frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 10 July 2019
doi: 10.3389/fnins.2019.00668

Identifying Imaging Markers for Predicting Cognitive Assessments Using Wasserstein Distances Based Matrix Regression

Jiexi Yan[1], *Cheng Deng*[1*], *Lei Luo*[2], *Xiaoqian Wang*[2], *Xiaohui Yao*[3], *Li Shen*[3] *and Heng Huang*[2*]

[1] School of Electronic Engineering, Xidian University, Xi'an, China, [2] Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, United States, [3] Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Alzheimer's disease (AD) is a severe type of neurodegeneration which worsens human memory, thinking and cognition along a temporal continuum. How to identify the informative phenotypic neuroimaging markers and accurately predict cognitive assessment are crucial for early detection and diagnosis Alzheimer's disease. Regression models are widely used to predict the relationship between imaging biomarkers and cognitive assessment, and identify discriminative neuroimaging markers. Most existing methods use different matrix norms as the similarity measures of the empirical loss or regularization to improve the prediction performance, but ignore the inherent geometry of the cognitive data. To tackle this issue, in this paper we propose a novel robust matrix regression model with imposing Wasserstein distances on both loss function and regularization. It successfully integrate Wasserstein distance into the regression model, which can excavate the latent geometry of cognitive data. We introduce an efficient algorithm to solve the proposed new model with convergence analysis. Empirical results on cognitive data of the ADNI cohort demonstrate the great effectiveness of the proposed method for clinical cognitive predication.

Keywords: Alzheimer's disease, cognitive assessment, Wasserstein distance, matrix regression, feature selection

OPEN ACCESS

Edited by:
Dongdong Lin,
Mind Research Network (MRN),
United States

Reviewed by:
Heung-Il Suk,
University of North Carolina at Chapel Hill, United States
Yong Liu,
Institute of Automation (CAS), China

****Correspondence:***
Cheng Deng
chdeng.xd@gmail.com
Heng Huang
henghuanghh@gmail.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *24 January 2019*
Accepted: *11 June 2019*
Published: *10 July 2019*

Citation:
Yan J, Deng C, Luo L, Wang X, Yao X, Shen L and Huang H (2019) Identifying Imaging Markers for Predicting Cognitive Assessments Using Wasserstein Distances Based Matrix Regression. Front. Neurosci. 13:668. doi: 10.3389/fnins.2019.00668

1. INTRODUCTION

Alzheimer's disease (AD), the most common form of dementia, is a Central Nervous System (CNS) chronic neurodegenerative disorder with progressive impairment of learning, memory and other cognitive function. As an incurable disease which severely impacts human thinking and behavior, Alzheimer's disease is the 6th cause of death in the United States (Alzheimer's Association, 2018). Along with the rapid progress in high-throughput genotype and brain image techniques, neuroimaging has been developed to effectively predict the progression of AD or cognitive performance in plentiful research (Ewers et al., 2011; Wang et al., 2011b), which benefits for early diagnosis and explorition of brain function associated with AD (Petrella et al., 2003; Avramopoulos, 2009). The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005; Jack et al., 2008) provides neuroimaging and cognitive measurement of normal aging, mild cognitive impairment as well as AD samples, which provides a wealth of resources for the study of Alzheimer's diagnosis, treatment and prevention.

Until now, numerous studies (Eskildsen et al., 2013; Moradi et al., 2015) have utilized neuroimaging techniques to detect pathology associated with AD. Among them, structural magnetic resonance imaging (MRI) is the most extensively used imaging modality in AD related studies because of its completely non-invasive nature, high spatial resolution, and high availability. Thus, researchers have extracted plentiful MRI boimarkers in classifying AD patients in different disease over the past few years (Duchesne et al., 2008; Eskildsen et al., 2013; Guerrero et al., 2014). And these abundant MRI boimarkers have been used to many AD related studies, such as AD status prediction and MCI-to-AD conversion prediction. Despite of great efforts, we still cannot identify informative AD-specific biomarkers for the early diagnosis and prediction of disease progression. The reason for this is that the number of clinical status of AD is small, which makes it difficult to observe and understand the cognitive progression.

Consequently, many studies use clinical cognitive tests to measure cognitive assessment. Recently, several clinical tests have been presented to access individual's cognitive level, such as Trail making test (TRAILS) and and Rey Auditory Verbal Learning Test (RAVLT) (Schmidt, 1996). Through predicting the cognitive scores with MRI biomarks, we can explore the association between imaging biomarkers and AD and find informative AD-specific biomarkers. Therefore, a wide range of machine learning approaches have been proposed to predict the cognitive scores and uncover the pathology associated with AD (Wang et al., 2011a, 2016; Moradi et al., 2017).

In the current study of predicting cognitive scores with longitudinal phenotypic markers extracted from MRI data, regression method has been demonstrated as a effective way to excavate the correlation between cognitive measures. To modify the traditional regression model, recent methods proposed to integrate novel regularization term (such as sparse regularization and low-rank regularization) into the traditional regression model (Obozinski et al., 2010; Jie et al., 2015; Moradi et al., 2017). In fact, the intrinsic idea of the study mentioned above is utilizing different matrix norm or the combination of matrix norms as the similarity measures of the empirical loss or regularization to fit the prior assumption of neuroimaging markers. Though the effectiveness of specific matrix norm as regularization, these matrix norms simply meet the assumption rather than make full use of the inherent geometry of the data. Thus, it is easy to achieve a suboptimal solution for these models.

To tackle this problem, in this paper we consider Wasserstein distance as distance metric for regression model. Different from L_p distances ($p \geq 0$) (Luo et al., 2017) or Kullback-Leibler (Csiszár and Shields, 2004) and other f-divergences (Ali and Silvey, 1966), Wasserstein distance is well-defined between any pair of probability distributions over a sample space equipped with a metric. Thus, it provides a meaningful notion of distance for distributions supported on non-overlapping low dimensional manifolds. For better performance of cognitive score predication, we propose to substitute Wasserstein distance for matrix norm.

