Adaptive structured sparse multiview canonical correlation analysis for multimodal brain imaging association identification

Lei DU[1*], Huiai WANG[1], Jin ZHANG[1], Shu ZHANG[2], Lei GUO[1],
Junwei HAN[1] & the Alzheimer's Disease Neuroimaging Initiative[†]

[1]*School of Automation, Northwestern Polytechnical University, Xi'an 710072, China;*
[2]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*

Abstract Multimodal brain imaging data can be obtained conveniently through rapidly advancing neuroimaging techniques. These multimodal data, which characterize the brain from distinct perspectives, offer a rare opportunity to comprehensively understand the neuropathology of complex brain disorders. Thus, identifying hidden relationships among multimodal brain imaging data is essential and meaningful. The pairwise correlation between two imaging modalities has been extensively studied. However, the multi-way association among more than three modalities remains unclear and is highly challenging. The difficulty and indeterminacy are largely due to the loss imbalances caused by multiple modalities fusion and the lack of reasonable consideration of the relationship implicated in different brain areas. To address both issues, we propose a structured sparse multiview canonical correlation analysis (SMCCA) with adaptive loss balancing and a novel graph-group penalty. The adaptive loss balancing technique encourages SMCCA to fairly optimize each sub-objective. The graph-group constraint penalizes the brain's regions of interest (ROIs) hierarchically with different regularizations at different levels. We derive an efficient algorithm and present its convergence. Experimental results on synthetic and real neuroimaging data confirm that, compared with state-of-the-art methods, our method is a better alternative as it identifies higher or comparable correlation coefficients and better canonical weights. Importantly, delivered by the canonical weights, the identified ROIs of each modality show a high correlation to each other and brain disorders, which demonstrates the potential of our method for untangling the intricate relationship among multimodal brain imaging data.

Keywords multimodal brain imaging correlation, sparse multiview canonical correlation analysis, task imbalance, adaptive loss balancing, graph-group penalty

1 Introduction

Alzheimer's disease (AD) is a severe chronic neurodegenerative disease, and its pathogenesis is still under way [1]. Neuroimaging techniques have opened the door to investigate AD and other brain disorders non-invasively. Owing to this advantage, we have acquired many neuroimaging manifestations of AD, including the atrophic brain structure, declined metabolic rates of cerebral glucose, and accumulation of extracellular plaques of amyloid beta (Aβ) [2,3]. In practice, the use of each characterization alone is generally insufficient to accurately diagnose an at-risk individual because many of them might not be AD-only characteristics [4]. For example, in addition to AD, hippocampal atrophy also happens to patients with frontotemporal lobar degeneration (FTLD) [5], and Aβ deposition is also associated with Parkinson's

disease [6]. Reasonably, a single marker is insufficient for the diagnosis or treatment of a specific disease. Hence, jointly using multiple imaging markers, especially those of multiple brain imaging modalities, can take full advantage of the cross-complementary information of different modalities. It can also provide rich knowledge of brain structure and function and is useful for developing deep neurophysiological characteristics. Taking AD as an example, structural brain atrophy, brain glucose metabolic rate, and β-amyloid extracellular plaque accumulation are all very sensitive markers, and they are carried by different brain imaging data. Evidently, a combination of them can help accurately identify an individual under risk. Therefore, jointly using the characteristics of multiple modalities other than those of a unimodal one could be beneficial to understand the neuropathology of AD and other disorders [7–13].

In general, there are two distinct strategies to take advantage of the complementary information of multimodal brain imaging data. The first kind, which offers the largest quantity of methods, is roughly divided into the feature-level fusion and result-level fusion [2]. As their names suggest, the feature-level fusion concatenates different imaging features of multimodal data as a whole, and the result-level fusion jointly analyzes the independent result obtained from every single modality. These methods have shown great success in case-control studies, such as classification, cognition ability prediction, and blind source separation [2, 7, 14]. Essentially, they mainly focus on prediction tasks and thus ignore the complex relationship embedded in multimodal data.

The second kind of multimodal method intends to exploit the relationship between heterogeneous imaging features of multiple modalities. This strategy has widened and deepened our knowledge of AD [15–18]. Examples include, but are not limited to, the association between Aβ burdens and white matter hyperintensities (WMHs) and that between brain atrophy and hypometabolism. The relationship between Aβ burdens and WMHs reveals that they affect cognition differently, and the structural atrophy is in agreement with the hypometabolism spanning the whole brain [18]. Importantly, this between-modality relationship could be a reasonable hallmark to accurately diagnose a suspected individual because a compositional characterization of multimodal data can be better than that of a unimodal one. As a result, accurately identifying the association between multimodal brain imaging data would be helpful in revealing the pathology of AD. Finally, it could help reduce the false-positive rate, determine the spectrum subtype of disorders, and lead to personalized medicine and precision medicine [19].

The above-mentioned methods are usually designed to mine the pairwise relationship between two imaging modalities. Hence, they cannot be directly applied to three or more modalities of imaging data, which dramatically limits their application in brain science, as currently, multiple types of brain imaging data from distinct perspectives carry rich information. To better understand AD, a critical and meaningful challenge is to seek the multi-way association embedded in heterogeneous multimodal brain imaging data. Therefore, designing novel methods for identifying the complex multi-way association from multimodal imaging data is an urgent need.

Sparse multiview canonical correlation analysis (SMCCA) [20, 21] could be an alternative, but two significant drawbacks keep us from using it. First, the pairwise associations of different modality-pairs could be different due to heterogeneous and diverse changing patterns in multiple modalities. This condition incurs the biased optimization issue for multiple sub-objectives which is called the loss imbalance [22]. Second, SMCCA either employs the ℓ_1-norm or fused lasso penalty and thus cannot reasonably reveal the hierarchical organization of the brain [23]. The human brain exhibits a marked symmetry across the sagittal plane, indicating a grouping structure between the left and right hemispheres [23]. Meanwhile, the brain's left and right hemispheres are not identical to each other, so the left and right components could have their own structure and function [24]. Thus, identifying relevant brain areas at different levels would be helpful in understanding the hierarchical structure of the human brain.

One may argue that a popular strategy exists in medical imaging, which uses multimodal information to obtain a good-quality brain imaging scan [25–28]. For example, Feng et al. [25] transferred multi-scale features from a target modality to the auxiliary modality, by which magnetic resonance (MR) imaging could be accepted for follow-up usage. They also designed a multi-stage integration network to extract the associations among hierarchical stages of different contrasts for super-resolution. As we aim to investigate AD's imaging characteristics under different modalities, these imaging techniques could be distinct and cannot address the issue raised above.

In this paper, we propose an adaptive structured SMCCA (i.e., gradient normalization-based SMCCA (GradNorm SMCCA)) with the gradient normalization-based loss balancing technique and a new graph-group penalty to overcome the aforementioned drawbacks. The gradient normalization technique dynamically adjusts the importance of each sub-objective to prompt an overall optimization. The graph-

group penalty function penalizes brain regions of interest (ROIs) hierarchically by first grouping the ROIs of the left and right hemispheres and then networking ROIs spanning the whole brain. We provided an effective algorithm to optimize our model and its convergence. We compared our method with SMCCA [20], Adaptive SMCCA [29], and graph-guided pairwise group lasso (GGL) SMCCA [30] using synthetic datasets and real neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ablation experimental results demonstrated that GradNorm SMCCA has flexible and good bi-multivariate association identification and feature selection capabilities. In particular, our method estimates the best overall association and more clear and meaningful imaging features of multiple modalities. Hence, it is very promising in untangling complex associations among multimodal brain imaging data.

2 Method

In this paper, we present the scalars as italic letters, column vectors as boldface lowercase letters, and matrices as boldface capitals. $\boldsymbol{X}_i \in \mathbb{R}^{n \times p}$ $(i = 1, \ldots, M)$ denotes the i-th modality of the brain imaging data with n subjects and p imaging quantitative traits (QTs), where M is the number of imaging modalities. $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm of the vector $\boldsymbol{x}$.

2.1 Sparse multiview canonical correlation analysis

SMCCA can be directly applied to analyze the pairwise association of more than three data matrices (modalities). According to [20], SMCCA is defined as follows:

$$\min_{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M} \sum_{i,j=1; i<j}^{M} -\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j \boldsymbol{u}_j + \sum_{i=1}^{M} \lambda_i \|\boldsymbol{u}_i\|_1, \tag{1}$$
$$\text{s.t.} \quad \|\boldsymbol{X}_i \boldsymbol{u}_i\|_2^2 = 1, \quad i = 1, \ldots, M,$$

which can be equivalently rewritten as

$$\min_{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M} \sum_{i,j=1; i<j}^{M} \|\boldsymbol{X}_i \boldsymbol{u}_i - \boldsymbol{X}_j \boldsymbol{u}_j\|_2^2 + \sum_{i=1}^{M} \lambda_i \|\boldsymbol{u}_i\|_1, \tag{2}$$
$$\text{s.t.} \quad \|\boldsymbol{X}_i \boldsymbol{u}_i\|_2^2 = 1, \quad i = 1, \ldots, M.$$

$\boldsymbol{u}_i \in \mathbb{R}^{p \times 1}$ denotes the canonical weight corresponding to $\boldsymbol{X}_i$, i.e., the i-th modality of imaging QTs. $\|\boldsymbol{u}_i\|_1$ is the ℓ_1-norm. λ_i is a tradeoff parameter used to control the model sparsity.

Nonetheless, SMCCA has two drawbacks. First, SMCCA treats every sub-objective equally. For multimodal problems, the correlations of different modality-pairs are usually different due to the heterogeneous and diverse changing patterns for distinct modalities. Therefore, treating all sub-objectives equally could be suboptimal because this simple fusion strategy might focus on easy sub-objectives while inadequately optimizing the hard ones. Second, SMCCA only selects imaging features at the individual level. As analyzed earlier, the human brain exhibits a complex organization of brain areas, and the lack of consideration of the hierarchical organization of the brain might be inefficient.

