Network regularization in imaging genetics improves prediction performances and model interpretability on Alzheimers's disease

Nicolas Guigui, C. Philippe, A. Gloaguen, S. Karkar, V. Guillemot, T. Löfstedt, V. Frouin

NETWORK REGULARIZATION IN IMAGING GENETICS IMPROVES PREDICTION PERFORMANCES AND MODEL INTERPRETABILITY ON ALZHEIMER'S DISEASE

N. Guigui[†⋆], C. Philippe[†], A. Gloaguen[†⋆], S. Karkar[†], V. Guillemot[‡], T. Löfstedt[§], V. Frouin[†]

[†] CEA, Neurospin, Gif-sur-Yvette, 91191, France
[⋆] CentraleSupelec, Gif-sur-Yvette, 91190, France
[‡] Pasteur Institute, Bioinformatics and Biostatistics Hub, C3BI, USR 3756 IP CNRS, Paris, France
[§] Department of Radiation Sciences, Umeå University, Umeå, Sweden

ABSTRACT

Imaging-genetics is a growing popular research avenue which aims to find genetic variants associated with quantitative phenotypes that characterize a disease. In this work, we combine structural MRI with genetic data structured by prior knowledge of interactions in a Canonical Correlation Analysis (CCA) model with graph regularization. This results in improved prediction performance and yields a more interpretable model.

***Index Terms*—** Imaging-Genetics, Networks, Structured constraints, Generalized Canonical Correlation Analysis

1. INTRODUCTION

The joint study of brain images and genetic data is of growing interest to find associations between genetic variants and disease-related features that can be measured on brain MRI. The rationale behind this research avenue is that imaging endophenotypes stand as proxies to disease status and evolution, that can be measured with non-invasive methods to facilitate early diagnosis and follow-up. However, genetic information is expected to provide insights into the underlying disease mechanisms and valuable tracks to finding therapies.

Existing studies seek pairwise associations between a set of genetic variants and brain regions or at the whole genome and whole brain scale [1]. In this work, we demonstrate how combining multiple sources and layers of information, i.e. imaging and genetic experimental data and prior knowledge, with data integration methods brings supplemental information, both in terms of performance and interpretability compared to univariate approaches. The aim of data integration methods is to account for known interactions between variables and data from each modality.

Nonetheless, they raise many methodological challenges as both data types are very high-dimensional (10^4 variables) [2]. Canonical Correlation Analysis (CCA), and its regularized generalization to multiple blocks, RGCCA [3], is a framework that allows to model and quantify interactions within and between blocks of variables of heterogeneous types, and with respect to the disease status. Variants of RGCCA and CCA allow to select the variables that are the most relevant in the model [4, 5, 6].

However, biomarker identification from genotyping data usually fails to provide biologically relevant associations. This is because, different alterations in the same biological pathway often lead to the same pathological phenotype, but they will all fail to pass significance thresholds. Moreover, empirical evidences show that many segregating variants affect multiple traits and a precise estimation of the proportion of such variants remains elusive [7]. Therefore including genetic interaction networks as prior knowledge to wisely concentrate the information from genome-wide disseminated single-nucleotide polymorhpisms (SNPs) is a promising approach.These can be used in a data pre-processing step or to compute a regularization constraint [8], or both as in [9]. Azencott [8] gives a review of statistical methods to regularize variable selection with graphs, however the focus is mainly on univariate regression.

Following the framework formulated in [10], we present here the use of the GraphNet penalty on a multiblock CCA model at a genome-wide scale. We use the graph Pathway Commons (PC) (`http://www.pathwaycommons.org`) as prior knowledge to predict the disease status and illustrate our method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (`http://adni.loni.usc.edu/`). In Section 2, we present the method and evaluation scheme, then in section 3 we present the data used to assess the method. In section 4 we present the results and discuss them in section 5.

2. METHOD

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Z} \in \mathbb{R}^{n \times q}$ represent two blocks of variables, for n individuals with respectively p and q features (e.g. imaging and genetics data). We denote by $\mathbf{y} \in \{0, 1\}^n$ the disease status or target. $\mathbf{X}$ and $\mathbf{Z}$ are scaled respectively by $\sqrt{p}$ and $\sqrt{q}$ to balance for the different number of features [11] and we assume that $\mathbf{X}, \mathbf{Z}, \mathbf{y}$ have been centred and scaled to unit variance.

