Structured and Sparse Canonical Correlation Analysis as a Brain-Wide Multi-Modal Data Fusion Approach

Ali-Reza Mohammadi-Nejad, Gholam-Ali Hossein-Zadeh,
and Hamid Soltanian-Zadeh,* *Senior Member, IEEE*

***Abstract*— Multi-modal data fusion has recently emerged as a comprehensive neuroimaging analysis approach, which usually uses canonical correlation analysis (CCA). However, the current CCA-based fusion approaches face problems like high-dimensionality, multi-collinearity, unimodal feature selection, asymmetry, and loss of spatial information in reshaping the imaging data into vectors. This paper proposes a structured and sparse CCA (ssCCA) technique as a novel CCA method to overcome the above problems. To investigate the performance of the proposed algorithm, we have compared three data fusion techniques: standard CCA, regularized CCA, and ssCCA, and evaluated their ability to detect multi-modal data associations. We have used simulations to compare the performance of these approaches and probe the effects of non-negativity constraint, the dimensionality of features, sample size, and noise power. The results demonstrate that ssCCA outperforms the existing standard and regularized CCA-based fusion approaches. We have also applied the methods to real functional magnetic resonance imaging (fMRI) and structural MRI data of Alzheimer's disease (AD) patients ($n = 34$) and healthy control (HC) subjects ($n = 42$) from the ADNI database. The results illustrate that the proposed unsupervised technique differentiates the transition pattern between the subject-course of AD patients and HC subjects with a p-value of less than 1×10^{-6}. Furthermore, we have depicted the brain mapping of functional areas that are most correlated with the anatomical changes in AD patients relative to HC subjects.**

***Index Terms*— ADNI, canonical correlation analysis (CCA), magnetic resonance imaging (MRI), multi-modal data fusion, multivariate analysis.**

I. Introduction

By ADJUSTING the imaging protocols and pulse sequence parameters, magnetic resonance imaging (MRI) generates a variety of image contrasts that provide a wealth of information about the anatomy and physiology of the brain. In the fusion terminology, a "modality" is defined as a single image contrast. The main goal of gathering and analyzing multiple modalities is to utilize the common as well as unique information from complementary modalities to understand the problem at hand, which cannot be done by independent analysis of individual modalities.

In data integration approaches, data from different modalities are usually analyzed through separate pipelines and the results combined at the interpretation level to yield decision level fusion [1], [2]. This approach may use information from one modality to improve the overall result of other modalities. However, there should be benefits to fuse the data in earlier steps, in particular, after pre-processing but before statistical analysis to yield feature-level fusion [3]. Unlike data integration methods, data fusion techniques incorporate all features from different modalities into a combined analysis and allow for true interaction between different data modalities while characterizing between-subject variability in a data-driven and exploratory manner.

Canonical correlation analysis (CCA) [4] is a classic tool in multivariate data fusion which provides optimal projections of two data modalities (with different scales, resolutions, and dimensionalities) in such a way that the correlation between the projections is maximized. Similar to [5], a CCA-based fusion approach may be used at the feature level. Recently, several extensions of CCA have been proposed to fuse different modalities at the feature level, including multi-modal CCA (mCCA) [6] and mCCA + joint independent component analysis (jICA) [7].

Typically, the number of variables (voxels) in the MRI datasets is much larger than the number of observations (subjects). Due to this high dimensionality and high noise

Manuscript received December 19, 2016; revised February 27, 2017; accepted March 3, 2017. Date of publication March 14, 2017; date of current version June 29, 2017. This work was supported in part by the Iranian Cognitive Science and Technologies Council Research under Grant 108. MRI acquisition was done by the Alzheimer's Disease Neuroimaging Initiative (ADNI), partially supported by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, under Grant U01 AG024904 and the Department of Defense, under Award W81XWH-12-2-0012. *Asterisk indicates corresponding author.*

A.-R. Mohammadi-Nejad is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, and also with the Medical Image Analysis Laboratory, Department of Radiology, Henry Ford Health System, Detroit, MI, USA (e-mail: mohammadi_nejad@ut.ac.ir).

G.-A. Hossein-Zadeh is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, and also with the School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran (e-mail: ghzadeh@ut.ac.ir).

*H. Soltanian-Zadeh is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, also with the School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran, and also with the Medical Image Analysis Laboratory, Department of Radiology, Henry Ford Health System, Detroit, MI, USA (e-mail: hszadeh@ut.ac.ir; hsoltan1@hfhs.org).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TMI.2017.2681966

level, variable selection is critical to avoid over-fitting of the data [6]. The standard CCA, however, does not perform variable selection and hence unimodal feature selection is performed for each modality separately [8], which is blind to the next modality and therefore may limit biological interpretability of the results.

To overcome the above limitation, sparse CCA (sCCA) has been proposed [9], [10]. This approach includes a built-in procedure for variable selection. In sCCA, a sparsity penalty function such as an *l1* penalty is often imposed as a regularization to identify sparse sets of associated variables that are highly correlated. Through imposed sparsity, parsimonious multivariate methods increase the interpretability of the output and potentially improve the generalizability of the produced model. In addition, in neuroimaging, the observed data (images) are positive. Because existence of negative weights in the canonical coefficients makes the interpretations difficult, it seems rational to add a simple non-negativity restriction as a regularization term to the standard CCA [11], [12].

To apply CCA on the imaging data, it is essential to reshape the image data into vectors. Such reshaping breaks down the spatial structure and dependencies of the image data to its local neighborhood voxels. CCA and sCCA do not make any assumptions about the spatial smoothing, dependency, or structural relationships of the input variables. This limits their performances on complex and high-dimensional biological imaging data in real problems. Furthermore, adjacent pixels/voxels in a homogeneous region of the image are typically correlated. Therefore, canonical coefficients associated with these voxels should have similar magnitudes to reflect the underlying association [12], [13]. A limited number of previous CCA-based fusion approaches investigated this spatial smoothness [14]–[18]. For example, in [17], Lin *et al.* developed a group sparse CCA approach for fusing functional and genetic datasets. They assumed that all voxels in a region of interest (ROI) have almost the same canonical coefficients. In this paper, we add two constraints of non-negativity and smoothness to sCCA. The groupness constraint in [17] is a special case of our graph-based smoothing constraint.

Finally, some of the previous data fusion methods are applicable when only one of the datasets is high-dimensional (number of features ≫ number of samples) [11], [16]. These methods are not symmetric, so by changing data modalities from phenotypic or behavioral to neuroimaging or genetic, for example, they may not work properly.

To overcome the above problems, we present a symmetric, structured, and smooth extension of the sCCA, which we refer to as structured and sparse CCA (ssCCA). Using some targeted biological heuristics, we develop a physiologically interpretable multivariate data fusion technique that finds the interaction between different modalities in an exploratory manner. The proposed method directly takes two sets of extra-large multi-modal data (e.g., an anatomical and a functional dataset) and preserves spatial structure and smoothness of the image data in the calculation of the canonical coefficients. The proposed optimization framework does a flexible combination of sparsity and smoothness in a unified framework for big datasets with strong convergence guaranties. The main objective of the proposed method is to combine related datasets to find the best canonical variates (CVs) that fit the subject-course (a vector of subject weights, a scalar value per subject) and at the same time to find physiologically interpretable and informative features that describe the hidden phenomenon in these datasets.

The proposed method and its implementation are introduced in the next section. The experimental dataset and the pipeline of pre-processing of each modality are described in Section 3. In Section 4, we present our evaluation and comparison studies of the proposed method and another CCA-based fusion approaches, using simulated multi-modal datasets as well as real datasets of Alzheimer's disease patients and age-matched healthy control subjects that include anatomical and functional MRI data. Finally, in Section 5, we conclude the paper with a summary of the presented materials and discussion of the future work.

By convention, matrices are denoted by boldface capital letters, vectors are denoted by boldface lowercase letters, and scalars are denoted by lowercase letters.