2.2 GradNorm SMCCA

To attain the overall optimization and account for the brain's hierarchical structure, we propose an adaptive structured SMCCA (GradNorm SMCCA) based on the gradient normalization loss balancing technique and a new graph-group penalty. We first define GradNorm SMCCA as follows:

$$\min_{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M} \sum_{i,j=1; i<j}^{M} s_{ij} \|\boldsymbol{X}_i \boldsymbol{u}_i - \boldsymbol{X}_j \boldsymbol{u}_j\|_2^2 + \sum_{i=1}^{M} \left(\lambda_{i1} \Omega_{g2}(\boldsymbol{u}_i) + \lambda_{i2} \|\boldsymbol{u}_i\|_2\right), \tag{3}$$
$$\text{s.t.} \quad \|\boldsymbol{X}_i \boldsymbol{u}_i\|_2^2 = 1, \quad i = 1, \ldots, M.$$

$\boldsymbol{X}_i$ and $\boldsymbol{u}_i$ are the same as those of SMCCA. s_{ij} here is a loss balancing parameter. λ_{i1} and λ_{i2} are nonnegative tuning parameters, which are usually obtained via the grid search strategy. $\Omega_{g2}(\boldsymbol{u}_i)$ is the

penalty identifying desirable imaging features in accordance with the brain's hierarchical structure. $\|\boldsymbol{u}_i\|_2$ is used to ensure model stability.

In summary, our method has two advantages. First, the adaptive weight s_{ij} can dynamically adjust the importance of each sub-objective and thus enables a comprehensive optimization for all objectives. Second, our method holds the grouping effect at two different levels, i.e., grouping the ROIs of the left and right hemispheres and highly correlated ROIs spanning the whole brain, indicating a flexible and meaningful feature selection capability. Next, we will introduce the gradient normalization method and graph-group penalty in detail.

2.2.1 *Gradient normalization-based loss balancing*

To ensure a comprehensive optimization, the balancing weight s_{ij} should be different for different sub-objectives. A naive strategy is to find s_{ij}'s through cross-validation. However, in practice, this is a very inefficient method due to the unsearchable candidate set and high computational effort. Therefore, an adaptive s_{ij}, which is time-varying in accordance with the sub-objective's importance during the optimization, is preferable.

Furthermore, in optimization, the convergence speed is largely determined by the gradient amplitude of the loss function. Hence, the biased optimization issue of a model (3) could be caused by the distinct gradients of its multiple sparse canonical correlation analysis (SCCA) sub-objectives. On this account, s_{ij} should be a function of the gradient of the i-th loss $\mathcal{L}_{ij}$, i.e.,

$$s_{ij} = f(\nabla \mathcal{L}_{ij}). \tag{4}$$

In addition, we utilize two ratios, e.g., the longitudinal ratio and the cross-sectional ratio, to control the optimization of the proposed SMCCA. The longitudinal ratio measures the optimization rate of each sub-objective at time t, where t is the iteration number. The cross-sectional ratio is the relative inverse training rate of each sub-objective. Clearly, the longitudinal ratio indicates the optimization rate of a single sub-objective at different iterations, while the cross-sectional one measures the optimization rate of multiple sub-objectives at the same iteration. Therefore, using both ratios in combination can effectively balance multiple sub-objectives and finally achieve the overall optimization [22].

Specifically, the longitudinal ratio is defined as

$$r_{ij}^{\texttt{long}}(t) = \frac{\mathcal{L}_{ij}(t)}{\mathcal{L}_{ij}(0)},$$

and the cross-sectional ratio is

$$r_{ij}^{\texttt{cs}}(t) = \frac{r_{ij}^{\texttt{long}}(t)}{E\left(\mathcal{L}_{ij}(t)\right)},$$

where $E\left(\mathcal{L}_{ij}(t)\right)$ is the average ratio of all sub-objectives.

We then define the loss function of s_{ij} as

$$L_{\texttt{grad}}(t; s_{ij}(t)) = \sum_{i,j=1,i \leqslant j}^{M} \left| G_{ij}(t) - \bar{G}(t) \times [r_{ij}^{\texttt{cs}}(t)]^\alpha \right|, \tag{5}$$

where $G_{ij}(t)$ denotes the ℓ_2-norm of the gradient of $s_{ij}\mathcal{L}_{ij}$ at the t-th iteration and $\bar{G}(t)$ is the average value of all $G_{ij}(t)$'s. α is a hyperparameter determining the strength of the restoring force, which pulls sub-objectives back to a common optimization rate [22].

Taken together, the gradient, longitudinal ratio, and cross-sectional ratio can prompt a comprehensive optimization for all sub-objectives.

2.2.2 *Graph-group penalty*

Because the human brain exhibits group and graph structures, a model that takes into account this hierarchical structure is preferable. To achieve this, we introduce a graph-group penalty $\Omega_{g2}(\cdot)$. Suppose there are $2q$ brain ROIs spanning the whole brain with each hemisphere having q ones, and then the k-th ($k \in [1, q]$) ROI is composed of two parts, i.e., the left and right i-th ROIs. For simplicity, the left and
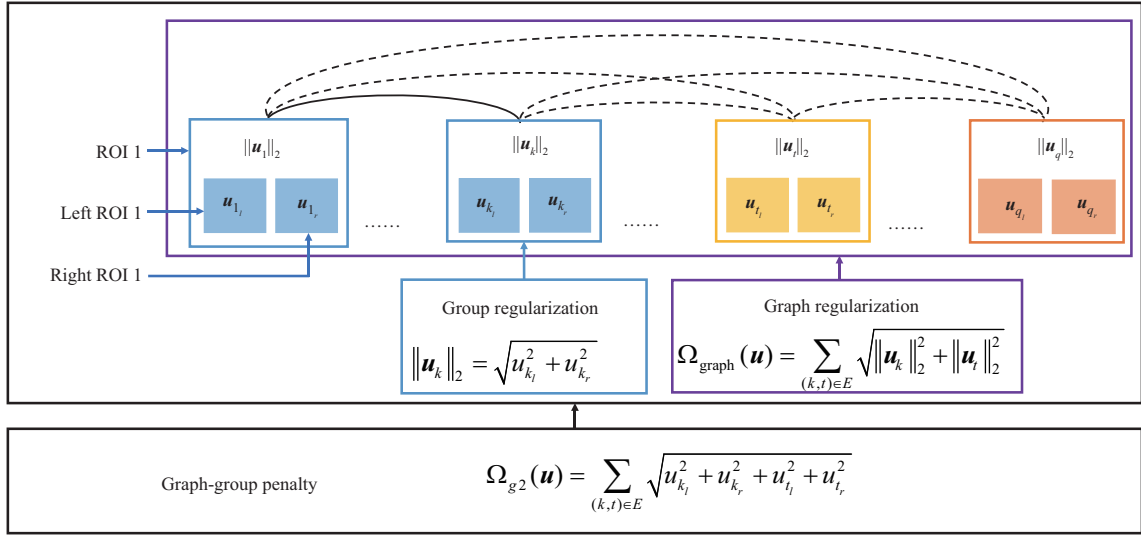
Figure 1 (Color online) Graph-group penalty. u_{k_l} and u_{k_r} denote the weights of the left and right parts of the k-th ROI, respectively. The left and right parts of each ROI first form a group, and then all ROIs form a graph with the same color, indicating a sub-graph. The dotted lines indicate that no assumption is made for the group structure, and thus every two ROIs are connected at the beginning of the algorithm. The solid line indicates that there indeed is a connection between the two ROIs.

right k-th ROIs are denoted by k_l and k_r, respectively. By mapping all ROIs to a graph where vertices have a one-to-one correspondence with ROIs, we define the graph-group penalty as follows:

$$\Omega_{g2}(\boldsymbol{u}) = \sum_{(k,t)\in E} \sqrt{u_{k_l}^2 + u_{k_r}^2 + u_{t_l}^2 + u_{t_r}^2}, \tag{6}$$

where E is the edge set of the graph and $\boldsymbol{u}$ is the weight vector corresponding to ROIs.

Clearly, $\Omega_{g2}(\cdot)$ jointly penalizes the left and right parts of the k-th and t-th ROIs. Figure 1 is the illustration of the graph-group penalty. According to Theorem 2 in [31], the difference between the estimated weight values of u_{k_l}, u_{k_r}, u_{t_l}, and u_{t_r} is upper-bounded. This grouping effect accommodates two advantages. First, the left and right parts of the same ROI could be grouped together. Second, taking the left and right parts of an ROI as a whole, two ROIs across the brain with high correlation will also be grouped together. Of note, the two connected ROIs do not need to be geographical neighbors. To fully take advantage of the brain imaging data, we use the graph-group penalty in the data-driven mode, indicating that every two ROIs are assumed to be connected at the beginning, and our model can automatically prune those unconnected ROI pairs. In summary, this novel penalty could simultaneously reveal the group structure and network structure in accordance with the organizational structure of the human brain.

2.2.3 Optimization algorithm

The proposed model (3) is non-convex and thus cannot be directly solved. However, this objective will be convex in each $\boldsymbol{u}_i$ if we fix the remaining canonical weights as constants. Here, taking $\boldsymbol{u}_i$ as an example, we present how to solve the GradNorm SMCCA model via the alternative convex search method [32].

