

PARTIAL LEAST SQUARES MODELLING FOR IMAGING-GENETICS IN ALZHEIMER’S DISEASE: PLAUSIBILITY AND GENERALIZATION

Marco Lorenzi^{*}, Boris Gutman[†], Derrek P. Hibar[†], Andre Altmann^{*}, Neda Jahanshad[†],
Paul M. Thompson[†] and Sebastien Ourselin^{*}

^{*} Translational Imaging Group, CMIC, University College London, London, UK

[†] Imaging Genetics Center, University of Southern California, Marina del Rey, CA, USA

ABSTRACT

In this work we evaluate the ability of PLS in generalizing to unseen clinical cohorts when applied to the analysis of the joint variation between genotype and phenotype in Alzheimer’s disease (AD). The model is trained on single-nucleotide polymorphisms (SNPs) and brain volumes obtained from the ADNI database for a large cohort of healthy individuals and AD patients, and validated on the ADNI MCI and ENIGMA cohorts. The experimental results confirm the ability of PLS in providing a meaningful description of the joint dynamics between brain atrophy and genotype data in AD, while providing important generalization results when tested on clinically heterogeneous cohorts.

Index Terms— GWA, imaging-genetics, genotype, phenotype, Alzheimer’s disease, machine learning

1. INTRODUCTION

Imaging genetics is a central scientific field for the discovery of the mechanisms linking genotype to the phenotypical traits observable in neuroimaging [1]. Classical genome-wide association (GWA) studies are usually performed by investigating the relationship between multiple genetic variants and candidate phenotypical traits (such as brain regional volumes) by means of independent univariate analysis [2]. These studies typically require very large samples in order to detect meaningful associations, since genetic traits usually account for a small fraction of the phenotype variance. Moreover, since mass univariate testing does not account for potential gene-gene interactions, it is highly prone to multiple comparison problems, and might lead to underpowered discoveries of significant associations.

Recent methodological advances in imaging genetics introduce multivariate approaches to capture meaningful

genotype-phenotype interactions [3]. Although based on different statistical assumptions, most of these approaches rely on simultaneous regression and dimensionality reduction strategies, such as partial least squares (PLS) [4], or independent component analysis (ICA) [5]. In particular, PLS is an appealing approach due to the parsimonious description of the multivariate correlation patterns, and for the relative simplicity of the implementation [6]. Multivariate models applied to the large dimensional genotype data easily run into overfitting problems, and the generalization of the related findings to unseen data is usually problematic. Even though different cross-validation and regularization strategies have been introduced in order to mitigate this crucial issue, it still remains to be verified whether the findings obtained with high-dimensional multivariate models do generalize to unseen and heterogeneous clinical cohorts.

In this work we evaluate the ability of PLS in generalizing to unseen clinical cohorts when applied to the analysis of the joint variation between genotype and phenotype in Alzheimer’s disease (AD). The analysis proposed in this work is based on the identification of the meaningful features associated to the largest PLS weights estimated in a cohort of AD and healthy individuals (Section 2), and subsequently validated i) on an ADNI group of subjects affected by mild cognitive impairment (MCI) (Section 5.2.1), and ii) on the ENIGMA cohort [7] composed by very heterogeneous clinical populations (Section 5.2.2). The experimental results confirm the ability of PLS in providing a meaningful description of the joint dynamics between brain atrophy and genotype data in AD, while providing important generalization results.

2. OVERVIEW OF PARTIAL LEAST SQUARES

Let $(x_i)_{i=1}^N$ and $(y_j)_{j=1}^M$ be distinct sets of features, and let \mathbf{X} and \mathbf{Y} be respectively the $N \times K$, and $M \times K$ matrices of (normalized) samples for K distinct subjects. PLS models the relationship between (x_i) and (y_j) by finding a latent space for which the projections of \mathbf{X} and \mathbf{Y} have maximal covariance. In practice, PLS can be implemented through the singular value decomposition (SVD) of the cross-product ma-

This study was funded in part by NIH ENIGMA Center grant U54 EB020403, supported by the Big Data to Knowledge (BD2K) Centers of Excellence program, by the NIH R BRC UCLH/UCL High Impact Initiative-BW.mn.BRC10269, by the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), and by the EPSRC grants EP/H046410/1, EP/J020990/1, and EP/K005278. AA holds an MRC Medical Bioinformatics Career Development Fellowship, funded from award MR/L016311/1.

