HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2019 April 05.

Published in final edited form as:

IEEE Trans Med Imaging. 2019 April ; 38(4): 909–918. doi:10.1109/TMI.2018.2874964.

Infant Brain Development Prediction with Latent Partial Multi-View Representation Learning

Changqing Zhang,
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA and College of Intelligence and Computing, Tianjin University, Tianjin, China, (zhangchangqing@tju.edu.cn).

Ehsan Adeli,
Department of Psychiatry and Behavioral Sciences, Stanford University, California, USA, (eadeli@stanford.edu).

Zhengwang Wu,
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA, (wuzhengwang1984@gmail.com).

Gang Li,
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA, (gang_li@med.unc.edu).

Weili Lin, and
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA, (weilLlin@med.unc.edu).

Dinggang Shen
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA, and also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea, (dgshen@med.unc.edu).

Abstract

The early postnatal period witnesses rapid and dynamic brain development. However, the relationship between brain anatomical structure and cognitive ability is still unknown. Currently, there is no explicit model to characterize this relationship in the literature. In this paper, we explore this relationship by investigating the mapping between morphological features of the cerebral cortex and cognitive scores. To this end, we introduce a *multi-view multi-task* learning approach to intuitively explore complementary information from different time-points and handle the missing data issue in longitudinal studies simultaneously. Accordingly, we establish a novel model, *Latent Partial Multi-View Representation Learning.* Our approach regards data from different time-points as different views and constructs a latent representation to capture the complementary information from incomplete time-points. The latent representation explores the complementarity across different time-points and improves the accuracy of prediction. The minimization problem is solved by the Alternating Direction Method of Multipliers (ADMM).

Correspondence to: Dinggang Shen.

Experimental results on both synthetic and real data validate the effectiveness of our proposed algorithm.

Keywords

Infant brain development; Longitudinal analysis; Cognitive ability; Multi-view learning

I. INTRODUCTION

RESEARCH on infant brain development [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] has received significant attention recently. To better understand early brain development, exploring quantitative relationship between cognitive ability and structural or functional development of infant cerebral cortex is of immense importance, as it may lead to improved health and well-being of children. However research in this area is scarce. With the advancement of magnetic resonance imaging (MRI) and image processing techniques, we now can quantitatively measure the morphology of cerebral cortex during early brain development, a characteristic that is highly correlated with human cognitive ability [11].

In this work, we propose a novel method to predict several scores related to the cognitive development of infant brains in a longitudinal study. To this end, we use longitudinal data from a cohort of infants, scanned at birth, every 3 months in the first year, every 6 months in the second year, and once a year after the second year. In this dataset, cognitive development scores were measured for each subject at the age of four years (48 months of age). Specifically, the cognitive ability of each infant was estimated using the Mullen Scales of Early Learning (MSEL) [12], including the visual reception scale (VRS), fine motor scale (FMS), receptive language scale (RLS), expressive language scale (ELS), and early learning composite scale (ELC). To build the prediction model from longitudinal MR images to the cognitive scores, we extract several types of morphological features from MR images for characterizing the structure of cerebral cortex (similar to [13], [14], [15]). Then, we build a quantitative mapping between the longitudinal morphological features of cerebral cortex and the five cognitive scores.

Given the limited amount of data that can be gathered for such a study together with the longevity and duration of data gathering, there are several challenges associated with this study: (1) samples are often very limited and building predictive machine learning models can be tricky due to the Small-Sample-Size (SSS) problem [16], [4]; (2) missing data at certain time-points are unavoidable in longitudinal studies due to various reasons (e.g., no show-up or dropouts) [16]; (3) unlike the single output regression, our problem comprises multiple outputs (scores) that are naturally interrelated [12]. To address all these challenges, we propose a method based on convex optimization techniques to recover a latent representation for each subject and simultaneously predict multiple cognitive scores given this latent representation. Our proposed method can effectively learn the subject-specific representation, regardless of the existence of missing data in any time-points.

To address the missing data problem, different approaches have been used in the literature. One straightforward approach is to learn one model based on the available data at each time-

point and then integrate the outputs of these models, as shown in Fig. 1(a). Although this strategy is simple, the complementary information of different time-points is not well exploited. To exploit multiple data sources, some other methods (such as [17], [18]) manually group samples according to the availability of data sources, and then learn one model for each group, as shown in Fig. 1(b). However, both of the aforementioned types of approaches make the SSS problem even more serious. This is because the number of samples available at one single time-point or the number of samples in one group may be much smaller than the set of all samples. As an alternative approach, data imputation methods like matrix completion [19], [20] usually recover missing values with a low-rank constraint, and then learn a model based on the completed data, as shown in Fig. 1(c). To be able to utilize the low-rank assumption, these approaches assume that the data are uniformly and randomly missing, which is not the case for our application, since the data are usually missing in blocks instead of missing uniformly [16]. Accordingly, we uncover a latent representation for each subject and learn a unified model based on all subjects, as shown in Fig. 1(d). Our approach does not assume any uniformity or other constraints on the missing data and only leverages the time-points available from each subject to build the latent representation.

The longitudinal MRI data comprises multiple data sources from multiple time-points that describe subjects from multiple views. Note that, for each time-point, the data corresponding to a subset of subjects are missing, as shown in Fig. 2. To build the relationship between the incomplete multi-view data and multiple cognitive scores, we propose a novel partial multi-view multi-task regression method, termed as *Latent Partial Multi-view Representation Learning*. Our model seeks a comprehensive and compact latent representation for each subject from the observed data at multiple time-points. Accordingly, a prediction model is learned based on the inferred latent representation, as shown in Fig. 3. The proposed model has two primary advantages: 1) Unlike most existing multi-view methods (e.g., [21], [22]) that learn models directly on the original noisy features, our model exploits the complementarity among different time-points and effectively improves the prediction accuracy. 2) Our regression model is learned based on all subjects, while existing methods [17], [18] learn multiple regression models based on different subsets of subjects and thus are not applicable for the small-sample-size problems.