Considering $\boldsymbol{u}_j$ $(j \neq i)$ as constants, we first write the Lagrangian function of (3) with respect to $\boldsymbol{u}_i$, i.e.,

$$\min_{\boldsymbol{u}_i} \sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\boldsymbol{u}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \left(\lambda_{i1} \sum_{(k,t)\in E} \sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2} + \lambda_{i2} \|\boldsymbol{u}_i\|_2^2 + \gamma \|\boldsymbol{X}_i\boldsymbol{u}_i\|_2^2 \right), \tag{7}$$

where u_{i,k_l} denotes the left k-th ROI of the canonical weight of the i-th imaging modality. Then, taking the derivative of (7) with respect to $\boldsymbol{u}_i$ and letting the gradient be zero, we can arrive at

$$\boldsymbol{u}_i = \left(\sum_{j=1,j\neq i}^{M} s_{ij}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i + \gamma\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i + \lambda_{i1}\boldsymbol{D}_{g2} + \lambda_{i2}\boldsymbol{I} \right)^{-1} \left(\sum_{j=1,j\neq i}^{M} s_{ij}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{u}_j \right). \tag{8}$$

$\boldsymbol{D}_{g2}$ is a block diagonal matrix, and the k-th block element is $(\sum_{t=1,t\neq k}^{q} \frac{1}{2\sqrt{u_{i,k_l}^2+u_{i,k_r}^2+u_{i,t_l}^2+u_{i,t_r}^2}})\boldsymbol{I}_k$, where

$\boldsymbol{I}_k$ is an identity matrix that has the same size as the k-th group. In particular, $\boldsymbol{I}_k$ equals $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ because each ROI includes the left and right components.

Next, we continue to update s_{ij} based on the GradNorm algorithm [22]. Similarly, we take the derivative of (5) with respect to s_{ij} and set it to zero:

$$\frac{\partial L_{\text{grad}}(s_{ij})}{\partial s_{ij}} = 0. \tag{9}$$

Then we can easily update each s_{ij} accordingly.

Finally, we present the pseudocode in Algorithm 1, where every $\boldsymbol{u}_i$ and s_{ij} are alternately calculated until convergence. In this algorithm, the parameters are generally pre-tuned using the cross-validation or holdout method, and thus they are treated as inputs here. Each $\boldsymbol{u}_i$ ($i = 1, \ldots, M$) is initialized at the beginning of the procedure. Then, a nested loop is called to find the optimal solution. In the internal loop, each $\boldsymbol{u}_i$ is alternatively updated based on the closed-form equation. The external loop is used to update the gradient of each sub-objective.

Algorithm 1 GradNorm SMCCA algorithm

Require:
  $\boldsymbol{X}_i \in \mathcal{R}^{n \times p}$, λ_{i1}, λ_{i2}, γ, $i = 1, \ldots, M$.
Ensure:
  Output $\boldsymbol{u}_i$, $i = 1, \ldots, M$.
1: Initialize $\boldsymbol{u}_i \in \mathcal{R}^{p \times 1}$, $s_{ij} = 1$, $i = 1, \ldots, M$, $i \neq j$; ▷ Initialization.
2: **while** not convergence **do**
3:   **for** $i = 1, \ldots, M$ **do**
4:     Calculate the diagonal matrix $\boldsymbol{D}_{g2}$; ▷ Compute the current sub-gradient for the graph-group penalty.
5:     Update $\boldsymbol{u}_i$ according to (8) and scale $\boldsymbol{u}_i$ by $\boldsymbol{u}_i = \frac{\boldsymbol{u}_i}{\|\boldsymbol{X}_i\boldsymbol{u}_i\|_2}$; ▷ Compute the current weight for the i-th modality.
6:   **end for**
7:   Update each s_{ij} according to (9); ▷ Compute the current gradient for each sub-objective.
8: **end while**

2.2.4 *Convergence analysis*

We now prove the convergence of Algorithm 1.

Theorem 1. The objective value of (3) is decreased in each iteration of Algorithm 1.

Proof. Let $\bar{\boldsymbol{u}}_i$ denote the updated $\boldsymbol{u}_i$ for ease of presentation. According to (8), we have

$$\sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\bar{\boldsymbol{u}}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \lambda_{i1}\bar{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{D}_{g2}\bar{\boldsymbol{u}}_i + \lambda_{i2}\bar{\boldsymbol{u}}_i^{\mathrm{T}}\bar{\boldsymbol{u}}_i + \gamma\bar{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\bar{\boldsymbol{u}}_i$$
$$\leqslant \sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\boldsymbol{u}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \lambda_{i1}\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{D}_{g2}\boldsymbol{u}_i + \lambda_{i2}\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_i + \gamma\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{u}_i, \tag{10}$$

which can be rewritten as

$$\sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\bar{\boldsymbol{u}}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \lambda_{i2} \|\bar{\boldsymbol{u}}_i\|_2^2 + \gamma \|\boldsymbol{X}_i\bar{\boldsymbol{u}}_i\|_2^2 + \lambda_{i1}\sum_{k=1}^{q}\sum_{t=1;t\neq k}^{q} \frac{\bar{u}_{i,k_l}^2 + \bar{u}_{i,k_r}^2}{2\sqrt{u_{i,k_l}^2+u_{i,k_r}^2+u_{i,t_l}^2+u_{i,t_r}^2}}$$
$$\leqslant \sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\boldsymbol{u}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \lambda_{i2} \|\boldsymbol{u}_i\|_2^2 + \gamma \|\boldsymbol{X}_i\boldsymbol{u}_i\|_2^2$$
$$+ \lambda_{i1}\sum_{k=1}^{q}\sum_{t=1;t\neq k}^{q} \frac{u_{i,k_l}^2 + u_{i,k_r}^2}{2\sqrt{u_{i,k_l}^2+u_{i,k_r}^2+u_{i,t_l}^2+u_{i,t_r}^2}}, \tag{11}$$

and further

$$\sum_{j=1,j\neq i}^{M} s_{ij} \|\boldsymbol{X}_i\bar{\boldsymbol{u}}_i - \boldsymbol{X}_j\boldsymbol{u}_j\|_2^2 + \lambda_{i2} \|\bar{\boldsymbol{u}}_i\|_2^2 + \gamma \|\boldsymbol{X}_i\bar{\boldsymbol{u}}_i\|_2^2 + \lambda_{i1}\sum_{(k,t)\in E} \frac{\bar{u}_{i,k_l}^2 + \bar{u}_{i,k_r}^2 + \bar{u}_{i,t_l}^2 + \bar{u}_{i,t_r}^2}{2\sqrt{u_{i,k_l}^2+u_{i,k_r}^2+u_{i,t_l}^2+u_{i,t_r}^2}}$$

$$\leqslant \sum_{j=1,j\neq i}^{M} s_{ij} \left\| \boldsymbol{X}_i \boldsymbol{u}_i - \boldsymbol{X}_j \boldsymbol{u}_j \right\|_2^2 + \lambda_{i2} \left\| \boldsymbol{u}_i \right\|_2^2 + \gamma \left\| \boldsymbol{X}_i \boldsymbol{u}_i \right\|_2^2$$

$$+ \lambda_{i1} \sum_{(k,t)\in E} \frac{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}{2\sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}}. \tag{12}$$

On the basis of the inequality in Lemma 1 of [33]: $\|\boldsymbol{a}\|_2 - \frac{\|\boldsymbol{a}\|_2^2}{2\|\boldsymbol{a}_0\|_2} \leqslant \|\boldsymbol{a}_0\|_2 - \frac{\|\boldsymbol{a}_0\|_2^2}{2\|\boldsymbol{a}_0\|_2}$, we easily have

$$\sqrt{\bar{u}_{i,k_l}^2 + \bar{u}_{i,k_r}^2 + \bar{u}_{i,t_l}^2 + \bar{u}_{i,t_r}^2} - \frac{\bar{u}_{i,k_l}^2 + \bar{u}_{i,k_r}^2 + \bar{u}_{i,t_l}^2 + \bar{u}_{i,t_r}^2}{2\sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}}$$

$$\leqslant \sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2} - \frac{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}{2\sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}}. \tag{13}$$

Then, we repeatedly apply (13) to (11) with respect to the weight of each ROI, and we arrive at

$$\sum_{j=1,j\neq i}^{M} s_{ij} \left\| \boldsymbol{X}_i \bar{\boldsymbol{u}}_i - \boldsymbol{X}_j \boldsymbol{u}_j \right\|_2^2 + \lambda_{i2} \left\| \bar{\boldsymbol{u}}_i \right\|_2^2 + \gamma \left\| \boldsymbol{X}_i \bar{\boldsymbol{u}}_i \right\|_2^2 + \lambda_{i1} \sum_{(k,t)\in E} \sqrt{\bar{u}_{i,k_l}^2 + \bar{u}_{i,k_r}^2 + \bar{u}_{i,t_l}^2 + \bar{u}_{i,t_r}^2}$$

$$\leqslant \sum_{j=1,j\neq i}^{M} s_{ij} \left\| \boldsymbol{X}_i \boldsymbol{u}_i - \boldsymbol{X}_j \boldsymbol{u}_j \right\|_2^2 + \lambda_{i2} \left\| \boldsymbol{u}_i \right\|_2^2 + \gamma \left\| \boldsymbol{X}_i \boldsymbol{u}_i \right\|_2^2$$

$$+ \lambda_{i1} \sum_{(k,t)\in E} \sqrt{u_{i,k_l}^2 + u_{i,k_r}^2 + u_{i,t_l}^2 + u_{i,t_r}^2}. \tag{14}$$

Writing this inequality into the matrix form, we obtain

$$\sum_{j=1,j\neq i}^{M} s_{ij} \left\| \boldsymbol{X}_i \bar{\boldsymbol{u}}_i - \boldsymbol{X}_j \boldsymbol{u}_j \right\|_2^2 + \lambda_{i1} \sum_{(k,t)\in E} \Omega_{g2}(\bar{\boldsymbol{u}}_i) + \lambda_{i2} \left\| \bar{\boldsymbol{u}}_i \right\|_2^2 + \gamma \left\| \boldsymbol{X}_i \bar{\boldsymbol{u}}_i \right\|_2^2$$

$$\leqslant \sum_{j=1,j\neq i}^{M} s_{ij} \left\| \boldsymbol{X}_i \boldsymbol{u}_i - \boldsymbol{X}_j \boldsymbol{u}_j \right\|_2^2 + \lambda_{i1} \sum_{(k,t)\in E} \Omega_{g2}(\boldsymbol{u}_i) + \lambda_{i2} \left\| \boldsymbol{u}_i \right\|_2^2 + \gamma \left\| \boldsymbol{X}_i \boldsymbol{u}_i \right\|_2^2. \tag{15}$$

Now, we can deduce that the objective decreases when updating $\boldsymbol{u}_i$. Likewise, we can draw the same conclusion with respect to the remaining weights of $[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_M]$. Moreover, $L_{\text{grad}}(\boldsymbol{s})$ is convex in each s_{ij}, and thus updating s_{ij} will not break the convergence of Algorithm 1. Combining these conclusions together, the proof is completed.