	Training Data		Testing Data	
	Healthy	AD	MCI _s	MCI _c
# individuals	409	248	460	228
Sex (% females)	48	56	39	38
MMSE	29.09 ± 1.1	23.22 ± 2	28.02 ± 1.7	26.95 ± 1.8
Education (years)	16.37 ± 2.7	15.33 ± 2.9	16 ± 2.8	15.7 ± 2.8
APOE4 (% 0,1,2)	72, 26, 2	31, 48, 21	56, 35, 9	33, 51, 16

Table 1. Summary socio-demographic, clinical and genetic information. MCI_s: stable MCI subjects, MCI_c: MCI subjects converted to AD during the observational time of the study. MMSE: mini-mental state examination.

trix $\mathbf{C} = \mathbf{X}\mathbf{Y}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. The columns of the orthogonal matrices \mathbf{U} and \mathbf{V} are the eigen-components which characterize the common variation of \mathbf{X} and \mathbf{Y} respectively, and can be used to explore the modes of joint variability between the sets of features, while the diagonal matrix $\mathbf{\Lambda}$ is composed by the eigen-values which quantify the amount of common variability explained by each component. Finally, for any test observations (\mathbf{x}, \mathbf{y}) , the projections $P_{\mathbf{x}} = \mathbf{x}^T\mathbf{U}$ and $P_{\mathbf{y}} = \mathbf{y}^T\mathbf{V}$ are a low-dimensional representation of the data on the latent PLS space.

2.1. Efficient Estimation of PLS on high dimensional data

When the dimension of the feature space is large, such as when analyzing genotype data, the naive storage and analysis of the associated correlation matrix is usually computationally prohibitive. Fortunately, as illustrated in [8], the SVD computation of the large matrix $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$ can be derived from the eigen-problem associated to the transposed matrices:

$$(\mathbf{X}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y})\mathbf{A} = \mathbf{A}\mathbf{L}, \quad (1)$$

and by subsequently computing $\mathbf{U} = \mathbf{X}(\mathbf{Y}^T\mathbf{Y}\mathbf{L}^{-\frac{1}{2}})$, and $\mathbf{V} = \mathbf{Y}\mathbf{B}$, with $\mathbf{B} = \mathbf{A}(\mathbf{A}^T\mathbf{Y}^T\mathbf{Y}\mathbf{A})^{-\frac{1}{2}}$, and $\mathbf{\Lambda} = \mathbf{L}^{\frac{1}{2}}$.

3. DATA

We selected genotype and phenotype data available in the ADNI1/2 datasets for 409 healthy individuals, 248 patients affected by AD, 460 patients affected by mild cognitive impairment (MCI), and 228 MCI patients subsequently converted to AD during the observational time of the study. Summary socio-demographic, clinical and genetic information are available in Table 1.

The phenotype features consist in the individual’s regional volumes reported in ADNI1/2 for whole brain, ventricles, and average bilateral hippocampi, entorhinal cortex and mid-temporal lobes. The volumes were normalized by covarying for age, total intracranial volume, and sex, and subsequently standardized by group-wise mean and standard deviation of the healthy and AD individuals.

The genotype features consist in the individual’s minor allele counts for each of the 1,167,126 single-nucleotide poly-

morphisms (SNPs) in chromosomes 1 to 22 available in the study. SNP data was downloaded from the ADNI website and preprocessed with plink [9] by filtering SNPs which did not meet the quality control criteria (Minor Allele Frequency (MAF) < 0.01, Genotype Call Rate < 95%, Hardy-Weinberg Equilibrium < 1×10^{-6}). Finally, the genotyped SNPs passing QC were imputed to the HapMap III reference panel and further QCed to keep only high quality imputed SNPs (i.e. MAF > 0.01 and $RSQR > 0.3$), and the missing individual SNPs were replaced by the group-wise median. The resulting allele counts were finally standardized by group-wise mean and standard deviation of the healthy and AD individuals.