II. RELATED WORK

Longitudinal Analysis of Infant Brain.

There has been intensive research conducted on infant brain development. The first line of research mainly focuses on studying the longitudinal development of cortical features [6] or the growth model [7] of the infant brain. The research in [6] studies the longitudinal development of regional cortical thickness (CT) and surface area (SA) in healthy infants from term birth to 2 years of age, revealing heterogeneous growth patterns of CT and SA. The work in [7] proposes a computational growth model for simulating the dynamic development of the cerebral cortex for term infants. In this model, the cerebral cortex is modeled as a deformable elastoplasticity surface driven via a growth model. The second line of research aims to predict the longitudinal postnatal development of cortical features (e.g.,

cortical thickness maps) [8], [9] or white matter fibers after term birth [23]. On the one hand, all the above mentioned models focus on modeling the longitudinal dynamic development of infant brain MR images after term birth, rather than relating the infant brain development scores (e.g., these five cognitive scores mentioned above) and the longitudinal neuroimages. On the other hand, Smyser et al. [3] and Kersbergen et al. [5] focus on the analysis of preterm infant development. Specifically, the work in [3] aims to identify the earliest forms of cerebral functional connectivity and characterize their development based on functional MRI instead of using structural MRI. Kersbergen et al. [5] investigate third-trimester extrauterine brain growth and correlate this with clinical risk factors in the neonatal period. Although longitudinal data are involved in the studies conducted by [3] and [5], in contrast to ours, they do not leverage such longitudinal data for prediction.

Multi-view Learning.

Many real-world applications usually involve multi-view learning, since data usually can be obtained from multiple sources or represented with multiple types of features. Due to the effectiveness of exploring the complementarity among multiple views, multi-view learning has attracted close attention recently. Some methods try to minimize the disagreement between different views under the co-training framework [24], [25], [26]. Furthermore, the work in [27] provides theoretical analyses to support the success and appropriateness of co-training-based methods. Multiple kernel learning (MKL) [28] uses a predefined set of kernels from multiple views and learns the optimized weights for kernels to integrate these views. Recently, some methods advocate for the learning of a latent common subspace across different views, typically, based on canonical correlation analysis (CCA) [25], [29]. Although promising performance has been achieved by these methods, most of them are not applicable for data with incomplete views. Several previous methods (e.g., [30], [31], [32], [17]) also take advantage of multi-modal imaging data for disease diagnosis. For instance, Gray et al. [30] integrate the similarities from multiple neuroimaging and biological measures to generate an embedding, based on which the classifier is learned. Singanamalli et al. [31] extend the canonical correlation analysis (CCA) as supervised multiview canonical correlation analysis (sMVCCA), to find a common representation for multi-modal data. Recently, Liu et al. [32] extract the common features of multiple image modalities under the framework of deep de-noising autoencoder. Similar to our method, Yuan et al. [17] also propose a technique to handle the data with missing modalities. They divide samples according to the availability of data sources, and then classifiers are learned based on each group of samples. Unlike our approach, this strategy cannot scale well for problems where the number of data sources is large, or the number of samples is small.

Multi-task learning.

Our problem belongs to the category of multi-task learning (MTL) problems, since we aim to predict multiple scores simultaneously. Naively, multitask learning problems can be reduced to multiple singletask learning (STL) problems, in which each task is solved independently. However, with this setting, the correlations among different tasks cannot be properly explored. Plenty of empirical studies have proven that exploiting the relationship among multiple related tasks (in the context of MTL) can generally provide superior predictive performance compared to the case of learning each task independently [33], [34],

[35]. Furthermore, there are some works providing theoretical foundations for the success of multi-task learning [36], [37], especially for the small-sample-size issue in each task.

III. Material and Preprocessing

Material.

In our study, T1-weighted and T2-weighted MR images from 23 infant subjects were collected and each infant was scheduled to have longitudinal scans at 9 different time-points (i.e., 1, 3, 6, 9, 12, 18, 24, 36 and 48 months). As can be inferred from Fig. 2, most subjects did not show up for all scheduled time-points, thus causing the missing data issue. This is typical in longitudinal studies. Five Mullen cognitive scores [12], i.e., Visual Reception Scale (VRS), Fine Motor Scale (FMS), Receptive Language Scale (RLS), Expressive Language Scale (ELS), and Early Learning Composite (ELC), are measured for each subject at the 48^{th} month. The 5^{th} score (i.e., ELC) can be interpreted as the composite of the other four scores. Therefore, the 5^{th} cognitive score correlates with the other four [12], [38].

Image Processing.

All infant MR images are preprocessed by an established infant-specific computational pipeline [39]. Briefly, the pipeline includes intensity inhomogeneity correction, skull stripping, cerebellum removal, tissue segmentation, separation of left/right hemispheres, topology correction, inner and outer surface reconstruction. Then, for each vertex on the inner or outer cortical surface, 7 types of morphological features are computed, i.e., cortical thickness, local gyrification index, mean curvature, vertex area, sulcal depth measured in Euclidean distance, sulcal depth measured in string distance, and vertex volume [6], [13], [40], [41], [42]. Interested readers can refer to [43], [39] for the details of different types of morphological features. These morphological features jointly measure brain anatomical structure and are shown to be highly correlated with the cognitive abilities [44]. However, using the morphological features of each vertex to predict cognitive scores is redundant and computationally expensive. Therefore, the Region-Of-Interest (ROI) based analysis is adopted. To obtain a meaningful ROI definition, we warp the well-accepted FreeSurfer parcellation [45] onto each individual cortical surface. For each ROI, we can compute its representative morphological features by either averaging or summing up the corresponding features over all vertices belonging to that ROI. Notably, for the vertex area and vertex volume, we use their sum over the entire ROI vertices as the area and volume of each ROI, respectively. For the other 5 types of features, the mean value of each feature in each ROI is calculated as the morphological feature.

