ENFORCING CO-EXPRESSION IN MULTIMODAL REGRESSION FRAMEWORK

PASCAL ZILLE[1], VINCE D. CALHOUN[2] and YU-PING WANG[1,*]

[1] *Biomedical Engineering Department, Tulane University.*
[2] *The Mind Research Network, University of New Mexico.*
[*] *E-mail: wyp@tulane.edu*

We consider the problem of multimodal data integration for the study of complex neurological diseases (e.g. schizophrenia). Among the challenges arising in such situation, estimating the link between genetic and neurological variability within a population sample has been a promising direction. A wide variety of statistical models arose from such applications. For example, Lasso regression and its multitask extension are often used to fit a multivariate linear relationship between given phenotype(s) and associated observations. Other approaches, such as canonical correlation analysis (CCA), are widely used to extract relationships between sets of variables from different modalities. In this paper, we propose an exploratory multivariate method combining these two methods. More Specifically, we rely on a 'CCA-type' formulation in order to regularize the classical multimodal Lasso regression problem. The underlying motivation is to extract discriminative variables that display are also co-expressed across modalities. We first evaluate the method on a simulated dataset, and further validate it using Single Nucleotide Polymorphisms (SNP) and functional Magnetic Resonance Imaging (fMRI) data for the study of schizophrenia.

Keywords: Multimodal Analysis, Collaborative Regression, CCA, Sparse Models, Schizophrenia.

1. Introduction

An increasing amount of high-dimensional biomedical data such as micro arrays (mRNA, SNP) or brain imaging sequences (MRI, PET) is collected every day. Classical unimodal analysis often ignore the potential joint effects that may exist, for example, between genes and specific brain regions for diseases such as Schizophrenia, Alzheimer, etc. By harnessing these joint effects across modalities, we might be able to identify new mechanisms that uni-modal methods may fail to capture. Imaging genomics is an emerging field whose aim is precisely to leverage the wealth of biomedical knowledge provided by genomic and brain imaging data. Integrating such multimodal data sets is critical to extract meaningful bio-markers, improve clinical outcome prediction or identify potential associations across modalities. Unfortunately, as mentioned by Lin[1], such studies using genomic and brain imaging data often run into two limitations: The first one is an average small sample size, which may result in over fitting issues. In order to address such constraint, many authors relied on the use of sparse models. One classical method introduced by Tibshirani[2] is the Lasso regression. The second limitation is poor biomarker reproducibility across studies. Although this issue remains an open problem, one may hope that using appropriate priors over the solution will lead to an improved consistency of the result across different studies.

1.1. *Motivation: the study of Schizophrenia*

Schizophrenia is a serious neurological disorder that affects around 1% of the general population. It is regarded as the result of various factors including genetic variants, brain development abnormalities and environmental effects. Identifying critical genes or SNPs related to schizophrenia[3,4] has been a challenging issue. Many studies relied as well on brain imaging techniques[5,6] to pinpoint functional abnormalities in brain regions for schizophrenia patients. Multimodal analysis (e.g. using both genomic and brain imaging) often improve generalization in situations in which many irrelevant features are present. In their recent paper, Cao et al.[7] proposed a sparse representation based variable selection (SRVS) algorithm relying on sparse regression model to integrate both SNP and fMRI in order to perform biomarker selection for the study of schizophrenia. Lin[8] proposed a group sparse canonical correlation analysis (CCA) method based on SNP and fMRI data to extract correlation between genes and brain regions. Le Floch et al.[9] combined univariate filtering and Partial Least Squares (PLS) to identify SNPs covarying with various neuroimaging phenotypes. It appears that both regression and CCA methods display promising behaviors when combining SNP and fMRI data for the study of schizophrenia. In this work, we will try to merge these two methods in order to make the most out of both formulations.

The rest of this paper is organized as follows: we introduce in Section 2 some of the relevant methods as well as the motivation for this work. A novel approach to multivariate regression problems is then proposed in Section 3. Such method is then evaluated on both synthetic and real datasets in Section 4, followed by some discussions and concluding remarks in Section 5.

2. Methods

2.1. *Learning with L_1 penalty*

We consider $M \in \mathbb{N}^+$ distinct (i.e. from different modalities) datasets with n samples and $p_m \in \mathbb{N}^+$ $(m = 1, .., M)$ variables each. The m-th dataset is represented by a matrix $\mathbf{X}_m \in \mathbb{R}^{n \times p_m}$. Additionally, each sample is assigned a class label (e.g. case/controls) $y_i \in \{-1, 1\}$, $i = 1, .., n$. Our goal is to look for a linear link between those class labels and the M data matrices. Let us consider the following regression model:

$$\min_{\beta} \sum_{m=1}^{M} \|\mathbf{y} - \mathbf{X}_m \beta_m\|_2^2 + \lambda \|\beta\|_1 \tag{1}$$

