What PLS can still do for Imaging Genetics in Alzheimer's disease

Federica Cruciani
Dept. of Computer Science
University of Verona
Verona, Italy
federica.cruciani@univr.it

Andre Altmann
Dept. of Medical Physics
and Biomedical Engineering, UCL
London, United Kingdom
a.altmann@ucl.ac.uk

Marco Lorenzi
Epione Research Project
Université Côte d'Azur, Inria Sophia Antipolis
Sophia-Antipolis, France
marco.lorenzi@inria.fr

Gloria Menegaz
Dept. of Computer Science
University of Verona
Verona, Italy
gloria.menegaz@univr.it

Ilaria Boscolo Galazzo
Dept. of Computer Science
University of Verona
Verona, Italy
ilaria.boscologalazzo@univr.it

***Abstract*—In this work we exploited Partial Least Squares (PLS) model for analyzing the genetic underpinning of grey matter atrophy in Alzheimer's Disease (AD). To this end, 42 features derived from T1-weighted Magnetic Resonance Imaging, including cortical thicknesses and subcortical volumes were considered to describe the imaging phenotype, while the genotype information consisted of 14 recently proposed AD related Polygenic Risk Scores (PRS), calculated by including Single Nucleotide Polymorphism passing different significance thresholds. The PLS model was applied on a large study cohort obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database including both healthy individuals and AD patients, and validated on an independent ADNI Mild Cognitive Impairment (MCI) cohort, including Early (EMCI) and Late MCI (LMCI). The experimental results confirm the existence of a joint dynamics between brain atrophy and genotype data in AD, while providing important generalization results when tested on a clinically heterogeneous cohort. In particular, less AD specific PRS scores were negatively correlated with cortical thicknesses, while highly AD specific PRSs showed a peculiar correlation pattern among specific subcortical volumes and cortical thicknesses. While the first outcome is in line with the well known neurodegeneration process in AD, the second could be revealing of different AD subtypes.**

***Index Terms*—Partial Least Squares, Imaging Genetics, grey matter atrophy, Polygenic Risk Scores**

I. Introduction

Alzheimer's Disease (AD) is the most common cause of dementia, affecting 46.8 million people worldwide [1]. The pathophysiology of AD and its genetic drivers have been widely studied in recent years. On the imaging side, studies based on structural magnetic resonance imaging (MRI) data have consistently observed both global and local atrophic changes during early stages of AD, mainly localised in the medial temporal lobe structures including the amygdala, hippocampus, entorhinal cortex, and parahippocampal gyrus [2]. However, recent studies have proved the involvement of other brain regions, such as basal ganglia, in the disease progression [3], while others were able to identify AD subtypes showing distinct atrophy patterns, starting and spreading across different areas [4]. On the genetic side, Polygenic Risk Scores (PRS) are gaining popularity since they represent a single (or few) score(s) combining the effects of multiple independent genetic variants in a subject's genome derived from a large genome-wide association study (GWAS) study. The PRS are informative about the individual overall genetic disease risk enabling the associations between genetic profiles and imaging features on smaller cohorts. This in particular is the target of Imaging Genetics (IG) which aims at investigating the effects of genetic variations on brain function and structure and in which our work is framed. Such methods, applied in particular to AD onset, allowed a better understanding of the genetic underpinnings on brain modulations [5]. PRS for AD have been shown to be associated with clinical diagnosis and disease progression [6], cognitive decline [7] and imaging biomarkers [7]–[9] both on healthy and cognitive impaired patients. Previous studies have generally focused on the hippocampal volume solely to evaluate its association with PRS for AD in cognitive impaired cohorts, considering its central role in AD pathophysiology [6], [7]. A wider range of brain morphometric features was investigated in association with PRS for AD in clinically normal cohorts [9]. To the best of our knowledge, the interaction between PRS for AD and a complete set of brain structural imaging phenotypes, such as cortical thickness and subcortical volumes, has not been deeply investigated in a cognitive impaired cohort. Typically, univariate models have been applied to characterize IG associations, however such methods do not account for potential cross features interactions and are highly prone to multiple comparison problems leading to underpowered discoveries of significant associations [10]. Multivariate methods, on the other hand, can address such limitations. Latent variable and

This work was partially supported by Fondazione CariVerona (Bando Ricerca Scientifica di Eccellenza 2018, EDIPO project - reference number 2018.0855.2019).