For the entire cerebral cortex, the FreeSurfer parcellation includes a total of 70 anatomically meaningful ROIs [43], and for each ROI, we obtain 7 types of morphological features. Accordingly, a 490-dimensional feature vector can be obtained for each subject at each time-point. Then, we learn a regression model between the 490-dimensional vectors and the 5 cognitive scores.

IV. INFANT BRAIN DEVELOPMENT PREDICTION WITH LATENT PARTIAL MULTI-VIEW REPRESENTATION LEARNING

In this section, we introduce a novel multi-view multitask learning method, which does not hold the limitations of discarding or completing incomplete data in advance, and thus can fully take advantage of all the observed data based on a latent multi-view representation for each subject.

A. Formulation

We denote the data at multiple time-points as $\{\mathbf{X}_1,\ldots,\mathbf{X}_T;\mathbf{Y}\}$, where $\mathbf{X}_t \in \mathbb{R}^{D\times N}$ is the data matrix at the t^{th} time-point and $\mathbf{Y} \in \mathbb{R}^{C\times N}$ is the score matrix, where D is the dimensionality of the original feature space, C is the dimensionality of output, and N is the number of samples. In our model, we formulate the learning task as a multi-task (C scores) multi-view learning problem with each view comprising data from one of the T time-points. We aim to uncover a multi-view latent representation which holds the reconstruction ability for the data at different time-points. Specifically, the reconstruction ability indicates the degree of the information from different time-points encoded into the latent representation. Accordingly, the formulation is

$$\min_{\mathbf{H}} \sum_{t=1}^{T} \mathscr{V}\left(\mathscr{F}_t(\mathbf{H}), \mathbf{X}_t\right), \quad (1)$$

where $\mathscr{V}(\cdot,\cdot)$ measures the reconstruction loss and $\mathscr{F}_t(\cdot)$ indicates the underlying mapping from the latent representation $\mathbf{H}$ to the observations at the t^{th} time-point, i.e., $\mathbf{X}_t$, $\forall t \in \{1,\ldots,T\}$. $\mathscr{V}(\cdot,\cdot)$ is defined as

$$\mathscr{V}\left(\mathscr{F}_t(\mathbf{H}), \mathbf{X}_t\right) = \left\|\left(\mathscr{F}_t(\mathbf{H}) - \mathbf{X}_t\right)\right\|_{2,1}, \quad (2)$$

where $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-norm of the residual encouraging some rows of the matrix to be zero. The underlying assumption is that the noises are feature-specific, i.e., a few features are noisy, hence we do not need to consider their reconstruction loss. This loss leads to a level of robustness against feature noise. For the mapping $\mathscr{F}_t(\cdot)$, we employ a linear projection in our model, which is a simple but effective technique especially for the high-dimensional data. Accordingly, we have

$$\mathscr{V}\left(\mathscr{F}_t(\mathbf{H}), \mathbf{X}_t\right) = \left\|\mathbf{P}_t\mathbf{H} - \mathbf{X}_t\right\|_{2,1}. \quad (3)$$

Based on the learned latent representation $\mathbf{H}$ from multiple views, we can define the following multi-task regression term to predict the five cognitive scores as

$$\min_{\mathbf{W}} \mathscr{L}(\mathbf{W}, \mathbf{H}, \mathbf{Y}) = \|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_1. \quad (4)$$

This ℓ_1-norm constrained loss function leads to a robust loss [46]. Note that, the learned model $\mathbf{W}$ is learned based on all the N samples regardless of the missing status.

Since the early learning composite scale (ELC) is correlated with other scales, for the learned model $\mathbf{W}$ we introduce the well-known low-rank regularization

$$\mathscr{R}(\mathbf{W}) = \|\mathbf{W}\|_*, \quad (5)$$

where $\|\cdot\|_*$ is the matrix nuclear-norm. Putting these terms in a unified optimization problem, our objective function is induced as

$$\min_{\Omega} \underbrace{\|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_1}_{\text{prediction error}} + \underbrace{\alpha \sum_{t=1}^{T} \omega_t^r \left\| \mathscr{P}_{\mathbf{O}_t}(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t) \right\|_{2,1}}_{\text{reconstruction error}} + \underbrace{\beta\|\mathbf{W}\|_*}_{\text{task correlation}} \quad (6)$$

$$s.t. \sum_{t=1}^{T} \omega_t = 1, \omega_t \geq 0; \mathbf{P}_t^\top \mathbf{P}_t = \mathbf{I}, t = 1, \cdots, T.$$

For convenience, we denote $\Omega = \left\{\mathbf{W}, \mathbf{H}, \{\mathbf{P}_t\}_{t=1}^T, \{\omega_t\}_{t=1}^T\right\}$ as the variable set to be optimized, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_T)$ is the weight vector for multiple time-points. $r > 1$ for ω_t is used to avoid a trivial solution that only considers one of the T time-points and adjusts the complementarity of multiple time-points [47]. The constraint $\mathbf{P}_t^\top \mathbf{P}_t = \mathbf{I}$ is introduced, since without this constraint $\mathbf{P}_t$ can be pushed arbitrarily close to zero only by re-scaling $\mathbf{P}_t/s$ and $\mathbf{H}s$ ($s > 0$) while preserving the same loss. Moreover, our model can be efficiently solved with the constraint (see $\mathbf{P}_t$-subproblem in optimization part). $\mathscr{P}_{\mathbf{O}_t}(\cdot)$ is a filter function to handle the incomplete data for the t^{th} time-point. Let o_t^s be an indicator variable showing the existence of data for subject s at time-point t, i.e., $o_t^s = 1$ if the data is available, and a very small scalar $\epsilon > 0$ otherwise. $\mathbf{o}_t$ will then be defined as the indicator vector from all indicator variables of training samples. Accordingly, we can define a diagonal matrix $\mathbf{O}_t = \text{diag}(\mathbf{o}_t)$, denoted as the filter matrix of the t^{th} time-point, and hence $\mathscr{P}_{\mathbf{O}_t}(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t) = (\mathbf{P}_t\mathbf{H} - \mathbf{X}_t)\mathbf{O}_t$. Note that $\epsilon > 0$ is a small value to optimize $\mathbf{H}$-subproblem in next subsection.

