genetic contributions to brain fiber microstructure.

Key Words—multivariate statistics, DTI, twin studies, genetics

1. INTRODUCTION

In recent years, twin studies have become increasingly popular in cognitive neuroscience and medical imaging as a means to determine genetic influences on brain structure and function, the ultimate goal being the discovery of specific genes that influence brain development and disease.

To estimate the genetic and environmental contributions to a phenotype, several models have been established that compare correlations in monozygotic twins (MZs, who share 100 % of their genes) to those found between dizygotic twins (DZs, who share 50 % of their genes on average). One common measure is the ANOVA-based (Pearson) intraclass correlation (ICC), which quantifies the resemblance within twin pairs. From these ICC measures for MZ and DZ pairs, a simple heritability estimate (Falconer's estimate) can be computed [16]. Another more recent model is the A/C/E model, which uses information from both twin types to distinguish sources of variance that are attributable to additive genetic factors (A), common environment (C) and environmental factors unique to each individual (E) [16].

Many prior studies examined genetic influences on aspects of brain anatomy such as regional gray and white matter volumes [4], fiber structure [5], [13], and cortical thickness [23]. Some subcortical structures, such as the thalamus and the basal ganglia, were found to be jointly influenced by a single genetic factor in a large-scale pediatric twin study [21]. Even so, multidimensional signals, such as diffusion tensors, have not been thoroughly studied.

Diffusion tensor imaging (DTI) offers a means to understand the genetics of brain fiber architecture. DTI measures the multidirectional profile of water diffusion in tissue. This method provides vital information on brain architecture and composition; the resulting estimates of fiber integrity are correlated with intellectual performance [5]. Higher-dimensional models of the diffusion signal have been proposed, although most DTI studies compute and analyze a diffusion tensor (DT) at each voxel whose eigenvectors represent the three orthogonal principal directions of the diffusion, and its eigenvalues represent the magnitude of diffusion along these axes. Several tractography methods estimate white matter connectivity from the principal eigenvector field, or orientation distribution function (ODF)-based analogs [14].

To better understand fiber characteristics, several scalar and multivariate measures may be derived from the DTs. Fractional anisotropy (FA), mean diffusivity (MD), and the tensor's eigenvalues are among the most common measures. More recently, several more sophisticated quantities have been used; the geodesic anisotropy (GA) [3], [14] measures the geodesic distance between tensors on the positive-definite symmetric tensor manifold. In a DTI study of blind subjects [12], we also found that a multivariate statistical analysis of the full diffusion tensor outperformed derived scalar signals in detecting group differences.

In prior work [13], we published a new multivariate formulation of the ICC to assess correlation between twin pairs using a distance on the tensor manifold, and computed A/C/E models using this distance on diffusion tensors. Although our previous methods to compute ICC were correct for large sample sizes, they were biased for small sample sizes, and the level of bias was unknown. Here, we present an unbiased negative log-likelihood based multivariate variance component model to estimate resemblances between the MZ and DZ twin pairs. From these multivariate variance components, we compute unbiased estimates of genetic (A) and

environmental (C, E) influences on multidimensional signals, in this case DTI.

The first step in identifying specific genes influencing brain structure is to search for heritable measures in images [7]. We therefore computed 3D maps of genetic and environmental effects from imaging data from 100 healthy young adult twins. We performed all statistical computations in the Log-Euclidean framework [2] because the full DT does not form a vector subspace of the vector space of 3x3 matrices, as the matrices must always be positive-definite and symmetric. The Log-Euclidean framework allows for simple computations on the DT manifold.

2. METHODS

2.1. Intraclass correlation for univariate measures

The standard approach to measure the resemblance between twin pairs is to use the intraclass correlation (ICC):

$$ICC_{uni} = \frac{MS_{between} - MS_{within}}{MS_{between} + MS_{within}}, \quad (1)$$

where $MS_{between}$ and MS_{within} are the mean square differences between pairs and within pairs. When a small sample size is used, ICC values, computed from eq (1), may be negative due to the variability in the sample. Adding more twin pairs to the study will not affect the within-pair variance, however, it will affect the distribution of the means if there are differences between twin pairs. With increased numbers of twin pairs, the estimated ICC becomes positive if the trait is heritable.