II. Proposed Method

A. Canonical Correlation Analysis (CCA)

Let us consider two random vectors $\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)$ and $\mathbf{y} = \left(y_1, y_2, \ldots, y_q\right)$, which contains $p-$ and $q-$ dimensional vectors of voxels, respectively. Suppose that n i.i.d. samples are measured (from n subjects) of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times q}$, respectively. Assume that the measurements for each variable $\mathbf{x}_i \in \mathbb{R}^n$ (or $\mathbf{y}_i \in \mathbb{R}^n$) has been standardized to have zero mean and unit variance. Prior biological knowledge of these data allows us to hypothesize that there is a meaningful correlation between the two datasets, i.e., there exists a reciprocal relationship between the $\mathbf{X}$ variables and the $\mathbf{Y}$ variables. The aim of CCA is to find two projection vectors (canonical correlation coefficients, CCCs) $\mathbf{w}_1 \in \mathbb{R}^p$ and $\mathbf{w}_2 \in \mathbb{R}^q$ to maximize the correlation between the first pair of canonical variates $\mathbf{u}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{v}_1 = \mathbf{Y}\mathbf{w}_2$

$$\begin{aligned}\rho = \ & \arg\max_{\mathbf{w}_1,\mathbf{w}_2} \mathbf{w}_1^T\mathbf{C}_{xy}\mathbf{w}_2 \\ & s.t.\ \mathbf{w}_1^T\mathbf{C}_{xx}\mathbf{w}_1 = 1 \ \text{ and } \ \mathbf{w}_2^T\mathbf{C}_{yy}\mathbf{w}_2 = 1 \end{aligned} \tag{1}$$

where $\mathbf{C}_{xx}$, $\mathbf{C}_{yy}$, and $\mathbf{C}_{xy}$ are covariance and cross-covariance matrices, respectively. In practice, $\mathbf{C}_{xy}$, $\mathbf{C}_{xx}$, and $\mathbf{C}_{yy}$ are replaced by the observed sample cross-covariance and covariance matrices as $\mathbf{X}^T\mathbf{Y}$, $\mathbf{X}^T\mathbf{X}$, and $\mathbf{Y}^T\mathbf{Y}$, respectively [16].

B. Sparsity Constraint

Due to the small number of samples but high-dimensional variables/features in the neuroimaging datasets, (1) faces the overfitting problem. Therefore, sparse penalties such as *l1*-norm are imposed on $\mathbf{w}_1$ and $\mathbf{w}_2$ in the sCCA analysis [10]:

$$\begin{aligned} & \max_{\mathbf{w}_1,\mathbf{w}_2} \ \mathbf{w}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{w}_2 \\ & s.t.\ \mathbf{w}_1^T\mathbf{X}^T\mathbf{X}\mathbf{w}_1 = 1, \quad \mathbf{w}_2^T\mathbf{Y}^T\mathbf{Y}\mathbf{w}_2 = 1, \\ & \quad \|\mathbf{w}_1\|_1 \le c_1, \quad \|\mathbf{w}_2\|_1 \le c_2 \end{aligned} \tag{2}$$

where $c_1 > 0$ and $c_2 > 0$ control the level of sparsity. Setting small values for c_1 and c_2 increases the penalty and forces more CCCs to 0.

C. Collinearity Problem

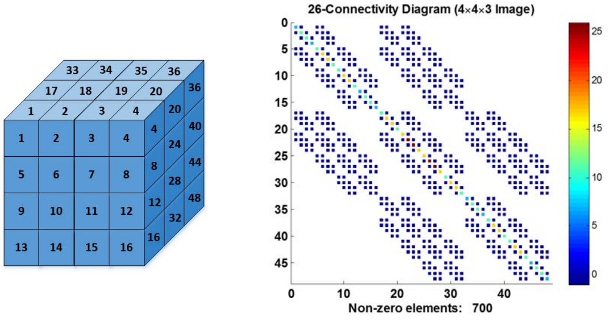
When $n \ll p$ or $n \ll q$, the features in $\mathbf{X}$ and $\mathbf{Y}$ tend to be highly collinear (linearly dependent). This phenomenon is analogous to the multi-collinearity problem in regression analysis. This makes the ill-conditioned matrices $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ singular and the inverse operations on them, i.e., $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and $\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}$, lead to unreliable results in the computation of standard CCA. The condition placed on the data to guarantee that $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ will be invertible is $n \geq p+q+1$. However, this condition is not usually met in neuroimaging.

Treating the covariance matrix as an identity matrix helps us overcome this problem [9], [10], [19]. Therefore, $\mathbf{w}_1^T\mathbf{X}^T\mathbf{X}\mathbf{w}_1$ and $\mathbf{w}_2^T\mathbf{Y}^T\mathbf{Y}\mathbf{w}_2$ in (2) can be replaced by the square of the $l2$-norms of $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively (i.e., $\|\mathbf{w}_1\|_2^2 = 1$ and $\|\mathbf{w}_2\|_2^2 = 1$). Notably, when the estimated covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ are identity matrices, CCA equals two-block Mode A partial least squares (PLS) [20]. The function maximized by PLS is the covariance between the two CVs (latent variables) $\mathbf{u}_1$ and $\mathbf{v}_1$, which can be written as $max_{\|\mathbf{w}_1\|_2=\|\mathbf{w}_2\|_2=1} corr\left(\mathbf{X}\mathbf{w}_1, \mathbf{Y}\mathbf{w}_2\right)\sqrt{var\left(\mathbf{X}\mathbf{w}_1\right)}\sqrt{var\left(\mathbf{Y}\mathbf{w}_2\right)}$ [20]. An interesting feature of PLS is that it tries to construct CVs that explain their own modality via $max\left(var(.)\right)$ and meanwhile are well correlated with the corresponding CVs in the other modality via $max\left(corr(.)\right)$. Next, based on [10], we replace the equality constraint of the $l2$-norm (i.e., $\|\mathbf{w}_1\|_2^2 = 1$ and $\|\mathbf{w}_2\|_2^2 = 1$) with $\leq$ operator. This means that we replace the non-convex $l2$-norm penalty with its convex version, which includes the equality constraint on the $l2$-norm.

D. Non-Negativity Constraint

In the CCA, the canonical coefficient of each feature represents its contribution in the final linear combination. The most prominent features are expected to have large weights and the redundant features have zero weights. However, in standard CCA and its extensions, due to the absence of any restriction on the sign of the extracted CCC vectors, the canonical coefficient of each feature could be positive or negative. Therefore, the negative features (voxels) in the reconstructed map of the brain cannot be interpreted as weights (negative weights are not interpretable). In this condition, a constant cannot be added to all features (voxels) to make them positive because the zero-valued features (voxels) lose their interpretation as no contribution to the canonical variate. Similarly, the absolute value of negative weights cannot be used because positive and negative weights are expected to be different. Therefore, we restrict the elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ in (2) to be non-negative and add the constraints $w_{1_i} \geq 0$, $\forall i = 1, \ldots, p$ and $w_{2_j} \geq 0$, $\forall j = 1, \ldots, q$ to this equation.

E. Smoothing Constraint



(a) (b)

Fig. 1. (a) A 3D labeled volume with dimensions of 4×4×3. (b) Laplacian matrix using the nearest neighbors in a 3×3×3 neighborhood (i.e., 26-connectivity). The horizontal and vertical axes in (b) reflect the voxel labels and color-coded dots show relative weights of the voxels.

In a pre-processing step before applying a CCA-based method to the imaging datasets, 2D and 3D data are reshaped into 1D vectors. This reshaping destroys the spatial information/relation and dependent structure of the variables to their local neighborhoods. To preserve such essential information after vectorization, we add a structural smoothing penalty term to the objective function. This approach is less restrictive than making all voxel weights equal but more restrictive than the unconstrained method where the weights are independent. To form such a penalty, consider a network represented by an undirected weighted graph. The vertices of this graph correspond to the voxels (e.g., p voxels) and the edges indicate the link between the voxels in the graph. Also, $\mathbf{L}$ is the matrix weight of the edges where l_{ij} denotes the weight of the edge $e = \{i \sim j\}$. Next, for a given adjacency matrix $\mathbf{A}$, we define $\mathbf{D} = \text{diag}\left(d_1, d_2, \ldots, d_p\right)$ where $d_j = \sum_{k=1}^{P} a_{jk}$ and the associated Laplacian matrix as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. This matrix is a sparse and symmetric connectivity matrix that encourages similar weights for the neighboring voxels. For a given weight vector $\mathbf{u}$, it can be shown that [21]:

$$\mathbf{u}^T\mathbf{L}\mathbf{u} = \sum_{1 \leq j < k \leq p} a_{jk}\left(u_j - u_k\right)^2 \tag{3}$$

This constraint generates a heavy penalty if the neighboring voxels have dissimilar weights. It displays a local smoothing effect by encouraging the variables that are linked (as represented by $\mathbf{L}$) to have relatively similar coefficients. The biological motivation of this penalty is that the neighboring voxels that are linked to a predefined structure are expected to have similar weights. Therefore, these voxels should have smooth coefficients. Fig. 1 shows an example of the produced Laplacian matrix for a 3D volume of $4 \times 4 \times 3$ voxels. In this figure, the number of non-zero elements in the Laplacian matrix is 700 [sparsity rate = 700/(48×48) = 30.38%].