Model properties.—To summarize, we highlight that the proposed latent partial multi-view representation enjoys the following merits: (1) Our regression model $\mathbf{W}$ is learned by

simultaneously utilizing all the subjects and time-points, which is especially important for the small-sample-size case. (2) The latent multi-view representation could depict data themselves more comprehensively than each single view individually, which makes the prediction model more accurate and robust. (3) The inter-task correlations, feature-specific corruption, and output (score) noise are explicitly encoded in our model to jointly guarantee the robustness of the model.

B. Optimization

Our objective function in Eq. (6) simultaneously seeks a latent representation from multiple views and learns a multitask prediction model with respect to the latent representation. Since the objective function is not jointly convex with respect to all the variables $\mathbf{P}_t$, $\mathbf{H}$ and $\mathbf{W}$, we employ Augmented Lagrange Multiplier (ALM) with Alternating Direction Method (ADM) strategy [48], [49]. To adopt ADM strategy to our problem, we need to make our objective function separable. Therefore, we introduce auxiliary variables $\mathbf{J}$, $\{\mathbf{E}_t\}_{t=1}^{T}$ and $\mathbf{E}$, and then we have the following equivalent problem

$$\begin{aligned} &\min_{\mathbf{\Omega}} \|\mathbf{E}\|_1 + \alpha \sum \omega_t^r \|\mathscr{P}_{\mathbf{O}_t}(\mathbf{E}_t)\|_{2,1} + \beta \|\mathbf{J}\|_* \qquad (7)\\ &s.t. \sum_{t=1}^{T} \omega_t = 1, \omega_t \geq 0;\ \mathbf{P}_t^{\top}\mathbf{P}_t = \mathbf{I};\ \mathbf{J} = \mathbf{W};\\ &\mathscr{P}_{\mathbf{O}_t}(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t) = \mathscr{P}_{\mathbf{O}_t}(\mathbf{E}_t);\quad \mathbf{WH} = \mathbf{Y} + \mathbf{E}. \end{aligned}$$

The augmented Lagrangian function of Eq. (7) is given as

$$\begin{aligned} &\mathscr{L}(\Omega) = \left\|\mathbf{E}\right\|_1 + \alpha \sum \omega_t^r \left\|\mathscr{P}_{\mathbf{O}_t}(\mathbf{E}_t)\right\|_{2,1} + \beta \left\|\mathbf{J}\right\|_* \qquad (8)\\ &+ \Phi(\mathscr{G}_B, \mathbf{J} - \mathbf{W}) + \sum_{t=1}^{T} \Phi\left(\mathscr{G}_t, \mathscr{P}_{\mathbf{O}_t}(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t - \mathbf{E}_t)\right) + \Phi(\mathscr{G}_A, \mathbf{WH} - \mathbf{Y} - \mathbf{E})\\ &s.t. \sum_{t=1}^{T} \omega_t = 1, \omega_t \geq 0;\ \mathbf{P}_t^{\top}\mathbf{P}_t = \mathbf{I},\ t = 1, \cdots, T, \end{aligned}$$

where Ω is the set of all variables to be optimized and we define $\Phi(\mathscr{G}, \mathbf{Z}) = \frac{\mu}{2}\left\|\mathbf{Z}\right\|_F^2 + \left\langle \mathscr{G}, \mathbf{Z} \right\rangle$ for simplicity, with $\langle \cdot, \cdot \rangle$ being the matrix inner product. $\{\mathscr{G}_t\}_{t=1}^{T}$, $\mathscr{G}_A$ and $\mathscr{G}_B$ are Lagrangian multipliers along with the constraints, and $\mu > 0$ is a penalty hyperparameter.

Below, we provide the optimization for each sub-problem:

H-subproblem.—For solving this sub-problem, by fixing other variables except $\mathbf{H}$, we should solve the following problem:

$$\min_{\mathbf{H}} \sum \Phi\left(\mathscr{G}_t, \mathscr{P}_{\mathbf{O}_t}\left(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t - \mathbf{E}_t\right)\right) + \Phi\left(\mathscr{G}_A, \mathbf{WH} - \mathbf{Y} - \mathbf{E}\right). \quad (9)$$

Taking the derivative with respect to **H** and setting it to zero, we have

$$\begin{aligned} &\mathbf{AH} + \mathbf{HB} = \mathbf{C} \quad (10)\\ &\text{with } \mathbf{A} = \mathbf{W}^\top\mathbf{W}, \mathbf{B} = \sum \mathbf{s}_t,\\ &\mathbf{C} = \sum\left(\mathbf{P}_t^\top\mathbf{X}_t\mathbf{S}_t + \mathbf{P}_t^\top\mathbf{E}_t\mathbf{S}_t - \mathbf{P}_t^\top\mathscr{G}_t\mathbf{O}_t/\mu\right) + \mathbf{W}^\top\left(\mathbf{Y}^\top + \mathbf{E} - \mathscr{G}_A/\mu\right). \end{aligned}$$

The above equation is a Sylvester equation [50]. We set o_t^s as a very small number ϵ for missing time-points, instead of zero, to ensure the unique solution of the Sylvester Equation of (10). Specifically, in this way, the matrix **B** in the equation (10) will be positive-definite, which makes Proposition 4.1 provable. If we set o_t^s to zero, there is no guarantee for a unique solution, and the numerical instability will also rise [51].

Proposition 4.1: The Sylvester equation (10) has a unique solution.