The model described by Eq. 1 performs both variable selection and regularization. It often improves the prediction accuracy and interpretability of the results compared to the use of classical ℓ_2 norm regularization terms, especially when the number of variables is far greater than the number of observations. In some situations, we have several output vectors $\mathbf{y}_m, \forall m = 1, .., M$ and the m datasets are from the same modality: multi-task Lasso[10] was proposed to capture shared structures among the various regression vectors. We consider the following model:

$$\min_{\beta} \sum_{m=1}^{M} \|\mathbf{y}_m - \mathbf{X}_m \beta_m\|_2^2 + \lambda \sum_{p=1}^{P} \|\beta^p\|_2 \tag{2}$$

where P is the dimension of the problem and β^p is the p-th row of the matrix such that $\beta = [\beta_1, .., \beta_m]$ (i.e. the β_m are stacked horizontally). Such norm is also referred to as the ℓ_1/ℓ_2 norm, and is used to both enforce joint sparsity across the multiple β_m and estimate only a few non-zero coefficients. Enforcing regularity within a modality[11,12] (and across tasks) has been an active aspect of regression models, and has proven to increase reliability and results. However, since often pair-wise closeness is looked for in the common subspace, such methods will often fail to capture relationships across modalities.

2.2. *Collaborative learning*

Collaborative (or Co-regularized) methods[13] are based on the optimization of measures of agreement and smoothness across multi-modal datasets. Smoothness across modalities is enforced through a joint regularization term. Their general model can be expressed as follows:

$$J(\beta) = \sum_{m=1}^{M} \|\mathbf{y} - \mathbf{X}_m\beta_m\|_2^2 + \gamma \sum_{m,q=1}^{M} \|\mathbf{U}_m\beta_m - \mathbf{U}_q\beta_q\|_2^2 + \lambda\|\beta\|_1 \tag{3}$$

where the $\mathbf{U}_m,\ m = 1, .., M$ are arbitrary matrices whose roles are to control the cross-view joint regularization between each pair of vectors $(\beta_m, \beta_q),\ m, q = 1, .., M$. Scalar parameter $\gamma \geq 0$ controls the influence of such cross-regularization term. Notice that if $\gamma = 0$, we fall back on the original Lasso formulation. Collaborative learning is an interesting extension of Eq.1 allowing the user to explicitly enforce regularization across modalities. In this work, we rely on a special case of collaborative methods (introduced later in section 3) to address the following aspects: (i) Extend the regularization idea across modalities; (ii) Assume that relationships between variable are not available as a prior knowledge (as opposed, e.g., to Xin[11]); (iii) Define links between components using correlation measure. To do so, we first briefly introduce in the next section some of the classical methods to extract meaningful relationships between variables across modalities.

2.3. *Extracting relationship between datasets*

A wide variety of problems amount to the joint analysis of multimodal datasets describing the same set of observations. Often, a mean to perform such analysis is to learn projection subspaces using paired samples such that structures of interest appear more clearly. Some of these methods are for example: Canonical correlation analysis[14] (CCA), Partial least squares[9] (PLS) or cross-modal factor analysis (CFA). Among them, CCA is probably the most widely used. Its goal is to extract linear combinations of variables with maximal correlation between two (or more) datasets. Using similar notations as in the previous section,and assuming $M = 2$, one formulation of CCA is expressed as follow:

$$\underset{\beta_1,\beta_2}{\operatorname{argmin}}\ J_{cca}(\beta_1, \beta_2) \ = \ \|\mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2\|^2 \tag{4}$$

to which a constraint on the norm of canonical vectors β_1, β_2 is added to avoid the trivial null solution. In recent years, CCA has been widely applied to genomic data analysis. As a consequence, many studies on sparse versions of CCA (sCCA) have been proposed[8,15–18] to

cope with the high dimension but low sample size problem. In the next section, we will rely on a CCA term to measure co-expression between variables from different modalities.

3. Enforcing cross-correlation in regression problems

3.1. *MT-CoReg formulation*

As discussed in Section 1, several methods have been proposed to: (*i*) Associate a phenotype and datasets while enforcing prior over solution; (*ii*) Extract relationships between coupled or co-expressed datasets. In the present study, we propose to associate both the regression and CCA frameworks in the case of $M = 2$ datasets. Our motivation is to extract informative features that also display a significant amount of correlation across modalities. A simple way to combine Lasso and sparse CCA would be a weighted combination of Eq.(1) and Eq.(4):

$$\min_{\beta} J(\beta) = (1-\gamma) \sum_{m=1}^{2} \|\mathbf{y} - \mathbf{X}_m \beta_m\|_2^2 + \gamma \|\mathbf{X}_1 \beta_1 - \mathbf{X}_2 \beta_2\|^2 + \lambda \|\beta\|_1 \tag{5}$$

where $\gamma \in [0, 1]$ is a weight parameter. Notice that Eq.(5) can be expressed within the collaborative framework introduced in Section 2.2. If we take a look at Eq.(3) with $M = 2$, $\mathbf{U}_1 = \mathbf{X}_1$ and $\mathbf{U}_2 = \mathbf{X}_2$, we fall back on Eq.(5). Let us call this model CoReg for *Collaborative Regression*. Interestingly, a similar model has been considered before by Gross[19] to perform prediction using breast cancer data. However, to our opinion, such formulation might prove to be too constraining. It essentially amounts to force each component of the β_m's to fit both the regression term and the CCA one. We illustrate such behaviour using a toy dataset later in Section 3.4. Since our goal is to perform feature selection, we may allow the model to be slightly more flexible. We thus propose an alternative formulation by first duplicating each β_m into two components such that:

$$\beta_m = [\alpha_m, \theta_m] \ , \ \forall m = 1, 2 \tag{6}$$

where α_m, θ_m are vectors from $\mathbb{R}^{p_m}$. As a consequence, the β_m's are now matrices such that $\beta_m \in \mathbb{R}^{p_m \times 2}$ $\forall m = 1, 2$. We then propose the following MT-CoReg formulation:

$$\min_{\beta} J(\beta) = (1-\gamma) \sum_{m=1}^{2} \|\mathbf{y} - \mathbf{X}_m \alpha_m\|_2^2 + \gamma \|\mathbf{X}_1 \theta_1 - \mathbf{X}_2 \theta_2\|^2 + \lambda \sum_{m=1}^{2} \sum_{i=1}^{p_m} \|\beta_m^i\|_2 \tag{7}$$

where β_m^i is the i-th row of β_m, i.e. $\beta_m^i = [\alpha_m(i), \theta_m(i)] \in \mathbb{R}^2$. The third term of Eq.(3.3) is simply the ℓ_1/ℓ_2 norm of each of the β_m. As we can observe from looking at Eq.(3.3), each 'component' (i.e. column of β_m) will be involved in separate parts of the functional J: (*i*) components α_m are the fit to the regression term of Eq.(3.3); (*ii*) components θ_m are the fit of the CCA term of Eq.(3.3). Each pair (α_m, θ_m) and $m = 1, 2$ is coupled through the use of the ℓ_1/ℓ_2 norm from the third term in Eq.(3.3). Although their values are different, shared sparsity patterns are encouraged within each pair (α_m, θ_m). As a consequence, we allow the method to be significantly more flexible in terms of solutions: different values can be taken to simultaneously fit the Regression and CCA parts. We hope that such framework will encourage the selection of features that are discriminative (via the regression part) but also co-expressed across modalities (via the CCA part). Note that when $\gamma = 0$, criterion (3.3)

essentially reduces to the initial regression problem of Eq.(1), while setting $\gamma = 1$ amounts to solving a conventional sparse CCA problem. A schematic view of the MT-CoReg pipeline can be seen in Fig.(1). In the next section, we briefly explain how to solve the problem described in Eq.(3.3).

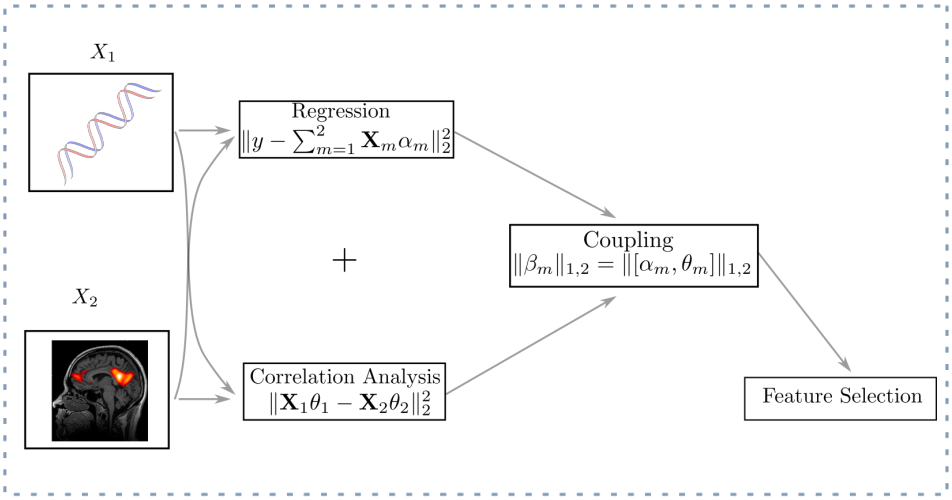

Fig. 1. Schematic view of the MT-CoReg pipeline. From two different datasets $X1$ and $X2$ from different modalities (here SNP and fMRI respectively), we fit both a regression and CCA terms and couple the resulting components (α_m, θ_m) using the $\ell 1/\ell_2$ norm denoted $\|\cdot\|_{1,2}$ here. The ultimate goal is to find discriminative SNP and brain regions that are also co-expressed across modalities.

3.2. *Optimization*

We solve the problem from Eq.(3.3) by optimizing the β_m's alternatively over iterations until convergence, in a similar fashion to Wilms[20] et al. formulation of sCCA. Suppose we have an initial value β_1^* for β_1, and want to estimate β_2. Updating matrix β_2 can be recast into a problem of the following form:

$$\min_{\tilde{\beta}_2} J(\tilde{\beta}_2|\beta_1^*) = \|\tilde{\mathbf{y}}_2 - \tilde{\mathbf{X}}_2\beta_2\|_F^2 + \lambda \sum_{i=1}^{p_2} \|\beta_2^i\|_2 \tag{8}$$

where

$$\tilde{\mathbf{y}}_2 = [\sqrt{(1-\gamma)}\mathbf{y}, \sqrt{\gamma}\mathbf{X}_1\theta_1^*], \quad \tilde{\mathbf{X}}_2 = [\sqrt{(1-\gamma)}\mathbf{X}_2, \sqrt{\gamma}\mathbf{X}_2] \tag{9}$$