Although successfully applied to image retrieval (Rubner et al., 2000), contour matching (Grauman and Darrell, 2004), cancer detection (Ozolek et al., 2014), super-resolution (Kolouri and Rohde, 2015), and many other problems, there is an intrinsic limitation of Wasserstein distances. In fact, Wasserstein distances are defined only between measures having the same mass, which makes it difficult to applied Wasserstein distance into cognitive score prediction. To overcome such a limitation, many existed study (Piccoli and Rossi, 2014, 2016; Kondratyev et al., 2016), have been proposed. However, these methods are all based on distributions or histogram features of data. As we know, in cognitive score prediction, we usually use the original features rather histogram features to learn the regression model parameters. Additionally, most of these methods use traditional matrix norm to characterize model parameters in Wasserstein distance loss minimization problem. This often leads to suboptimal results since matrix norm is usually sensitive to real noise.

To perfectly integrate Wassterstein distance into regression model for better performance of cognitive score prediction, in this paper we propose a novel efficient and robust Matrix Regression method to employ Joint Wasserstein distances minimization on both loss function and regularization (JWMR for short). Different from the existing methods, which need to extract histogram features of data in the preprocessing stage and then calculate Wasserstein distances based on them, our method considers histogram operator as an important component of objective function and uses it to constrain loss term and the estimated model parameters which are generated by original data features. This is the first time for exploiting Wasserstein distance as loss and regularization terms. As a result, our method is more reliable and applicable than traditional regression method using ℓ_p-norm regularizer. We derive an efficient algorithm based on a relaxed formulation of optimal transport, which iterates through applications of alternating optimization. We provide the convergence analysis of our algorithm and describe a statistical bound for the proposed new model. We apply our method on cognitive data of the ADNI cohort and obtain promising results.

Our main contributions are three-fold: (1) The proposed robust matrix regression via joint Wasserstein distances minimization to circumvent the natural limitation of matrix norms in regression model; (2) The proposed model is suitable for revealing the relationship between cognitive measures and neuroimaging markers; (3) Because our method not only includes composition of $W(\cdot,\cdot)$, but also the computations of Wasserstein distances with regard to different terms, we derive an efficient algorithm to solve this problem with convergence analysis.

2. STUDY OF COGNITIVE SCORE PREDICTION

2.1. Notations

We summarize the notations and definitions used in this paper. Matrices are written as boldface uppercase letters. $\|\cdot\|_F$ and $\|\cdot\|_*$ denote Frobenius norm and nuclear norm, respectively. $\langle\cdot,\cdot\rangle$ is the inner product operation. $\mathbf{e} \in \mathbb{R}^m$ is a column vector of ones. $\mathbf{0} \in \mathbb{R}^m$ is a column vector of zeros. For vector $\mathbf{m} \in \mathbb{R}^m$, its i-th element is denoted by $m_{(i)}$. For matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its i-th row,

j-th column and (i,j)-th element are denoted by $\mathbf{m}^i$, $\mathbf{m}_j$, and m_{ij}. The $\ell_{2,1}$-norm of $\mathbf{M}$ is defined as

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{m}^i\|_2, \tag{1}$$

where $\|\mathbf{m}^i\|_2$ denotes the ℓ_2-norm of the vector $\mathbf{m}^i$. We define the Kullback-Leibler (KL) divergence between two positive vectors by

$$KL(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y}) \rangle + \langle \mathbf{y} - \mathbf{x}, \mathbf{e} \rangle, \tag{2}$$

where / denotes the element-wise division.

2.2. Matrix Regression for Cognitive Score Prediction

In the association study of predicting cognitive scores from imaging markers, a wide range of work has employed regression models to uncover the relationship between neuroimaging data and cognitive test scores and predict cognitive score. Given the imaging feature matrix $\mathbf{A} \in \mathbb{R}^{m \times l}$ and the cognitive score matrix $\mathbf{Y} \in \mathbb{R}^{l \times n}$, a common paradigm for regression to predict cognitive score is to minimize the penalized empirical loss:

$$\min_{\mathbf{Z}} \mathcal{L}(\mathbf{Y} - \mathbf{A}^T\mathbf{Z}) + \lambda\Omega(\mathbf{Z}), \tag{3}$$

where $\lambda > 0$ is the balance parameter, $\mathbf{Z} \in \mathbb{R}^{m \times n}$ is the weight matrix, which is estimated from the imaging feature matrix $\mathbf{A}$ and the cognitive score matrix $\mathbf{Y}$ to capture the relevant features for predicting the cognitive scores, $\mathcal{L}(\mathbf{Y} - \mathbf{A}^T\mathbf{Z})$ is the empirical loss on the training set, and $\Omega(\mathbf{Z})$ is the regularization term that encodes imaging feature relatedness. Different assumptions on the loss $\mathcal{L}(\mathbf{Y} - \mathbf{A}^T\mathbf{Z})$ and variate $\mathbf{Z}$ lead to different models. The representative model include:

Least Squares Regression (LSR) (Lu et al., 2012):

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{A}^T\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_F^2, \tag{4}$$

Low Rank Representation (LRR) (Liu et al., 2010):

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{A}^T\mathbf{Z}\|_1 + \lambda\|\mathbf{Z}\|_*, \tag{5}$$

Feature Selection Based on $\ell_{2,1}$-norm (Nie et al., 2010):

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{A}^T\mathbf{Z}\|_{2,1} + \lambda\|\mathbf{Z}\|_{2,1}. \tag{6}$$

2.3. Feature Selection for Informative Imaging Marker Identification

Due to the progress and prosperity of brain imaging and high-throughput genotyping techniques, a large amount of brain imaging data is available and a great quantity of imaging markers is alternative to predict cognitive score. However, not all of them are related to the pathological changes specific to AD, namely some imaging markers are redundancy for the prediction task. A forthright method to tackle this problem is to perform feature selection, which aims to choose a subset of informative features for improving prediction.

Feature selection has been demonstrated as a efficient way to reflect the correlation between cognitive measures after removing the non-distinctive neuroimaging markers. Regression techniques with specific regularization can also used to identify discriminative imaging markers. For instance, sparse regression models have been extensively utilized to select discriminative voxels for AD study in previous works (Guerrero et al., 2014; Liu et al., 2014; Xu et al., 2017). Many sparse-inducing norm have been iterated into the spare regression model: ℓ_1 shrinkage methods such as LASSO can identify informative longitudinal phenotypic markers in the brain that are related to pathological changes of AD (Liu et al., 2014); group LASSO with a $\ell_{2,1}$-norm can select the most informative imaging markers related to all participants including AD, mild cognitive impairment (MCI) and healthy control (HC) by imposing structured sparsity on parameter matrix (Jie et al., 2015); $\ell_{1,1}$-norm regularization term can achieve both structured and flat sparsity (Wang et al., 2011a).