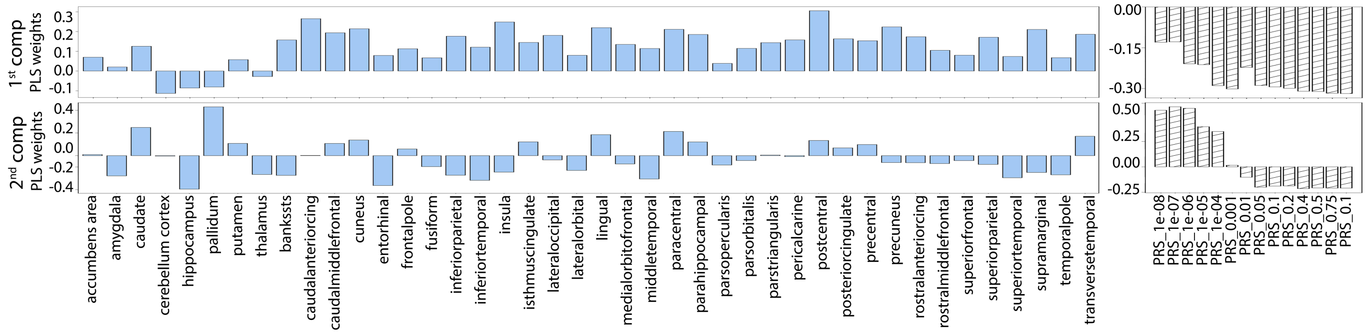

Fig. 1. First and second PLS components weights (rows) for the phenotype and the genotype features (columns).

multi-view models, for example, aim at finding a latent low dimensional space by the optimization of a target function such that the projections of the features hold some maximized joint properties. Partial Least Squares (PLS) maximizes the covariance between the latent projections, further addressing features collinearity which generally affects both imaging and genetics derived features. PLS is increasingly being exploited in IG studies, particularly in imaging transcriptomics aiming at investigating the association between imaging phenotypes and gene expression values in brain disorders [11]. Moreover, relying on different genetic features, such as Single Nucleotide Polymorphism (SNPs), Lorenzi *et al.* [10] exploited PLS to uncover the genetic underpinnings of brain atrophy in AD. Despite these promising results, the potentialities of a classical statistical model as PLS in the AD domain are still under investigated, though could help to disambiguate the associations between different feature sets considering its inherent ability to provide a straightforward explanation of the outcomes, which is not always the case for complex deep models.

The objective of our work was the characterization of the different stages of AD in the PLS latent space representation, which is indeed generated by meaningful associations found between brain morphometric features and PRS in AD. Moreover, in order to assess the generalization capability of our model computed on AD and healthy controls, an unseen cohort of subjects affected by mild cognitive impairment (MCI) was used for testing.

II. Materials and Methods

Phenotypes and genotypes used in this study were derived from the AD Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The full cohort comprehended 826 subjects from the ADNI-1,ADNI-2 and ADNI-GO phases including 243 Controls (CN), 289 Early MCI (EMCI), 179 Late MCI (LMCI) and 115 AD patients (age: 72.9 ± 6.2, 71.2 ± 7.2, 71.9 ± 7.7, and 74.8 ± 7.9; females/males: $131/112$, $126/163$, $79/100$, and $47/68$). AD and CN subjects were considered as the discovery cohort, while EMCI and LMCI were kept for testing. The considered imaging features were region-based morphometric descriptors derived from T1-w MRI images extracted by UCSF using FreeSurfer version 5.1 and accessed through the ADNI website (date accessed 18/02/2022). 84 anatomical regions of interest (ROIs) were included. The average thickness and the volume were considered for cortical and subcortical ROIs, respectively. The subcortical volumes were normalized by the intracranial volume of the respective subject. 42 features were finally obtained by averaging left and right hemispheres and were considered as phenotype. The genetic information was represented by the 14 PRS proposed in [6]. Briefly, each PRS was calculated by including all independent SNPs passing a p-value threshold in the most recent GWAS [12]. The thresholds adopted were $1e-08$, $1e-07$, $1e-06$, $1e-05$, $1e-04$, 0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.5, 0.75, 1. The related PRS will be named as PRS*_threshold_value*. SNPs in the extended APOE locus were excluded from the PRS construction to enable investigations of risk independent from APOE. We refer to [6] for further details on PRS computation. A standardization to reach zero mean and unitary standard deviation was applied to the feature sets. Age was then regressed out from the image-derived features, while the first two principal components, describing the genetic information of the whole population on which the PRS were calculated, were regressed out from the genetic features, following [6]. PLS was finally applied in order to model the joint variation between phenotype and genotype observed in healthy and AD individuals, following [8], [10]. Then, the generalization capability of our model was assessed on an unseen cohort of MCI subjects. The data variability explained by each component was calculated, and the number of components was chosen in order to allow to represent at least the 60% of it. A permutation test based on the obtained singular values was finally performed to assess the significance of the model In brief, the test checked whether the singular values obtained by the model were higher than the ones obtained by randomly permuting all rows of the phenotype matrix ($10e4$ permutations were used). The Mann Whitney non-parametric U-test was performed to assess the significance of the latent space projection difference across groups. Finally, the generalization of the PLS model was tested on the MCI group by statistically assessing the ability of the estimated PLS components in splitting EMCI and LMCI subjects, through group-wise comparison of the projections in the latent space.