For this study, we use the restricted maximum likelihood (REML) method, which gives an unbiased ICC estimate. We define x_1 and x_2 as scalar values of measures in twin 1 and twin 2. N is the total number of pairs. This labeling is interchangeable as there is no "first" and no "second" member of the pair. We also give a set of transformed variables: $\underline{a}_i = \frac{k_i'(\underline{x}_1 - \underline{x}_2)}{\sqrt{2}}$, $\underline{b}_i = \frac{k'(\underline{x}_1 + \underline{x}_2)}{\sqrt{2}}$ and $\underline{c} = \frac{k_0'(\underline{x}_1 - \underline{x}_2)}{\sqrt{2}}$ where k_0 is a vector and $k_1,...,k_{N-1}$ are vectors of length one, orthogonal to each other and orthogonal to k_0. The variance of $\underline{a}_i$ and $\underline{c}$ are $\sigma^2 - \omega^2$ and $\underline{b}_i$ is $(\sigma^2 - \omega^2)$. Thus for REML, $\frac{1}{2}\overline{(\sigma^2 + \omega^2)} = \frac{1}{2(N-1)}\sum_{i=1}^{N-1} b_i^2 = \frac{1}{N-1} MS_{between}$ while $\frac{1}{2}\overline{(\sigma^2 - \omega^2)} = \frac{1}{2N}(c^2 \sum_{i=1}^{N-1} a_i^2) = \frac{1}{N} MS_{within}$. The non-negative REML formula to estimate ICC from univariate measures is

$$ICC_{uni,REML} = \max\left(0, \frac{\frac{N}{N-1} MS_{between} - MS_{within}}{\frac{N}{N-1} MS_{between} + MS_{within}}\right) \quad (2)$$

2.2. Intraclass correlation for multivariate measures

Eq. (1) and (2) are univariate formulations, and their extension to multivariate data is not straightforward. In [12,13], we applied a multivariate version of equation (1) to the genetic analysis of a dataset of 92 twins. Here we set out to generalize Eq. (2) to use REML.

We first start by briefly describing the multivariate generalization of Eq. (1). We define t_1 and t_2 as 6-dimensional random vectors, which represent each subject's deviation from the mean of the overall sample [8]. In the example presented in this work, t_1 and t_2 are both 6-dimensional vectors defined at each voxel, containing the deviation of the DTs of twin 1 and twin 2, respectively from the mean diffusion tensor of the sample (after nonlinear image registration [13]). The multivariate intraclass correlation matrix [18] is defined as follows:

$$\Gamma = \Sigma^{-1/2} \Omega \Sigma^{-1/2}, \quad (3)$$

where Σ, Ω are the expected values of $(t_1 - \mu)(t_1 - \mu)^T$ and $(t_2 - \mu)(t_1 - \mu)^T$ respectively. Here, μ is the sample mean of all of the t_1 and t_2 vectors. This implies that $\Sigma - \Omega$, which basically estimates the within pair correlation, is positive semi-definite. The maximum eigenvalue of this ICC matrix Γ is considered to be the multivariate ICC value.

We define two new random vectors $u = \frac{1}{2}(t_1 - t_2)$ and $v = \frac{1}{2}(t_1 + t_2)$. If we have N independent realizations (u_i, v_i) of (u, v), then the deviance (twice the negative log-likelihood, except for irrelevant constants) is [19, 22]:

$$D(\Delta, \Xi, \mu) = n\log(\det(\Delta)) + n\log(\det(\Xi)) + \sum_{i=1}^{N} u_i^T \Delta^{-1} u_i + \sum_{i=1}^{N} (v_i - \mu)^T \Xi^{-1} (v_i - \mu), \quad (4)$$

where $\Delta = \frac{1}{2}(\Sigma - \Omega)$, the within-pair variance, and $\Xi = \frac{1}{2}(\Sigma + \Omega)$, the between-pair variance. Thus, the maximum likelihood estimate of μ is $\hat{\mu} = \frac{1}{2}(\bar{t}_1 + \bar{t}_2)$, where $\bar{t}$ is mean of t derived from the full sample, i.e., the mean of all 2N vectors.