Obviously, Eq.(8) is a classical group-lasso regression problem[10] (cf. Eq.(2)). It is easy to show that updating β_1 reduces to solving a similar problem. As a consequence, solving our mixed Lasso/CCA problem from Eq.(3.3) can be briefly summarized as:

1 Initialization: estimate initial values for α_1, β_1, α_2, β_2 using ridge regression and ridge CCA.
2 Assume β_1's value fixed, and update β_2 using Eq.(8).
3 Assume β_2's value fixed, and update β_1 using the adapted version of Eq.(8).
4 Go back to step 2. until convergence

3.3. *Parameter selection*

Solving problem from Eq.(3.3) requires the estimation of two parameters, λ and γ, which respectively control the weights of the sparsity and the co-expression regularization terms.
The choice of sparsity parameter λ for this type of problems is known to display a high sensitivity[21]. In order to make the searching process more robust, we chose to let the sparsity level of the solution control the tuning parameter value[22,23]. Consider a column vector $\beta \in \mathbb{R}^p$ (e.g. a column of β from Eq.): let us denote $|\beta|_\kappa$ the κ-th ($\kappa \in \mathbb{N}^+$) largest absolute magnitude of β. We can define a correspondence between λ and κ by making sure that for each iteration, we have $\lambda \in [|\beta|_\kappa, |\beta|_{\kappa+1}]$. The selection can be looked for around the sample size (i.e. $\kappa = n$ for the entire estimation process), which helps drastically stabilize the estimation process in practice.
As for the estimation of γ, we chose to rely on a technique introduced by Sun et al.[24] based on variable selection stability. Its main goal is to select a given tuning parameter so that the associated variable selection method (in our case, the model from Eq.(3.3)) is stable in terms of the features it selects. In this framework, the training set is split in two halves using resampling (bootstrap resampling in our case). The variable selection method is then applied to each of the subsamples along a grid of candidate values for the parameter. Kappa selection criterion[25] is then used to measure the degree of agreement between the two sets of variables obtained for a given parameter value. This process is then repeated a number of times, and an approximated measure of selection consistency is derived. The parameter value for which this consistency is the highest (after correction for the number of non-zeros elements retained) is the one kept for the estimation.

3.4. *MT-CoReg VS. CoReg*

As mentioned earlier in Section 3.1, in their CoReg model from Eq.(5) Gross et al.[19] did not separate the solution vectors β_m into two components. We then propose to illustrate the behavior of both models (Eq.(5) and Eq.(3.3)) on a toy dataset.
We generated $M = 2$ data matrices $\mathbf{X}_1, \mathbf{X}_2$ such that $p_1 = p_2 = 30$ and $n = 50$ observations. We used a latent variable model to simulate cross-correlated components so that columns $p = [1, ..5] \cup [10, ..15]$ of $\mathbf{X}_1, \mathbf{X}_2$ are mutually co-expressed. We further use columns $p = [10, ..15] \cup [20, ..25]$ to generate a phenotype vector $\mathbf{y}$ such that $\mathbf{y}_i \in \{-1; 1\}$. With such setup, columns $p = [10, ..15]$ correspond to both non-zeros values in the true regression and canonical coefficients. Furthermore, let us point out that these non-zero values are diferent (canonical coefficients' amplitude is lower than the regression ones). This setup can be seen in the first row of Fig.(2, *Truth*), where the blue and red curves are the values taken by the canonical and regression coefficients respectively. Resulting estimates for sCCA, Lasso, CoReg[19] as well as proposed method MT-CoReg can also be seen in Fig.(2). In such scenario, while CoReg model assumes that regression and canonical coefficients have identical values, MT-CoReg has a wider scope and allows a finer joint estimation of both components types.

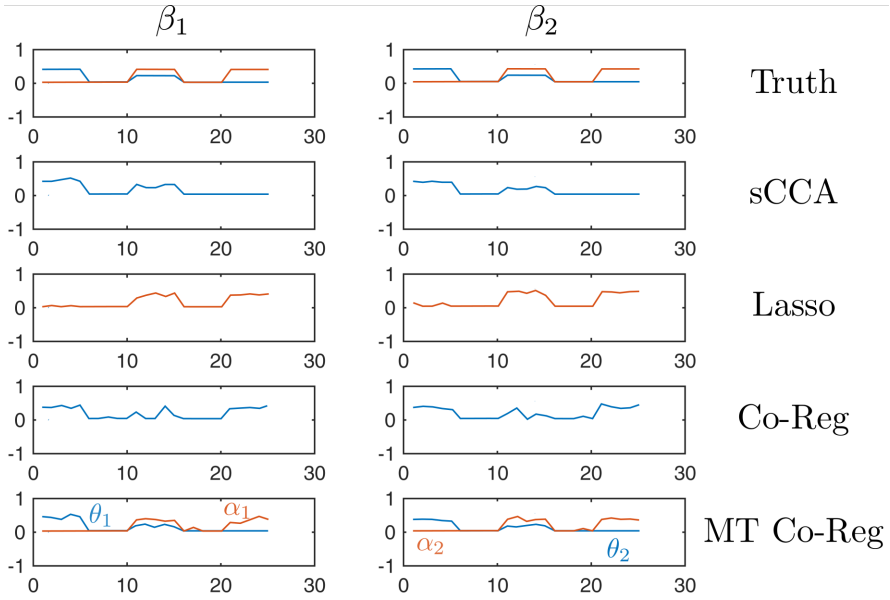

Fig. 2. Resulting estimates β_1, β_2 on the toy dataset. (*Truth*) blue and red curves are the values taken by the true canonical and regression coefficients respectively. Solutions obtained with sCCA, Lasso, CoReg[19] and proposed method MT-CoReg are displayed. Notice that columns columns $p = [10, ..15]$ correspond to both non-zeros values in the true regression and canonical coefficients, although their amplitudes are different. By relaxing the assumption that regression and canonical coefficients have identical values, MT-CoReg allows a finer joint estimation of both components types compared to CoReg.