F. ssCCA

Finally, after considering the solution for the multi-collinearity problem in the neuroimaging data and adding non-negativity and smoothing terms for both of the modalities, the proposed ssCCA with the tuning parameters

$c_1 > 0, c_2 > 0, c_3 > 0, c_4 > 0$ is formulated as

$$\begin{aligned}
&\max_{\mathbf{w}_1,\mathbf{w}_2} \ \mathbf{w}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{w}_2\\
&s.t. \ \ \|\mathbf{w}_1\|_2^2 \le 1, \quad \|\mathbf{w}_2\|_2^2 \le 1,\\
&\qquad \|\mathbf{w}_1\|_1 \le c_1, \quad \|\mathbf{w}_2\|_1 \le c_2,\\
&\qquad \mathbf{w}_1^T\mathbf{L}_{w_1}\mathbf{w}_1 \le c_3, \quad \mathbf{w}_2^T\mathbf{L}_{w_2}\mathbf{w}_2 \le c_4,\\
&\qquad w_{1_i} \ge 0 \quad \forall i = 1,\ldots,p, \quad w_{2_j} \ge 0 \quad \forall j = 1,\ldots,q \qquad (4)
\end{aligned}$$

where c_3 and c_4 are penalty parameters that control the level of smoothness and $\mathbf{L}_{w_1}$ and $\mathbf{L}_{w_2}$ represent the semi-positive definite Laplacian matrices of two modalities (corresponding to $\mathbf{w}_1$ and $\mathbf{w}_2$), respectively. The tuning parameters $C = (c_1, c_2, c_3, c_4)$ control the model complexity and have to be tuned. To facilitate computation, we write constraints on $\mathbf{w}_1$ and $\mathbf{w}_2$ in Lagrangian form as [10], [16]:

$$\begin{aligned}
&\mathop{max}_{\mathbf{w}_1,\mathbf{w}_2} \ \mathbf{w}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{w}_2 - \lambda_1\|\mathbf{w}_1\|_1 - \lambda_2\|\mathbf{w}_2\|_1\\
&\qquad -\frac{1}{2}\mathbf{w}_1^T\left(\mathbf{I} + \alpha_1\mathbf{L}_{w_1}\right)\mathbf{w}_1\\
&\qquad -\frac{1}{2}\mathbf{w}_2^T\left(\mathbf{I} + \alpha_2\mathbf{L}_{w_2}\right)\mathbf{w}_2\\
&s.t. \ w_{1_i} \ge 0 \quad \forall i = 1,\ldots,p, \quad w_{2_j} \ge 0 \quad \forall j = 1,\ldots,q \qquad (5)
\end{aligned}$$

where the regularization parameters $\lambda_1, \lambda_2, \alpha_1$, and α_2 have the same roles as c_1, c_2, c_3, and c_4, respectively.

The objective function $\mathbf{w}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{w}_2$ in (5) is bilinear in $\mathbf{w}_1$ and $\mathbf{w}_2$: that is, with $\mathbf{w}_1$ fixed, it is linear in $\mathbf{w}_2$ and vice versa. The above problem is also a biconvex problem, meaning that by fixing $\mathbf{w}_1$ ($\mathbf{w}_2$), (5) is convex with respect to $\mathbf{w}_2$ ($\mathbf{w}_1$). This property suggests an iterative algorithm for finding $\mathbf{w}_1$ and $\mathbf{w}_2$, i.e., optimizing $\mathbf{w}_1$ with a fixed $\mathbf{w}_2$ and then, optimizing $\mathbf{w}_2$ with the $\mathbf{w}_1$ found in the previous step (details given in Supplementary Material 1, Section A). The complete algorithm is described in Algorithm I (details are given in Supplementary Material 1, Section B).

In the above problem, *l1*-norm is used as a sparsity constraint which is a convex penalty term. Therefore, the steps 3(a-d) in Algorithm I monotonically converge to the global solution of (5) for $\mathbf{w}_1$ and $\mathbf{w}_2$ (details are given in Supplementary Material 1, Section A along with its differences from [22]).

The iterative algorithm starts from an initial point and alternately approximates and updates $\mathbf{w}_1$ and $\mathbf{w}_2$ to minimize the cost function. The iterations stop when the convergence criterion is met and the resulting $\mathbf{w}_1$ and $\mathbf{w}_2$ vectors are taken as the optimal solution. To find the next CCCs vectors and CVs, the above stages are executed after the contributions of the first CVs are regressed out (deflated) from $\mathbf{X}$ and $\mathbf{Y}$ (step 7). After that, the algorithm is repeated for the residual matrices to obtain the remaining pairs of CVs. In each iteration, the extracted CVs are orthogonal to the previous pair of CVs. This process can be repeated until the residual matrices contain no more information or until the decision is made that further analysis is no longer useful.

To initialize $\mathbf{w}_1$ and $\mathbf{w}_2$ as the inputs of the proposed algorithm (Algorithm I), we use the singular value decomposition (SVD) of $\mathbf{X}^T\mathbf{Y}$. Since $\mathbf{X}^T\mathbf{Y}$ may have a very large dimension ($p \times q$) and its SVD cannot be computed on a personal computer, we use a QR decomposition algorithm. According to this decomposition, the left and right singular vectors of $\mathbf{X}^T\mathbf{Y}$ are the same as the original matrices manipulated using the traditional SVD, up to a sign change (details are given in Supplementary Material 1, Section C).

Algorithm I. The iterative algorithm of structured and sparse CCA

```
Definitions:
$prox_P(\mathbf{y},\lambda) \triangleq \operatorname*{argmin}_{\mathbf{x}}\left\{\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda P(\mathbf{x})\right\}$   $P_w(\mathbf{w}) \triangleq \|\mathbf{w}\|_1$
$\lambda_{max}(\mathbf{A}) \triangleq$ largest eigenvalue of $\mathbf{A}$   $\|\mathbf{X}\|_{\mathbf{A}} \triangleq \sqrt{\mathbf{X}^T\mathbf{A}\mathbf{X}}$
$\mathbf{S}_{\mathbf{w}_i} \triangleq \mathbf{I} + \alpha_i\mathbf{L}_{\mathbf{w}_i}, \forall i = 1,2$   $T_{w_i} \triangleq \lambda_{max}(\mathbf{S}_{w_i}), \forall i = 1,2$

Input: two datasets $\mathbf{X}$ and $\mathbf{Y}$, $T_{w_1}$, and $T_{w_2}$.
Output: corresponding CVs and canonical coefficients of two datasets.

1: Initialize $\mathbf{w}_1$ and $\mathbf{w}_2$ by QR version of SVD of $\mathbf{X}^T\mathbf{Y}$, with and $\|\mathbf{w}_1\|_2^2 = 1$ and $\|\mathbf{w}_2\|_2^2 = 1$.
2: Use two steps 5-fold cross validation to obtain the optimal tuning parameters.
3: Solve $\mathbf{w}_1^{(k)}, \mathbf{w}_2^{(k)}$ using the following iterations until it convergence:
    a) Estimate $\hat{\mathbf{w}}_1$:
        $\mathbf{w}_1^{(k+1)} = prox_{p_{\mathbf{w}_1}}\left(\mathbf{w}_1^{(k)} + \frac{1}{T_{\mathbf{w}_1}}\left(\mathbf{X}^T\mathbf{Y}\mathbf{w}_2^* - \mathbf{S}_{\mathbf{w}_1}\mathbf{w}_1^{(k)}\right), \frac{\lambda_{\mathbf{w}_1}}{T_{\mathbf{w}_1}}\right)$
    b) $\mathbf{w}_1^* = \begin{cases}\hat{\mathbf{w}}_1/\|\hat{\mathbf{w}}_1\|_{\mathbf{S}_{\mathbf{w}_1}}, & \|\hat{\mathbf{w}}_1\|_{\mathbf{S}_{\mathbf{w}_1}} > 0\\ 0, & otherwise\end{cases}$
    c) Estimate $\hat{\mathbf{w}}_2$:
        $\mathbf{w}_2^{(k+1)} = prox_{p_{\mathbf{w}_2}}\left(\mathbf{w}_2^{(k)} + \frac{1}{T_{\mathbf{w}_2}}\left(\mathbf{Y}^T\mathbf{X}\mathbf{w}_1^* - \mathbf{S}_{\mathbf{w}_2}\mathbf{w}_2^{(k)}\right), \frac{\lambda_{\mathbf{w}_2}}{T_{\mathbf{w}_2}}\right)$
    d) $\mathbf{w}_2^* = \begin{cases}\hat{\mathbf{w}}_2/\|\hat{\mathbf{w}}_2\|_{\mathbf{S}_{\mathbf{w}_2}}, & \|\hat{\mathbf{w}}_2\|_{\mathbf{S}_{\mathbf{w}_2}} > 0\\ 0, & otherwise\end{cases}$
4: Output: $\mathbf{w}_1 = \mathbf{w}_1^*/\|\mathbf{w}_1^*\|_2$, $\mathbf{w}_2 = \mathbf{w}_2^*/\|\mathbf{w}_2^*\|_2$.
5: Calculate the CVs as $\mathbf{u} = \mathbf{X}\mathbf{w}_1$ and $\mathbf{v} = \mathbf{Y}\mathbf{w}_2$.
6: Do a permutation analysis to check if the extracted CVs are statistically significant. If the correlation is significant, go to Step 7. Otherwise, there is no significant correlation between the current two datasets.
7: Deflate $\mathbf{X}$ and $\mathbf{Y}$ by subtracting the effects of the CVs $\mathbf{u}$ and $\mathbf{v}$ from the data space: $\mathbf{X} = \mathbf{X} - \mathbf{u}\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\mathbf{u}^T\mathbf{X}$, $\mathbf{Y} = \mathbf{Y} - \mathbf{v}\left(\mathbf{v}^T\mathbf{v}\right)^{-1}\mathbf{v}^T\mathbf{Y}$.
8: Do this iterative algorithm to find the next CVs.
```