Nevertheless, matrix norms such as ℓ_1-norm, $\ell_{2,1}$-norm, and $\ell_{1,1}$-norm have the natural limitation that they can not take the inherent geometry of the data into account. On this account, we need to select a new distance metric to measure the empirical loss and regularization term. In this paper, we choose the smoothed Wassersetein distance as the distance metric.

2.4. Smoothed Wasserstein Distance

Wasserstein distance, originally introduced in Monge (1781), is a powerful geometrical tool for comparing probability distributions. It is derived form the optimal transport theory and is intrinsically the optimal solution of transportation problem in linear programming (Villani, 2008).

In a more formal way, given access to two sets of points $\mathcal{X}_S = \left\{\mathbf{x}_i^S \in \mathbb{R}^d\right\}_{i=1}^{N_S}$ and $\mathcal{X}_T = \left\{\mathbf{x}_i^T \in \mathbb{R}^d\right\}_{i=1}^{N_T}$, we construct two empirical probability distributions as follows

$$\hat{\mu}_S = \sum_{i=1}^{N_S} p_i^S \delta_{\mathbf{x}_i^S} \quad \text{and} \quad \hat{\mu}_T = \sum_{i=1}^{N_T} p_i^S \delta_{\mathbf{x}_i^T}, \tag{7}$$

where p_i^S and p_i^T are probabilities associated to $\mathbf{x}_i^S$ and $\mathbf{x}_i^T$, respectively, and $\delta_{\mathbf{x}}$ is a Dirac measure that can be interpreted as an indicator function taking value 1 as the position of $\mathbf{x}$ and 0 elsewhere. For these two distribution, the polytope of transportation plans between $\mathcal{X}_S$ and $\mathcal{X}_T$ is defined as follows:

$$\mathcal{U}_{\hat{\mu}_S, \hat{\mu}_T} = \left\{ \mathbf{P} \in \mathbb{R}_+^{N_S \times N_T} \text{s.t.} \left| \begin{array}{l} \mathbf{P}\mathbf{e} = \mathbf{p}^S \\ \mathbf{P}^T\mathbf{e} = \mathbf{p}^T \end{array} \right. \right\}. \tag{8}$$

Given a ground metric matrix $\mathbf{C} \in \mathbb{R}_+^{N_S \times N_T}$, the optimal transport consists in finding a probabilistic coupling defined as a joint probability measure over $\mathcal{X}_S \times \mathcal{X}_T$ with marginals $\hat{\mu}_S$ and $\hat{\mu}_T$ that minimize the cost of transport

$$\min_{\mathbf{P} \in \mathcal{U}_{\hat{\mu}_S, \hat{\mu}_T}} \langle \mathbf{C}, \mathbf{P} \rangle, \tag{9}$$

where $\mathbf{P} = \left\{p(i,j), i = 1, \cdots, N_S, j = 1, \cdots, N_T\right\}$ is the flow-network matrix, and $p(i,j)$ denotes the amount of earth moved from the source $\mathcal{X}_S$ to the target $\mathcal{X}_T$. This problem admits a unique solution $\mathbf{P}^*$ and defines a metric on the space of probability measures (called the Wasserstein distance) as follows:

$$W(\hat{\mu}_S, \hat{\mu}_T) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathcal{U}_{\hat{\mu}_S,\hat{\mu}_T}} \langle \mathbf{C}, \mathbf{P} \rangle . \tag{10}$$

Optimizing Wasserstein distance problem requires several costly optimal transport problems. Specialized algorithm can solve it with $\mathcal{O}((N_S+N_T)\log(N_S+N_T)^2+N_SN_T(N_S+N_T)\log(N_S+N_T))$ (Orlin, 1993). To solving the computational problem, recent works have proposed novel method to accelerate the calculation procedure. Furthermore, as a minimum of affine functions, the Wasserstein distance itself is not a smooth function of its arguments. To overcome the above problems, Cuturi (2013) proposed to smooth the optimal transport problem with an entropy term:

$$W_\gamma(\hat{\mu}_S, \hat{\mu}_T) = \min_{\mathbf{P}\in\mathcal{U}_{\hat{\mu}_S,\hat{\mu}_T}} \langle \mathbf{C}, \mathbf{P} \rangle - \gamma e(\mathbf{P}), \tag{11}$$

where $\gamma > 0$ and $e(\cdot)$ is the entropy function:

$$e(\mathbf{P}) = -\left\langle \mathbf{P}, \log(\mathbf{P}) \right\rangle . \tag{12}$$

With the entropy term, we can use Sinkhorn-Knopp matrix scaling algorithm to solve the optimal transport problem (Sinkhorn and Knopp, 1967).

3. MATRIX REGRESSION BASED ON JOINT WASSERSTEIN DISTANCE

In the above formulations, the loss term and estimated variate are characterized via the simple matrix norm. Thus, these models can be easily solved by conventional convex optimization methods [e.g., ADMM (Liu et al., 2010), gradient based methods (Bubeck et al., 2015), and reweighted iterative methods (Nie et al., 2010)]. However, they do not take into account the geometry of the data through the pairwise distances between the distributions' points. Accordingly, these models often achieve the suboptimal results in cognitive score predication.

3.1. Joint Wasserstein Matrix Regression

Comparing with matrix norm, Wasserstein distance can circumvent the above limitation. Therefore, in this paper we propose to use Wasserstein distance to jointly characterize loss term and estimated variate $\mathbf{Z}$, which is formulated as

$$\min_{\mathbf{Z}} \sum_{i=1}^{l} W_\gamma((h(\mathbf{A}^T\mathbf{Z})^i), h(\mathbf{Y}^i)) + \lambda \sum_{i=1}^{m} W_\gamma(h(\mathbf{Z}^i), \mathbf{0}), \tag{13}$$

where $h(\cdot)$ and $\mathbf{Y}^i$ denote the histogram operator and ith row of matrix $\mathbf{Y}$, respectively. It should be noted that we use the histogram operator to constrain each variable in model (13).