III. Results

Two latent components were needed to explain at lest the 60% of data variability, accounting for 54% and 18% of data variability, respectively. The PLS weights of phenotype and genotype in the first and second latent component of the model are shown in Fig. 1. The PLS model associates a weight to each input feature reflecting its relevance in shaping the latent space, that is in the association between genotype and phenotype. The first component revealed a widespread negative correlation between phenotype and genotype. The five most relevant brain regions were postcentral gyrus, caudalanterior cingulate, insular cortex, lingual gyrus and cuneus. On the genetics side, the less AD specific PRS, hence the ones having a less stringent p-value threshold for SNPs inclusion, showed the highest weights. Moving to the second component, pallidum, hippocampus, caudate, entorhinal and inferiortemporal appeared as the most relevant regions. More in detail, pallidum was anticorrelated to the hippocampus volume, and entorhinal and inferiortemporal thicknesses, while it appeared to be correlated with the caudate volume. On the genetic side, this component highlighted the most AD specific PRS, hence the ones including SNPs peculiar for AD. These were positively correlated with pallidum and caudate volumes, while a negative correlation was found with hippocampus. Moreover, the permutation test confirmed the significance of our model resulting in $p = 0.0428$. The latent space representation of AD and CN groups is shown in Fig. 2 for the two PLS components. Both showed a separation between the two classes, particularly evident in the second one. The projection in the latent space led to significant group-wise differences for the phenotype on both PLS components, reaching $p < 1e{-}12$ on the first and $p < 1e{-}17$ for the second one. Conversely, a trend towards significance was found for the AD vs CN difference in the genotype latent space projection, with $p = 0.086$ and $p = 0.121$ for the first and second component respectively.

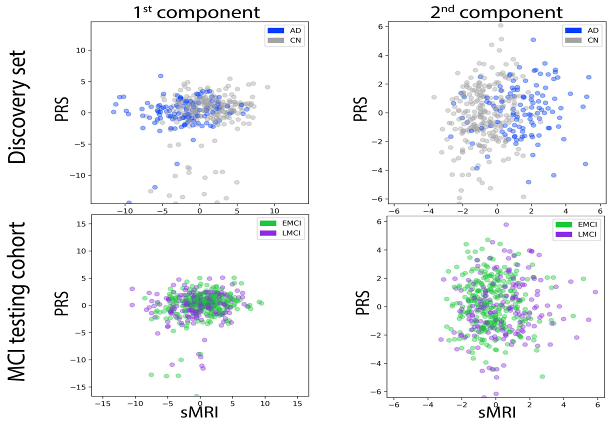

Fig. 2. Latent representation of the discovery set and MCI cohort validation set (rows) on the first two PLS components (columns) (AD: blue, CN: grey, EMCI: green, LMCI: violet).

Fig. 2 proves also the model generalization capability by showing the projection of the MCI independent set on the latent space generated by AD and CN subjects. While the first component showed a major overlapping between EMCI and LMCI, the second one allowed a clearer separation, with the LMCI being distributed in the same latent space region as the AD and the EMCI being more central.

Finally, Fig. 3 summaries the PLS latent space projections scores for the MCI group on both components, separately for genotype and phenotype. A significant difference was found for phenotype in both components, $p = 0.042$ and $p = 0.007$, respectively. The genotype differences did not reach the significance, though a moderate trend towards significance was present in the second component ($p = 0.130$).