Structured Sparse CCA The SGCCA [4] optimization problem in the context of two blocks and a target is the following:

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^p \\ \mathbf{z}\in\mathbb{R}^q}} \mathcal{C}(\mathbf{x},\mathbf{z}) = \mathbf{x}^T\mathbf{X}^T\mathbf{y} + \mathbf{z}^T\mathbf{Z}^T\mathbf{y} + \mathbf{x}^T\mathbf{X}^T\mathbf{Z}\mathbf{z} \quad (1)$$

$$\text{s.t. } \|\mathbf{x}\|_2^2 = \|\mathbf{z}\|_2^2 = 1 \text{ and } \|\mathbf{x}\|_1 \leq s_1,\ \|\mathbf{z}\|_1 \leq s_2 \quad (2)$$

The first two terms aim at maximizing the covariance between the two latent variables representing each block and the target, while the third term is the interaction between the two blocks. The problem is constrained with an ℓ_1-norm with parameters s_1 and s_2, in order to obtain sparse loadings $\mathbf{x}$ and $\mathbf{z}$, to select the key variables in the model.

We propose to add a structured penalty to the objective function of SGCCA in order to select variables that are known to interact. Let $\mathcal{G} = (V, E)$ be a graph with q vertices (e.g. the genes), representing the known interactions between the q variables. Let $\mathbf{A} \in \mathbb{R}^{q\times q}$ be its adjacency matrix, and $\mathbf{L} \in \mathbb{R}^{q\times q}$ be its Laplacian matrix. The penalty is composed of two parts, the first enforces the weights of variables that are connected in $\mathcal{G}$ to be close to one another. The second part is an ℓ_2-norm penalty for isolated nodes, as they would be more likely to be selected otherwise. Let $1 \leq d < q$ be the number of isolated variables in $\mathcal{G}$, and suppose variables are ordered so that these nodes are the last d columns of $\mathbf{Z}$. Let $\mathbf{J_d}$ be the $q \times q$ diagonal matrix with the last d coefficients equal to 1, all others to 0:

$$P_L^{\lambda,\gamma}(\mathbf{z}) = \lambda\frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q} A_{ij}(z_i - z_j)^2 + \gamma \sum_{j=q-d}^{q} z_j^2 \quad (3)$$

$$= \mathbf{z}^T(\lambda\mathbf{L} + \gamma\mathbf{J_d})\mathbf{z} \quad (4)$$

When used in combination with the ℓ_1 norm, this penalty is known as GraphNet [12]. The final optimization problem is thus:

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^p \\ \mathbf{z}\in\mathbb{R}^q}} f(\mathbf{x},\mathbf{z}) = \mathcal{C}(\mathbf{x},\mathbf{z}) - P_L^{\lambda,\gamma}(\mathbf{z}) \quad (5)$$

$$\text{s.t. } \|\mathbf{x}\|_2^2 \leq 1, \|\mathbf{z}\|_2^2 \leq 1 \text{ and } \|\mathbf{x}\|_1 \leq s_1,\ \|\mathbf{z}\|_1 \leq s_2 \quad (6)$$

where $\lambda, \gamma, s_1, s_2$ are regularization hyper-parameters. We henceforth call this model GraphNet-GCCA (GN-GCCA). Notice that the ℓ_2 constraints (2) are changed to inequality constraints in (6) in order to make the feasible set convex. For certain choices of parameters s_1 and s_2, the ℓ_2 constraints will be active at the optimum solution as discussed in [5].

Model fitting The problem does not have an analytical solution and is non-convex, but it is multi-convex in the sense that the loss function $-f$ is convex when considered as a function of each individual block loading, while the others are kept fixed. Minimizing such functions under convex constraints can be achieved with a so called block relaxation algorithm. It consists in alternating between the blocks to minimize the partial loss w.r.t each vector, as in the procedure described in [10, alg. 2]. Let $\mathcal{P} \subset \mathbb{R}^p$ and $\mathcal{Q} \subset \mathbb{R}^q$ be the convex sets where the constraints (6) are satisfied, and $\text{proj}_\mathcal{P}$ and $\text{proj}_\mathcal{Q}$ the orthogonal projections on those sets. Details to compute these projections are given in [10]. Each partial constrained minimization problem is solved with a few FISTA iterations [13] with step sizes α_x and α_z. Algorithm 1 describes the full procedure. This algorithm was implemented using the pylearn-parsimony package (`https://github.com/neurospin/pylearn-parsimony`).

