Hippocampal Surface Mapping of Genetic Risk Factors in AD via Sparse Learning Models

Jing Wan[1,2,⋆], Sungeun Kim[1,⋆], Mark Inlow[1,3], Kwangsik Nho[1], Shanker Swaminathan[1], Shannon L. Risacher[1], Shiaofen Fang[2], Michael W. Weiner[4], M. Faisal Beg[5], Lei Wang[6], Andrew J. Saykin[1,⋆⋆], Li Shen[1,2,⋆⋆], and ADNI

[1] Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
[2] Computer and Information Science, Purdue University Indianapolis, IN, USA
[3] Mathematics, Rose-Hulman Institute of Technology, IN, USA
[4] Radiology, Medicine and Psychiatry, UC San Francisco, CA, USA
[5] School of Engineering Science, Simon Fraser University, BC, Canada
[6] Psychiatry and Behavioral Sciences, Northwestern University, IL, USA

Abstract. Genetic mapping of hippocampal shape, an under-explored area, has strong potential as a neurodegeneration biomarker for AD and MCI. This study investigates the genetic effects of top candidate single nucleotide polymorphisms (SNPs) on hippocampal shape features as quantitative traits (QTs) in a large cohort. FS+LDDMM was used to segment hippocampal surfaces from MRI scans and shape features were extracted after surface registration. Elastic net (EN) and sparse canonical correlation analysis (SCCA) were proposed to examine SNP-QT associations, and compared with multiple regression (MR). Although similar in power, EN yielded substantially fewer predictors than MR. Detailed surface mapping of global and localized genetic effects were identified by MR and EN to reveal multi-SNP-single-QT relationships, and by SCCA to discover multi-SNP-multi-QT associations. Shape analysis identified stronger SNP-QT correlations than volume analysis. Sparse multivariate models have greater power to reveal complex SNP-QT relationships. Genetic analysis of quantitative shape features has considerable potential for enhancing mechanistic understanding of complex disorders like AD.

1 Introduction

Recent advances in brain imaging and high throughput genotyping techniques enable new approaches to study the influence of genetic variation on brain structure and function. Existing imaging genetics studies employ summary statistics (e.g., volume, thickness) [7] and detailed voxel-wise measures [8] as phenotypes to discover genetic risk factors. Genetic mapping of hippocampal shape, an

[⋆] Equal contribution by J Wan (wanjing@iupui.edu) and S Kim (sk31@iupui.edu).

[⋆⋆] Correspondence to L Shen (shenli@iupui.edu) or AJ Saykin (asaykin@iupui.edu). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (U01 AG024904, http://adni.loni.ucla.edu). This project was also supported in part by NIBIB R03 EB008674, NIA 1RC 2AG036535, CTSI-IUSM/CTR(RR025761), NIA P30 AG10133, and NIA R01 AG19771.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part II, LNCS 6892, pp. 376–383, 2011.

Table 1. Participant characteristics

Category	HC	MCI	AD	p-value
Gender (M/F)	91/75	184/103	68/61	0.041
Baseline Age (years; Mean±STD)	76.18±4.91	74.99±7.21	75.36±7.78	0.198
Education (years; Mean±STD)	16.20±2.63	15.71±2.98	15.07±3.04	< **0.005**
Handedness (R/L)	155/11	260/27	121/8	0.411

under-explored area, has strong potential as a neurodegeneration biomarker for Alzheimer's disease (AD) and mild cognitive impairment (MCI). The present study investigates genetic effects of top candidate single nucleotide polymorphisms (SNPs) on hippocampal shape features in a large cohort.

Massive univariate analyses are often used in imaging genetics [7,8], and can quickly identify important associations between individual SNPs and imaging quantitative traits (QTs). However, it treats SNPs and QTs as independent units, and overlooks relationships in which multiple SNPs jointly effect multiple QTs. In this work, two multivariate sparse models, the elastic net and sparse canonical correlation analysis, are used to study genetic effects on hippocampal shape and are expected to have greater power to reveal complex SNP-QT relationships. These models could enable discovery of a small set of relevant features which may provide potential surrogate biomarkers for therapeutic trials.