Next, we add the non-negativity constraint, as mentioned in (5), to the proposed ssCCA problem. To this end, we replace the proximal operator in step 3(a) and (c), using the positive proximal operator. This operator for the *l1*-norm (i.e., $P(\mathbf{w}) = \|\mathbf{w}\|_1$) has the simple closed form solution $prox^+(\mathbf{y}, \lambda) = (\mathbf{y} - \lambda)_+$ [22], i.e., the positive soft-thresholding operator.

G. Simulated Data

Two 3D whole brain datasets were simulated with two imaging modalities for two groups, HC and patients. For the simulated datasets, we generated features (voxels) using multivariate normal distributions with mean and variance parameters obtained from experimental data. Using T1-weighted MNI152 1 mm data and FreeSurfer software version 5.3.0 (http://www.surfer.nmr.mgh.harvard.edu), we generated a label map with $G = 190$ anatomical ROIs. Then, we estimated the mean intensity mu_i, $i = 1,\ldots,G$ and standard deviations $\sigma_i, i = 1,\ldots,G$ of the ROIs. Next, for creating the two modalities, we normalized the extracted label map to 2 mm and 3 mm isotropic voxel sizes using the FLIRT module in FSL software version 5.0.9 (www.fmrib.ox.ac.uk/fsl) for the whole brain volume of each simulated individual and each modality. Then, the intensities of the voxels in each ROI were generated by a vector drawn

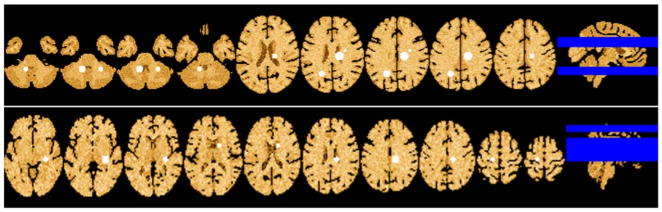


Fig. 2. Multi-slice view of the 3D simulated brain using the mean and standard deviation values extracted from 190 ROIs parcellated from T1-weighted MNI152 1 mm data. Each modality is generated in a resolution that mimics experimental data: (1^{st} row) 2 mm and (2^{nd} row) 3 mm isotropic voxels. For clear demonstration of the ROIs (in the 2^{nd} group of subjects), these locations are shown in white.

from the multivariate normal distribution with parameters $(mu_i, \sigma_i^2 \boldsymbol{I})$, $i = 1, \ldots, G$. The voxel intensities can be interpreted as baseline measurements for that voxel [bl in (6)]. The dimensions of the simulated brain images are $91\times108\times91$ and $60\times72\times60$ voxels, respectively for the first ($\mathbf{X}$) and second ($\mathbf{Y}$) modalities. After that, we extracted the in-brain voxels of each modality. This made the dimensions of the first and second modalities $p = 189,625$ and $q = 53,840$, respectively. The number of non-zero elements of the corresponding Laplacian matrices of the first and second modalities was 4,672,415 (sparsity rate $= 0.01\%$) and 1,282,724 (sparsity rate $= 0.04\%$), respectively.

To correlate the simulated datasets, a latent model similar to [9] was applied. To this end, in the second group (patients), a small subset of variables (voxels) in $\mathbf{X}$, i.e., p_c number, were correlated with a subset of variables (voxels) in $\mathbf{Y}$, i.e., q_c number, while the rest of the variables were independent. The locations of these latent variables for the two modalities are shown in Fig. 2. The correlated variables between the two modalities are located in spherical clusters with different sizes. To produce an association between these two modalities, we randomly selected 9 locations (5 for the first and 4 for the second modality) as the centers of the spherical regions. These centers were chosen such that all of the spheres were entirely inside the brain. We randomly selected the radii of the spheres. In the first modality, they were 8, 6, 4, 8, and 10 mm and for the second one, they were 6, 9, 6, and 9 mm (details are given in Supplementary Material 2, Table S2.I). This resulted the number of correlated voxels in the two modalities as $p_c = 1107$ and $q_c = 240$, respectively.

To generate an association between the two modalities, we first set a latent variable μ_j, $j = 1, \ldots, n$, to play the role of the samples (subjects) in the two modalities. Using this variable, we simulated the same samples in different modalities. This variable was produced from a normal distribution $\mu_j \sim \mathrm{N}\left(0, \sigma_\mu^2\right)$. In addition, we defined two variables δ_i, $i = 1, \ldots, p_c$ and β_i, $i = 1, \ldots, q_c$ to play the role of imaging features that were correlated such that $\sum_{i=1}^{p_c} \delta_i = \sum_{i=1}^{q_c} \beta_i = 1$. For the correlated voxels, we simulated the data using the following strategy [9]:

$$\begin{aligned} x_{ji} &= bl + \delta_i \mu_j + e_{x_{ji}} \quad \forall j = 1, \ldots, n, i = 1, \ldots, p_c \\ y_{ji} &= bl + \beta_i \mu_j + e_{y_{ji}} \quad \forall j = 1, \ldots, n, i = 1, \ldots, q_c \end{aligned} \tag{6}$$

For the uncorrelated features (voxels), $\mu_j = 0$, $j = 1, \ldots, n$, and therefore, the features were the sum of the baseline (bl) and a zero-mean white Gaussian noise with a standard deviation of σ_e.

TABLE I
Demographic Information of the Participants Involved in This Study

	HC	AD
No. of subjects	42	34
Age (mean ± SD)	73.45 ± 5.60	76.04 ± 8.00
Gender (M/F)	18 / 24	17 / 17
Years of education (mean ± SD)	16.79 ± 2.23	15.44 ± 2.39
Handedness (R/L)	37 / 5	32 / 2

III. Real Data Collection and Analysis

A. Real Data Collection

We downloaded the T1-weighted and resting state functional MRI (rs-fMRI) data of 76 subjects (34 Alzheimer's disease – AD – and 42 elderly healthy controls – HC) from ADNI.[1] The two groups were matched for their age, sex, and education, with demographic data shown in Table I. All subjects underwent whole-brain MRI scanning on 3.0 T Philips Medical Systems scanners, on at least one of two occasions (baseline and 6 months later) from ADNI-2. The parameters of the T1-weighted MP-RAGE sequence were: acquisition matrix = 256×256; voxel size = 1.2 mm × 1.0 mm × 1.0 mm; TR = 6.78 ms; TE = 3.16 ms; flip angle = 9°; and 170 sagittal slices. The rs-fMRI data included 140 image volumes acquired while the subjects were at rest in the scanner, using a gradient-echo EPI pulse sequence with the following parameters: acquisition matrix = 64×64; voxel size = 3.31 mm × 3.31 mm × 3.31 mm; 48 axial slices; TR = 3000 ms; TE = 30 ms; flip angle = 80°.

B. Pre-Processing

The T1-weighted images were pre-processed as follows. First, the intensities of the T1-weighted images were bias corrected using the N3 v1.1 package in the FreeSurfer. Then, skull stripping was performed using the hybrid watershed algorithm in FreeSurfer. Next, the brain-extracted volumes were spatially normalized into the standard MNI space (MNI152 3 mm) using the FLIRT. Then, we produced a global binary brain mask by multiplying all individual binary masks. Finally, we applied this global mask to individual brain extracted T1-weighted images. The final dimensions of the structural images were 60×72×60 and contained 53,328 intra-cerebral voxels.