Algorithm 1 Block relaxation algorithm for GN-GCCA, with desired precision ϵ

Initialize: $\mathbf{x}^{(0)}, \mathbf{z}^{(0)}$
repeat
  for $k = 1..\text{max\_iter}$ **do**
    $\text{grad}_x \leftarrow \mathbf{X}^T(\mathbf{y} + \mathbf{Z}\mathbf{z}^{(t)})$
    $\mathbf{x}^{(t+1)} \leftarrow \text{proj}_\mathcal{P}(\mathbf{x}^{(t)} + \alpha_x \cdot \text{grad}_x)$
  end for
  for $k = 1..\text{max\_iter}$ **do**
    $\text{grad}_z \leftarrow \mathbf{Z}^T(\mathbf{y} + \mathbf{X}\mathbf{x}^{(t+1)}) - \tilde{\mathbf{L}}\mathbf{z}^{(t)}$
    $z^{(t+1)} \leftarrow \text{proj}_\mathcal{Q}(\mathbf{z}^{(t)} + \alpha_z \cdot \text{grad}_z)$
  end for
until $f(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - f(\mathbf{x}^{(t+1)}, \mathbf{z}^{(t+1)}) \leq \epsilon$
return $\mathbf{x}^{(t+1)}, \mathbf{z}^{(t+1)}$

Model selection and evaluation The previous model has four regularization hyperparameters: s_1 and s_2 control the sparsity, while λ controls the *roughness* of $\mathbf{z}$ over $\mathcal{G}$, and γ controls the shrinkage for isolated features because of its interaction with λ, this parameter was set manually . To compare our method with SGCCA, we set aside a test set and a training set. The prediction performance is evaluated by fitting a Linear Discriminant Analysis (LDA) model on the projections $t_1 = \mathbf{X}\mathbf{x}$ and $t_2 = \mathbf{Z}\mathbf{z}$. This allows to seek the best hyperparameters, assessing them in a 5-fold cross-validation (CV) scheme, on the training set. We then choose the combination of parameters that achieved the highest area under the ROC curve (AUC) on the classification task with target $\mathbf{y}$, and train the model on the whole training set with this choice of parameters. Finally the AUC is computed for this model on the test set.

3. DATA AND EXPERIMENTS

We used brain structural MRI data and genotype data from the ADNI database. The individuals are classified in four categories: Healthy Control (HC), Mild Cognitive Impairment (MCI), MCI who converted to AD in the 24 months following screening (MCIC), and Alzheimer's Disease (AD).

The Imaging data We used the computed mean cortical thickness for 75 ROI for each hemisphere and subcortical volumes for 63 brain regions with FreeSurfer on

the Aparc2009 atlas, resulting in $p = 213$ features. Images for 406 individuals were available. We corrected the features for age and sex with a linear regression (the model is fitted separately on each fold of the CV).

The genetic data SNPs data were pruned for missing genotypes, Hardy-Weinberg test and minor allele frequency, and discarded when on chromosome Y or mitochondrial chromosomes, resulting in 491616 loci. As in PC, vertices are genes and not SNPs, we computed a score for each gene to reflect the burden of SNPs contained in this gene and its flanking regions (40kb upstream and downstream), thus accounting for *cys*-regulating SNPs. For an individual i, the state of the SNP s_k is encoded as $\alpha_{ik} \in \{0, 1, 2\}$ for the number of alternative alleles, w.r.t the reference genome. A multiple-SNP risk score for each gene is then computed as a weighted sum of the allelic counts α_{ik} of each SNP:

$$X_{ig} = \sum_{k=1}^{p_g} \beta_k \alpha_{ik}$$

for individual i and gene g, where β is the log odds ratio estimated in the IGAP meta-analysis [14]. p_g is the number of SNPs mapped to gene g. Genetics data where available for 757 patients and SNPs were mapped to $q = 19217$ genes.