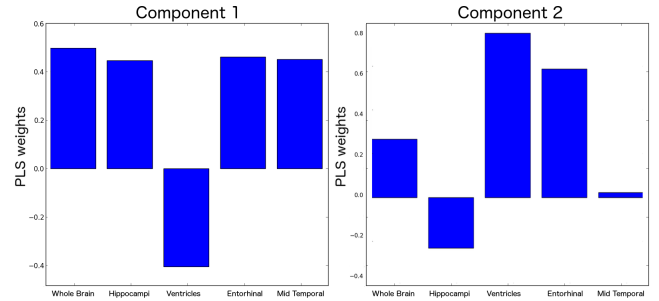


Fig. 1. Main PLS eigen-component of \mathbf{V} for the phenotype features. Component 1: ventricles volume is anti-correlated with respect to the volume of the other brain areas. Component 2: hippocampal volume is anti-correlated with respect to the other brain structures.

4. STATISTICAL ANALYSIS

PLS was applied in order to model the joint variation between phenotype and genotype observed in healthy and AD individuals. We denote by \mathbf{X} the matrix of dimension $1,167,130 \times 657$ of genotype features for the selected groups, and by \mathbf{Y} the associated phenotype matrix of dimension 5×657 . PLS was performed by following Section 2.1 in order to identify the eigen-components \mathbf{U} and \mathbf{V} of common genotype-phenotype variation modelled in $\mathbf{C} = \mathbf{X}\mathbf{Y}^T$.

The generalization of the PLS model was tested on the ADNI MCI data by statistically assessing the ability of the estimated PLS components in providing clinically relevant separation between MCI subjects converted to AD and subjects that remain stable during the observation period, through group-wise comparison of the projections in the latent space spanned by \mathbf{U} and \mathbf{V} . We also statistically assessed the group-wise differences between AD and healthy controls with respect to the set of most informative SNPs associated to the eigen-component \mathbf{U} . The resulting reduced set of SNPs leading to significant differences was finally tested on the ADNI MCI and ENIGMA cohort.

5. RESULTS

5.1. Model training and estimated components

The first eigen-components of variation between genotype and phenotype modelled by PLS in AD and healthy individuals are illustrated in Figures 1 and 2. The analysis of the respective eigen-values showed that these components alone accounted for 60% of the total variation in the data (40% for component 1, and 20% for component 2).

The first component of phenotype variation is identified by the anti-correlation between ventricles volume and the volume of the other brain areas (Figure 1). Figure 2 shows the SNPs with largest absolute weights in the genotype eigen-components, grouped by chromosome location (for each eigen-component we identified 584 SNPs above the 99.95th percentile of the distribution of the absolute weights). The figure reports the percentage of informative SNPs per chromosome identified by the PLS model with respect to the number of SNPs available for each chromosome. We note that chromosome 19 is the most represented among the SNPs identified by eigen-component 1. The second component is characterized by anticorrelation between hippocampal volume and the other brain structures (principally ventricles and entorhinal cortex), while the largest represented chromosomes are the number 7, 15, and 14. The subsequent linkage disequilibrium (LD) analysis identified 210 and 232 independent SNPs for respectively component 1 and 2, uniformly distributed across the 22 chromosomes. The significance of the estimated PLS model was assessed through permutation test [10]: the eigenvalues estimated by the model (i.e. the total explained variability) were higher than the ones obtained by randomly permuting the rows of the phenotype matrix \mathbf{Y} with $1 - p = 0.0208$ (10e3 permutations).