When defining $G = \frac{1}{N}\sum_{i=1}^{N} u_i u^T$ and $H = \frac{1}{N}\sum_{i=1}^{N} \bar{v}_i \bar{v}^T$, the concentrated negative log-likelihood of eq. (4) is :

$$D_*(\Delta, \Xi) \overset{\Delta}{=} \min_{\mu} D(\Delta, \Xi, \mu) = n[\log(\det(\Delta)) + \log(\det(\Xi)) + tr\Delta^{-1}G + tr\Xi^{-1}H], \quad (5)$$

Maximum likelihood estimates of the variance components can be computed as

$$\hat{\Sigma} = \hat{\Xi} + \hat{\Delta} = H + G$$
$$\hat{\Omega} = \hat{\Xi} - \hat{\Delta} = H - G \quad (6)$$

Since $E(G) = \frac{1}{2}(\Sigma - \Omega)$ and $E(H) = \frac{1}{2}\frac{N-1}{N}(\Sigma + \Omega)$, we can compute unbiased estimates using:

$$\hat{\Sigma} = \frac{N}{N-1}H + G \qquad (7)$$

$$\hat{\Omega} = \frac{N}{N-1}H - G$$

These unbiased estimates are also the REML estimates.

2.3. Data and Preprocessing

2.3.1. Participant description and image acquisition

We acquired 3D structural brain MRI scans and DT-MRI scans from 100 subjects: 25 pairs of MZ twins (25.1±1.5SD years old) and 25 pairs of DZ twins (23.1±2.1 years) on a 4T Bruker Medspec MRI scanner with an optimized diffusion tensor sequence [6]. Imaging parameters were: 21 axial slices (5 mm thick), FOV = 23 cm, TR/TE 6090/91.7 ms, 0.5 mm gap, with a 128×100 acquisition matrix. 30 gradients were applied: three scans with no diffusion sensitization and 27 diffusion-weighted images for which gradient directions were evenly distributed on the hemisphere [10]. The reconstruction matrix was 128×128, yielding a 1.8x1.8 mm^2 in-plane resolution. Total scan time was 3.05 minutes.

2.3.2 Image Preprocessing and Registration

sMRI images were automatically skull-stripped using the Brain Surface Extraction software (BSE) [20] followed by manual editing. Each masked image was registered via 9-parameter linear transformation to a high-resolution single-participant brain template image, the Colin27 template, using the FLIRT software [9]. Linearly registered sMRI images were then registered to a Mean Deformation Template (MDT; created from the dataset) using a 3D fluid registration [11,15]. From the resulting deformation fields, Jacobian matrices were obtained.

From the DICOM DT-MR images, diffusion tensors (3x3 positive symmetric matrices) were computed and smoothed using Log-Euclidean tensor denoising to eliminate singular, negative definite, or rank-deficient tensors, using MedINRIA (http://www.sop.inria.fr/asclepios/software/MedINRIA).

Extracerebral tissues were manually deleted from one of the diagonal component images (D_{xx}), yielding a binary brain extraction mask (cerebellum included). Masked tensor images were registered by 9-parameter transformation to the corresponding sMRI images in the standard template space using FLIRT software [9].

The tensors at each voxel were rotationally reoriented using transformation parameters from linear and nonlinear registrations [1] to ensure that the multidimensional tensor orientations remained consistent with the anatomy after image transformation [1, 24]. Two separate algorithms are used to compute the tensor rotations: the Finite Strain (FS) and the preservation of principal direction (PPD) algorithms ([1, 24]).

2.4. Scalar Statistics in the Log-Euclidean space

As a scalar statistic to compare to our multivariate measures, we used the GA [13] - the manifold equivalent of the FA computed in the Log-Euclidean framework [2,11]. We renormalized GA by applying the hyperbolic tangent transformation to the GA values (tGA) as in [3], to create maps with a comparable range to the FA.

2.5. Statistical analysis for twins

We computed GA and tGA values as well as the matrix logarithms of the full diffusion tensors for each participant. Two sets of voxel-wise intraclass correlation matrices for the MZ pairs and DZ pairs were computed for all the univariate and multivariate measures detailed above.