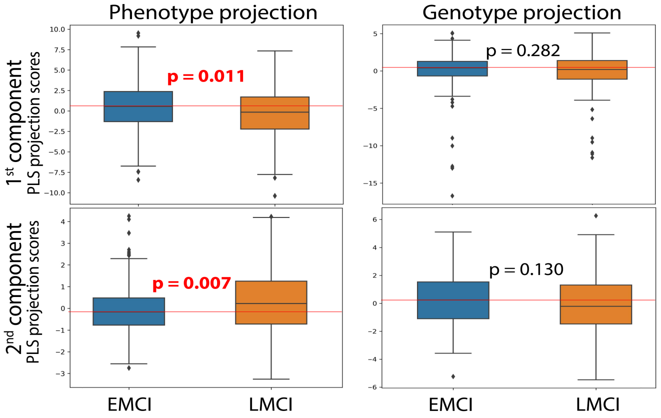

Fig. 3. Latent space projection scores of the MCI cohort on the first two PLS components. Significant differences between EMCI (blue) and LMCI (orange), as revealed by the Mann Whitney non-parametric U-test, are highlighted in red for both phenotype and genotype features.

IV. Discussion

In this work we modeled the relation between gray matter atrophy and PRS via joint multivariate statistical modeling in AD, showing a good generalization of the results by testing the model on an unseen cohort of MCI subjects. Results showed that two PLS components explained a sufficient amount of data variability ($> 60\%$). Both components showed a significant separation between AD and CN in the latent space, confirmed also in the MCI projection. Moreover, the latent spatial distribution observed between AD and CN was replicated by the distribution of EMCI and LMCI in the same space.

The association between PRS and brain atrophy has been mainly addressed in literature via general linear model regression. In Scelsi *et al.* [6], for example, the authors focused on the hippocampus volume and found a significant negative correlation between such measure and AD specific PRS in cognitively impaired subjects, in line with our findings. The PRS association with a series of cortical features was explored by Sabunco *et al.* [9] on an healthy cohort. They calculate PRS involving up to 26 independent common sequence variants associated with AD and showed a correlation between late-onset AD PRS and cortical thickness in several AD-specific regions such as entorhinal cortex, temporopolar cortex, lateral temporal cortex, inferior parietal cortex, inferior parietal sulcus, posterior cingulate cortex, and inferior frontal cortex.

The PLS model, on the other side, is a well established method for multivariate analysis and has been widely employed in IG studies. In the work by Lorenzi and colleagues [10], it was used to link brain atrophy to the complete set of SNPs from AD patients, uncovering a significant link between the TRIB3 gene and the stereotypical pattern of grey matter loss in AD. They relied on few structural MRI features for collecting IDPs, while on the full set of SNPs for the genotype. A similar approach was followed in [13], where they were able to stratify the early stages of AD in the PLS latent space by exploiting T1-w features and cerebrospinal fluid levels of t-tau, p-tau and amyloid-beta biomarkers.

Thanks to the straightforward PLS explainabilty, we were able to recover the features leading the correlation between imaging and genetic features. The analysis of the weights associated to each feature can indeed allow to compare their relative importance and directly evaluate the genotype/phenotype association, highlighting those having a higher impact on the latent space derivation. In our model, the first component represented the great majority of data variability (54%) revealing an anticorrelation between less specific PRS scores and cortical thicknesses, that is inline with the well-known neurodegeneration process in AD. Indeed, the PRS included in this study were associated with disease progression and diagnosis, with an increasing score being correlated with the worsening of the disease. The negative correlation with the phenotype hence could be associated to a decrease in cortical thickness, typical of AD progression [2].

The second component, even if it explained a smaller fraction of the full data variability (18%), showed the most significant separation ($p < 1e - 12$) between AD and CN, for the phenotype, that was well preserved in the independent MCI cohort ($p = 0.007$). The PRS having the highest associated weights were the ones showing low p-value cut-offs, namely PRS_1e-07, PRS_1e-06 and PRS_1e-08, indeed scores that include established AD risk variants. Such PRS showed an anticorrelation with hippocampus volume and entorhinal cortex thickness among the others, being among the well known most affected regions in AD [2]. Of interest, they were also correlated with pallidum and caudate volumes, with the former showing the highest associated weight. Singleton and colleagues [14] have shown that a significant difference in pallidum volume was present between two AD subtypes, namely typical AD and behavioural AD, with the former featuring an increased pallidum volume compared to the latter. Moreover, Chen *et al.* [4] found a difference between AD subtypes related to the starting site of atrophy, and were able to identify three AD subtypes: (i) typical, for which atrophy begins in hippocampus and amygdala, (ii) cortical, where atrophy starts in the temporal lobe, followed by cingulate and insula and (iii) subcortical with atrophy beginning in pallidum, putamen and caudate. Therefore, we hypothesize that the second component obtained by our model could explain particular differences found across AD groups. In fact, it appears to explain data variability highly specific for AD, due to the high weights associated with the most conservative PRS. On the phenotype, at the same time, high importance was assigned to regions which have been demonstrated to play a role in AD subtypes identification. Further investigation is however needed to strengthen our hypothesis.