Single subject rs-fMRI datasets were pre-processed using FEAT in FSL. The first 5 EPI volume images were discarded to remove the initial transients and then slice timing correction was performed. Next, images underwent: motion

[1] Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from many other sources. Detailed ADNI Acknowledgements information is available in http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Manuscript_Citations.pdf.

correction using MCFLIRT; removal of non-brain tissue using BET; spatial smoothing using a 5 mm full-width-at-half-maximum (FWHM) Gaussian kernel; mean intensity normalization to force each volume to have the same mean intensity; and high-pass temporal filtering (0.01 Hz). After preprocessing, the fMRI volumes were registered (affine registration algorithm) to the 3 mm isotropic standard space (MNI152) using T1-weighted anatomical images.

Eigenvector centrality mapping (ECM) of the rs-fMRI time series in the MNI-space was performed using the fast ECM (fECM) software [23], which yields a voxel-wise measure of the centrality of a node (voxel) to the functional brain network. The benefits of centrality mapping relative to the common techniques to study functional connectivity (FC) in the rs-fMRI data (i.e., ICA and seed-based correlations) is that it does not rely on the prior definition of ROIs and considers the brain as a large network, rather than dividing it into several sub-networks. ECM requires the computation of a voxel-wise connectivity matrix to calculate its eigenvectors [24]. A mask of the intra-cerebral voxels [after excluding the white matter (WM) regions] across all subjects' pre-processed datasets (i.e., in the intersection of all single-subject masks) was applied before the EC maps were computed. The output of the centrality mapping is a single map for each subject (contains 33,117 voxels), which is used as a voxel-wise FC map for our multi-modal fusion approach.

C. Data Fusion

The 3D structural image and centrality map of each subject were reshaped into one-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The dimensions of each matrix were [number of subjects] × [number of voxels] for each of the two modalities. It is noteworthy that the arrangement of the subjects in each of the two modalities has to be quite similar. In each modality, we first put the HC subjects randomly in the top rows and then put the AD subjects randomly in the remaining rows. Next, data normalization was done in such a way that each column of each matrix had zero mean and unit variance to account for the scale differences among the different datasets. In the next step, the Laplacian matrices for different modalities were prepared based on the predefined neighborhood connectivity where we used 26-connectivity. The dimensions of the Laplacian matrices of the first and second modalities were 53,328×53,328 and 33,117×33,117, respectively. The number of non-zero elements for these matrices were 1,270,664 (sparsity rate = 0.04%) and 601,623 (sparsity rate = 0.05%) for the first and second modalities, respectively. The Laplacian matrices are sparse; they are constructed once and used multiple times in the process. Next, the proposed method described in section II above was applied to two modalities. The CCC vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ were calculated in such a way that, each produced pairs of maximally correlated CVs were orthogonal to the previously extracted CVs vectors. The outcome of this stage was $l = min\{rank(\mathbf{X}), rank(\mathbf{Y})\}$ CVs vectors, produced by linear combinations of the original data using CCCs with different sparsity and structural constraints.

D. Reliability Evaluation

When the number of variables is large, it is probable that a random pair of variables will show a high correlation by chance. To make sure that the correlations obtained by ssCCA were statistically significant, we performed a non-parametric permutation analysis. At first, the canonical correlation was calculated for the original datasets. For the permutation test, the rows of one dataset ($\mathbf{X}$ or $\mathbf{Y}$) were randomly permutated for all features while keeping the other dataset intact. This process was repeated 10,000 times. The proposed ssCCA method was applied to each of these permutations. Then, the canonical correlation values of the extracted CVs in the permutated datasets were calculated and used to estimate the probability distribution of the canonical correlation of two CVs under the null hypothesis (pairs of CVs were correlated by chance). The p-value of the correlation of each pair of CVs was measured by the proportion of the correlations that were larger than the original correlation (real dataset). If the canonical correlation for the original datasets was small enough (less than 5%), then that pair of CVs were considered statistically significant.

E. Parameter Optimization

Optimization of the penalty parameters ($\lambda_1, \lambda_2, \alpha_1$, and α_2) for each pair of CVs was done by the $k-$fold cross-validation. The dataset was divided into k subsets (based upon subjects), of which $k-1$ subsets form the training set and the remaining subset forms the validation set. For each method, at each fold of the cross-validation, the estimation of the model (weight vectors) was done using the $k-1$ training samples and tested on the remaining sample. This was repeated k times, such that each subset functioned both as a validation set and as a part of the training set. Our cross-validation strategy had two steps [17]: (1) first, λ_1 and λ_2 were set to zero and the optimal values for α_1 and α_2 were found; and (2) using the optimal values of α_1 and α_2 obtained in the first stage, the optimal values of λ_1 and λ_2 were found. Based on the recommendation of [25] for selecting the tuning parameters, we minimized the mean absolute difference between the canonical correlation in the training and testing subsets in a cross-validation procedure. This measure is minimal when the test sample correlation is equal to the training sample correlation and thus its minimization increases the generalizability of the results and decreases the over-fitting of the model.

IV. Results

A. Simulation Study

To assess the performance of the proposed method, we simulated two correlated datasets to mimic real datasets. The aim of these simulation studies was to investigate the performance of the proposed method, over the traditional CCA-based fusion approaches such as standard CCA without regularization, i.e., $\lambda_1 = \lambda_2 = \alpha_1 = \alpha_2 = 0$ in (5) and sparse and regularized CCA [$l1 + l2$, i.e., $\alpha_1 = \alpha_2 = 0$ in (5)]. In all simulations, a 5-fold cross-validation was used to select the optimal parameters. Here, it should be noted that the traditional CCA and sCCA algorithms cannot be

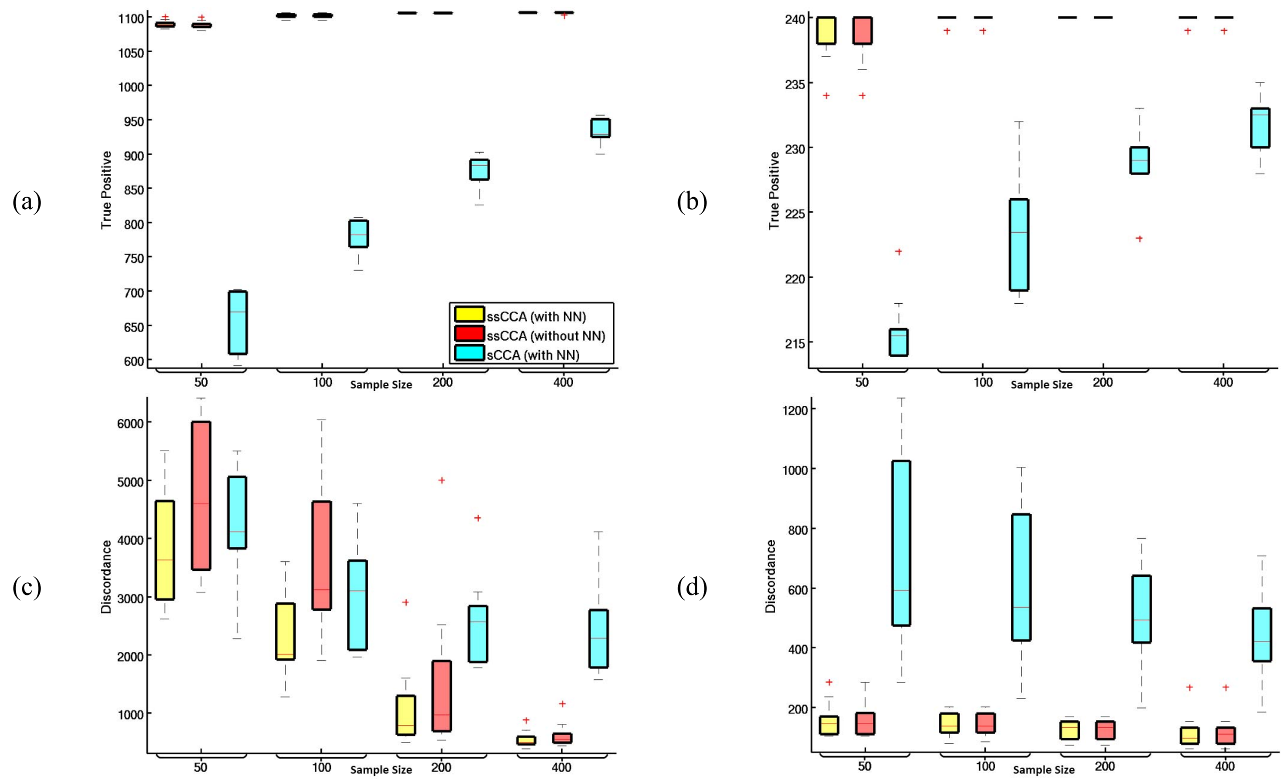


Fig. 3. Effects of sample size, dimensionality of features, and non-negativity constraint. The yellow, red, and blue boxplots show the results of ssCCA (with NN), ssCCA (without NN), and sCCA (with NN) using two different criteria: (a, b) true positive and (c, d) discordance. Plots (a, c) show the results for the first modality ($\boldsymbol{w}_1$) and (b, d) demonstrate the results for the second one ($\boldsymbol{w}_2$). The numbers of the truly correlated variables in the first and second modalities are 1107 and 240, respectively. The numbers of the uncorrelated variables in the two modalities are 188,518 and 53,600, respectively.

applied to a high dimensional multi-modal dataset without a unimodal feature selection (like PCA) as a pre-processing step. The proposed SVD framework (used in the general ssCCA framework) overcomes this problem.