3.2. Optimization Algorithm

Solving problem (13) is extremely challenging since it not only includes the composition of $h(\cdot)$ and $W_\gamma(\cdot,\cdot)$, but also the computations of Wasserstein distance with regard to different terms. Some existing (Genevay et al., 2016; Rolet et al., 2016) algorithms are only suitable for solving Wasserstein distance loss minimization with matrix norm regularizer. To cope with this challenge, we relax the marginal constraints $\mathcal{U}_{\hat{\mu}_S,\hat{\mu}_T}$ in (11) using a Kullback-Leibler divergence from the matrix to target marginals $\hat{\mu}_S$ and $\hat{\mu}_T$ (Frogner et al., 2015; Chizat et al., 2016), i.e., (11) is converted as

$$\begin{aligned} W_\gamma(\hat{\mu}_S, \hat{\mu}_T) \;=\; & \min_{\mathbf{P}\in\mathcal{U}_{\hat{\mu}_S,\hat{\mu}_T}} \gamma KL(\mathbf{P}|\mathbf{K}) + \mu KL(\mathbf{P}\mathbf{e}|\hat{\mu}_S) \\ & + \mu KL(\mathbf{P}^T\mathbf{e}|\hat{\mu}_T), \end{aligned} \tag{14}$$

where $\mathbf{K} = \exp(-\mathbf{C}/upgamma)$.

Algorithm 1: Optimization Algorithm of our proposed method.

```
Input: the given ADNI data $\mathbf{A}$ and related cognitive score
    matrix $\mathbf{Y}$ and parameter $\lambda$
Output: model parameter $\mathbf{Z}$
Initialization: $\mathbf{P}^0$ and $\hat{\mathbf{P}}^0$
repeat
  for $t = 1$ to $m$ do
    Update each $\mathbf{Z}^i$ with proximal coordinate descent
  end for
  Update $\mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots, \hat{\mathbf{P}}_{(m)}$ via Sinkhorn iteration
until convergence
```

Let

$$f_{\hat{\mu}_S,\hat{\mu}_T}(\mathbf{P}) = \gamma KL(\mathbf{P}|\mathbf{K}) + \mu KL(\mathbf{P}\mathbf{e}|\hat{\mu}_S) + \mu KL(\mathbf{P}^T\mathbf{e}|\hat{\mu}_T), \tag{15}$$

where parameters $\gamma,\ \mu \geq 0$. Then model (11) ultimately becomes the following form

$$\begin{aligned} \min \quad & J(\mathbf{Z}; \mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots, \hat{\mathbf{P}}_{(m)}) \\ & = \sum_{i=1}^{l} f_{(\mathbf{A}^T\mathbf{Z})^i,\mathbf{Y}^i}(\mathbf{P}_{(i)}) + \gamma \sum_{i=1}^{m} f_{\mathbf{Z}^i,\mathbf{0}}(\hat{\mathbf{P}}_{(i)}) \\ \text{s.t.} \quad & \mathbf{Z}^i \geq 0,\ \forall i = 1, 2, \cdots, m \end{aligned} \tag{16}$$

where $\mathbf{P}$ and $\hat{\mathbf{P}}$ denote the flow-network matrix of $W_\gamma((h(\mathbf{A}^T\mathbf{Z})^i), h(\mathbf{Y}^i))$ and $W_\gamma(h(\mathbf{Z}^i), \mathbf{0})$, respectively, and $\mathbf{Z}^i \geq 0$ means all the elements in $\mathbf{Z}^i$ is greater than or equal to 0.

Due to the relax operation in (14), we can straightly utilize the original data $\mathbf{A}^T\mathbf{Z}$, $\mathbf{Y}$, and $\mathbf{Z}$ in model (16). Thus, we do not need to extract the histogram features of data in the preprocessing stage, which makes it suitable for the prediction task in neuroimaging data.

TABLE 1 | Numbers of participants in the experiments using two different types of imaging markers.

	#Total	#AD	#pMCI	#sMCI	#HC
FreeSurfer	805	186	167	226	226
VBM	805	186	167	226	226

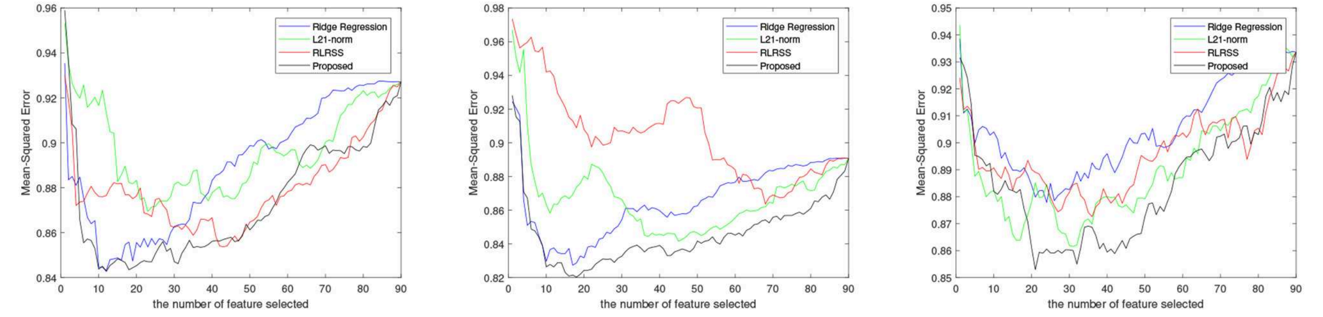

FIGURE 1 | RMSE of four regression methods for VBM memory score prediction.

Strong convexity of model (16) is given by the entropy terms $KL(\mathbf{P}|\mathbf{K})$. Thus, we propose to solve (16) by block coordinate descent, alternating the minimization with respect to the parameters $\{\mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots, \hat{\mathbf{P}}_{(m)}\}$ and each $\mathbf{Z}_i$, which can be updated independently and therefore in parallel. This is summarized in Algorithm 1. We now detail the two steps of the procedure.

Updating coefficient matrix $\mathbf{Z}$. Minimizing with respect to one $\mathbf{Z}_i$ while keeping all other variables fixed to their current estimate yields the following problem

$$\min_{\mathbf{Z}^i} KL(\mathbf{P}_{(i)}\mathbf{e}|(\mathbf{A}^T\mathbf{Z})^i) + \lambda KL(\mathbf{P}_{(i)}\mathbf{e}|\mathbf{Z}^i). \tag{17}$$

Recalling the definition (2), it is easy to calculate the gradient of objective (17) with regard to each $\mathbf{Z}_i$. Thus, we can use accelerated gradient descent (Bubeck et al., 2015) to optimize problem (17).