2 Materials and Methods

Magnetic resonance imaging (MRI) and genotype data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [7]. ADNI is a landmark investigation sponsored by the NIH and industrial partners designed to collect longitudinal neuroimaging, biological and clinical information from over 800 participants that will track the neural correlates of memory loss from an early stage. Further information can be found at www.adni-info.org. 582 non-Hispanic Caucasian participants (166 Healthy Control (HC), 287 MCI, 129 AD participants) with segmented hippocampal data and quality controlled (QC) genotype data were included in this study (Table 1).

Hippocampal Shape: Hippocampi were segmented from the baseline MRI scans by applying probabilistic-based FreeSurfer and Large Deformation Diffeomorphic Metric Mapping (FS+LDDMM) [3]. This fully-automated segmentation pipeline first uses FreeSurfer subcortical labeling to provide information for initialization, and then employs LDDMM to generate a diffeomorphic transformation so that anatomical structures can be mapped consistently and smoothly. To remove size effect, total intracranial volume (ICV) was adjusted to a constant (i.e., mean ICV of all HCs) and each hippocampus was scaled accordingly. Rigid body transformation was then applied to register each hippocampus to a template (defined as the mean of all HCs) in a least square fashion. Surface signals were extracted as the deformation along the surface normal direction of the template, and were adjusted for baseline age, gender, education, and handedness using the regression weights derived from the HC participants (Table 1).

Candidate SNPs: The SNP data were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA). We focused on top AD genetic risk factors, including top 23 SNPs from the AlzGene database [1] as of 09/01/2010, and a SNP from the TOMM40 gene adjacent to the APOE gene. The TOMM40 SNP was included because it was unclear whether the SNP played a unique role in AD or served solely as an APOE marker. Four SNPs were excluded due to failed imputation or quality check. Among the remaining 20 SNPs (Fig. 1(a)), 10 SNPs were available from the ADNI data and 10 SNPs were successfully imputed using MACH v1 [4] and IMPUTE v2 [6] software packages. The QC criteria for the SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. The selected 20 SNPs were numerically coded to test additive genetic effect, i.e., dose dependent effect of the minor allele.

Overall Strategy: For comparative analysis, *multiple regression models* were fit using all 20 SNPs to predict the hippocampal volume (mean of left and right, covaried for age, gender, education, handedness and ICV) and, in addition, the surface signal at each location or vertex on the hippocampal surface. The *elastic net regression* was then applied to identify a small set of relevant SNPs for each surface location. Finally, *sparse canonical correlation analysis* was used to examine more complex relationships between SNP sets and surface regions.

Multiple Regression: Under the additive model, the surface signals are linearly related to the number of minor alleles. This implies, assuming no interactions between SNPs, the multiple regression model $S_{i,j} = \beta_{0,j} + \beta_{1,j}\mathrm{SNP}_{i,1} + \cdots + \beta_{20}\mathrm{SNP}_{i,20} + \epsilon_{i,j}$, where $S_{i,j}$ is the surface signal at vertex j for subject i. The model utility F test was used to test the null hypothesis of no relationship between S_j and the 20 SNPs for the $j = 1, \ldots, 13222$ vertices. Gaussian random field theory (RFT) methods [13], implemented in SurfStat [12], were used to ensure the family-wise error rate did not exceed 0.05. While this procedure can detect any linear relationship between S_j and the SNPs this flexibility comes at the cost of reduced power to detect a relationship between a specific SNP and S_j. Sparse regression methods, which seek to accurately predict the response variable using a minimal number of predictors, address this and other regression shortcomings by integrating variable selection and model estimation.