Proof 4.1: The Sylvester equation **AH** + **HB** = **C** has a unique solution for **H** exactly when there are no common eigenvalues of **A** and -**B** [50]. Since $\mathbf{S}_t = \mathbf{O}_t^{\mathrm{T}}\mathbf{O}_t$ is strictly positive-definite due to the introduced ϵ, **B** is a positive-definite matrix, and all of its eigenvalues are positive: $\beta_i > 0$. Since **A** is a positive semi-definite matrix, all of its eigenvalues are nonnegative: $\alpha_i \geq 0$. Hence, for any eigenvalues of **A** and **B**, $\alpha_i + \beta_j > 0$. Accordingly, the Sylvester equation (10) has a unique solution.

$\mathbf{P}_t$-subproblem.—By fixing other variables except $\mathbf{P}_t$, we should solve the following problem:

$$\min_{\mathbf{P}_t} \Phi\left(\mathscr{G}_t, \mathscr{P}_{\mathbf{O}_t}\left(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t - \mathbf{E}_t\right)\right) \quad s.t.\ \mathbf{P}_t^\top\mathbf{P}_t = \mathbf{I}. \quad (11)$$

The optimization with orthogonality constraints can be efficiently solved with the way of constraint-preserving update formula and corresponding curvilinear search algorithms [52].

W-subproblem.—By fixing other variables except **W**, we should solve the following problem:

$$\min_{\mathbf{W}} \Phi\left(\mathscr{G}_B, \mathbf{J} - \mathbf{W}\right) + \Phi\left(\mathscr{G}_A, \mathbf{WH} - \mathbf{Y} - \mathbf{E}\right), \quad (12)$$

which has the following closed-form solution

$$\mathbf{W} = \left(\mathbf{I} + \mathbf{H}\mathbf{H}^T\right)^{-1}\left(\mathbf{J} + \mathscr{G}_B/\mu + \left(\mathbf{Y} + \mathbf{E} - \mathscr{G}_A/\mu\right)\mathbf{H}^T\right). \quad (13)$$

J-subproblem.—By fixing other variables except **J**, we should solve the following problem:

$$\min_{\mathbf{J}} \gamma \left\|\mathbf{J}\right\|_* + \Phi\left(\mathscr{G}_B, \mathbf{J} - \mathbf{W}\right). \quad (14)$$

The above problem can be efficiently solved by the singular value thresholding operator:

$$\mathbf{J} \leftarrow \mathscr{D}_\tau(\mathbf{A}) \text{ with } \mathbf{A} = \mathbf{W} - \mathscr{G}_B/\mu, \quad (15)$$

where $\tau = \gamma/\mu$ is the thresholds of the spectral soft-threshold operation $\mathscr{D}_\tau(\mathbf{A}) = \mathbf{U}\max(\mathbf{S} - \tau, 0)\mathbf{V}^T$ with $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ being the Singular Value Decomposition (SVD) of **A** and the max operation being taken element-wise.

$\mathbf{E}_t$-subproblem.—By fixing other variables except E_t, we should solve the following problem:

$$\min_{\mathbf{E}_t} \alpha\omega_t^r \|\mathscr{P}_{\mathbf{O}_t}(\mathbf{E}_t)\|_{2,1} + \Phi\left(\mathscr{G}_t, \mathscr{P}_{\mathbf{O}_t}\left(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t - \mathbf{E}_t\right)\right) = \min \frac{\alpha\omega_t^r}{\mu}\|\mathscr{P}_{\mathrm{O}_t}(\mathbf{E}_t)\|_{2,1} + \frac{1}{2}$$
$$\|\mathscr{P}_{\mathbf{O}_t}\left(\mathbf{E}_t - \left(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t + \mathscr{G}_t/\mu\right)\right)\|_F^2. \quad (16)$$

This subproblem can be efficiently solved by Lemma 3.2 in [53].

E-subproblem.—For solving this sub-problem, by fixing other variables except **E**, we should solve the following problem:

$$\min_{\mathbf{E}} \|\mathbf{E}\|_1 + \Phi\left(\mathscr{G}_A, \mathbf{W}\mathbf{H} - \mathbf{Y} - \mathbf{E}\right) = \min \frac{1}{2}\|\mathbf{E} - \left(\mathbf{W}\mathbf{H} - \mathbf{Y} + \mathscr{G}_A/\mu\right)\|_F^2 + \frac{1}{\mu}\|\mathbf{E}\|_1. \quad (17)$$

This step involves minimization of the ℓ_1-norm of a matrix, which can be optimally obtained using soft thresholding operator or the proximal operator for the ℓ_1-norm [54].

For the weight vector $\boldsymbol{\omega}$, by using a Lagrange multiplier it can be updated by the following rule:

$$\omega_t \leftarrow \left(1/\left\|\mathbf{E}_t\right\|_{1,2}\right)^{1/r-1} / \sum_{t=1}^{T} (1/\left\|\mathbf{E}_t\right\|_{1,2})^{1/r-1}, \quad (18)$$

since $\|\mathbf{E}_t\|_{1,2} > 0$ in practice, $\omega_t > 0$ is guaranteed. Additionally, for the multipliers, we have the following update rule:

$$\begin{aligned} &\mathcal{G}_t \leftarrow \mathcal{G}_t + \mu\left(\mathcal{P}_{\mathbf{O}_t}\left(\mathbf{P}_t\mathbf{H} - \mathbf{X}_t - \mathbf{E}_t\right)\right), \quad t = 1, \cdots, T; \quad (19)\\ &\mathcal{G}_A \leftarrow \mathcal{G}_A + \mu(\mathbf{WH} - \mathbf{Y} - \mathbf{E}),\ \mathcal{G}_B \leftarrow \mathcal{G}_B + \mu(\mathbf{J} - \mathbf{W}). \end{aligned}$$

The penalty parameter μ is updated as $\mu \leftarrow \min(\mu\rho, 10^6)$. These subproblems are iteratively solved to update the variables involved.

In optimization, we initialize the weights of different time-points $\boldsymbol{\omega}$ by $\omega_1 = \ldots \omega_T = 1/T$, and all the other variables are set to zeros.