4. Experiments

In this section, we evaluate the proposed estimator from Eq.3.3. Performances will be assessed in terms of feature selection relevance on both simulated and real data.

4.1. *Results on synthetic data*

For our first test, we simulate both fMRI and SNP datasets. Similar to the toy dataset from Section 3.4, we start by generating explanatory variables $\alpha_1^*, \alpha_2^* \in \mathbb{R}^{900}$ for both genomic and brain imaging data. The first 100 components of α_1^*, α_2^* are drawn from Normal distribution, while the rest is set to zero. The total number of observations is set to $n = 200$. Genomic values are coded as 0 (no minor allele), 1 (one minor allele), and 2 (two minor allele). We first define a minor allele frequency η drawn from a uniform distribution $\mathcal{U}([0.2, 0.4])$. The i-th SNP is then generated from a binomial distribution $\mathcal{B}(2, \eta_i)$. For the imaging data, voxels values were drawn from a Gaussian distribution $\mathcal{N}(0, I_p)$. Finally, binary phenotype $\mathbf{y}$ data are generated from $\mathcal{B}(1, d_i)$, where $d_i = \dfrac{exp(5\sum_{m=1}^{M} \mathbf{X}_m \alpha_m^*)}{1 + exp(5\sum_{m=1}^{M} \mathbf{X}_m \alpha_m^*)}$. Furthermore, we add 100 additional variables to the problem that will play the role of cross-correlated variables. Two canonical vectors $\theta_1^*, \theta_2^* \in \mathbb{R}^{100}$ are drawn from Normal distribution. Cross-correlated SNP are drawn from $\mathcal{B}(2, \text{logit}^{-1}(-a_i + \text{logit}(\eta_i)))$ where a is issued from $\mathcal{N}(\theta_1^* \mathbf{y}, I_{100})$, while cross-correlated voxels are drawn from $\mathcal{N}(\theta_2^* \mathbf{y}, I_{100})$. The final dataset is made of $n = 200$ observations of $p = 1000$ variables for both SNP and fMRI. Each of these datasets is made of explanatory and cross-

correlated components. A common way to assess the performance of a model when it comes to feature selection is to measure the true positive rate (TPR) and false positive rate (FPR). TPR reflects the proportion of variables that are correctly identified, while FDR reflects the proportion of variables that are incorrectly selected by the model. We apply MT-CoReg to 100 random generation of the dataset described above. The tuning parameter γ from Eq.(3.3) that weights the CCA term against the regression one is optimized through a grid search over $\{[0] \cup [10^{-1+\ell/20}]; \ell = 0, .., 20\}$. We plotted TPR values against FDR ones in Fig.(3) for two different cases. In the first (left) subfigure are displayed TPR/FDR values relative to non-zero components of α_1^*, α_2^* for $\gamma = 0$ (i.e. classical Lasso), $\gamma = \gamma$(C.S.) where the weight value is determined using consistency selection (C.S.) scheme described in Section 3.3, and $\gamma = 1$ (i.e. classical sCCA). We can observe that although classical regression seems to perform slightly better for really low FDR values, MT-CoReg is quickly catching up around $FDR \approx 0.15$. sCCA, on the other hand, has a low selection power. The second (bottom) figure displays TPR/FDR values relative to non-zero components of θ_1^*, θ_2^*, i.e. the cross-correlated components. We can observe that MT-CoReg performs as well as sCCA, while Lasso is unable to properly select the components of interest. It is encouraging to see that MT-CoReg takes the best of both methods and seems to properly select the components we are interested in. It seems to confirm our hypothesis that using a mix of both terms may lead to an improved feature selection accuracy. In the next section, we apply the same method to a real dataset of fMRI and SNP data.