In the simulation studies, the ground truth is known and the performance of the methods can be evaluated by computing the true positive (TP) (sensitivity), false positive (FP) ($1-$ specificity), i.e., the number of noise variables with non-zero weights in the CCC vector, false negative (FN), i.e., the number of correlated variables that have zero weight and are not selected, and discordance (FP + FN), which reflects the number of incorrectly identified variables. In all simulations, 50 replications were generated and the averages of TP and discordance over these 50 replications are reported.

The effect of sample size, dimensionality of the features, non-negativity constraint, and the standard deviation of the noise for different methods were investigated through the first and second simulation studies, respectively. Also in [26], we probed the performance of the fusion methods using a low dimensional simulated dataset.

B. Simulation 1: Effect of Sample Size, Dimensionality of Features, and Non-Negativity Constraint

The sample size and dimensionality of the available multi-modal datasets have a significant influence on the accuracy of the fusion methods. To demonstrate these effects, we varied the sample size from 25 to 200 in four levels of 25, 50, 100, and 200 samples per group. In addition, the simulated datasets have a different number of features, i.e., 189,625 and 53,840. In the simulated datasets, the range of intensity values for the simulated brain was between 797 and 9454. The standard deviation used for simulation of the latent variable μ and the added white Gaussian noise to the simulated brain were $\sigma_\mu = 1000$ and $\sigma_e = 7.0$, respectively. The average peak signal to noise ratio (PSNR) for the simulated images was -16.9 dB and SNR was 15.2 dB. Fig. 3 illustrates the effect of sample size and non-negativity (NN) constraint on TP and discordance of ssCCA and sCCA. This figure shows that, based on the TP criteria (Fig. 3. a-b), the proposed ssCCA (with and without NN) has a relatively accurate and stable TP pattern in different sample sizes for both of the modalities, in comparison with the sCCA (with NN) approach. The proposed method generates acceptable results even in low sample size scenarios.

In Fig. 3.c, the results demonstrate that, for high-dimensional data and low sample sizes, the non-negativity constraint is even more important than the smoothing constraint. However, by increasing the sample size, the result of ssCCA (without NN) converges to the result of ssCCA (with NN). The reason for these observations is that, in ssCCA (without NN), when the sample size increases to 400, automatically the CCC weights of all of the 50 replications become non-negative. Therefore, in this condition, there is no difference between ssCCA (with NN) and ssCCA (without NN). In addition, based on the discordance metric, for the first modality (Fig. 3.c), when the number of samples increases, the difference between the discordance of ssCCA (with and without NN) and sCCA also increases.

Finally, in Fig. 3.d, the results show that there is a small difference between the results of ssCCA (with NN) and ssCCA (without NN). The justification for these observations for the second modality is that the CCC vectors extracted from ssCCA (without NN) in all of the sample sizes and in all of the replications are non-negative. As a result, this constraint does not change the final results considerably and the TP and discordance of the ssCCA (with NN) and ssCCA (without NN) are almost the same. Fig. 3.d demonstrates that for the second modality, the discordance of ssCCA approach in different sample sizes is almost the same. The results suggest that when the dimensionality of the data decreases from 189,625 (first modality) to 53,600 (second modality), the performance of the proposed fusion approach improves. Based on the results of this section, in the following simulation and experimental studies, we have added the non-negativity constraint to ssCCA.

C. Simulation 2: Effect of Noise

To examine the effect of SNR on the accuracy of recovering truly correlated variables in each modality, the TP and discordance of the simulated data as a function of the standard deviation of the additive Gaussian white noise for the sCCA and ssCCA approaches are evaluated. In this simulation, the sample size was 50 (25 per group), σ_μ was 1000, and the standard deviation of the noise was 7.0, 8.0, and 9.0. For these simulated data, the PSNR varied from -16.9 to -19.1 dB and the SNR varied from 15.2 to 12.8 dB. Fig. S2.1 (in Supplementary Material 2) shows the effect of the noise variance on TP and discordance. The results show that, for a wide range of noise power, the ssCCA approach is superior to the sCCA approach. In addition, ssCCA is not very sensitive to the noise power.

D. Experimental Results

We used our proposed ssCCA method to analyze the correlation between structural and functional datasets of 76 subjects, which were randomly divided into two subsets: training and testing. The optimal parameters were obtained from the training data by 5-fold cross validation. The models were estimated as well as the features were selected from the training data using the optimal parameters. Then, these estimated models were applied to the testing data to predict the correlation between the two datasets.

Here and for both of the modalities we extracted three CVs, which all of them were statistically significant. However, we only used the first pair of CVs (Fig. 4) with a correlation of 85% (for comparison, the first CVs of sCCA is shown in Supplementary Material 2, Fig. S2.2). The first pair of CVs extracted the hidden pattern in the current population (42 HC vs. 34 AD) and the transition pattern between the AD and HC groups is clearly seen in the extracted subject-course. We performed a two-sample t-test on the first pair of CVs to test if the CVs were significantly different between the two groups in each modality. To this end, 100,000 random permutations were performed on the first pairs of CVs for each modality and after each permutation, the t-statistics was manipulated. After performing the 100,000 permutations, we counted the number of permutations whose t-statistics were higher than the t-statistics of the original CV. This fraction was used to determine the p-value of the t-test. The result shows that the p-value for the first pair of CVs is less than 1×10^{-6} between the AD and HC subjects for both modalities. To correct for the unbalance of the available dataset (34 AD vs. 42 HC), we used inverse probability weighting technique in doing the t-statistics analysis.

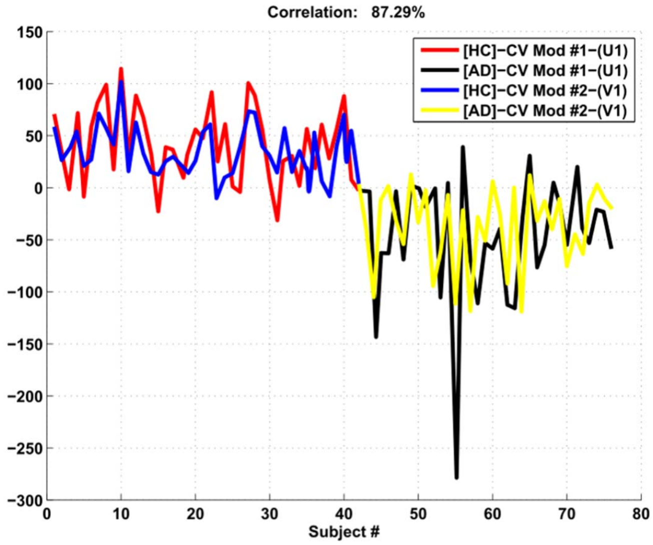


Fig. 4. The first CVs produced by the proposed ssCCA method where the correlation between the two CVs is 85.27%.

Note that the proposed method is an unsupervised learning process that finds these CVs without any knowledge of the status of each subject, and only by using the exploratory informative feature extraction, predicts the true subject-course of the data. Fig. 5 displays the ssCCA results on the brain in a manner similar to the traditional voxel-based analysis. This spatial map is produced based on the back-reconstruction of the CCC vectors for each modality to the 3D brain volume. Fig. 5 (2nd row) shows the brain mapping of functional area ($\mathbf{w}_2$) correlated with the anatomical ($\mathbf{w}_1$) changes in Fig. 5 (1st row) for the AD and HC subjects that are used to produce these CVs. There are brain ROIs in the first pair of CVs that are correlated between the structural to functional and functional to structural modalities, respectively (details are given in Tables S2.II and S2.III in Supplementary Material 2). To find the labels of the regions, we used the cortical and sub-cortical Harvard-Oxford Atlases and Cerebellar Atlas in the MNI152 space after normalization with FLIRT using FSL.