Testing phase: To obtain the regression scores for any novel test subject $\mathbf{x}$, a 2-step procedure is conducted as follows:

$$\begin{aligned} &\text{step 1.} \min_{\mathbf{h}} \sum_{t=1}^{T} \omega_t^r \|\mathbf{P}_t\mathbf{h} - \mathbf{x}_t\|^2 s_t + \lambda \sum_{t=2}^{T} \|\mathbf{P}_t\mathbf{h} - \mathbf{P}_{t-1}\mathbf{h}\|^2 \quad (20)\\ &\text{step 2.}\ \mathbf{y} = \mathbf{Wh}, \end{aligned}$$

where in the first step our model aims to uncover the latent representation according to the observed data and constrained with the temporal smooth term for the new coming subject, and then projects it onto the output space in the second step.

C. Complexity and Convergence

Our method is composed of multiple sub-problems. For updating $\mathbf{H}$, the classical algorithm for the Sylvester equation is the Bartels Stewart algorithm [50], whose complexity is $O(N^3)$, where N is the number of samples. The complexity of updating $\mathbf{J}$ (the nuclear norm proximal operator) and $\mathbf{W}$ are $O(N^3)$ and $O(K^3)$, where K is the dimension of latent representation. For updating $\mathbf{E}$, the main complexity is the matrix multiplication, which is $O(DKN)$, where D is the dimension of the original feature space. Overall, the total complexity is $O(K^3 + N^3 + DKN)$ for each iteration. Under the condition $K \ll D$ and $K \ll N$, the total complexity is basically $O(N^3 + DN)$. It is difficult to generally prove the convergence for our algorithm. Fortunately, empirical evidence on both the synthetic data the real data suggests that the proposed algorithm has very strong and stable convergence.

D. Discussion

One of the challenges for the proposed formulation is that a low-dimensional latent representation may not be enough to reconstruct an entire morphological map. It is important to note that here, our method operates on the ROI-based features (i.e., a relatively small-sized feature vector), rather than maps. Hence, there is no need to reconstruct an entire map from the latent representation. Furthermore, we only intend to encode the intrinsic information for the prediction task (and not to actually reconstruct the maps). Therefore, a low-dimensional latent representation usually suffices, which is empirically validated by our experiments. In addition, we formulate the problem in a linear setting, i.e., $\mathbf{P}_t\mathbf{H} - \mathbf{X}_t$, which is a special case for a mapping in our general model of Eq. (1). This mapping can be extended in the future by addressing nonlinear relationships with kernel methods or neural networks.

V. Experiments

We conduct experiments on both the synthetic and real infant brain data to evaluate our method. The performance is measured with Root Mean Squared Error (RMSE). All the hyperparameters are tuned from the set $\{10^{-3},10^{-2},0.1,1,10,10^2,10^3\}$ through a nested leave-one-out cross-validation. The number of dimensions for latent representation is set as $K = 10$ for the infant brain data and $K = 30$ for the synthetic data.

A. Experiments on Synthetic Data

On the synthetic data, we evaluate the effectiveness of our model from the following aspects: (1) the ability to explore multiple views; (2) the robustness against missing data; (3) the convergence property in practice. The latent representation matrix $\mathbf{H}$, model $\mathbf{W}$ and projections $\{\mathbf{P}_t\}_{t=1}^{T}$ are randomly generated with each element independently sampled from a uniform distribution on the [0, 1] interval. To simulate the correlation among multiple outputs, we process $\mathbf{W}$ by $\mathbf{w}_i = \sum_{j \neq i} \lambda_j \mathbf{w}_j \left(\text{i.e., } i = 5 \text{ and } \lambda_j = \frac{1}{C-1} \right)$, which is similar to the prediction of the five cognitive scores of the infant brain data. Then, the output matrix $\mathbf{Y}$ is obtained by $\mathbf{Y} = \mathbf{WH} + \mathbf{E}$, and the observations are generated with $\mathbf{X}_t = \mathbf{P}_t\mathbf{H} + \mathbf{E}_t$. The noise matrices, i.e., $\mathbf{E}$ and $\mathbf{E}_t$, are generated by randomly corrupting 10% samples, respectively. The dimension of the output space is $C = 5$, which is similar to our real application. We have $N = 23$ samples, which is equal to the size of real dataset. The setting is mainly used to validate the effectiveness of our approach for small-sample-size problem. There are 10 views generated to investigate effect of using a varying numbers of views.

As shown in Fig. 4(a), with the help of more views, the performance is sensibly improved. In Fig. 4(b), we investigate the effect of missing data rate with different numbers of views. It is observed that the degradation is mild with the rise of missing rate especially when the missing rate is not larger than 0.25 or with more views. The convergence conditions are all reached within less than 60 iterations as shown in Fig. 4(c).

B. Infant Brain Development Prediction

Performance with Different Numbers of Time-Points.—We first run our method with the data from different numbers of time-points. According to the Table I, our model can well leverage the data of different time-points for promising performance. We conduct the paired *t*-test between the results before using data at the 18^{th} month (i.e., only using the data from the 0^{th} to the 12^{th} month) versus the results by including the data at the 18^{th} month (i.e., using the data from the 0^{th} to the 18^{th} month). The *p*-value is generally large (i.e., 0.664) which indicates that the improvement after including scans acquired at the 18^{th} month is not significant. Similar cases are also observed for the results before and after including data at the 24^{th} and the 36^{th} months. One possible reason is that data at these later time-points are much more severely incomplete. The second possible reason is from the law of diminishing marginal returns, which is also consistent with the experimental results on the synthetic data as shown in Fig. 4(a).