The objective value of (3) is lower-bounded by 0. This, coupled with Theorem 1, indicates that Algorithm 1 will converge to the global or a local optimum.

3 Experiments and results

3.1 Experiments setup

We carefully chose four related methods, namely, SMCCA, Adaptive SMCCA, GGL-SMCCA [30], and g2SMCCA, for comparison. Because we focused on analyzing imaging data after scanning, the imaging methods which aim to obtain imaging scans [25–28] were not included in the comparison. The decomposed SMCCA [21] was also excluded because it was a specific form of SMCCA.

- SMCCA [20]. SMCCA directly combines multiple SCCA objects without accounting for the task imbalance issue. Thus, it suffers from a severe optimization bias under the task imbalance circumstance.
- Adaptive SMCCA [29]. This method introduces adaptive weights to SMCCA to balance among multiple sub-objectives. However, its task-balancing technique and additional assumption on covariance break the Pearson correlation coefficient and might incur unknown risks [34].

- GGL-SMCCA [30]. SMCCA and Adaptive SMCCA only use the ℓ_1-norm as the penalty and thus are insufficient to identify the complex organization of the brain. GGL-SCCA [30] can determine the graph structure but cannot balance among multiple sub-objectives. Therefore, extended to the multi-modal mode, GGL-SMCCA can be a good alternative to access the performance of our task-balancing mechanism.

- g2SMCCA. This is a reduced version of our proposed method that wipes off the balancing weight s_{ij}.

The experiments were performed via the nested fivefold cross-validation method where the inner loop was in charge of tuning the parameters. The parameters with the highest task average canonical correlation coefficients (CCCs) were regarded as the optimal parameters, where $\mathrm{CCC} = \frac{u_i^{\mathrm{T}} X_i^{\mathrm{T}} X_j u_j}{\sqrt{u_i^{\mathrm{T}} X_i^{\mathrm{T}} X_i u_i}\sqrt{u_j^{\mathrm{T}} X_j^{\mathrm{T}} X_j u_j}}$ with X_i's being centered. The candidate interval was first set to 10^i, $i = -3, -2, -1, 0, 1, 2, 3$. Once we obtained the winner parameters, denoted by Θ, we narrowed the interval, i.e., $\Theta \pm 0.5$, with a small step of 0.1. The external loop calculated and collected the experimental results of the optimal parameters. The hyperparameter α in GradNorm was given as 0.5 because a lower value could be suitable for the symmetric GradNorm SMCCA [22]. All the methods were run on the same software platform and data partition. The stopping condition was set to $\max_i |(u_i)^{t+1} - (u_i)^t| \leqslant 10^{-5}$, $i = 1, 2, 3$.

3.2 Results on synthetic data

3.2.1 *Data source*

To efficiently present the performance of our method, we generated six synthetic datasets with graph-group properties, different dimensions, and different levels of balance. The data generation procedure was similar to that used in [29]. The latent variables μ_1–$\mu_4 \in \mathbb{R}^{n \times 1}$, $n = 120$ were generated from a standard normal distribution, and the correlations between tasks were accommodated therein. Specifically, we designed a group structure by assigning a pair of variables with the same value. Similarly, the variables in the same graph were assigned the same value. This setup enabled the variables in the same group and graph to have the same or similar contribution. Data 1–4 $\in \mathbb{R}^{n \times q_1}$ were low dimensional with $q_1 = 150$. Data 5 $\in \mathbb{R}^{n \times q_2}$ and Data 6 $\in \mathbb{R}^{n \times q_3}$ simulated a relatively high dimensional problem where $q_2 = 500$ and $q_3 = 1000$, respectively. Moreover, Data 1 simulated a task-balanced problem, and the remaining datasets were imbalanced ones. Distinct imbalances were created by different noise intensities. Specifically, we introduced nine intermediate vectors α_1–α_9 and generated the data matrices as follows:

$$X_1 = \mu_1 \alpha_1 + \mu_2 \alpha_2 + \mu_4 \alpha_3 + \epsilon,$$
$$X_2 = \mu_1 \alpha_4 + \mu_3 \alpha_5 + \mu_4 \alpha_6 + \epsilon,$$
$$X_3 = \mu_2 \alpha_7 + \mu_3 \alpha_8 + \mu_4 \alpha_9 + \epsilon,$$

where ϵ was the noise signal. Thus, canonical weights were the summation of the intermediate vectors, i.e., $u_1 = \alpha_1 + \alpha_2 + \alpha_3$, $u_2 = \alpha_4 + \alpha_5 + \alpha_6$, $u_3 = \alpha_7 + \alpha_8 + \alpha_9$. By changing the noise intensities of intermediate vectors, we can generate different levels of imbalances. The details of each dataset are presented below.

- Data 1–4: We first introduced two vectors $\beta_1 = (\underbrace{2, \ldots, 2}_{10}, \underbrace{1, \ldots, 1}_{10}, \underbrace{-1, \ldots, -1}_{12})^{\mathrm{T}}$ and $\beta_2 = (2, 2, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, -1, -1, 0, 0)^{\mathrm{T}}$. Then the other three vectors $\alpha_3 = (\underbrace{0, \ldots, 0}_{40}, \beta_2, \beta_2, \beta_2, \beta_2, \underbrace{0, \ldots, 0}_{46})^{\mathrm{T}}$, $\alpha_6 = (\underbrace{0, \ldots, 0}_{30}, \beta_1, \beta_2, \beta_2, \underbrace{0, \ldots, 0}_{56})^{\mathrm{T}}$, and $\alpha_9 = (\underbrace{0, \ldots, 0}_{50}, \beta_1, \beta_1, \underbrace{0, \ldots, 0}_{36})^{\mathrm{T}}$. In Data 1, all the remaining intermediate vectors followed the normal distribution $N(0, 0.1)$, and thus all its sub-tasks were balanced. In Data 2, α_1 and α_4 followed $N(0, 0.2)$, and thus it was a task-imbalanced problem. Similarly, in Data 3, α_2 and α_7 were generated from $N(0, 0.2)$. In Data 4, its α_5 and α_8 followed $N(0, 0.2)$. Therefore, Data 2–4 were unbalanced datasets.

- Data 5: $\beta_3 = (\underbrace{2, \ldots, 2}_{20}, \underbrace{1, \ldots, 1}_{20}, \underbrace{-1, \ldots, -1}_{24})^{\mathrm{T}}$, $\beta_4 = (\underbrace{2, \ldots, 2}_{8}, \underbrace{0, \ldots, 0}_{8}, \underbrace{1, \ldots, 1}_{8}, \underbrace{0, \ldots, 0}_{8}, \underbrace{-1, \ldots, -1}_{8}, \underbrace{0, \ldots, 0}_{8})^{\mathrm{T}}$. $\alpha_3 = (\underbrace{0, \ldots, 0}_{170}, \beta_4, \beta_4, \beta_4, \beta_4, \underbrace{0, \ldots, 0}_{138})^{\mathrm{T}}$, $\alpha_6 = (\underbrace{0, \ldots, 0}_{150}, \beta_3, \beta_3, \beta_4, \beta_4, \underbrace{0, \ldots, 0}_{126})^{\mathrm{T}}$, $\alpha_9 = (\underbrace{0, \ldots, 0}_{200},$

Table 1 CCCs (mean±std) estimated from synthetic data sets[a]