The Network In this work, we used a network of physical interactions from Pathways Commons (PC), keeping only undirected, highly confident edges i.e. interactions supported by PubMed references with at least 2 different sources. It was extracted with XGR (eXploring Genomic Relations) [15].

4. RESULTS

Performance To probe the model's performance in learning early signs of the disease, we used the individuals with status HC and AD to train the models ($n = 214$ individuals), and individuals with status MCI or MCIC as test set ($n = 157$). The predictive power of the genetics data alone, and of the imaging data alone were assessed with an ℓ_1, ℓ_2 regularized logistic regression (we refer to these models as LR Gen and LR Im). We also assessed the impact of the GraphNet penalty with a logistic regression and ℓ_1 regularization (LR GN Gen). The mean performance in the CV of the selected model and the performance on the test set of the best model trained on the whole training data are reported for the five models in Table 4. We also reported the number of selected features (signature). s_2 clearly stood out as a key parameter for performance and we observed better CV-mean AUC for smaller values. λ was searched in a range of small values ($10^{-3} - 10^2$) to avoid $z = 0$ from being an optimal but uninteresting solution.

Selected brain regions The brain ROI selected by GN-GCCA on the structural MRI data are represented in Figure 1.

Model	**Signature**	**CV Score** (Acc/AUC)	**Test Score** (Acc/AUC)
LR Gen	19200 genes	0.53/0.56	0.52/0.55
LR GN Gen	35 genes	0.61/NC	0.57/0.62
LR Im	42 ROI	0.88/0.94	0.68/0.72
SGCCA	2 genes 23 ROI	0.90/0.95	0.68/0.74
GN-GCCA	196 genes 18 ROI	0.86/0.94	0.70/0.75

Table 1. Results of the experiments for models with genetics data only (LR Gen, LR GN Gen), imaging only (LR Im) and both (SGCCA, GN-GCCA).

The regions selected consist in areas where atrophy is classically observed in the early stage of AD: hippocampus and temporal lobes, but not the exhaustive set of involved regions [16]. These might be the regions where AD related atrophy is linked to a specific molecular mechanism.

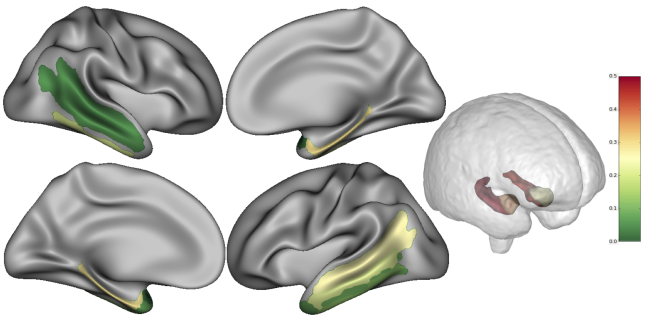

Fig. 1. Visualization of the brain ROI selected by GraphNet-GCCA

Selected gene network The subgraph formed by the 196 genes selected by GN-GCCA is represented Figure 2. Among these, five are linked to AD by previous publications (APOC1, IFNA4, NDUFA5, SDHC, TOMM40). There are nine connected components (165 singletons). The biggest one comprises seven genes. A functional analysis using DAVID (`https://david.ncifcrf.gov/`) shows that three genes belong to the Reactome Pathway R-HSA-977225 Amyloid fiber formation (underlined in Fig 2).

5. DISCUSSION AND CONCLUSION

We demonstrated a significant gain in prediction performance when integrating imaging and genetic data to build predictive models of AD. The gain is even more important when adding structured prior knowledge to the genetic data. Du *et al.* in [6] and references therein investigated graph regularized CCA on ADNI on up to a few hundred candidate SNPs, and a few hundred imaging ROIs while our method is tested on the whole genome. Moreover, the graphs they used were weighted by sample correlations and did not leverage prior