Elastic Net Regression: The ability of sparse regression methods to detect and model genetic relationships was investigated by estimating the above model at each hippocampal location using elastic net (EN). EN produces sparse solutions by adding a coefficient magnitude penalty to the least squares objective function [14]. More specifically, the EN coefficient estimates minimize the penalized least squares objective function $\mathrm{ElNet}_j(\beta_0, \beta_1, ..., \beta_{20}) = \sum_{i=1}^{n}(S_{i,j} - \hat{S}_{i,j})^2 + \lambda P_\alpha(\beta_1, \ldots, \beta_{20})$, in which $\hat{S}_{i,j} = \beta_{0,j} + \beta_{1,j}\mathrm{SNP}_{i,1} + \cdots + \beta_{20,j}\mathrm{SNP}_{i,20}$

and the penalty $\hat{P}_\alpha(\beta_1, \ldots, \beta_{20}) = \alpha \sum_{k=1}^{20} |\beta_k| + (1-\alpha) \sum_{k=1}^{20} \beta_k^2$ is a convex combination of the L_1 lasso and L_2 ridge penalties. This objective function has two parameters: λ controls the amount of shrinkage; and α adjusts the trade-off between lasso and ridge to capitalize on their strengths and minimize their weaknesses. The preceding regression analysis was duplicated using the Glmnet [2,9] implementation of EN with $\alpha = 0.5$ and λ chosen using 10-fold cross-validation.

Sparse Canonical Correlation Analysis: The surface signals represent samples of a smooth function defined on the hippocampus. Methods which capitalize on the resulting correlation between surface signals at neighboring vertices by modeling the joint relationship between multiple surface signals and SNPs should provide increased power to detect any relationships present [10]. To investigate this possibility for linear relationships, sparse canonical correlation analysis (SCCA) was used. Let $X_i = (\mathrm{SNP}_{i,1}, \mathrm{SNP}_{i,2}, \ldots, \mathrm{SNP}_{i,20})'$ be the vector of the 20 SNPs for subject i and $Y_i = (S_{i,1}, S_{i,2}, \ldots, S_{i,m})'$ be the vector consisting of the surface signals at the $m = 13,222$ vertices. Canonical correlation analysis (CCA) produces linear combinations (canonical variates) $U_j = A_j'Y$ and $V_j = B_j'X$, $j = 1, \ldots, 20$, such that the correlation between U_j and V_j is maximized subject to orthogonality constraints. Two major weaknesses of CCA are that it requires the number of observations n to exceed the combined dimension of Y and X (here 13,242) and that it produces nonsparse A_j and B_j which are difficult to interpret. The SCCA method employed here ameliorates these weaknesses using the penalized matrix decomposition approach [11]. This method maximizes the correlation between U and V subject to the coefficient vector constraints $P_1(A) \le c_1$ and $P_2(B) \le c_2$. Here the L_1 penalty $P(A) = \sum_{k=1}^{p} |A(k)|$ was used for both P_1 and P_2. Values for c_1 and c_2 were chosen using Witten and Tibshirani's permutation tuning procedure. The SCCA analyses were computed using the R package PMA (Penalized Multivariate Analysis v.1.0.7.1).

3 Results

In the volumetric analysis of 20 SNPs, only APOE SNP (rs429358) has a significant (p $\le$ 0.0004) effect on the hippocampal volume. The Pearson correlation coefficient between the APOE SNP and hippocampal volume was -0.159.

Fig. 2(a) shows the map of F-statistics of multiple regression (MR). Regions with $F \ge 3.0$ and spatial extent ≥ 2.4 resels have a random field theory adjusted p-value ≤ 0.05. Fig. 2(b) shows the mean of the absolute residuals (fitted errors) over all subjects. The residual map of elastic net (EN) is almost identical to Fig. 2(b), showing similar predictive power between EN and MR.

However, the predictors selected by EN are much more sparse than those of MR (see Fig. 1(a-c)). Combining Fig. 1(c) with (a) and (b), we can extract the coefficient map for a specific SNP and examine localized genetic effects on the surface. Shown in Fig. 1(d-g) are examples of the APOE and TOMM40 SNPs, which elucidate the benefit of sparsity achieved in EN compared to MR. While MR indicates a global effect on the surface (f,g), EN identifies localized regional effects (d,e) and yields useful information for biomarker discovery.