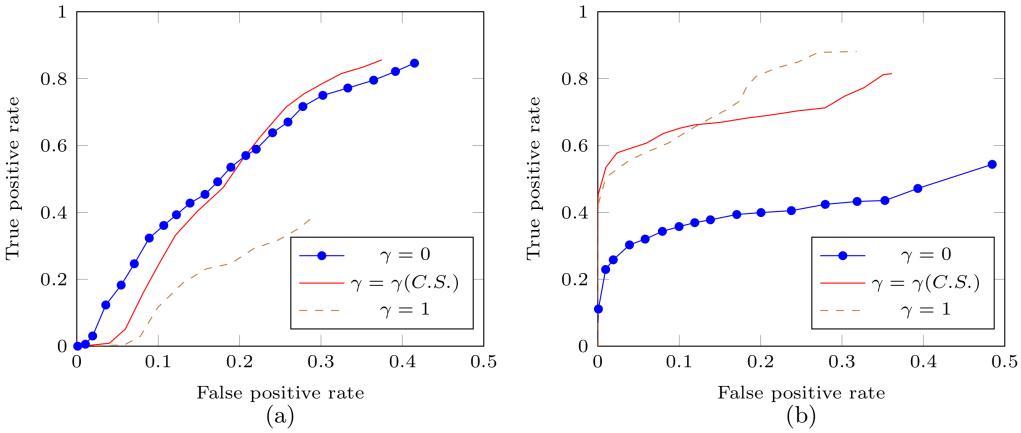

Fig. 3. TPR against FDR values averaged over 100 simulations for different γ values. Fixing $\gamma = 0$ amounts to using Lasso regression, while $\gamma = 1$ is equivalent to classical sparse CCA. γ(C.S.) is the ROC curve obtained while using consistency selection (C.S.) scheme described in section 3.3 to automatically estimate γ. (a) values for the selection of first 100 components (i.e. the explanatory components) only (b) values for the selection of the last 100 components (i.e. the cross-correlated components). It can be seen that a non-trivial weight combinaison for γ seems to be taking the best of the two methods that are Lasso ($\gamma = 0$) and CCA ($\gamma = 1$).

4.2. *Results on real imaging genetics data*

4.2.1. *Data acquisition*

Both SNP and fMRI acquisition were conducted by the Mind Clinical Imaging Consortium (MCIC) for 214 subjects, including 92 schizophrenia patients (age: 34 ± 11, 22 females) and 116 controls (age 32 ± 11, 44 females). Schizophreniac were diagnosed based on DSM-IV-TR citeria. Controls were free of any medical, neurological of psychiatric illnesses.

fMRI were acquired during a sensor motor task with auditory simulation. Data were pre-processed with SPM5, spatially normalized and resliced, smoothed, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as regressors. For each patient, a stimulus-on vs. stimulus-off contrast image was extracted. 116 ROIs were extracted based on the aal brain atlas, which resulted in 41236 voxels left for analysis. SNP data were obtained from blood sample using Illumina Infinium HumanOmni1-Quad array covering 1,140,419 SNP loci. After standard quality control procedures using PLINK software package [a], a final dataset spanning $777,635$ SNP loci was available. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 (no minor allele), 1 (one minor allele) and 2 (two minor alleles). SNPs with $> 20\%$ missing data were deleted and missing data were further imputed. SNPs with minor allele frequency $< 5\%$ were removed. This procedure yielded a final set of $129,145$ SNPs.

4.2.2. *Significance analysis*

In order to achieve a stable feature selection process, we follow Lin[8] and perform $N = 100$ random samplings out of the 214 total subjects, where for each time 80% are used for training and parameter selection, while the remaining 20% are used for evaluation. At the $k-th$ random sampling, we can calculate a set of solution vectors $\hat{\beta}_m^k, m \in \{1, 2\}$. It is then possible to define a measure of relevance p_m^i for the i-th feature in the m-th dataset such that: $p_m^i = \frac{1}{N}\sum_{k=1}^{N} I(\hat{\beta}_m^k(i) \neq 0)$ where $i = 1, .., d_m$ is the feature index and $I(\cdot)$ is the indicator function. We can then rank each SNP and voxel based on their associated relevance measure and apply a cut-off threshold of 0.3 (c.f. Lin[8]). After applying this significance test, we were left with a subset of 43 SNP spanning 30 genes and 6 ROI with a number of selected voxels over 5.
We display in Table.1 the list of each of the 43 selected SNP, as well as their associated genes. Some of them have been identified by other similar studies[8,26,27] such as CNTNAP2, GLI2, GRIK3, NOTCH4, SUCLG2, GABRG2. Others have been identified from well-known databases[28] such as GRIK4 or HTR4. We display in Table.2 the list of the selected ROI as well as the corresponding voxel count for each one of them. ROI for which less than 5 voxels were selected where dismissed. Once again, it is encouraging to note that each of the selected ROI (3, 7, 11, 40, 51, 100 from aal.) have been identified in similar studies[8,29] on the same dataset. Other studies pointed out both functionnal or structural differences in the middle occipital gyrus[30] and the parahippocampal gyrus[31] for schizophrenic patients. Finally, a detailled slice view of the selected voxels can be seen in Fig.(4).

[a] http://pngu.mgh.harvard.edu/ purcell/plink

Table 1. List of selected SNP and their associated genes.

SNP ID	Gene name	SNP ID	Gene name	SNP ID	Gene name	SNP ID	Gene name
rs3856465	ATP6V1C2	rs11607732	GRIK4	rs815533	CACNA2D3	rs10748732	HPSE2
rs12333931	CNTNAP2	rs12332417	HTR4	rs2373347	CNTNAP2	rs13359903	HTR4
rs2407264	CYSLTR2	rs7725785	HTR4	rs9535112	CYSLTR2	rs11875988	LIPG
rs6567629	DHRSX	rs12454370	LIPG	rs858341	ENPP1	rs9787820	LRRC4C
rs16842460	EPHB1	rs17819648	MAML2	rs11927660	FGF12	rs3134797	NOTCH4
rs17599845	FHIT	rs3134799	NOTCH4	rs10926254	FMN2	rs394657	NOTCH4
rs4659573	FMN2	rs1009708	PDE2A	rs11060822	FZD10	rs7111188	PDE2A
rs12824777	FZD10	rs17016738	RARB	rs2963094	GABRG2	rs12101383	SMAD6
rs10831614	GALNTL4	rs7030433	SMARCA2	rs7602673	GLI2	rs573700	SPRY2
rs6753202	GPD2	rs9849270	SUCLG2	rs1392744	GRIK3	rs1105880	UGT1A6
rs10502240	GRIK4	rs17863787	UGT1A6				

Table 2. List of selected ROI (from aal.) and associated voxel count.