Updating parameter set $\{\mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots, \hat{\mathbf{P}}_{(m)}\}$. For fixed $\mathbf{Z}$, the update of each $\mathbf{P}_{(i)}$ and $\hat{\mathbf{P}}_{(i)}$ boils down to an OT problem, which can be solved via Sinkhorn iteration (Cuturi, 2013). These steps are summarized in Algorithm 2, where we list the detailed iteration process for each $\mathbf{P}_{(i)}$. For each $\hat{\mathbf{P}}_{(i)}$, we need to replace $(\mathbf{A}^T\mathbf{Z})^i$ and $\mathbf{Y}^i$ with $\mathbf{Z}^i$ and $\mathbf{0}$.

3.3. Convergence Analysis

Following Sandler and Lindenbaum (2011), we can derive the theorem as follow.

Theorem 1. *Algorithm 1 converges to a local minimum.*

Proof: Algorithm 1 is the alternative iteration with two iteration stage. In the first stage, we can use gradient descent to solve the convex problem (17). Thus it is obvious that it has a feasible solution. And in the second stage, the problem is a sequence of linear programming processes. As shown in (Sandler and Lindenbaum, 2011), there is a feasible solution for every one of them. To sum up, a feasible solution for (16) exists.

$J(\mathbf{Z}; \mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots \hat{\mathbf{P}}_{(m)})$ is convex, so applying (17) can derive globally optimal $\mathbf{Z}^k$ when given a $\{\mathbf{P}_{(1)}, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_{(1)}, \cdots \hat{\mathbf{P}}_{(m)}\}^{k-1}$, where k denotes the iteration time. Besides, linear programming minimizes the flow-network matrix $\mathbf{P}$ and $\hat{\mathbf{P}}$. Thus, we can find global optimal $\mathbf{P}^k$ and $\hat{\mathbf{P}}^k$ for a give $\mathbf{Z}^{k-1}$. Furthermore, the accelerated gradient descent used to update $\mathbf{Z}$ and the Sinkhorn Iteration used to update $\mathbf{P}, \hat{\mathbf{P}}$ both have been proven converge.

TABLE 2 | Prediction performance measured by RMSE with top 10 features.

		RR	$\ell_{2,1}$	RSR	RLRSS	Proposed
VBM	FLUENCY	0.8446	0.9166	0.9044	0.8564	**0.8437**
	RAVLT	0.8376	0.8636	0.8742	0.8943	**0.8263**
	TRAILS	0.9040	0.8823	0.8865	0.8886	**0.8820**
FreeSurfer	FLUENCY	0.8136	0.8387	0.8536	0.8686	**0.8122**
	RAVLT	0.7833	0.8051	0.8337	0.8132	**0.7815**
	TRAILS	0.8416	**0.8181**	0.8433	0.8379	0.8626

The bold values indicate the minimal value in the raw (i.e., the best performance among these methods).

TABLE 3 | Prediction performance measured by RMSE with top 30 features.

		RR	$\ell_{2,1}$	RFS	RLRSS	Proposed
VBM	FLUENCY	0.8627	0.8815	0.8879	0.8503	**0.8471**
	RAVLT	0.8543	0.8663	0.8741	0.8736	**0.8327**
	TRAILS	0.8826	0.8618	0.8903	0.8743	**0.8603**
FreeSurfer	FLUENCY	0.8351	0.8323	0.8517	0.8322	**0.8186**
	RAVLT	0.8136	0.7903	0.8154	0.8051	**0.7788**
	TRAILS	0.8295	0.8677	0.8579	0.8335	**0.8274**

The bold values indicate the minimal value in the raw (i.e., the best performance among these methods).

Since the objective in these two stage is the same, $J(\mathbf{Z}^k; \{\mathbf{P}, \hat{\mathbf{P}}\}^{k-1}) \leq J(\mathbf{Z}^{k-1}; \{\mathbf{P}, \hat{\mathbf{P}}\}^{k-1})$, and $J(\mathbf{Z}^k; \{\mathbf{P}, \hat{\mathbf{P}}\}^k) \leq J(\mathbf{Z}^k; \{\mathbf{P}, \hat{\mathbf{P}}\}^{k-1})$.

In above, every iteration of Algorithm 1 monotonically decreases $J(\mathbf{Z}; \mathbf{P}_1, \cdots, \mathbf{P}_{(l)}, \hat{\mathbf{P}}_1, \cdots \hat{\mathbf{P}}_{(m)})$. This objective is lower bounded, and therefore the algorithm converges. □

4. EXPERIMENTAL RESULTS

In this section, we evaluate the prediction performance of our proposed method by applying it to the Alzheimer's Disease Neuroimaging Initiative (ANDI) database (adni.loni.usc.edu), where a plenty of imaging markers measured over a period of