V. Conclusion

The presented PLS model confirms that there exists a joint variation between grey matter atrophy and PRS in AD, spreading over all the regions considered in the study. Moreover, we were able to capture volumetric modulations that possibly relate to different AD subtypes.

References

[1] M. J. Prince, F. Wu, Y. Guo, L. M. G. Robledo, M. O'Donnell, R. Sullivan, and S. Yusuf, "The burden of disease in older people and implications for health policy and practice," *The Lancet*, vol. 385, no. 9967, pp. 549–562, 2015.

[2] M. N. Braskie, A. W. Toga, and P. M. Thompson, "Recent advances in imaging alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 33, no. s1, pp. S313–S327, 2013.

[3] H. Cho, J.-H. Kim, C. Kim, B. S. Ye, H. J. Kim, C. W. Yoon, Y. Noh, G. H. Kim, Y. J. Kim, J.-H. Kim *et al.*, "Shape changes of the basal ganglia and thalamus in alzheimer's disease: a three-year longitudinal study," *Journal of Alzheimer's disease*, vol. 40, no. 2, pp. 285–295, 2014.

[4] H. Chen, E. de Silva, C. H. Sudre, J. Barnes, A. L. Young, N. P. Oxtoby, F. Barkhof, D. C. Alexander, and A. Altmann, "What do data-driven alzheimer's disease subtypes tell us about white matter pathology and clinical progression?" *Alzheimer's & Dementia*, vol. 17, p. e054028, 2021.

[5] T. Steckler and G. Salvadore, "Neuroimaging as a translational tool in animal and human models of schizophrenia," in *Translational Neuroimaging*. Elsevier, 2013, pp. 195–220.

[6] M. A. Scelsi, R. R. Khan, M. Lorenzi, L. Christopher, M. D. Greicius, J. M. Schott, S. Ourselin, and A. Altmann, "Genetic study of multimodal imaging alzheimer's disease progression score implicates novel loci," *Brain*, vol. 141, no. 7, pp. 2167–2180, 2018.

[7] E. C. Mormino, R. A. Sperling, A. J. Holmes, R. L. Buckner, P. L. De Jager, J. W. Smoller, M. R. Sabuncu, A. D. N. Initiative *et al.*, "Polygenic risk of alzheimer disease is associated with early-and late-life processes," *Neurology*, vol. 87, no. 5, pp. 481–488, 2016.

[8] H. Elshatoury, F. Cruciani, F. Zumerle, S. F. Storti, A. Altmann, M. Lorenzi, G. Anbarjafari, G. Menegaz, and I. B. Galazzo, "Disentangling the association between genetics and functional connectivity in mild cognitive impairment," in *2021 IEEE EMBS BHI*, 2021, pp. 1–4.

[9] M. R. Sabuncu, R. L. Buckner, J. W. Smoller, P. H. Lee, B. Fischl, R. A. Sperling, and A. D. N. Initiative, "The association between a polygenic alzheimer score and cortical thickness in clinically normal subjects," *Cerebral cortex*, vol. 22, no. 11, pp. 2653–2661, 2012.

[10] M. Lorenzi, A. Altmann, B. Gutman *et al.*, "Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics," *Proceedings of the National Academy of Sciences*, vol. 115, no. 12, pp. 3162–3167, 2018.

[11] A. Arnatkeviciute, B. D. Fulcher, M. A. Bellgrove, and A. Fornito, "Imaging transcriptomics of brain disorders," *Biological Psychiatry Global Open Science*, 2021.

[12] B. W. Kunkle, B. Grenier-Boley, R. Sims *et al.*, "Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing," *Nature genetics*, vol. 51, no. 3, pp. 414–430, 2019.

[13] A. Casamitjana, P. Petrone, J. L. Molinuevo *et al.*, "Projection to latent spaces disentangles pathological effects on brain morphology in the asymptomatic phase of alzheimer's disease," *Frontiers in neurology*, vol. 11, p. 648, 2020.

[14] E. H. Singleton, Y. A. Pijnenburg, C. H. Sudre, C. Groot, E. Kochova, F. Barkhof, R. La Joie, H. J. Rosen, W. W. Seeley, B. Miller *et al.*, "Investigating the clinico-anatomical dissociation in the behavioral variant of alzheimer disease," *Alzheimer's research & therapy*, vol. 12, no. 1, pp. 1–12, 2020.