|		Training canonical correlation coefficient					Testing canonical correlation coefficient				
		SMCCA	GGL SMCCA	Adaptive SMCCA	g2SMCCA	GradNorm SMCCA	SMCCA	GGL SMCCA	Adaptive SMCCA	g2SMCCA	GradNorm SMCCA
Data 1	Task1	0.78±0.01	0.78±0.01	0.78±0.01	0.81±0.01	0.84±0.01	0.66±0.08	0.68±0.08	0.66±0.09	0.74±0.06	0.75±0.05
	Task2	0.83±0.01	0.84±0.01	0.83±0.01	0.85±0.01	0.87±0.01	0.77±0.09	0.78±0.08	0.77±0.09	0.79±0.08	0.80±0.07
	Task3	0.85±0.02	0.86±0.01	0.85±0.02	0.88±0.02	0.91±0.01	0.76±0.08	0.78±0.08	0.75±0.09	0.82±0.07	0.82±0.07
	Average	0.82	0.83	0.82	0.84	**0.87**	0.73	0.74	0.73	0.78	**0.79**
Data 2	Task1	0.74±0.02	0.66±0.01	0.74±0.02	0.73±0.01	0.81±0.01	0.57±0.17	0.48±0.20	0.57±0.15	0.66±0.04	0.71±0.09
	Task2	0.75±0.03	0.68±0.01	0.76±0.02	0.78±0.02	0.86±0.01	0.60±0.12	0.52±0.15	0.62±0.13	0.68±0.10	0.74±0.10
	Task3	0.78±0.04	0.72±0.05	0.80±0.04	0.84±0.02	0.91±0.01	0.66±0.11	0.56±0.15	0.69±0.11	0.77±0.09	0.81±0.05
	Average	0.76	0.69	0.77	0.79	**0.86**	0.61	0.52	0.63	0.70	**0.76**
Data 3	Task1	0.80±0.02	0.72±0.03	0.80±0.02	0.81±0.01	0.86±0.01	0.67±0.17	0.53±0.22	0.67±0.17	0.70±0.14	0.76±0.11
	Task2	0.79±0.04	0.73±0.06	0.80±0.04	0.81±0.03	0.86±0.01	0.64±0.12	0.48±0.21	0.64±0.14	0.69±0.13	0.73±0.11
	Task3	0.84±0.02	0.81±0.02	0.85±0.02	0.88±0.01	0.93±0.01	0.74±0.08	0.66±0.11	0.76±0.08	0.80±0.11	0.86±0.07
	Average	0.81	0.75	0.82	0.83	**0.88**	0.68	0.56	0.69	0.73	**0.78**
Data 4	Task1	0.75±0.01	0.68±0.04	0.75±0.01	0.77±0.02	0.82±0.01	0.63±0.14	0.52±0.18	0.62±0.14	0.68±0.10	0.73±0.06
	Task2	0.77±0.01	0.73±0.02	0.78±0.01	0.79±0.01	0.84±0.01	0.66±0.07	0.50±0.18	0.68±0.06	0.74±0.03	0.75±0.03
	Task3	0.80±0.02	0.74±0.03	0.81±0.02	0.84±0.02	0.91±0.01	0.65±0.09	0.51±0.15	0.68±0.09	0.76±0.06	0.80±0.05
	Average	0.77	0.72	0.78	0.80	**0.86**	0.65	0.51	0.66	0.73	**0.76**
Data 5	Task1	0.89±0.03	0.84±0.04	0.91±0.02	0.94±0.01	0.97±0.01	0.73±0.13	0.58±0.14	0.79±0.11	0.85±0.07	0.87±0.06
	Task2	0.89±0.02	0.84±0.04	0.91±0.01	0.94±0.01	0.97±0.01	0.76±0.08	0.64±0.05	0.81±0.09	0.84±0.10	0.87±0.09
	Task3	0.91±0.01	0.87±0.03	0.92±0.01	0.95±0.01	0.98±0.01	0.78±0.08	0.67±0.08	0.83±0.07	0.88±0.05	0.90±0.05
	Average	0.90	0.85	0.91	0.94	**0.97**	0.76	0.63	0.81	0.86	**0.88**
Data 6	Task1	0.93±0.01	0.90±0.02	0.94±0.01	0.95±0.01	0.97±0.01	0.86±0.07	0.79±0.09	0.89±0.06	0.91±0.05	0.93±0.04
	Task2	0.93±0.02	0.91±0.03	0.95±0.01	0.96±0.01	0.98±0.01	0.85±0.07	0.80±0.08	0.88±0.06	0.90±0.04	0.93±0.03
	Task3	0.95±0.01	0.93±0.02	0.96±0.01	0.97±0.01	0.98±0.01	0.88±0.05	0.83±0.06	0.91±0.04	0.92±0.03	0.95±0.02
	Average	0.94	0.92	0.95	0.96	**0.98**	0.86	0.81	0.89	0.91	**0.94**

a) The best average value was shown in bold.

$\beta_3, \beta_3, \beta_3, \underbrace{0, \ldots, 0}_{108})^{\mathrm{T}}$. The intermediate vectors α_1 and α_4 followed $N(0, 0.2)$, where the other vectors followed $N(0, 0.1)$.

• Data 6: $\beta_5 = (\underbrace{2, \ldots, 2}_{40}, \underbrace{1, \ldots, 1}_{40}, \underbrace{-1, \ldots, -1}_{40})^{\mathrm{T}}$, $\beta_6 = (\underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{-1, \ldots, -1}_{10},$
$\underbrace{0, \ldots, 0}_{10})^{\mathrm{T}}$. $\alpha_3 = (\underbrace{0, \ldots, 0}_{350}, \underbrace{\beta_6, \ldots, \beta_6}_{6}, \underbrace{0, \ldots, 0}_{290})^{\mathrm{T}}$, $\alpha_6 = (\underbrace{0, \ldots, 0}_{300}, \beta_5, \beta_5, \beta_6, \beta_6, \underbrace{0, \ldots, 0}_{340})^{\mathrm{T}}$, $\alpha_9 = (\underbrace{0, \ldots, 0}_{400},$
$\beta_5, \beta_5, \beta_5, \underbrace{0, \ldots, 0}_{240})^{\mathrm{T}}$. The intermediate vectors α_1 and α_4 followed $N(0, 0.2)$, and the remaining vectors were generated from $N(0, 0.1)$. Thus, Data 5 and 6 were unbalanced datasets with relatively high dimensions.

3.2.2 *Multi-way association identification*

In this study, we used the CCCs as the first evaluation criterion. A higher CCC score stands for better performance. Table 1 shows the training and testing CCCs of all methods and the average value across three tasks. Clearly, GradNorm SMCCA obtained the best average CCCs for all the datasets, indicating its outperformed overall performance. g2SMCCA was inferior to GradNorm SMCCA but was better than the remaining benchmarks. This finding demonstrated that using the novel graph-group penalty can enhance the identification power. In particular, all the methods performed well on Data 1. Hence, under a balanced situation, with or without the loss-balancing mechanism can handle it. However, on Data 2–6, where different sub-objectives were imbalanced, GradNorm SMCCA significantly outperformed the other competitors. Moreover, in all unbalanced datasets, Adaptive SMCCA performed better than SMCCA. This finding revealed the necessity of equipping with the task-balancing mechanism. Moreover, on all six datasets, g2SMCCA performed better than SMCCA and GGL SMCCA and thus indicated that the graph-group penalty could help improve the association by identifying hierarchical imaging features. In summary, GradNorm SMCCA performed the best in terms of CCC owing to its task-balancing technique
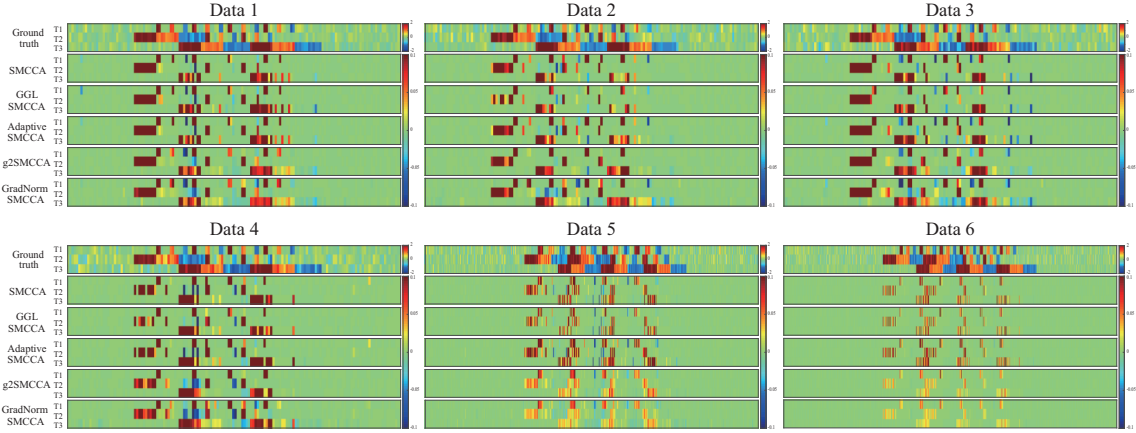
Figure 2 (Color online) Comparison of canonical weights in terms of each task for six synthetic datasets. The top row exhibits the ground truth, and the other rows correspond to the SMCCA methods: (1) SMCCA; (2) GGL SMCCA; (3) Adaptive SMCCA; (4) g2SMCCA; (5) GradNorm SMCCA. Moreover, there are three rows associated with three tasks (T1, T2, T3) in each panel.

and graph-group penalty.

3.2.3 *Feature selection*

In addition to CCCs, feature selection is also a very important criterion. We exhibited the feature selection results in Figure 2. Again, all methods performed successfully identified the true signals on Data 1 because it was an easy problem. On Data 2–6, where multiple sub-objectives were unequal, our method performed the best because the comparison methods lost several features for some sub-tasks. For example, for the third sub-task, GradNorm SMCCA recovered the truest signals. Compared to g2SMCCA, GradNorm SMCCA recognized comprehensive nonzero features for most tasks. For instance, in Data 4, GradNorm SMCCA identified more signals missed by g2SMCCA, especially in task 1. Hence, the task-balancing technique could also improve the overall feature subset selection. Moreover, owing to the graph-group penalty, the features selected by g2SMCCA and GradNorm SMCCA showed a clear group structure. While the heatmap is somewhat qualitative, a quantitative evaluation criterion could be desirable. Therefore, we further employed the sensitivity and specificity [35] for a quantitative comparison. The sensitivity was defined as $\frac{\text{true positive in the top } K \text{ selected features}}{K}$, and the specificity was calculated according to $\frac{K}{\text{selected features required to cover the ground truth}}$, where K denotes the number of nonzero features in line with the ground truth. In both criteria, the higher, the better, as presented in Table 2. Evidently, GradNorm SMCCA outperformed the five methods with respect to the average value, even on task-balanced Data 1. Of note, g2SMCCA also performed better than the other four methods due to its well-designed graph-group penalty. Interestingly, our method did not obtain the best value for all sub-objectives as we were concerned with the overall performance instead of every single sub-objective. In multi-task learning, the optimum of a single task was not always the best on the whole [36]. On this account, our method obtained the best overall scores but was inferior to SMCCA for a few specific sub-tasks. Combining the results of GradNorm SMCCA, g2SMCCA, and Adaptive SMCCA, we could deduce that the use of either the task-balancing technique or graph-group penalty could be insufficient, which confirmed the advantage of GradNorm SMCCA.