In this study, the goal was to simultaneously extract the group discriminating brain voxels in fMRI contrasts as well as the abnormal voxels reflected in the structural T1-weighted maps. To this end, we were able to visualize an underlying function–structure association by their joint analysis revealing strong sMRI–fMRI links. Several unimodal structural analysis such as voxel-based morphometry, volumetric and cortical

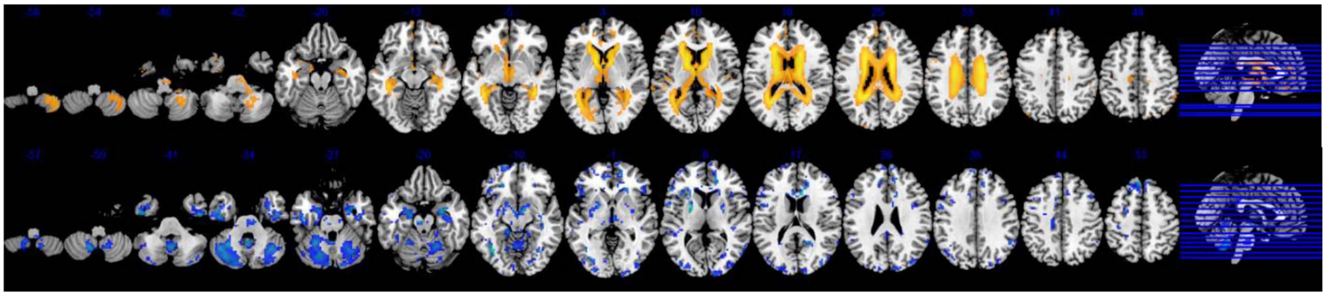


Fig. 5. The multi-slice view of the first canonical correlation coefficients of (top panel) structural ($\boldsymbol{w}_1$) and (bottom panel) functional ($\boldsymbol{w}_2$) modality, produced by ssCCA approach. These images show the most significant regions that are correlated between two modalities and can decode the true pattern (first pair of CVs) of difference between AD and HC subjects represented in Fig. 4.

thickness studies showed that all of these regions had significant atrophy in AD patients compared to HCs. They include caudate [27], hippocampus [1], lateral ventricle [28], thalamus [2], planum polare [29], planum temporale [29], heschl's gyrus [30], cerebral WM [31], accumbens [32], amygdala [1], [2], [29], paracingulate gyrus [33], and cerebellum [34].

In addition, several FC studies reported significant connectivity differences between the AD and HC groups in some areas. For example, in [35], it is shown that the AD group has more synchronization between sub-cortical regions such as pallidum and putamen and frontal cortices compared to the HC group. Also, in [36], it is shown that the FC between the amygdala and the default mode network (DMN) in the AD group is impaired. The AD group has also shown decreased FC between the left thalamus and inferior frontal gyrus [37] and between prefrontal lobe (such as frontal operculum cortex) and parietal lobes but increased positive correlations within the occipital lobe [38]. In addition, increased FC was observed between the bilateral thalamus and inferior temporal gyrus [37]. Furthermore, compared with the HCs, the AD patients have shown a significant amplitude of low-frequency fluctuations (ALFF) increases in the hippocampus [39]. Moreover, compared with the HCs, a decreased FC pattern was observed between the marginal division and the amygdala/parahippocampal region, the inferior frontal gyrus and the cerebellum for the AD patients [40].

V. Discussion

We have proposed a new CCA-based multi-modal data fusion approach that identified the unique and hidden pattern of subject variability in high dimensional structural and functional modalities of the AD patients. Most of the multivariate fusion methods in the literature rely on some form of a priori feature selection or feature extraction before invoking the final algorithm. In contrast, in our scalable algorithm, the feature selection method is not only spatially informed but also embedded, meaning that feature selection (which is done in other methods in a PCA step) is conducted together with a model fitting that leads to an improvement in prediction. Therefore, the proposed method can efficiently fit small-sample-large-variable problems. This method allows investigators to search in a data-driven fashion across the whole-brain for discovering correlated features across multiple modalities, which can potentially have completely different natures, signal- and contrast-to-noise ratios, voxel counts, and spatial smoothness.

In addition, the proposed method is a general, symmetric, and adaptive framework, based on the size, properties, and question at hand for the available data in each modality. Depending on the situation, one can turn the proposed constraints "on" or "off" for each of the modalities. In the full version of the proposed framework, different modalities are combined with equally important roles. We have shown that our approach finds the true patterns of the subjects-course between the AD patients and aged matched HC subjects with only the basic and minimal pre-processing steps, and that the extracted CVs are robust to population subsampling.

A. Interpretable Models for Whole-Brain Structural-Functional Fusion

The objective of ssCCA is to construct CVs that explain their own modality and at the same time are well correlated with the corresponding CVs in another modality. In other words, the first objective of ssCCA is to find the CVs that can explain a significant proportion of variance in each dataset. Its second objective is to find the CVs with relatively high correlation coefficients between the two datasets. Its third objective is that the first two objectives are obtained using the most interpretable and informative features. In contrast, the only objective of the standard CCA-based fusion is the construction of CVs that maximize their correlation coefficients with the CVs of another dataset. From this point of view, the CVs extracted by ssCCA are able to represent major information for individual modalities while the ones extracted by CCA may be trivial (e.g., noises with similar patterns) even if their correlation coefficient is maximum. Besides, ssCCA can handle high-dimensional and collinear data, which is often the case in real-world biological applications.

When the dimensions p and q are high, regularization is required in order to obtain a unique solution to the optimization problem. We propose a regularized version of sCCA-based fusion method that adds additional regularized terms that preserve local structure in the data while keeping the appealing properties of CCA. When the smoothness penalty is combined with the lasso penalty as in (4), the lasso penalty sets many of the CCCs to zero and for the remaining non-zero ones, the smoothing penalty encourages CCCs to take similar values by shrinking their differences toward zero.

Despite offering a sparse solution and automatic variable selection, there are several disadvantages to using *l1*-penalized methods like the lasso in practice. For example, lasso will typically select only a subset of "representative" predictors to include in the model [41]. This can make it difficult to interpret features (model will be too sparse) and sensitive to data resampling (e.g., during cross-validation) [42]. Moreover, lasso can select at most n non-zero coefficients out of p or q candidates [41], which may prove undesirable when the number of input features (p and q) exceeds the number of samples (n).

In this work, we presented a modification of the previously proposed methods that explicitly imposes structure on the model coefficients. This allows us to pre-specify constraints on the model coefficients (e.g., based on prior information like local smoothness), and then to tune these constraints. In other words, it helps us to construct an exploratory data-driven approach whose features are globally sparse but locally structured by the graph G (e.g., the Laplacian matrix). In addition, ssCCA promotes spatial contiguity, but instead of promoting sharp piecewise constant patches, it encourages the output clusters to appear in a smooth form.

B. General Comments on ssCCA

The number of CVs to be extracted from a CCA-based fusion approach is a very important parameter of the model. Although it is possible to extract it as the minimum rank of the datasets $\mathbf{X}$ and $\mathbf{Y}$, not all of them are generally useful. The main reason is that the measured data are never noise-free and some of the obtained CVs only describe the effect of noise, and it is common to ignore these CVs because of their small variation.

On a general note, it should be emphasized that interpretability of the results was one of our chief motivations in this work. As a result, the proposed method accommodates flexible constraints on model coefficients that give us the ability to detect a range of possible features, from smooth and localized to sparse and distributed. Beyond generalizability of the extracted features, a common approach for choosing the final set of features is to select the set that gives the highest prediction accuracy. As described in the Results Section, the proposed algorithm simultaneously generalizes the model by minimizing the difference between correlation in the training and testing subsets and constructs a classifier that yields classification accuracy (or goodness-of-fit) competitive with the state-of-the-art multivariate methods. However, the ultimate goal in our application is the discovery of informative and biologically inspired features (i.e., it should include relevant features while excluding nuisance features) and thus classification accuracy by itself is insufficient.

In addition, in this work, for the initialization of the ssCCA algorithm, we use the left and right singular vectors of $\mathbf{X}^T\mathbf{Y}$ using a new approach, i.e., QR decomposition instead of SVD. By this technique, we overcome the drawback of manipulating SVD in extra-large datasets. For example, in the simulation study, the dimensions of p and q are 188,518 and 53,600, respectively. Therefore, the dimensions of $\mathbf{X}^T\mathbf{Y}$ is 188,518×53,600 of floating numbers. Therefore, the required RAM for manipulating this matrix is about 38 GB, which is unavailable in personal computers and the computation is very time-consuming. However, by using the proposed techniques, the required memory is in the order of the number of the subjects, i.e., $n \times n$. Consequently, the datasets used in this study were analyzed in only 8 s using a desktop PC with Intel quad-core 2.93 GHz CPU and 8 GB of RAM. The computation times of ssCCA and sCCA are discussed in Supplementary Material 2.