The A/C/E model for MZ and DZ twins decomposes variation into genetic (A) and non-genetic (C/E) components. In the simplest case, we have, for the different types of twins

$$\Sigma_{MZ} = \Sigma_{DZ} = \theta_A + \theta_C + \theta_E \qquad (8)$$

$$\Omega_{DZ} = .5\theta_A + \theta_C \qquad (9)$$

$$\Omega_{MZ} = \theta_A + \theta_C \qquad (10)$$

The resulting unbiased estimates are:

$$\hat{\theta}_A = 2(\Omega_{MZ} - \Omega_{DZ}) = 2(\frac{N}{N-1}(H_{MZ} - H_{DZ}) - (G_{MZ} - G_{DZ})) \qquad (11)$$

$$\hat{\theta}_C = 2\Omega_{DZ} - \Omega_{MZ} = \frac{N}{N-1}(2H_{DZ} - H_{MZ}) - (2G_{DZ} - G_{MZ}) \qquad (12)$$

$$\hat{\theta}_E = \Sigma_{MZ} - \Omega_{MZ} = 2G_{MZ} \qquad (13)$$

3. RESULTS

The maximum eigenvalue of the REML multivariate intraclass correlation matrix measures the resemblance between twin pairs. Voxel-wise maps of these resemblances are shown in **Figure 1** (**Figure 1a** for MZ twins and **1b** for DZ twins). As expected, overall maps for the MZ twins display higher maximum eigenvalues than those for DZ twin pairs. **Figure 1c** and **1d** show the comparison between REML-based ICC and ICC with Pearson's correlation (i.e., REML ICC minus Pearson's ICC) for 6 MZ pairs (**1c**: small sample size) and 25 pairs of twins (**1d**: larger sample size). In smaller samples (**Figure 1c)** REML avoids a serious bias in the Pearson formula, although the difference between the estimators eventually tends to zero when samples are large.

Unbiased genetic (A) and shared environmental (C) contributions to brain morphological phenotypes are shown in **Figure 2** for the multivariate full DT. Limbic areas, the corpus callosum and some posterior white matter regions are shown to be under strong genetic control.

4. CONCLUSION

Here we demonstrated how to estimate the intraclass correlation using REML methods for multidimensional signals, and we applied the method to analyze DT images in a dataset of 100 twins. We first transformed the diffusion tensors to the log-Euclidean domain via matrix logarithm transformation. In the resulting space, the multivariate REML variance components

were computed on the diffusion tensor components of the individuals shifted by the mean of the whole sample. The A, C and E variance components in the A/C/E model were then computed using the derived multivariate variance components algorithm.

While we restricted ourselves to the DT in this study, our multivariate REML variance component model can be extended to any multivariate measure defined on a dataset from unordered pairs. In the future, this might include a parameterization of the ODF based on high angular and/or radial resolution *q*-space diffusion imaging. The algorithm to determine the A, C and E components can also be applied to other multidimensional measures in twin data, such as vectors of multiple traits. For example, the Jacobian determinants in tensor-based morphometry are only a scalar summary of the full deformation tensor, and as in [13], one could analyze the morphometric data in twins using a log-Euclidean distance on the associated strain tensors. Multivariate strategies may help to identify heritable measures in high-dimensional brain images, such as HARDI [5] or diffusion spectrum images.

6. ACKNOWLEDGMENTS

This study was supported by NIH grant R01 HD050735 and by NHMRC grant 496682, Australia.