3.3 Results on real neuroimaging data

3.3.1 *Data source*

We utilized the brain imaging data obtained from the ADNI database (adni.loni.usc.edu). The prime purpose of this initiative is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

In Table 3, we show the detailed information of 679 non-Hispanic Caucasian participants, including 164 healthy control (HC), 262 MCI, and 253 AD subjects. We used three modalities of brain imaging data containing the 18-F florbetapir PET (AV45) scans, fluorodeoxyglucose PET (FDG) scans, and structural

Table 2 Sensitivity and specificity estimated from synthetic datasets[a]

		Sensitivity					Specificity				
		SMCCA	GGL SMCCA	Adaptive SMCCA	g2SMCCA	GradNorm SMCCA	SMCCA	GGL SMCCA	Adaptive SMCCA	g2SMCCA	GradNorm SMCCA
Data 1	Task1	0.83±0.01	0.79±0.01	0.83±0.01	0.92±0.01	0.92±0.01	0.97±0.05	0.96±0.05	0.97±0.05	0.98±0.06	0.98±0.07
	Task2	0.86±0.01	0.86±0.01	0.86±0.01	0.86±0.02	0.86±0.02	0.94±0.03	0.94±0.03	0.94±0.03	0.94±0.05	0.94±0.04
	Task3	0.89±0.01	0.89±0.01	0.89±0.01	0.94±0.03	0.91±0.02	0.92±0.01	0.92±0.01	0.92±0.01	0.95±0.04	0.93±0.02
	Average	0.86	0.85	0.86	**0.91**	0.90	0.94	0.94	0.94	**0.96**	0.95
Data 2	Task1	0.79±0.01	0.75±0.01	0.79±0.01	0.83±0.01	0.88±0.01	0.96±0.04	0.95±0.06	0.96±0.04	0.97±0.03	0.98±0.06
	Task2	0.80±0.02	0.80±0.02	0.80±0.01	0.86±0.02	0.84±0.02	0.92±0.04	0.92±0.04	0.92±0.03	0.94±0.05	0.93±0.05
	Task3	0.91±0.03	0.91±0.02	0.91±0.03	0.88±0.01	0.86±0.02	0.93±0.04	0.93±0.03	0.93±0.04	0.91±0.02	0.90±0.02
	Average	0.83	0.82	0.83	**0.86**	**0.86**	**0.94**	0.93	**0.94**	**0.94**	**0.94**
Data 3	Task1	0.67±0.01	0.67±0.00	0.71±0.01	0.83±0.01	0.83±0.01	0.94±0.03	0.94±0.02	0.94±0.03	0.97±0.05	0.97±0.04
	Task2	0.84±0.02	0.82±0.02	0.86±0.01	0.89±0.01	0.84±0.02	0.93±0.05	0.92±0.04	0.94±0.03	0.95±0.03	0.93±0.04
	Task3	0.92±0.03	0.92±0.03	0.92±0.04	0.89±0.04	0.92±0.04	0.94±0.04	0.94±0.04	0.94±0.05	0.92±0.05	0.94±0.05
	Average	0.81	0.80	0.83	**0.87**	**0.87**	0.94	0.93	0.94	**0.95**	**0.95**
Data 4	Task1	0.75±0.01	0.71±0.01	0.75±0.01	0.83±0.01	0.83±0.01	0.95±0.03	0.94±0.07	0.95±0.07	0.97±0.05	0.97±0.04
	Task2	0.75±0.01	0.75±0.02	0.77±0.01	0.84±0.02	0.86±0.01	0.90±0.03	0.90±0.04	0.91±0.03	0.93±0.04	0.94±0.03
	Task3	0.81±0.02	0.78±0.03	0.81±0.02	0.88±0.03	0.80±0.03	0.86±0.03	0.84±0.04	0.86±0.03	0.91±0.04	0.85±0.04
	Average	0.77	0.75	0.78	**0.85**	0.83	0.90	0.89	0.91	**0.94**	0.92
Data 5	Task1	0.82±0.01	0.81±0.01	0.82±0.01	0.86±0.00	0.84±0.01	0.96±0.03	0.96±0.03	0.96±0.03	0.97±0.02	0.96±0.03
	Task2	0.84±0.02	0.82±0.02	0.84±0.02	0.86±0.01	0.85±0.02	0.91±0.04	0.90±0.04	0.91±0.03	0.93±0.03	0.92±0.03
	Task3	0.84±0.02	0.83±0.02	0.84±0.02	0.83±0.02	0.84±0.02	0.90±0.03	0.89±0.03	0.90±0.03	0.90±0.03	0.90±0.03
	Average	0.83	0.82	0.83	**0.85**	0.84	0.92	0.92	0.92	**0.93**	**0.93**
Data 6	Task1	0.79±0.00	0.79±0.00	0.79±0.00	0.81±0.00	0.82±0.00	0.95±0.02	0.95±0.02	0.95±0.02	0.96±0.02	0.96±0.02
	Task2	0.85±0.01	0.85±0.01	0.85±0.01	0.88±0.01	0.90±0.01	0.94±0.02	0.94±0.02	0.94±0.02	0.95±0.02	0.96±0.02
	Task3	0.88±0.01	0.88±0.01	0.88±0.01	0.89±0.01	0.90±0.00	0.93±0.01	0.93±0.02	0.93±0.01	0.94±0.02	0.94±0.01
	Average	0.84	0.84	0.84	0.86	**0.87**	0.94	0.94	0.94	**0.95**	**0.95**

a) The best values were shown in bold.

Table 3 Participant characteristics

	HC	MCI	AD
Num	164	262	253
Gender (M/F, %)	49.39/50.61	49.62/50.38	53.75/46.25
Handedness (R/L, %)	90.24/9.76	88.93/11.07	90.91/9.09
Age (mean±std.)	73.95±5.56	70.89±6.97	72.75±8.19
Education (mean±std.)	16.34±2.61	16.18±2.73	15.98±2.78

MRI (sMRI) scans. The FDG and AV45 scans were registered into the standard montreal neurological institute (MNI) space. We processed the sMRI data with voxel-based morphometry (VBM) through statistical parametric mapping (SPM) and aligned them to a T1-weighted template image, segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) maps. Then, the scans were also normalized to the standard MNI space and smoothed with an 8 mm^3 full width at half maximum (FWHM) kernel. All of the imaging data were aligned to each subject's same visit. Thereafter, 116 ROI level measurements were generated based on the MarsBaR automated anatomical labeling atlas. To remove the effects of the baseline age, gender, education, and handedness, the data were further adjusted by regression weights generated from HCs. We aim to examine the multi-way association between multimodal imaging data and select AD-relevant imaging QTs.

3.3.2 *Multi-way association identification and objective convergence*

Table 4 exhibits the training and testing CCCs of each sub-objective and their average value. Because there were three modalities of imaging data, each method had three pairwise CCCs corresponding to three sub-objectives. For simplicity, we use AV45-FDG to denote the CCC between AV45 and FDG, and so forth. As many important ROIs of the FDG and VBM overlapped and they were very different from those of AV45, the CCC of FDG-VBM was much higher than those of AV45-FDG and AV45-VBM. Clearly, except for FDG-VBM, GradNorm SMCCA obtained the best CCCs of all the comparison methods. In

Table 4 CCCs (mean±std.) estimated between three types of imaging QTs[a)]

	Method	AV45-FDG	AV45-VBM	FDG-VBM	Average
Training	SMCCA	0.35±0.07	0.30±0.07	**0.56±0.04**	0.40
	GGL SMCCA	0.33±0.04	0.28±0.04	0.51±0.10	0.37
	Adaptive SMCCA	0.37±0.06	0.31±0.07	0.55±0.05	0.41
	g2SMCCA	0.40±0.09	0.34±0.11	0.51±0.09	0.42
	GradNorm SMCCA	**0.45±0.13**	**0.39±0.15**	0.51±0.07	**0.45**
Testing	SMCCA	0.32±0.06	0.27±0.09	**0.52±0.06**	0.37
	GGL SMCCA	0.28±0.05	0.22±0.04	0.43±0.15	0.31
	Adaptive SMCCA	0.34±0.06	0.28±0.09	0.49±0.05	0.37
	g2SMCCA	0.38±0.06	0.31±0.09	0.47±0.08	0.39
	GradNorm SMCCA	**0.43±0.10**	**0.36±0.11**	0.47±0.07	**0.42**

a) The best values were shown in bold.
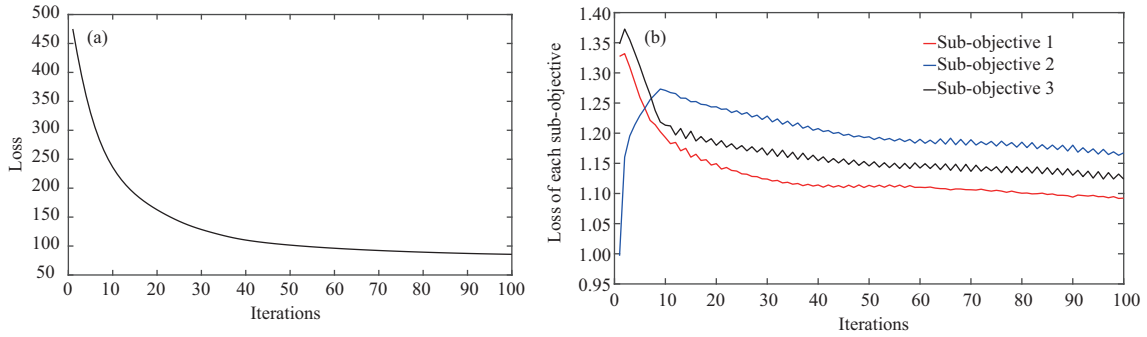
Figure 3 (Color online) Curves of (a) the overall loss function and (b) the loss of each sub-objective. Different sub-objectives follow different solution paths owing to the gradient balancing strategy. Without gradient balancing, sub-objective 2 will keep decreasing, and the overall objective will be suboptimal.

particular, SMCCA performed the best on FDG-VBM. This finding implies that SMCCA was focused on this sub-objective with those remaining ones insufficiently optimized, which confirmed the existence of loss imbalance and the dominance of the sub-objective of FDG-VBM for this real problem. Although our method did not achieve the highest CCC on FDG-VBM, the highest values were obtained on the remaining two sub-objectives and the overall average value. Hence, our method ensured comprehensive optimization. Adaptive SMCCA improved the performance of SMCCA but was still defeated by our method. GGL SMCCA and g2SMCCA did no consider loss balancing, and thus their losses were partially optimized as well. Moreover, all three CCCs of g2SMCCA were higher than those of SMCCA and GGL SMCCA, demonstrating that identifying reasonable graph-group structures could yield improved CCCs.