ROI ID (aal.)	ROI name	voxels nb.
51	Left middle occipital gyrus	13
7	Left middle frontal gyrus	11
11	Left middle frontal gyrus, orbital part	9
100	Right lobule VI of cerebellar hemisphere	9
3	Left superior frontal gyrus	8
40	Right parahippocampal gyrus	7

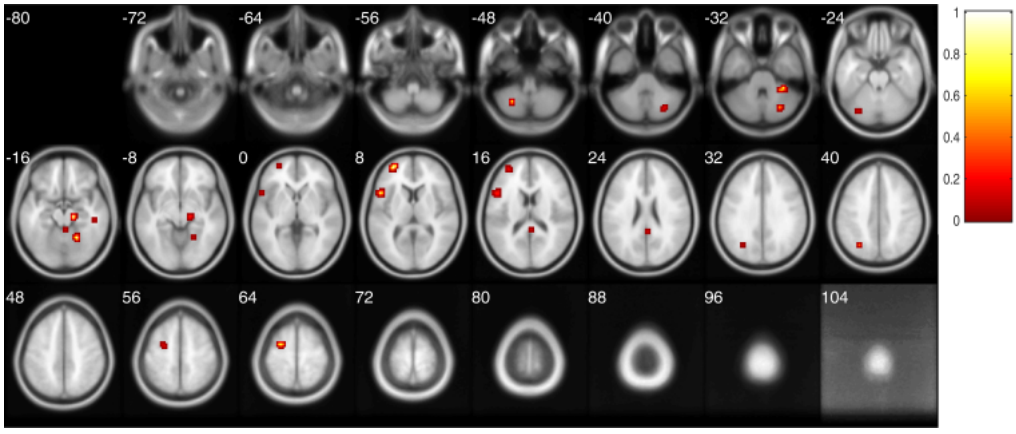

Fig. 4. Slice view of the selected voxels (without thresholding using cluster size) and their significance.

4.2.3. *Quantitative analysis*

In this section, we try to analyze the results of MT-CoReg using some quantitative metrics. We can first turn our attention to the Sum of Squared Errors (SSE) values obtained on the testing set during our tests. Histograms of SSE distributions for different γ values (i.e.

Lasso, MT-CoReg and sCCA) can be seen in Fig.(5,left): unsurprisingly, Lasso and MT-CoReg produce the lowest RSS values, while sCCA does not fit the phenotype. If we now look at Fig.(5,middle) where distributions of Pearson's correlation on the testing set are displayed for the same 3 strategies, we can see that MT-CoReg produces a better selection than Lasso in terms of cross-correlation. This seems to confirm our intuition that MT-CoReg makes the best of both Lasso and CCA by producing a solution that is good fit to the phenotype while selecting co-expressed features across modalities.
Distribution of γ values produced by the consistency selection scheme described in Section 3.3 can be seen in Fig.(5,right). Most of these values fall into the range $[0; 0.4]$, with a peak in $[0.2; 0.3]$. It does appear, at least in term of feature consistency selection, that a non-zero weight for the CCA term in Eq.(3.3) leads to improved performances.

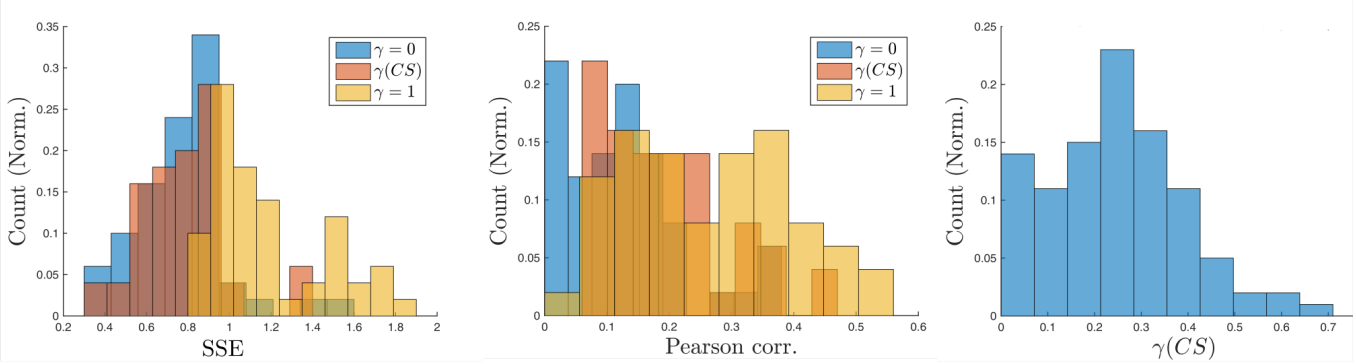

Fig. 5. Frequency distribution of RSS values (on the test set) for $N = 100$ sub-sampling of the original set of observations.