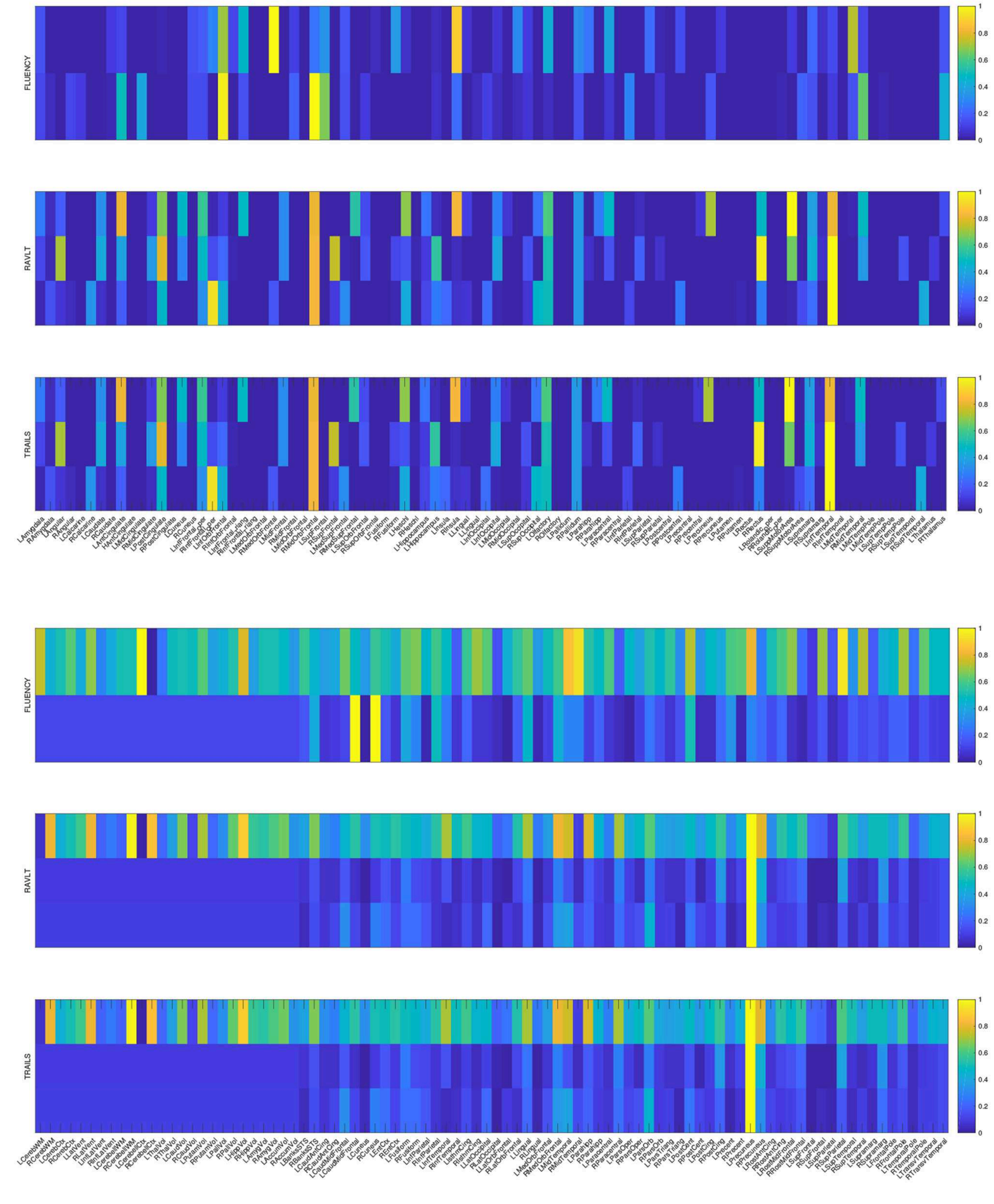

FIGURE 2 | Heat maps of our learned weight matrices on different cognitive assessment scores. The upper panel shows the weight matrices in VBM data and the lower panel shows the weight matrices in FreeSurfer data.

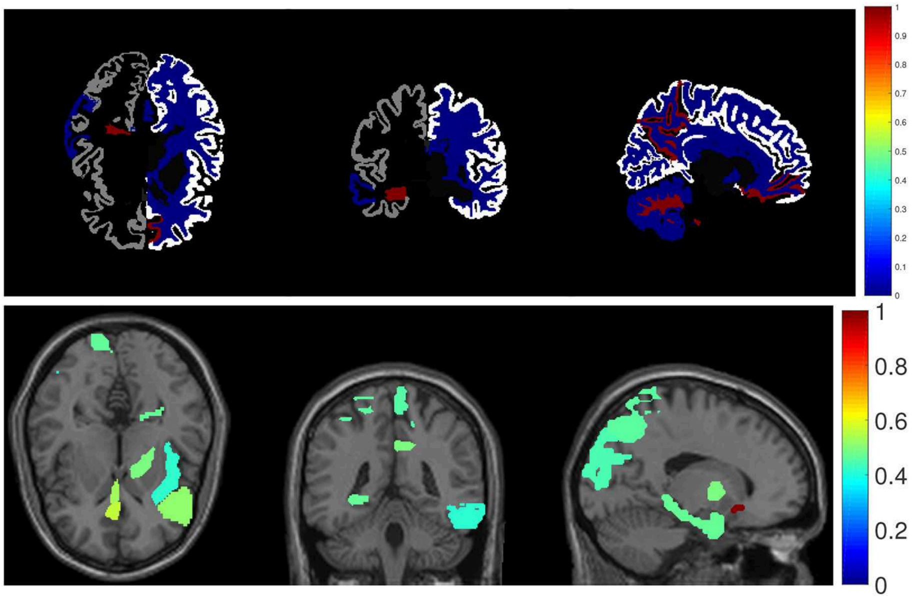

FIGURE 3 | Visualization of top identified imaging markers for RAVLT memory score prediction.

Algorithm 2: Sinkhorn Iteration.

```
Input: the given ADNI data $\mathbf{A}$ and coefficient matrix $\mathbf{Z}$
Output: $\{\mathbf{P}_1, \cdots, \mathbf{P}_l\}$
for $i = 1$ to $n$ do
    $\mathbf{K}_{(i)} = \exp(-\mathbf{C}_{(i)}/\gamma)$, where $C_{(i)}$ is the ground metric between $((\mathbf{AZ})^i)^T$ and $(\mathbf{Y}^i)^T$
    repeat
        $\mathbf{u}_i \leftarrow (((\mathbf{AZ})^i)^T/\mathbf{K}\mathbf{v}_i)$
        $\mathbf{v}_i \leftarrow ((\mathbf{Y}^i)^T/\mathbf{K}^T\mathbf{u}_i)$
    until convergence
    $\mathbf{P}_{(i)} \leftarrow (p_{(i)jt})_{n\times n}$, where the $(j,t)$-th element of $\mathbf{P}_{(i)}$ is $p_{(i)jt} = u_{i(j)}k_{(i)jt}v_{i(t)}$
end for
```

2 years are examined and associated to cognitive scores that are relevant to AD.

4.1. Data Description

The data used in the preparation of our work were obtained from the ADNI cohort. As we know, two widely employed automated MRI analysis techniques were used to process and extract imaging phenotypes form scans of ADNI participants (Shen et al., 2010). One is Voxel-Based Morphometry (VBM) (Ashburner and Friston, 2000), which was performed to define global gray matter (GM) density maps and extract local gray matter density values for 90 target regions. The other one is automated parcellation via FreeSurfer V4 (Fischl et al., 2002), which was conducted to define volumetric total intracranial volume (ICV). All these measures were adjusted for the baseline ICV using the regression weights derived from the healthy control (HC) participants. In this study, there are 805 participants, including 186 AD samples, progressive mild cognitive impairment (pMCI) samples, 167 stable mild cognitive impairment (sMCI) samples and 226 health control (HC) samples. In our work, we adopt FressSurfer markers and VBM markers as imaging phenotypes. Furthermore, the longitudinal scores were downloaded form three independent cognitive assessments including Fluency Test, RAVLT, and TRAILS. The details of these cognitive assessments can be found in the ADNI procedure manuals. The detailed information are shown in **Table 1**.