5.2. Model validation

5.2.1. Generalization to unseen MCI data

Figure 3 shows the differences between stable and converting MCI subjects in the projection on the latent space spanned by the first PLS eigen-components. The projection in the latent space leads to significant group-wise differences (Mann Whitney non-parametric U-test for difference in the median) for both genotype ($p = 0.042$, $U = 47467$, effect size = 0.45, with a value of 0.5 indicative of no effect) and phenotype ($p < 1e - 4$, $U = 33397$, effect size = 0.31) features.

The group-wise projections with respect to the second eigen-components were close to significance for the genotype component ($p = 0.087$), and significant for the phenotype one ($p < 0.001$) (not shown).

5.2.2. Investigation of individual SNPs

Each of the 584 informative SNPs identified by the first PLS eigen-component was independently tested in order to iden-

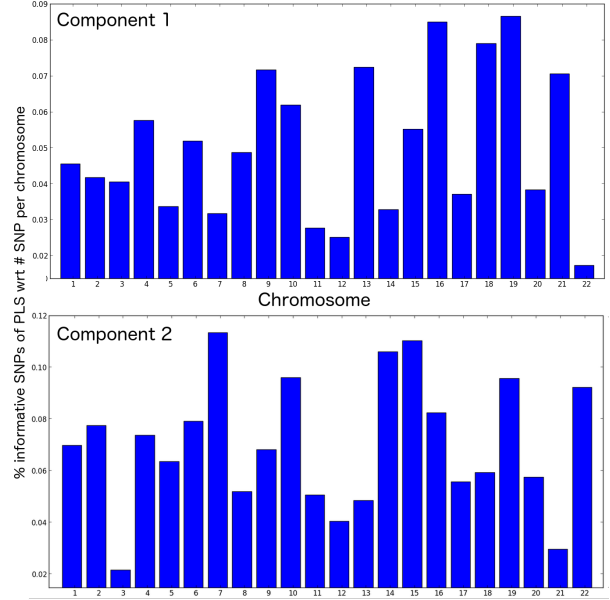


Fig. 2. Chromosome representativeness among the set of most informative SNPs associated to the main PLS eigen-component of \mathbf{U} . Chromosome 19 is the most represented among the SNPs identified by the first component.

tify group-wise differences between the clinical groups considered in the study. Table 2 shows the SNPs leading to the largest statistical evidence for the difference between healthy and AD individuals ($p < 0.05$, Bonferroni correction). The identified SNPs are highly ranked by the PLS model and have been already reported in previous GWAS studies in AD [11, 12]. The table also reports the statistical significance of the differences between ADNI MCI stables and converters with respect to the three identified SNPs. Finally, the last column reports the statistical association between the identified SNPs and the hippocampal volume in the ENIGMA cohort [7]. The identified SNPs consistently lead to significant statistical as-

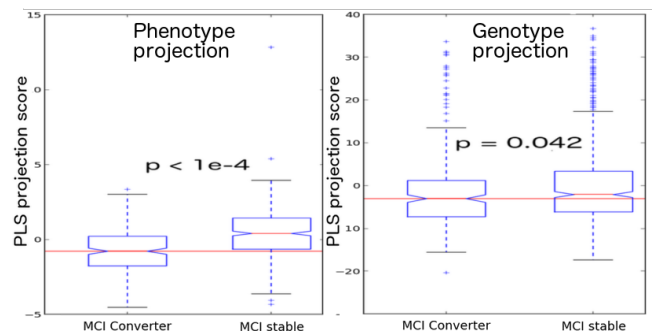


Fig. 3. The projection in the latent space of the first PLS component leads to significant differences between converting and stable MCIs for both phenotype and genotype features.

SNP	PLS ranking (comp. 1,2) out of 1,167,126 SNPs	Chromosome	Gene	healthy vs AD	MCI _s vs MCI _c	ENIGMA
rs157580	36th, -	19	TOMM40	1.2e-7	0.034	0.046
rs2075650	1st, 74th	19	TOMM40	1.3e-14	2.17e-6	0.007
rs157582	9th, -	19	TOMM40	5.4e-13	1.3e-8	0.01

Table 2. Identified SNPs with largest statistical evidence. Column 2: SNPs ranking relative to the absolute weights of the PLS eigen-components. Columns 5-7: p-value for the group-wise comparison (Mann Whitney non-parametric U-test), or for the correlation with respect to the hippocampal volume in the ENIGMA dataset.

sociation when tested on the clinically heterogeneous MCI ADNI and ENIGMA cohorts.