5. Conclusions

The main contributions of this paper can be summarized as follows. First, we proposed a novel variable selection approach using a CCA-like regularization term in order to enforce co-expression between modalities. Secondly, we present an efficient algorithm to solve this problem, as well as strategies to estimate the tuning parameters. On top of that, a series of experiments on both synthetical and real datasets were conducted, allowing us to evaluate the performances of the proposed method. We identified two sets of SNP and voxels in which a number of them have been previously reported to have potential relationship with the risk of schizophrenia. Further exploration of the optimization scheme (alternate estimations) as well as the selection of regularization parameter λ (see Section 3.3) will be needed in the future.

6. Acknowledgments

The authors wish to thank the NIH (NSF EPSCoR#1539067) for their partial support.

References

1. D. Lin, J. Zhang, J. Li, H. He, H.-W. Deng and Y.-P. Wang, *Multi-omic Data Integration* , p. 126 (2015).

2. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* , 267 (1996).
3. C. M. Lewis, D. F. Levinson, L. H. Wise, L. E. DeLisi, R. E. Straub, I. Hovatta, N. M. Williams, S. G. Schwab, A. E. Pulver, S. V. Faraone *et al.*, *The American Journal of Human Genetics* **73**, 34 (2003).
4. S. R. Sutrala, D. Goossens, N. M. Williams, L. Heyrman, R. Adolfsson, N. Norton, P. R. Buckland and J. Del-Favero, *Schizophrenia research* **96**, 93 (2007).
5. M. E. Shenton, C. C. Dickey, M. Frumin and R. W. McCarley, *Schizophrenia research* **49**, 1 (2001).
6. S. A. Meda, M. Bhattarai, N. A. Morris, R. S. Astur, V. D. Calhoun, D. H. Mathalon, K. A. Kiehl and G. D. Pearlson, *Schizophrenia research* **104**, 85 (2008).
7. H. Cao, J. Duan, D. Lin, Y. Y. Shugart, V. Calhoun and Y.-P. Wang, *Neuroimage* **102**, 220 (2014).
8. D. Lin, V. D. Calhoun and Y.-P. Wang, *Medical image analysis* **18**, 891 (2014).
9. É. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron *et al.*, *Neuroimage* **63**, 11 (2012).
10. M. Yuan and Y. Lin, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49 (2006).
11. B. Xin, Y. Kawahara, Y. Wang, L. Hu and W. Gao, *ACM Transactions on Intelligent Systems and Technology (TIST)* **7**, p. 60 (2016).
12. B. Jie, D. Zhang, B. Cheng and D. Shen, *Human brain mapping* **36**, 489 (2015).
13. U. Brefeld, T. Gartner, T. Scheffer and S. Wrobel, 137 (2006).
14. H. Hotelling, *Biometrika* **28**, 321 (1936).
15. D. M. Witten and R. J. Tibshirani, *Statistical applications in genetics and molecular biology* **8**, 1 (2009).
16. J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu and H. Li, *Biostatistics* **14**, 244 (2013).
17. L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, L. Shen, A. D. N. Initiative *et al.*, *Bioinformatics* , p. btw033 (2016).
18. Springer, *A novel structure-aware sparse learning algorithm for brain imaging genetics* 2014.
19. S. M. Gross and R. Tibshirani, *Biostatistics* **16**, 326 (2015).
20. I. Wilms and C. Croux, *Biometrical Journal* **57**, 834 (2015).
21. E. Parkhomenko, D. Tritchler and J. Beyene, *Statistical Applications in Genetics and Molecular Biology* **8**, 1 (2009).
22. J. Duan, J.-G. Zhang, H.-W. Deng and Y.-P. Wang, *PloS one* **8**, p. e59128 (2013).
23. Z. Xu, X. Chang, F. Xu and H. Zhang, *IEEE Transactions on neural networks and learning systems* **23**, 1013 (2012).
24. W. Sun, J. Wang and Y. Fang, *Journal of Machine Learning Research* **14**, 3419 (2013).
25. J. Cohen, *Psychological bulletin* **70**, p. 213 (1968).
26. D. Lin, H. He, J. Li, H.-W. Deng, V. D. Calhoun and Y.-P. Wang, 9 (2013).
27. J. Sun, P.-H. Kuo, B. P. Riley, K. S. Kendler and Z. Zhao, *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **147**, 1173 (2008).
28. P. Jia, J. Sun, A. Guo and Z. Zhao, *Molecular psychiatry* **15**, 453 (2010).
29. D. Lin, H. Cao, V. D. Calhoun and Y.-P. Wang, *Journal of neuroscience methods* **237**, 69 (2014).
30. S. Singh, S. Modi, S. Goyal, P. Kaur, N. Singh, T. Bhatia, S. N. Deshpande and S. Khushu, *Journal of biosciences* **40**, 355 (2015).
31. M. J. Escartí, M. de la Iglesia-Vayá, L. Martí-Bonmatí, M. Robles, J. Carbonell, J. J. Lull, G. García-Martí, J. V. Manjón, E. J. Aguilar, A. Aleman *et al.*, *Schizophrenia research* **117**, 31 (2010).