Generally, it is unreasonable to conduct training and testing with each single time-point in our task, considering the small-sample-size problem will be much more severe under the missing data condition. For example, there are only 10 subjects available at the 36^{th} month. Indeed, this is the motivation of our method (i.e., simultaneously utilizing data at all time-points to address the missing data issue). Nevertheless, we also conduct experiments on the other direction, i.e., adding data from the last time-point as shown in Fig. 5. The following observations can be drawn from the mentioned experiments: (1) With more time-points added, the performance becomes generally better. (2) The data at late time-points more strongly correlate with the cognitive scores than early ones. For example, the performance with the data at the 9^{th} time-point is better than that at the 0^{th} time-point. (3) We also note that the data at the 8^{th} time-point do not improve the prediction performance. The performance becomes worse when using both the 8^{th} and 9^{th} time-points, compared with the case of using only the data at the 9^{th} time-point. However, it is difficult to conclude that information at the 8^{th} time-point is not correlated with cognitive scores, since the missing data issue at this time-point is rather serious. Further investigations could be conducted with more data involved.

Performance Comparison.—Since existing methods are not applicable for our data, we adopt two strategies to process the data to make them suitable for the existing multi-task methods. Specifically, the first way is to complete the missing values simply with zero, and the second way is to fill the missing values with the averaged values of the observed ones. We compare our model with the following methods: 1) NN (nearest neighbour); 2) MtJFS (Multi-Task Learning with Joint Feature Selection) [55]; 3) RMTL (Robust Multi-Task Feature Learning) [34]; 4) TrMTL (Trace-Norm Regularized Multi-Task Learning) [56]. From Table II, it is observed that simply filling the missing values with zero is not reasonable since the performance tends to be relatively poor. Our method outperforms both TrMTL and RMTL that also constrain the prediction model to be low-rank, which validates the effectiveness of learning the regression model based on latent representation. The performance of MtJFS, which aims to jointly select a subset of features for all tasks, is also poor. The possible reasons are 1) its limitation of handling missing data, and 2) impropriety of sharing a common subset of features for the five cognitive scores.

In addition, we also conduct experiments by setting equal weights for all different time-points. Using all 9 time-points with equal weights leads to an average RMSE (Root Mean Squared Error) of 0.163, which is worse than the proposed weighting scheme (i.e., average RMSE = 0.158). The proposed weighting scheme leads to a relative improvement of about 3.8% in terms of RMSE. As shown in Fig. 6, we present the estimated weight vector and find that the weight corresponding to the 18^{th} month is relatively high. This indicates small reconstruction error and thus the information at the 18^{th} month is sufficiently encoded into the learned latent representation. However, since the number of samples at the 18^{th} month is relatively small, this may lead to over-fitting for the learned model.

C. Model Analysis

Parameter tuning.—We conduct experiments to investigate the impact of the number of dimensions for **H.** As shown in Fig. 7, it is observed that the prediction performance is relatively robust with respect to different numbers of dimensions for **H** within a broad range. Specifically, for the synthetic data, when the dimensionality K is larger than 32 (i.e., $K \geq 32$), the best performance is obtained; whereas, for the real data, when $32 \geq K \geq 8$, the performance is generally reasonable and the best performance is obtained in this range. As shown in the right subfigure of Fig. 7(a), a large dimensional **H** yields a model with unstable prediction performance and the results show that this model is prone to over-fitting. We also conduct experiments for evaluating the hyperparameter α in Fig. 8. On the real data, it can be observed that the best performance is obtained by setting α in the range [0.1 10]. However, for much larger α values (e.g., $\alpha > 100$), the prediction performance would suffer, due to overemphasis on the reconstruction loss. For the synthetic data, the performance is relatively stable with a large value for α.

VI. CONCLUSION AND DISCUSSION

In this work, we propose to explore the relationship between cognitive scores and morphological features of the cerebral cortex, and develop a novel multi-task multi-view regression model for this challenging problem. Based on the latent representation, our model effectively addresses the challenge of learning with incomplete longitudinal data. We also introduce an optimization algorithm for the proposed method and validated the effectiveness on both the synthetic and real data.

Our model is able to predict the cognitive scores even at the presence of missing longitudinal data, as we introduce the latent representation. However, there are several issues that require further clarifications and possible future investigations. First, it is difficult to analyze the correspondence between ROIs and cognitive scores after mapping the original features to the latent representation. Hence, preserving brain structural information in the latent representation may be helpful for analyzing which region(s) are critical in predicting specific score(s). Second, due to the cost and difficulty associated with longitudinal data collection, there are only 23 subjects with cognitive scores available, which is a relatively small dataset. More data should be acquired for better performance, and semi-supervised techniques could be utilized to leverage subjects without cognitive scores and enhance the generalization ability of our proposed method. Third, we linearly model the correlations, i.e., the

correlation between the latent representation and the cognitive scores as well as the correlation between different views. With more data involved in the future, non-linearity (e.g., deep networks) can be introduced to address more complex correlations. Finally, we use the FreeSurfer parcellation scheme to parcellate the cerebral cortex into different ROIs according to the gyral and sulcal patterns [45]. The rationality of using this scheme for infant brain parcellation lies in the fact that all major gyral and sulcal folds are established at term birth and are stable during postnatal brain development [57]. For example, this parcellation scheme has been successfully adopted in infant studies [23], [58]. However, leveraging infant-specific parcellation schemes could potentially further improve the performance.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61602337. This work was also supported by NIH under Grants CA206100, MH100217, MH107815, MH108914, MH110274, MH116225, MH117943, AA026762 and the BCP Grant 1U01MH110274.