In addition, we present the objective curves in Figure 3. The loss function converged to the minimum around the 50th iteration and remained stable after that. Thus, Theorem 1 and the convergence curve could demonstrate that our method converges fast for this real problem. In addition, the third sub-objective first increased and then decreased during the optimization. This trajectory ensures that the overall objective will be optimal, which is guaranteed by the gradient balancing strategy. All the results show that GradNorm SMCCA had the highest average CCC and outperformed the other methods on most sub-objectives with its well-designed graph-group penalty and smart loss balancing mechanism.

3.3.3 *Feature selection*

Figure 4 shows the heatmap regarding the selected imaging features. Each method could screen out several AD-related ROIs. Most highlighted ROIs were shared by all the methods, including the hippocampus and medial orbital frontal, proving the worth of identifying the multi-way association among multimodal data. In addition, g2SMCCA and GradNorm SMCCA found the parahippocampal, precuneus, and middle and inferior temporal in FDG, which were missed by the other methods. Clearly, GradNorm SMCCA, g2SMCCA, and GGL SMCCA reported group and network structures owing to their structured regularization. However, GradNorm SMCCA and g2SMCCA were better than GGL SMCCA, which could be evidence suggesting that the graph-group penalty is better than the GGL penalty. In addition, compared to g2SMCCA, GradNorm SMCCA identified more AD-related ROIs, again demonstrating the merit of
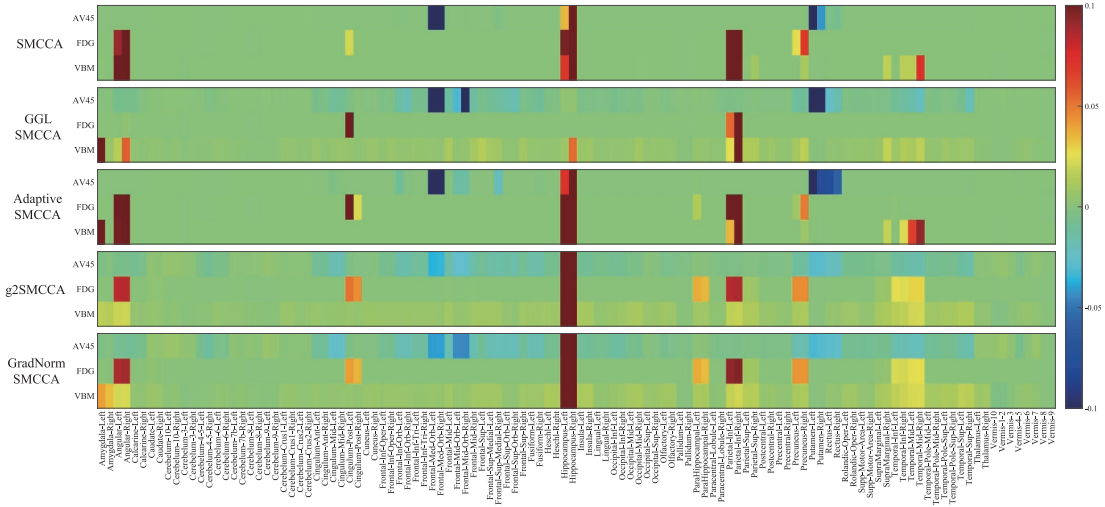
Figure 4   (Color online) Comparison of canonical weights in terms of each modality: AV45, FDG, and VBM (from top to bottom). Each row corresponds to an SMCCA method: (1) SMCCA; (2) GGL SMCCA; (3) Adaptive SMCCA; (4) g2SMCCA; (5) GradNorm SMCCA.
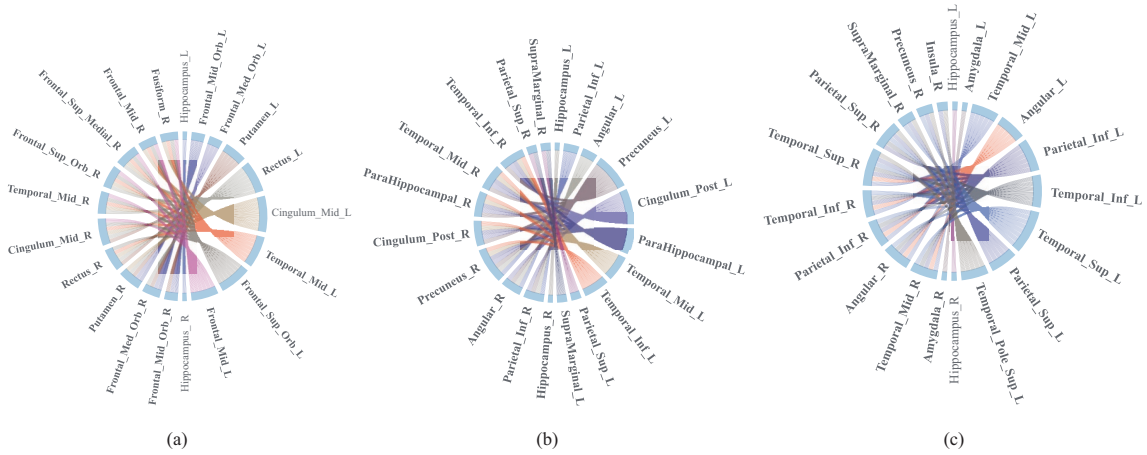
Figure 5   (Color online) Top 20 identified ROIs and their network connection of each modality. Each ROI is visualized as an arc, and ROIs of the same network are connected by lines of the same color. Each sub-figure corresponds to an imaging modality, i.e., (a) AV45, (b) FDG, and (c) VBM.

its loss-balancing strategy that fairly optimized all sub-objectives. In a word, these results demonstrate that the graph-group penalty and loss-balancing strategy endowed our method with a comprehensive and meaningful feature selection capability.

3.3.4   *Post analysis and interpretation*

To clearly exhibit the identified graph and group structures, Figure 5 visualizes the graph and group in terms of the selected ROIs of the three imaging modalities. Here, we show the network formed by the top 20 ROIs for each modality. In particular, if two ROIs were close enough, e.g., $D = (\frac{||u_{i_k}|-|u_{i_t}||}{|u_{i_k}|} + \frac{||u_{i_k}|-|u_{i_t}||}{|u_{i_t}|})/2 \leqslant \epsilon$, where ϵ is a desirable threshold, they belonged to the same sub-network, and an edge (line) connected both ROIs in the figure. Because the scans of different imaging technologies were different, we showed three networks associated with the three modalities. The network (circle) was composed of multiple ROIs (visualized as arcs) with the label annotated nearby. Clearly, our method identified several meaningful networks for AD, such as the amyloid deposition network formed by frontal areas and temporal areas in the AV45 scans. Besides, we also mapped these networks into the brain in Figure 6 [37]. These results show that our method identified group and network relationships among ROIs spanning the whole brain.