4.2. Performance Comparison on the ADNI Cohort

To evaluate the performance of our model, we compare it with the following related methods: **RR** (multivariate ridge regression), $\ell_{2,1}$ (robust feature selection based on $\ell_{2,1}$-norm), **RSR** (Regularized Self-Representation) (Zhu et al., 2015), and **RLRSS** (Robust Low-Rank Structured Sparse Model) (Xu et al., 2017). These comparing methods are all widely used in statistical learning and brain image analysis.

In the experiments, we use ridge regression for the prediction experiment after selecting the top related imaging markers. We tune the hyper-parameter of all models in the range of $\{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$ via nested five-fold cross-validation strategy, and report the best result of each method. To measure prediction performance, we compute the root mean square error (RMSE) between the predicted score and the ground truth.

The average results for each method are reported in **Figure 1**, while, we also list the RMSE using the top 10 and 30 imaging markers and reported in **Tables 2**, **3**. It can be seen that our ap proach obviously outperforms most of methods significantly. Different matrix norms fit different assumption of the cognitive measures, which makes it enable to uncover part of the correlation of cognitive measures. However, due to the natural limitation of the matrix norms, they fails to uncover the inherent geometry of the cognitive data. As for our proposed method, with the effectiveness of Wasserstein distance, it can well utilize the inherent geometry to reveal the underlying relationship between cognitive measures and neuroimaging markers.

4.3. Identification of Informative Markers

The primary goal of the proposed method is to identify the discriminative AD-specific imaging biomarkers which is crucial for early detection, diagnosis and prediction of AD. Therefore, we examine the neuroimaging markers selected by our method and show it in **Figure 2**. Visualizing the parameter weights shown in **Figure 2** can help us locate the informative markers which play important roles in the corresponding cognitive prediction tasks. As the heat map in **Figure 2** shows, different coefficient values are represented in different colors. The yellow polar means a significant effect of corresponding markers on cognitive score performance.

As the **Figure 2** shows, the extracted informative imaging biomarks are highly AD-specific and effective for related studies of AD, since it actually meets with the existing research findings. For example, among the top selected features, we found that hippocampal volume (HippVol) and middle temporal gyrus thickness (MidTemporal) are on the top, whose impact on AD have already been proved in the previous papers (Braak and Braak, 1991; West et al., 1994). Furthermore, it also confirms the important significance of the selected neuroimaging cognitive associations to uncover the relationships between MRI measures and cognitive levels.

4.4. Visualization of Top Identified Imaging Markers

As shown in **Figure 3**, we also visualize the top ten selected features for RAVLT memory score prediction on brain map as a demonstration. In the brainmap for FreeSurfer, the top 15 brain regions are (in descending order according to the ℓ_2-norm of feature weights): LPrecuneus, RCerebellWM, LHippVol, RCerebellCtx, RMedOrbFrontal, RLatVent, RCerebWM, RPrecuneus, LParahipp, LMidTemporal, LInfTemporal, RParacentral, LLingual, LPutamVol, RBanksSTS. In the brainmap for VBM, the top 15 brain regions are (in descending order according to the ℓ_2-norm of feature weights): LRectus, RAntCingulate, LInfFrontal_Triang, RMidCingulate, ROlfactory, RCalcarine, RAmygdala, RRectus, LParahipp, LPallidum, LInsula, RParacentral, LSupOccipital, LInfFrontal_Oper, RMidOrbFrontal.

5. CONCLUSION

To reveal relationship between neuroimaging data and cognitive test scores and predict cognitive score, we proposed a novel efficient matrix regression model which employs joint Wasserstein distances minimization on both loss function and regularization. To eliminate the natural limitation of the matrix norm in regression model, we utilize Wasserstein distance as distance metric. Wasserstein based regularizer can promote parameters that are close, according the OT geometry, which take into account a prior geometric knowledge on the regressor variables. Thus, our proposed method Furthermore, we provide an efficient algorithm to solve the proposed model. Extensive empirical studies on ADNI cohort demonstrate the effectiveness of our method.

AUTHOR CONTRIBUTIONS

JY, LL, CD, and HH designed the regression framework and implemented algorithm. JY wrote the manuscript and made the experiment. XW processed the data. JY and XW prepared figures and tables. CD, LL, XY, HH, and LS supervised study and revised the manuscript. All the authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

JY and CD were supported in part by the National Natural Science Foundation of China under Grant 61572388 and 61703327, in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02, 2017ZDCXL-GY-05-02, and 2018ZDXM-GY-176, and in part by the National Key R&D Program of China under Grant 2017YFE0104100. Detailed ADNI Acknowledgments information is available in http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Manuscript_Citations.pdf.

REFERENCES

Ali, S. M., and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* 131–142. doi: 10.1111/j.2517-6161.1966.tb00626.x

Alzheimer's Association (2018). 2018 alzheimer's disease facts and figures. *Alzheimer's Dement.* 14, 367–429. doi: 10.1016/j.jalz.2018.02.001

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry-the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Avramopoulos, D. (2009). Genetics of Alzheimer's disease: recent advances. *Genome Med.* 1:34. doi: 10.1186/gm34

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi: 10.1007/BF00308809

Bubeck, S., Lee, Y. T., and Singh, M. (2015). A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2016). Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*.

Csiszár, I., Shields, P. C. (2004). Information theory and statistics: a tutorial. *Found. Trends® Commun. Inform. Theory* 1, 417–528. doi: 10.1561/0100000004

Cuturi, M. (2013). "Sinkhorn distances: lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems* (Tahoe, CA), 2292–2300.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G. B., and Collins, D. L. (2008). MRI-based automated computer classification of probable AD versus normal controls. *IEEE Trans. Med. Imaging* 27, 509–520. Cuturi, 2013

Eskildsen, S. F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J. C., Collins, D. L., et al. (2013). Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage* 65, 511–521. doi: 10.1016/j.neuroimage.2012.09.058

Ewers, M., Sperling, R. A., Klunk, W. E., Weiner, M. W., and Hampel, H. (2011). Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends Neurosci.* 34, 430–442. doi: 10.1016/j.tins.2011.05.005

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). "Learning with a wasserstein loss," in *Advances in Neural Information Processing Systems* (Montreal, QC), 2053–2061.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). "Stochastic optimization for large-scale optimal transport," in *Advances in Neural Information Processing Systems* (Barcelona), 3440–3448.