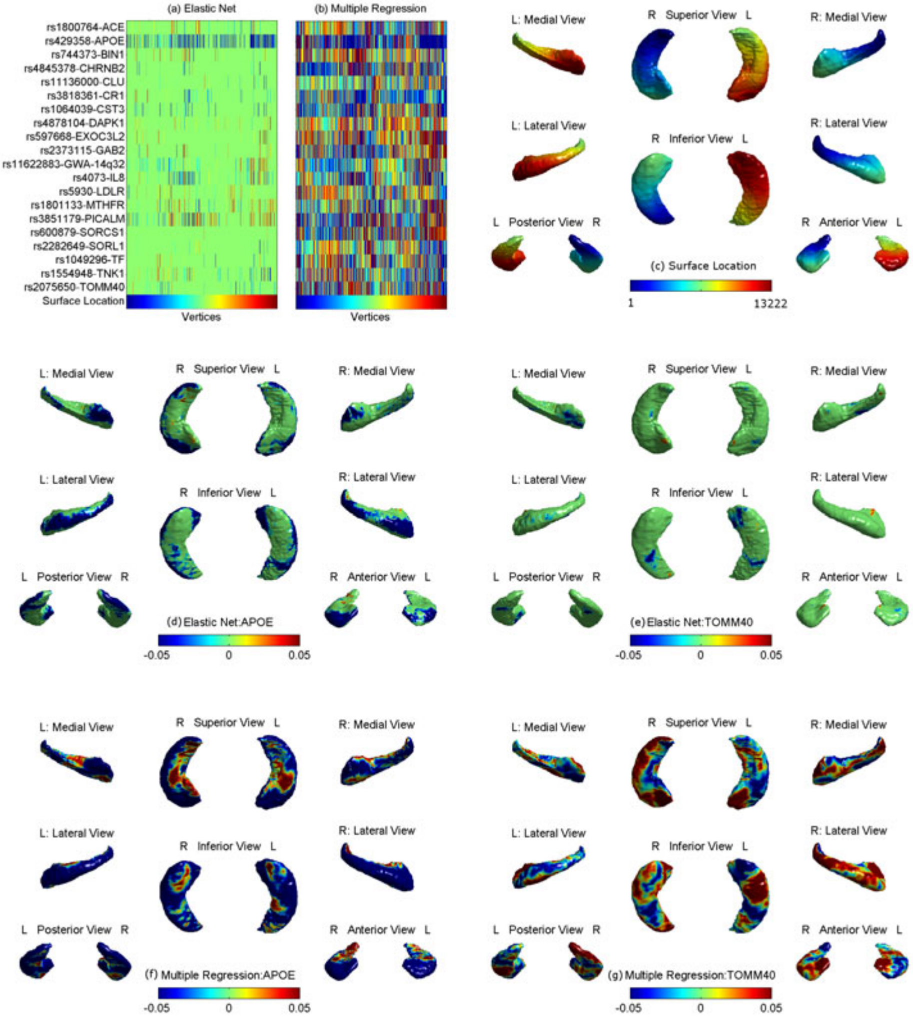

Fig. 1. (a-c) Heat maps of regression coefficients for elastic net (a) and multiple regression (b), where the hippocampal surface location (bottom row in (a,b)) is color-coded and mapped in (c). (d-e) Surface map of genetic effects of the APOE and TOMM40 SNPs estimated by elastic net (d,e) and multiple regression (f,g).

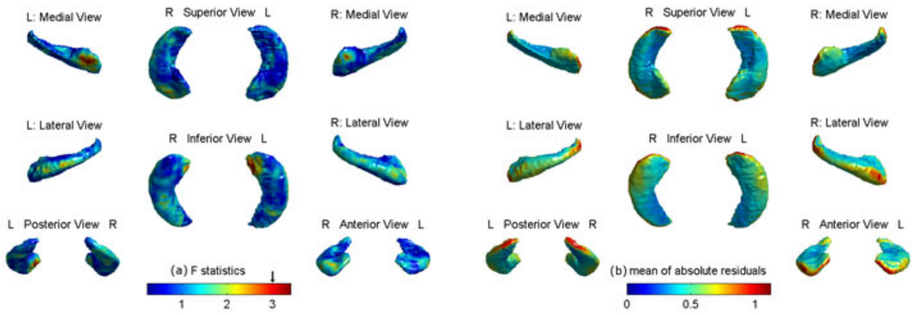

Fig. 2. (a) F-map of multiple regression. (b) Mean of absolution residuals.