6. CONCLUSIONS

PLS provides a valuable alternative to classical univariate analysis of GWA datasets for studying correlation patterns in large multidimensional genetic and structural data, by i) providing meaningful description of the joint dynamics between brain atrophy and genotype data in AD and, ii) generalizing to unseen clinically heterogeneous cohorts, such as the ENIGMA one. In this work the PLS model was trained on 657 ADNI subjects, which represent a tiny fraction ($\sim 5\%$) of the whole ENIGMA discovery cohort of 13,000 individuals. Although we don't expect a meaningful impact of the ADNI subset on the reported generalization of the PLS model in the ENIGMA cohort (last column of Table 2), we cannot exclude that this issue might be a potential source of bias of the resulting statistical result.

Importantly, PLS overcomes the classical multiple comparison problem of GWA studies by modeling the *joint* correlation of SNPs and brain features through the eigen-components. We note that this work is based on a standard implementation of PLS, as opposed to previously proposed sparse implementations based on the L^1 regularization of the feature weights [4]. However, the *posterior analysis* proposed in this work focuses on the investigation of the features associated to the top 0.05% PLS weights with respect to the L^1 norm. The proposed approach is thus based on a heuristic penalization of the PLS components, aimed to promote stability and sensitivity of the associated findings.

7. REFERENCES

- [1] K. L. Bigos and D. Weinberger, "Imaging genetics—days of future past," *NeuroImage*, vol. 53, no. 3, pp. 804–809, 2010.
- [2] D. P. Hibar, J. L. Stein, M. Renteria, et al., "Common genetic variants influence human subcortical brain structures," *Nature*, 2015.
- [3] J. Liu and V. Calhoun, "A review of multivariate analyses in imaging genetics," *Frontiers in neuroinformatics*, vol. 8, 2014.
- [4] É. Le Floch, V. Guillemot, V. Frouin, et al., "Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares," *NeuroImage*, vol. 63, no. 1, pp. 11–24, 2012.
- [5] J. Liu, G. Pearlson, Windemuth, et al., "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Human brain mapping*, vol. 30, no. 1, pp. 241–255, 2009.
- [6] C. Grellmann, S. Bitzer, J. Neumann, et al., "Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of mri and genetic data," *NeuroImage*, vol. 107, pp. 289–310, 2015.
- [7] N. M. Novak, J. L. Stein, S.E. Medland, et al., "EnigmaVis: Online Interactive Visualization of Genome-Wide Association Studies of the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium," *Twin Research and Human Genetics*, vol. 15, pp. 414–418, 6 2012.
- [8] K.J. Worsley, J.-I. Chen, J. Lerch, and A.C. Evans, "Comparing functional connectivity via thresholding correlations and singular value decomposition," *Philosophical Transactions: Biological Sciences*, vol. 360, no. 1457, pp. pp. 913–920, 2005.
- [9] S. Purcell, B. Neale, K. Todd-Brown, et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.
- [10] A. McIntosh and N. Lobaugh, "Partial least squares analysis of neuroimaging data: applications and advances," *NeuroImage*, vol. 23, pp. S250–S263, 2004.
- [11] B. Ferencz, S. Karlsson, and G. Kalpouzos, "Promising genetic biomarkers of preclinical Alzheimer's disease: the influence of APOE and TOMM40 on brain integrity," *International Journal of Alzheimers Disease*, vol. 2012, 2012.
- [12] M.I. Kamboh, M.M. Barmada, F.Y. Demirci, et al., "Genome-wide association analysis of age-at-onset in Alzheimer's disease," *Molecular psychiatry*, vol. 17, no. 12, pp. 1340–1346, 2012.