7. REFERENCES

[1] D. C. Alexander et al., "Spatial transformations of diffusion tensor magnetic resonance," *IEEE-TMI*, 20:1131-1139, 2001.
[2] V. Arsigny et al., "Fast and simple calculus on tensors in the log-Euclidean framework," in *Int Conf Med Image Comput Comput Assist Interv*. 8:115-22, *MICCAI* 2005.
[3] P. Batchelor et al., "A rigorous framework for diffusion tensor calculus", *Magn Reson Med*., 53:221–225, 2005.
[4] C.C. Brun et al., "Mapping the Regional Influence of Genetics on Brain Structure Variability - A Tensor-Based Morphometry Study", *NeuroImage*, 2009 Oct 15;48(1):37-49.
[5] M. C. Chiang et al., "Genetics of Brain Fiber Architecture and Intelligence", *Journal of Neuroscience*, 18;29(7):2212-24, 2009.
[6] G. Christensen et al., "*Deformable templates using large deformation kinematics*". IEEE Trans. Image Process. 5 1435–1447, 1996.
[7] D.C. Glahn et al., "*Neuroimaging Endophenotypes: Strategies for Finding Genes Influencing Brain Structure, Human Brain Mapping, Special Issue on Genomic Imaging*." HBM, 28(6):488-501, 2007.
[8] J. Hemelrijk. Underlining Random Variables. *Statistica Neerlandica*, 20:1–7, 1966.
[9] M. Jenkinson and S. Smith, "*A global optimization method for robust affine registration of brain images*," *Med Image Anal*, vol. 5, pp. 143-56, 2001.
[10] D. K. Jones et al., "Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging," *Magn Reson Med* 42: 515-25, 1999.
[11] P. Kochunov et al., "An optimized individual target brain in the Talairach coordinate system, Neuroimage 17: 922–927, 2003.
[12] A. D. Lee et al., "*Brain Differences Visualized in the Blind Using Tensor Manifold Statistics and Diffusion Tensor Imaging*". FBIT 2007: 470-476.
[13] A. D. Lee et al., "Tensor-based analysis of genetic influences on brain integrity using DTI in 100 twins". MICCAI2009, 967-974, Sept 2009.
[14] C. Lenglet et al, "*Mathematical Methods for Diffusion MRI Processing*". NeuroImage, 45(1 Suppl):S111-22, 2009.
[15] N. Lepore et al., "*Multivariate statistics of Jacobian matrices in Tensor Based Morphometry and their application to HIV/AIDS*," MICCAI, 2006, 191-198.
[16] M. C. Neale et al., *Mx: Statistical modeling*, (1999).
[17] A. Pfefferbaum et al., "Genetic regulation of regional microstructure of the corpus callosum in late life," *Neuroreport*, 12: 1677-81, 2001.
[18] C.R. Rao. Familial Correlations or the Multivariate Generalization of the Intraclass Correlation. *Current Science*, 14:66–67, 1945.
[19] D.C. Rao, G.P. Vogler, M. McGue, and J.M. Russell. Maximum Likelihood Estimation of Familial Correlations from Multivariate Quantitative Data on Pedigrees: A General Method and Examples. *American Journal of Human Genetics*, 41:1104–1116, 1987.
[20] D. W. Shattuck, R. M. Leahy. *BrainSuite: an automated cortical surface identification tool*, Medical Image Analysis 8, (202) 129–141, 2001.
[21] J. E. Schmitt et al., "A multivariate analysis of neuroanatomic relationships in a genetically informative pediatric sample." NeuroImage 35(1):70-82, March 2007.
[22] M.S. Srivastava, K.J. Keen, and R.S. Katapa. Estimation of Interclass and Intraclass Correlations in Multivariate Familial Data. *Biometrics*, 44:141–150, 1988.
[23] P. M. Thompson et al., "Genetic influences on brain structure," *Nat Neurosci*, 4:1253-8, 2001.
[24] H. Zhang et al., "Deformable registration of diffusion tensor MR images with explicit orientation optimization." Medical Image Analysis 10: 764-785, 2006.

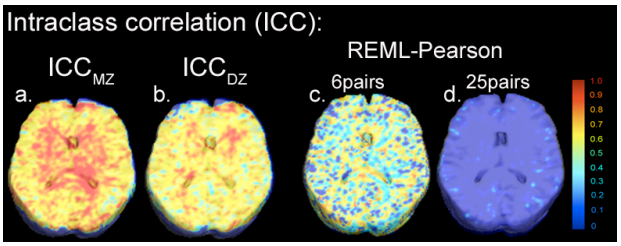

Figure 1. Maps show the intraclass correlation for MZ twins and DZ twins for the full diffusion tensor. (c) and (d) show the *difference* between REML-based ICC and ICC based on Pearson's correlation using 6 MZ twin pairs (c) and 25 MZ twin pairs (d) for the tGA measure. REML avoids bias when samples are smaller. Pearson's ICC underestimates the true REML ICC in small samples (c).

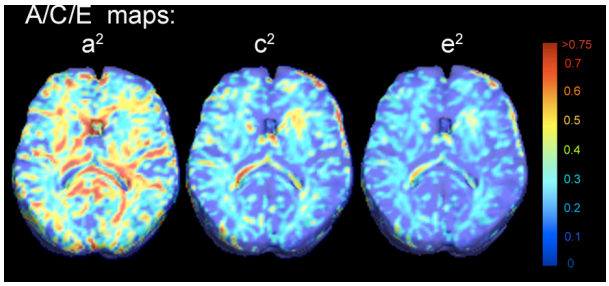

Figure 2. Maps show unbiased estimates of additive genetic (a^2), shared environmental (c^2) and unique environmental (e^2) proportions of variance for the full diffusion tensor (summing to 1).