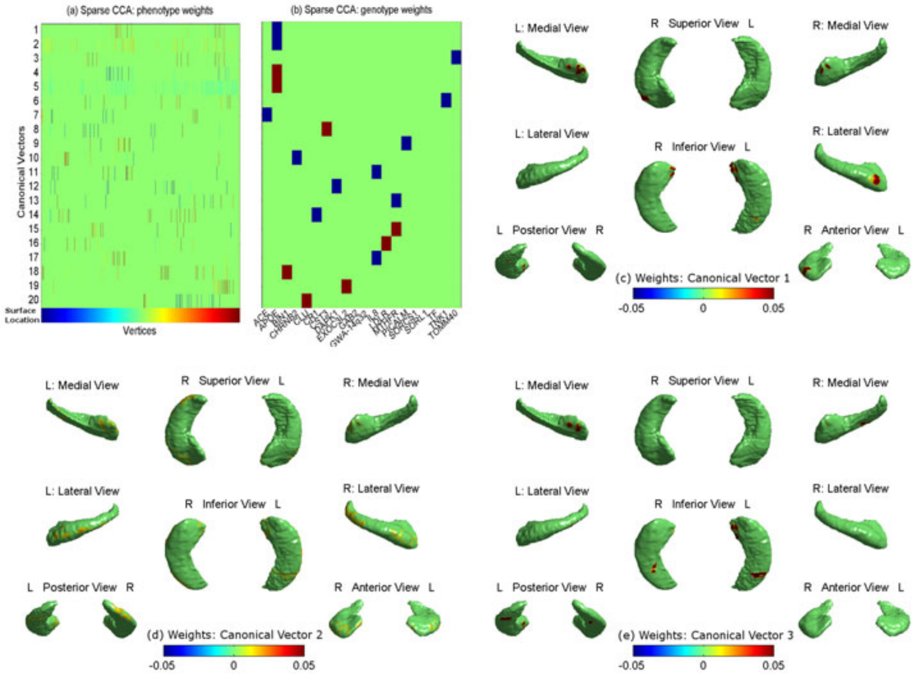

Fig. 3. (a-b) Weights of canonical vectors ordered by descending correlations between surface signals (a) and SNPs (b). (c-e) Surface maps of the top three canonical vectors: the first three rows in (a) are mapped onto the surface.

Fig. 3 shows the results of SCCA. Weights of 20 canonical vectors for vertex-based surface signals (a) and SNPs (b) were color-coded as heatmaps. The top three rows in (a) were mapped onto the hippocampal surface and shown in (c-e), respectively. In (a-b), canonical vector pairs (i.e., corresponding rows in (a-b)) were ordered by descending correlation between surface signals and SNPs; and the correlation coefficients of all 20 pairs ranged from 0.26 to 0.17 in descending order. This clearly demonstrated the increased power of shape analysis, since the strongest correlation between each of 20 SNPs and hippocampal volume in our volumetric analysis was between the APOE SNP and hippocampal volume with a magnitude of 0.159. This was corroborated by the fact that the maximum absolute correlation between the surface signal and APOE SNP was 0.20 among all vertices and was 0.19 among the vertices with $F \geq 3.0$.

In addition, the parameters for SCCA were automatically tuned by 100 permutations to increase the sparsity and smoothness. As a result, the identified surface locations, correlated with each SNP were more sparse than those for the same SNP from EN (see Fig. 3(a-b) vs Fig. 1(a)). Interestingly, the sparsity was maximized for SNPs, since each canonical SNP vector selected exactly one SNP (Fig. 3(b)), yielding a simple model easy to interpret (i.e., multi-SNP-multi-location associations became single-SNP-multi-location ones).