Moreover, for each diagnostic group, including HCs, MCIs, and ADs, we might be interested in identi-
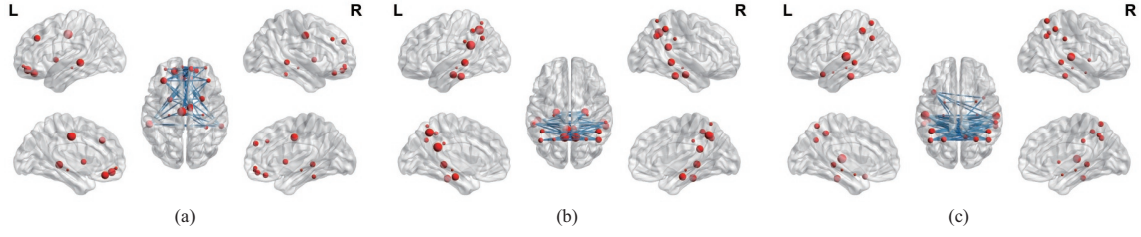
Figure 6 (Color online) Visualization of the identified ROIs and their network connection mapped on the brain. Each sub-figure corresponds to an imaging modality, i.e., (a) AV45, (b) FDG, and (c) VBM. Within each sub-figure, the left part (denoted by "L") is the lateral and medial views of the left hemisphere, the right side (denoted by "R") is those of the right hemisphere, and the middle is the dorsal view of the brain.
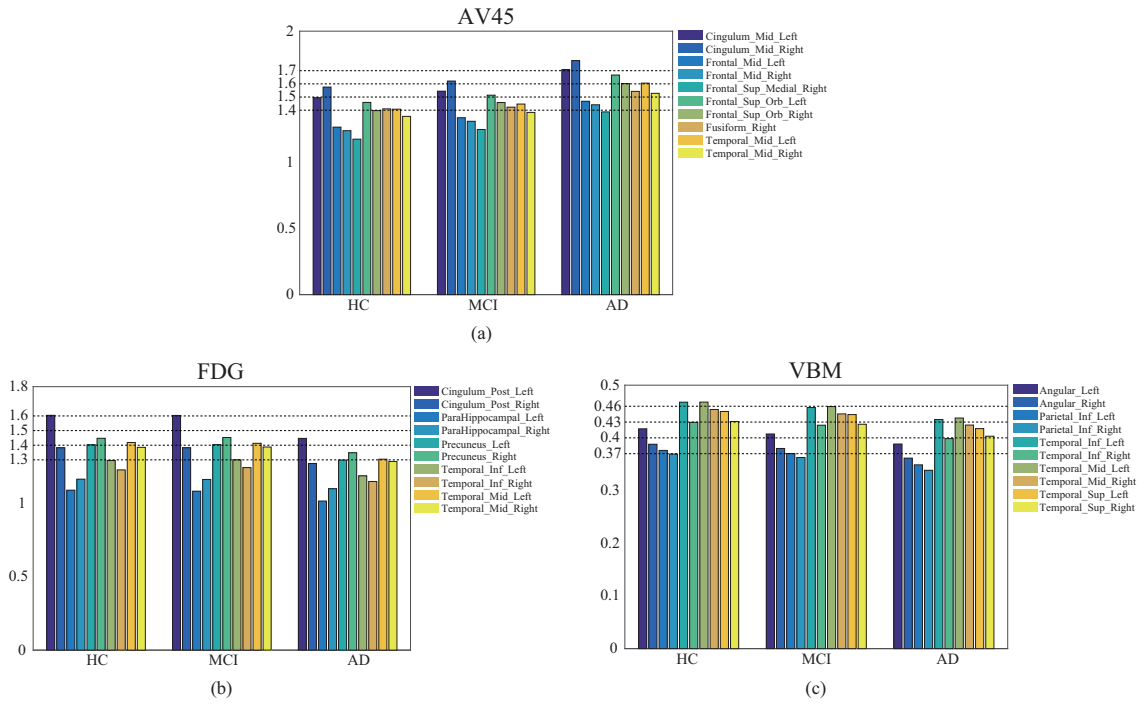
Figure 7 (Color online) Average value of the network ROI in each diagnosis category. Each sub-figure corresponds to an imaging modality, i.e., (a) AV45, (b) FDG, and (c) VBM. Within each sub-figure, the horizontal axis represents the category, and the vertical axis represents the average value. Different color bars denote different ROIs, and the labels and colors of ROIs are annotated next to the histogram. For clarity, we have marked several important values with dotted lines.

fying whether the identified networks were distinct among different groups. For each imaging modality, we chose the most significant network that was composed of the top selected ROIs within each modality. We showed their measurements in Figure 7. We also drew a few dotted lines to aid in distinguishing between different diagnostic groups. Clearly, prominent differences were observed for different groups. For AV45 scans, the patients with AD tended to gain elevated amyloid protein deposition than MCIs and HCs, and further MCIs had a more elevated deposition than HCs. Meanwhile, the network of the VBM scans showed an opposite pattern. Specifically, AD's network exhibited more atrophy than MCIs and HCs, and that of MCIs was more severe than that of HCs. These figures also convey that as dementia progresses, the brain atrophy gets more severe, the metabolic rates of cerebral glucose decline, and the extracellular amyloid deposition gradually elevates. These patterns were consistent with the neuroimaging manifestations of AD, which demonstrates that GradNorm can select valuable networks.

4 Conclusion

Seeking complicated relationships among multimodal brain imaging data is important and meaningful. Accordingly, we propose the GradNorm SMCCA method with adaptive loss balancing and graph-group penalty. The adaptive loss balancing adjusts the importance of each sub-objective by iteratively modifying

the gradient magnitudes. The graph-group regularization hierarchically penalizes ROIs at the graph and group levels. Therefore, GradNorm SMCCA can identify deep structural information and simultaneously accommodate an overall optimization. We also provide an efficient optimization algorithm and analyze its convergence. The experiment results on the synthetic and real ADNI data show that GradNorm SMCCA obtained higher average CCCs and better canonical weights. GradNorm SMCCA also identified meaningful networks for each imaging modality. These results demonstrate that GradNorm SMCCA was a powerful software tool and thus had a broad application prospect in aiding clinicians to determine suspicious imaging features and networks.

References

1 Goedert M, Spillantini M G. A century of Alzheimer's disease. Science, 2006, 314: 777–781
2 Calhoun V D, Sui J. Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. Biol Psychiatry-Cogn Neurosci Neuroimag, 2016, 1: 230–244
3 Rathore S, Habes M, Iftikhar M A, et al. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Neuroimage, 2017, 155: 530–548
4 Fjell A M, Westlye L T, Grydeland H, et al. Accelerating cortical thinning: unique to dementia or universal in aging? Cerebral Cortex, 2014, 24: 919–934
5 van de Pol L A, Hensel A, van der Flier W M. Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease. J Neurol Neurosurg Psychiatry, 2006, 77: 439–442
6 Lim E W, Aarsland D, Ffytche D, et al. Amyloid-β and Parkinson's disease. J Neurol, 2019, 266: 2605–2619
7 Zhang D, Wang Y, Zhou L, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage, 2011, 55: 856–867
8 Teipel S, Drzezga A, Grothe M J, et al. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. Lancet Neurol, 2015, 14: 1037–1053
9 Miller K L, Alfaro-Almagro F, Bangerter N K, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci, 2016, 19: 1523–1536
10 Du L, Huang H, Yan J, et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. Bioinformatics, 2016, 32: 1544–1551
11 Foley S F, Tansey K E, Caseras X, et al. Multimodal brain imaging reveals structural differences in Alzheimer's disease polygenic risk carriers: a study in healthy young adults. Biol Psychiatry, 2017, 81: 154–161
12 Du L, Liu F, Liu K, et al. Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method. IEEE Trans Med Imag, 2020, 39: 3416–3428
13 Du L, Zhang J, Liu F, et al. Identifying associations among genomic, proteomic and imaging biomarkers via adaptive sparse multi-view canonical correlation analysis. Med Image Anal, 2021, 70: 102003
14 Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage, 2012, 59: 895–907
15 Ball G, Aljabar P, Nongena P, et al. Multimodal image analysis of clinical influences on preterm brain development. Ann Neurol, 2017, 82: 233–246
16 Dähne S, Bieszmann F, Samek W, et al. Multivariate machine learning methods for fusing multimodal functional neuroimaging data. Proc IEEE, 2015, 103: 1507–1530
17 Lorenzi M, Simpson I J, Mendelson A F, et al. Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations. Sci Rep, 2016, 6: 1–8
18 Whitwell J L, Graff-Radford J, Tosakulwong N, et al. Imaging correlations of tau, amyloid, metabolism, and atrophy in typical and atypical Alzheimer's disease. Alzheimers Dement, 2018, 14: 1005–1014
19 Tognin S, van Hell H H, Merritt K, et al. Towards precision medicine in psychosis: benefits and challenges of multimodal multicenter studies-PSYSCAN: translating neuroimaging findings from research into clinical practice. Schizophrenia Bull, 2020, 46: 432–441
20 Witten D M, Tibshirani R J. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Appl Genet Mol Biol, 2009, 8: 1–27
21 Zhang J, Wang H, Zhao Y, et al. Identification of multimodal brain imaging association via a parameter decomposition based sparse multi-view canonical correlation analysis method. BMC Bioinf, 2022, 23: 128
22 Chen Z, Badrinarayanan V, Lee C-Y, et al. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of International Conference on Machine Learning, 2018. 794–803
23 Hugdahl K. Symmetry and asymmetry in the human brain. Eur Rev, 2005, 13: 119–133

24 Toga A W, Thompson P M. Mapping brain asymmetry. Nat Rev Neurosci, 2003, 4: 37–48

25 Feng C-M, Yan Y, Chen G, et al. Accelerated multi-modal MR imaging with transformers. 2021. ArXiv:2106.14248

26 Feng C-M, Fu H, Zhou T, et al. Multi-modal aggregation network for fast MR imaging. 2021. ArXiv:2110.08080

27 Feng C-M, Yan Y, Liu C, et al. Exploring separable attention for multi-contrast MR image super-resolution. 2021. ArXiv:2109.01664

28 Feng C-M, Fu H, Yuan S, et al. Multi-contrast MRI super-resolution via a multi-stage integration network. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021. 140–149

29 Hu W, Lin D, Cao S, et al. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia. IEEE Trans Biomed Eng, 2017, 65: 390–399

30 Du L, Zhang T, Liu K, et al. Identifying associations between brain imaging phenotypes and genetic factors via a novel structured SCCA approach. In: Proceedings of International Conference on Information Processing in Medical Imaging, 2017. 543–555

31 Du L, Liu K, Yao X, et al. Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. Med Image Anal, 2020, 61: 101656

32 Gorski J, Pfeuffer F, Klamroth K. Biconvex sets and optimization with biconvex functions: a survey and extensions. Math Meth Oper Res, 2007, 66: 373–407

33 Wang H, Nie F, Huang H, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics, 2012, 28: 229–237

34 Chen M, Gao C, Ren Z, et al. Sparse CCA via precision adjusted iterative thresholding. 2013. ArXiv:1311.6186

35 Du L, Liu K, Yao X, et al. Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics. IEEE ACM Trans Comput Biol Bioinf, 2021, 18: 227–239

36 Lin X, Zhen H-L, Li Z, et al. Pareto multi-task learning. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32: 60–70

37 Xia M, Wang J, He Y. BrainNet viewer: a network visualization tool for human brain connectomics. Plos One, 2013, 8: e68910