Grauman, K., and Darrell, T. (2004). "Fast contour matching using approximate earth mover's distance," in *CVPR 2004 Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1* (Washington, DC: IEEE).

Guerrero, R., Wolz, R., Rao, A., Rueckert, D., and Alzheimer's Disease Neuroimaging Initiative (2014). Manifold population modeling as a neuroimaging biomarker: application to ADNI and ADNI-GO. *NeuroImage* 94, 275–286. doi: 10.1016/j.neuroimage.2014.03.036

Jack, C. R. Jr, Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): mRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

Jie, B., Zhang, D., Cheng, B., Shen, D., and Initiative, A. D. N. (2015). Manifold regularized multitask feature learning for multimodality disease classification. *Hum. Brain Mapp.* 36, 489–507. doi: 10.1002/hbm.22642

Kolouri, S., and Rohde, G. K. (2015). "Transport-based single frame super resolution of very low resolution face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4876–4884.

Kondratyev, S., Monsaingeon, L., and Vorotnikov, D. (2016). A new optimal transport distance on the space of finite radon measures. *Adv. Differ. Equat.* 21, 1117–1164.

Liu, G., Lin, Z., and Yu, Y. (2010). "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa), 663–670.

Liu, M., Zhang, D., Shen, D., and Alzheimer's Disease Neuroimaging Initiative (2014). Identifying informative imaging biomarkers via tree structured sparse learning for AD diagnosis. *Neuroinformatics* 12, 381–394. doi: 10.1007/s12021-013-9218-x

Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., and Yan, S. (2012). "Robust and efficient subspace segmentation via least squares regression," in *European Conference on Computer Vision* (Florence: Springer), 347–360. doi: 10.1007/978-3-642-33786-4_26

Luo, L., Yang, J., Qian, J., Tai, Y., and Lu, G.-F. (2017). Robust image regression based on the extended matrix variate power exponential distribution of dependent noise. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 2168–2182. doi: 10.1109/TNNLS.2016.2573644

Monge, G. (1781). *Mémoire sur la Théorie des Déblais et des Remblais.* Paris: Histoire l'Académie Royale des Sciences de Paris.

Moradi, E., Hallikainen, I., Hänninen, T., Tohka, J., and Alzheimer's Disease Neuroimaging Initiative (2017). Rey's auditory verbal learning test scores can be predicted from whole brain MRI in Alzheimer's disease. *NeuroImage* 13, 415–427. doi: 10.1016/j.nicl.2016.12.011

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., and Alzheimer's Disease Neuroimaging Initiative (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* 104, 398–412. doi: 10.1016/j.neuroimage.2014.10.002

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., et al. (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's Dement.* 1, 55–66. doi: 10.1016/j.jalz.2005.06.003

Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 1813–1821.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.* 20, 231–252. doi: 10.1007/s11222-008-9111-x

Orlin, J. B. (1993). A faster strongly polynomial minimum cost flow algorithm. *Operat. Res.* 41, 338–350.

Ozolek, J. A., Tosun, A. B., Wang, W., Chen, C., Kolouri, S., Basu, S., et al. (2014). Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Med. Image Anal.* 18, 772–780. doi: 10.1016/j.media.2014.04.004

Petrella, J. R., Coleman, R. E., and Doraiswamy, P. M. (2003). Neuroimaging and early diagnosis of Alzheimer disease: a look to the future. *Radiology* 226, 315–336. doi: 10.1148/radiol.2262011600

Piccoli, B., and Rossi, F. (2014). Generalized wasserstein distance and its application to transport equations with source. *Arch. Ration. Mech. Anal.* 211, 335–358. doi: 10.1007/s00205-013-0669-x

Piccoli, B., and Rossi, F. (2016). On properties of the generalized wasserstein distance. *Arch. Ration. Mech. Anal.* 222, 1339–1365. doi: 10.1007/s00205-016-1026-7

Rolet, A., Cuturi, M., and Peyré, G. (2016). "Fast dictionary learning with a smoothed wasserstein loss," in *Artificial Intelligence and Statistics* (Cadiz), 630–638.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* 40, 99–121. doi: 10.1023/A:1026543900054

Sandler, R., and Lindenbaum, M. (2011). Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1590–1602. doi: 10.1109/TPAMI.2011.18

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A Handbook.* Los Angeles, CA: Western Psychological Services.

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., et al. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. *Neuroimage* 53, 1051–1063. doi: 10.1016/j.neuroimage.2010.01.042

Sinkhorn, R., and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Math.* 21, 343–348. doi: 10.2140/pjm.1967.21.343

Villani, C. (2008). *Optimal Transport: Old and New, Vol. 338.* Springer Science & Business Media. doi: 10.1007/978-3-540-71050-9

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A. J., et al. (2011a). "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *2011 IEEE International Conference on Computer Vision (ICCV)* (Barcelona: IEEE), 557–562.

Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A. J., Shen, L., et al. (2011b). "Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Toronto, ON: Springer), 115–123.

Wang, X., Shen, D., and Huang, H. (2016). "Prediction of memory impairment with mri data: a longitudinal study of Alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 273–281.

West, M. J., Coleman, P. D., Flood, D. G., and Troncoso, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *Lancet* 344, 769–772. doi: 10.1016/S0140-6736(94)92338-8

Xu, J., Deng, C., Gao, X., Shen, D., and Huang, H. (2017). "Predicting Alzheimer's disease cognitive assessment via robust low-rank structured sparse model," in *IJCAI: Proceedings of the Conference, Vol. 2017* (Melbourne, QC: NIH Public Access), 3880.

Zhu, P., Zuo, W., Zhang, L., Hu, Q., and Shiu, S. C. (2015). Unsupervised feature selection by regularized self-representation. *Pattern Recogn.* 48, 438–446. doi: 10.1016/j.patcog.2014.08.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.