Fig. 3(c-d) show surface regions related with the APOE SNP (rs429358) at different correlation levels. The correlated vertices in Fig. 3(c-d) have non-zero weights as in Fig. 1(d,f), but they are localized to smaller regions in Fig. 3(c-d). Fig. 3(e) shows surface regions related with the TOMM40 SNP (rs2075650). All vertices with non-zero weights in Fig. 3(e) also have non-zero weights in Fig. 1(e,g). However, compared to Fig. 1(e,g), vertices with non-zero weights in Fig. 3(e) are highly sparse and spatially localized to smaller areas. These two types of patterns are complimentary: the associations derived from EN are multi-SNP-single-location, while those found in SCCA are single-SNP-multi-location.

Five-fold cross-validation of SCCA yielded equally sparse SNP-QT patterns. The most consistent canonical component identified in all five trials is similar to the top finding using the entire data: the genetic vector contains only APOE, and the phenotype vector shows a pattern like Fig. 3(c). Training and testing correlation coefficients are 0.279 ± 0.017 (mean ± SD) and 0.175 ± 0.068, respectively, while the magnitudes of correlation coefficients between APOE and hippocampal volume in the same data are 0.159 ± 0.012 and 0.163 ± 0.056, respectively.

4 Discussion

Detailed surface mappings of localized genetic effects were identified from our hippocampal shape analysis. Different from existing massive univariate analyses [7,8], this study is among the first to simultaneously use multiple response variables with multiple predictors for analyzing real neurogenomic data [5,10] and may be the first for studying genetic influences on hippocampal morphometry using this paradigm. In our analyses, we combined two promising sparse multivariate models with a typical morphometric method. Investigation of other statistical models (e.g., [10]) and surface metrics, coupled with pathway analyses, will be important future topics to potentially yield new discoveries. As the best known AD genetic risk factor, APOE was the most prominent signal in all of our analyses, which to some extent validated the efficacy of our methods. Replication in independent large samples will be important to confirm the imaging genetic findings. Genetic analysis of quantitative shape features has considerable potential for examining disease mechanisms from a novel perspective that can inform selection of imaging biomarkers for early detection and therapeutic trials.

References

1. Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E.: Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. Nat. Genet. 39(1), 17–23 (2007)
2. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33(1), 1–22 (2010)
3. Khan, A.R., Wang, L., Beg, M.F.: Freesurfer-initiated fully-automated subcortical brain segmentation in mri using large deformation diffeomorphic metric mapping. Neuroimage 41(3), 735–746 (2008)
4. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R.: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34(8), 816–834 (2010)
5. Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N.I., Calhoun, V.: Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. Hum. Brain Mapp. 30(1), 241–255 (2009)
6. Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39, 906–913 (2007)
7. Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., Foroud, T., Pankratz, N., Moore, J.H., Sloan, C.D., Huentelman, M.J., Craig, D.W., Dechairo, B.M., Potkin, S.G., Jack Jr., C.R., Weiner, M.W., Saykin, A.J., ADNI: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. Neuroimage 53(3), 1051–1063 (2010)
8. Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M.J., Craig, D.W., Gerber, J.D., Allen, A.N., Corneveaux, J.J., Dechairo, B.M., Potkin, S.G., Weiner, M.W., Thompson, P.: Voxelwise genome-wide association study (vgwas). Neuroimage 53(3), 1160–1174 (2010)
9. Tibshirani, R.: Glmnet, http://www-stat.stanford.edu/~tibs/glmnet-matlab/
10. Vounou, M., Nichols, T.E., Montana, G.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage 53(3), 1147–1159 (2010)
11. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3), 515–534 (2009)
12. Worsley, K.J.: Surfst, http://www.math.mcgill.ca/keith/surfstat
13. Worsley, K.J., Andermann, M., Koulis, T., MacDonald, D., Evans, A.C.: Detecting changes in nonisotropic images. Hum. Brain Mapp. 8(2-3), 98–101 (1999)
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc. 67(2), 301–320 (2005)