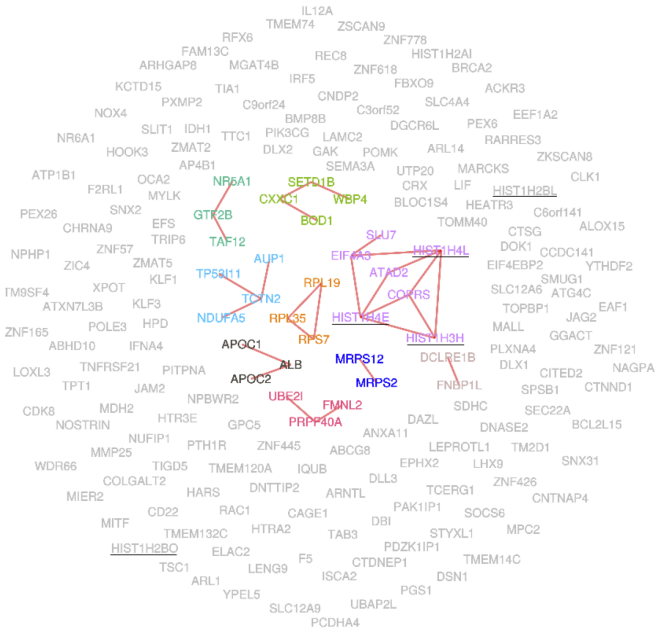

Fig. 2. Visualization of the subgraph formed by the genes selected by GN-GCCA. Colors show different connected components and genes that belong to HSA 977225 are underlined

knowledge of gene interactions. Their algorithm also considered the ℓ_1 norm to be smooth by adding a small constant to zero weights. Moreover their approach is not scalable to the whole-genome as it requires inverting a $p \times p$ matrix at each iteration. In comparison, our method uses an ℓ_1 projection that naturally results in sparse solutions. Lorenzi *et al.* [16] instead applied a hard threshold on many bootstrapped samples to select SNP-sets that are then mapped back to genes, and did not use graph constraints. None of these methods jointly use the target and both data types to compute the canonical loadings. Finally we identified several pathways that seem to play a role in the disease at an early stage. Note that a similar regularization could be applied to the imaging data where the graph would encode spatial proximity or connections between regions.

6. REFERENCES

[1] L. Shen *et al.*, "Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 183–207, 6 2014.

[2] F. S. Nathoo, L. Kong, and H. Zhu, "A Review of Statistical Methods in Imaging Genetics," 2018.

[3] A. Tenenhaus and M. Tenenhaus, "Regularized Generalized Canonical Correlation Analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, 4 2011.

[4] A. Tenenhaus *et al.*, "Variable selection for generalized canonical correlation analysis," *Biostatistics*, vol. 15, no. 3, pp. 569–583, 7 2014.

[5] D. M. Witten, R. J. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 7 2009.

[6] L. Du *et al.*, "Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method," *Bioinformatics*, vol. 32, no. 10, pp. 1544–1551, 5 2016.

[7] P. M. Visscher *et al.*, "10 Years of GWAS Discovery: Biology, Function, and Translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 7 2017.

[8] C.-A. Azencott, "Network-Guided Biomarker Discovery," in *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, A. Holzinger, Ed. Cham: Springer International Publishing, 2016, pp. 319–336.

[9] M. Hofree *et al.*, "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10, no. 11, pp. 1108–1115, 11 2013.

[10] T. Löfstedt *et al.*, "Structured Variable Selection for Regularized Generalized Canonical Correlation Analysis." Springer, Cham, 2016, pp. 129–139.

[11] M. Tenenhaus, A. Tenenhaus, and P. J. F. Groenen, "Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods," *Psychometrika*, vol. 82, no. 3, pp. 737–777, 9 2017.

[12] L. Grosenick *et al.*, "Whole-brain Sparse Penalized Discriminant Analysis for Predicting Choice," *NeuroImage*, vol. 47, p. S58, 7 2009.

[13] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 1 2009.

[14] V. Escott-Price *et al.*, "Common polygenic variation enhances risk prediction for Alzheimers disease," *Brain*, vol. 138, no. 12, pp. 3673–3684, 12 2015.

[15] H. Fang, B. Knezevic, K. L. Burnham, and J. C. Knight, "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits." *Genome medicine*, vol. 8, no. 1, p. 129, 2016.

[16] M. Lorenzi *et al.*, "Susceptibility of brain atrophy toTRIB3in Alzheimer's disease, evidence from functional prioritization in imaging genetics." *PNAS*, vol. 115, no. 12, pp. 3162–3167, 3 2018.