C. Future Directions

Most of the existing approaches focus on the fusion of two modalities but additional benefits may be obtained by combining more than two modalities in a single model and examining N-way data fusion [7]. Although we used two modalities for fusion but the procedure can be easily extended to multi-modal data by an extension of the objective function of CCA. There is also the possibility of applying this technique to non-imaging modalities. For example, MRI data and non-imaging modalities like behavioral or genetic data can be directly combined in a multi-modal ssCCA framework. It would be interesting to test ssCCA on datasets with other sources of variability (e.g., relatively homogeneous group of healthy subjects with large age-span) to determine strongly age-related features in different modalities. In such kind of studies, the proposed method has the capability to model or adjust the effect of covariates such as age and gender.

Here, we have four optimization parameters that should be optimized to find the minimum of (5). As mentioned in Parameter Optimization section, our optimization algorithm has two steps. In the first step, we assume that two parameters λ_1 and λ_2 are zero. Then we optimize the equation to find the best α_1 and α_2 in a 2D search plane. In the second step, we assume that α_1 and α_2 are constant (based on the first step) and find the best λ_1 and λ_2 to minimize the criterion. Therefore, this optimization procedure is sub-optimal; in our future work, we will improve the optimization algorithm to find all four parameters in a single step.

VI. Conclusion

We presented a method for voxel-level fusion of whole-brain multi-modal neuroimaging datasets. The proposed method can automatically detect informative and biologically-inspired features (voxels) based on some heuristic and targeted regularizations such as sparsity, non-negativity, and smoothness constraints. The benefit of our proposed CCA-based fusion approach is that, the resulting canonical coefficients explain their own modality and are well-correlated with the corresponding canonical coefficients in the other modality. The power of the proposed method was demonstrated in the simulation studies where its performance was unanimously better than standard CCA or sparse and regularized CCA. Finally, in a real dataset, we found correlated ROIs that most of them were previously reported in unimodal structural and functional studies of AD.

Acknowledgment

The authors thank Natalie Wiseman for the proof reading of the manuscript.

References

[1] E. Westman, J.-S. Muehlboeck, and A. Simmons, "Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion," *NeuroImage*, vol. 62, no. 1, pp. 229–238, Aug. 2012.

[2] C.-Y. Wee *et al.*, "Identification of MCI individuals using structural and functional connectivity networks," *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, Feb. 2012.

[3] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, Feb. 2011.

[4] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[5] N. M. Correa, Y. O. Li, T. Adali, and V. D. Calhoun, "Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 6, pp. 998–1007, Dec. 2008.

[6] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 39–50, Jun. 2010.

[7] J. Sui *et al.*, "Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia," *NeuroImage*, vol. 66, pp. 119–132, Feb. 2013.

[8] É. L. Floch *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, Oct. 2012.

[9] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statist. Appl. Genet. Molecular Biol.*, vol. 8, no. 1, pp. 1–34, 2009.

[10] D. M. Witten, R. 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, apr. 2009.

[11] B. B. Avants *et al.*, "Sparse canonical correlation analysis relates network-level atrophy to multivariate cognitive measures in a neurodegenerative population," *NeuroImage*, vol. 84, pp. 698–711, Jan. 2014.

[12] F. Deleus and M. M. Van Hulle, "Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis," *J. Neurosci. Methods*, vol. 197, no. 1, pp. 143–157, Apr. 2011.

[13] C. Baldassano, M. C. Iordan, D. M. Beck, and L. Fei-Fei, "Voxel-level functional connectivity using spatial regularization," *NeuroImage*, vol. 63, no. 3, pp. 1099–1106, Nov. 2012.

[14] 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, May 2016.

[15] L. Du *et al.*, "Structured sparse CCA for brain imaging genetics via graph OSCAR," *BMC Syst. Biol.*, vol. 10, no. 3, p. 68, Aug. 2016.

[16] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li, "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, vol. 14, no. 2, pp. 244–258, Oct. 2013.

[17] D. Lin, V. D. Calhoun, and Y.-P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.*, vol. 18, no. 6, pp. 891–902, Aug. 2014.

[18] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2012, pp. 199–207.

[19] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids, with applications to DNA microarrays," *Statist. Sci.*, vol. 18, no. 1, pp. 104–117, Feb. 2003.

[20] A. Tenenhaus and M. Tenenhaus, "Regularized generalized canonical correlation analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, Apr. 2011.

[21] X. He and P. Niyogi, "Locality preserving projections," *Neural Inf. Process. Syst.*, vol. 16, pp. 585–591, Dec. 2004.

[22] G. I. Allen. (Sep. 2013). "Sparse and functional principal components analysis." [Online]. Available: https://arxiv.org/abs/1309.2895

[23] A. M. Wink, J. C. de Munck, Y. D. van der Werf, O. A. van den Heuvel, and F. Barkhof, "Fast eigenvector centrality mapping of voxel-wise connectivity in functional magnetic resonance imaging: Implementation, validation, and interpretation," *Brain Connectivity*, vol. 2, no. 5, pp. 265–274, Oct. 2012.

[24] G. Lohmann *et al.*, "Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain," *PLoS One*, vol. 5, no. 4, p. e10232, Apr. 2010.

[25] S. Waaijenborg, P. C. de W. Hamer, and A. H. Zwinderman, "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis," *Statist. Appl. Genet. Molecular Biol.*, vol. 7, no. 1, pp. 1–29, Jan. 2008.

[26] A.-R. Mohammadi-Nejad, G.-A. Hossein-Zadeh, and H. Soltanian-Zadeh, "Discovering true association between multimodal data sets using structured and sparse canonical correlation analysis: A simulation study," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 820–823.

[27] S. Jiji, K. A. Smitha, A. K. Gupta, V. P. M. Pillai, and R. S. Jayasree, "Segmentation and volumetric analysis of the caudate nucleus in Alzheimer's disease," *Eur. J. Radiol.*, vol. 82, no. 9, pp. 1525–1530, Sep. 2013.

[28] L. G. Apostolova *et al.*, "Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (MCI), and Alzheimer Disease," *Alzheimer disease Assoc. disorders*, vol. 26, no. 1, pp. 17–27, 2012.

[29] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski, "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification," *Neurobiol. Aging*, vol. 32, no. 12, pp. 2322.e19–2322.e27, Dec. 2011.

[30] J. A. Harasty, G. M. Halliday, J. J. Kril, and C. Code, "Specific temporoparietal gyral atrophy reflects the pattern of language dissolution in Alzheimer's disease," *Brain*, vol. 122, no. 4, pp. 675–686, Apr. 1999.

[31] J. Li, P. Pan, R. Huang, and H. Shang, "A meta-analysis of voxel-based morphometry studies of white matter volume alterations in Alzheimer's disease," *Neurosci. Biobehavioral Rev.*, vol. 36, no. 2, pp. 757–763, Feb. 2012.

[32] M. Pievani *et al.*, "Striatal morphology in early-onset and late-onset Alzheimer's disease: A preliminary study," *Neurobiol. Aging*, vol. 34, no. 7, pp. 1728–1739, Jul. 2013.

[33] V. Doré *et al.*, "A surface based approach for cortical thickness comparison between PiB+ and PiB− healthy control subjects," *Proc. SPIE*, vol. 8314, p. 831413, Feb. 2012.

[34] K. Andersen, B. B. Andersen, and B. Pakkenberg, "Stereological quantification of the cerebellum in patients with Alzheimer's disease," *Neurobiol. Aging*, vol. 33, no. 1, pp. 197.e11–197.e20, Jan. 2012.

[35] E. J. Sanz-Arigita *et al.*, "Loss of 'small-world' networks in Alzheimer's disease: Graph analysis of fMRI resting-state functional connectivity," *PLoS One*, vol. 5, no. 11, p. e13788, Nov. 2010.

[36] H. Yao *et al.*, "Decreased functional connectivity of the amygdala in Alzheimer's disease revealed by resting-state fMRI," *Eur. J. Radiol.*, vol. 82, no. 9, pp. 1531–1538, Sep. 2013.

[37] B. Zhou *et al.*, "Impaired functional connectivity of the thalamus in Alzheimer's disease and mild cognitive impairment: A resting-state fMRI study," *Current Alzheimer Res.*, vol. 10, no. 7, pp. 754–766, Aug. 2013.

[38] K. Wang *et al.*, "Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study," *Human Brain Mapping*, vol. 28, no. 10, pp. 967–978, Oct. 2007.

[39] Z. Dai *et al.*, "Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3)," *NeuroImage*, vol. 59, no. 3, pp. 2187–2195, Feb. 2012.

[40] Z. Zhang *et al.*, "Altered functional connectivity of the marginal division in Alzheimer's disease," *Current Alzheimer Res.*, vol. 11, no. 2, pp. 145–155, Mar. 2014.

[41] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. B, Statist. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[42] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with GraphNet," *NeuroImage*, vol. 72, no. 2, pp. 304–321, May 2013.