REFERENCES

[1]. Mehler J, Jusczyk P, Lambertz G, Halsted N, Bertoncini J, and Amiel-Tison C, “A precursor of language acquisition in young infants”, Cognition, vol. 29, no. 2, pp. 143–178, 1988. [PubMed: 3168420]

[2]. Dubois J, Hertz-Pannier L, Cachia A, Mangin J, Le Bihan D, and Dehaene-Lambertz G, “Structural asymmetries in the infant language and sensori-motor networks”, Cerebral Cortex, vol. 19, no. 2, pp. 414–423, 2008. [PubMed: 18562332]

[3]. Smyser CD, Inder TE, Shimony JS, Hill JE, Degnan AJ, Snyder AZ, and Neil JJ, “Longitudinal analysis of neural network development in preterm infants”, Cerebral cortex, vol. 20, no. 12, pp. 2852–2862, 2010. [PubMed: 20237243]

[4]. Smyser CD, Dosenbach NU, Smyser TA, Snyder AZ, Rogers CE, Inder TE, Schlaggar BL, and Neil JJ, “Prediction of brain maturity in infants using machine-learning algorithms”, Neuroimage, vol. 136, pp. 1–9, 2016. [PubMed: 27179605]

[5]. Kersbergen KJ, Makropoulos A, Aljabar P, Groenendaal F, de Vries LS, Counsell SJ, and Benders MJ, “Longitudinal regional brain development and clinical risk factors in extremely preterm infants”, The Journal of pediatrics, vol. 178, pp. 93–100, 2016. [PubMed: 27634629]

[6]. Lyall AE, Shi F, Geng X, Woolson S, Li G, Wang L, Hamer RM, Shen D, and Gilmore JH, “Dynamic development of regional cortical thickness and surface area in early childhood”, Cerebral cortex, vol. 25, no. 8, pp. 2204–2212, 2014. [PubMed: 24591525]

[7]. Nie J, Li G, Wang L, Gilmore JH, Lin W, and Shen D, “A computational growth model for measuring dynamic cortical development in the first year of life”, Cerebral Cortex, vol. 22, no. 10, pp. 2272–2284, 2011. [PubMed: 22047969]

[8]. Meng Y, Li G, Rekik I, Zhang H, Gao Y, Lin W, and Shen D, “Can we predict subject-specific dynamic cortical thickness maps during infancy from birth?” Human Brain Mapping, vol. 38, no. 6, pp. 2865–2874, 2017. [PubMed: 28295833]

[9]. Rekik I, Li G, Lin W, and Shen D, “Predicting infant cortical surface development using a 4d varifold-based learning framework and local topography-based shape morphing”, Medical image analysis, vol. 28, pp. 1–12, 2016. [PubMed: 26619188]

[10]. Jha SC, Xia K, Ahn M, Girault JB, Li G, Wang L, Shen D, Zou F, Zhu H, Styner M et al., “Environmental influences on infant cortical thickness and surface area”, Cerebral Cortex, 2018.

[11]. Paterson SJ, Heim S, Friedman JT, Choudhury N, and Benasich AA, “Development of structure and function in the infant brain: Implications for cognition, language and social behaviour”, Neuroscience & Biobehavioral Reviews, vol. 30, no. 8, pp. 1087–1105, 2006. [PubMed: 16890291]

[12]. Bradley-Johnson S, “Mullen scales of early learning”, Psychology in the Schools, vol. 34, no. 4, pp. 379–382, 1997.

[13]. Li G, Wang L, Shi F, Lyall AE, Lin W, Gilmore JH, and Shen D, “Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age”, Journal of Neuroscience, vol. 34, no. 12, pp. 4228–4238, 2014. [PubMed: 24647943]

[14]. Li G, Nie J, Wang L, Shi F, Lin W, Gilmore JH, and Shen D, “Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age”, Cerebral cortex, vol. 23, no. 11, pp. 2724–2733, 2012. [PubMed: 22923087]

[15]. Li G, Nie J, Wang L, Shi F, Gilmore JH, Lin W, and Shen D, “Measuring the dynamic longitudinal cortex development in infants by reconstruction of temporally consistent cortical surfaces”, Neuroimage, vol. 90, pp. 266–279, 2014. [PubMed: 24374075]

[16]. Meng Y, Li G, Gao Y, Lin W, and Shen D, “Learning-based subject-specific estimation of dynamic maps of cortical morphology at missing time points in longitudinal infant studies”, Human Brain Mapping, vol. 37, no. 11, pp. 4129–4147, 2016. [PubMed: 27380969]

[17]. Yuan L, Wang Y, Thompson PM, Narayan VA, Ye J, and Initiative ADN, “Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data”, NeuroImage, vol. 61, no. 3, pp. 622–632, 2012. [PubMed: 22498655]

[18]. Li S-Y, Jiang Y, and Zhou Z-H, “Partial multi-view clustering”, in AAAI, 2014, pp. 1968–1974.

[19]. Cai J-F, Candes EJ, and Shen Z, “A singular value thresholding algorithm for matrix completion”, SIAM J Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[20]. Candes E and Recht B, “Exact matrix completion via convex optimization”, Communications of the ACM, vol. 55, no. 6, pp. 111–119, 2012.

[21]. Zheng S, Cai X, Ding C, Nie F, and Huang H, “A closed form solution to multi-view low-rank regression”, pp. 1973–1979, 2015.

[22]. Sharma A, Kumar A, Daume H, and Jacobs DW, “Generalized multiview analysis: A discriminative latent space” in CVPR, 2012, pp. 2160–2167.

[23]. Rekik I, Li G, Yap P-T, Chen G, Lin W, and Shen D, “Joint prediction of longitudinal development of cortical surfaces and white matter fibers from neonatal mri”, NeuroImage, vol. 152, pp. 411–424, 2017. [PubMed: 28284800]

[24]. Blum A and Mitchell T, “Combining labeled and unlabeled data with co-training”, in COLT. ACM, 1998, pp. 92–100.

[25]. Chaudhuri K, Kakade SM, Livescu K, and Sridharan K, “Multiview clustering via canonical correlation analysis”, in ICML, 2009, pp. 129–136.

[26]. Kumar A and Daume H, “A co-training approach for multi-view spectral clustering”, in ICML, 2011, pp. 393–400.

[27]. Wang W and Zhou Z-H, “Analyzing co-training style algorithms”, in ECML. Springer, 2007, pp. 454–465.

[28]. Zien A and Ong CS, “Multiclass multiple kernel learning”, in ICML, 2007, pp. 1191–1198.

[29]. Kakade SM and Foster DP, “Multi-view regression via canonical correlation analysis”, in COLT, 2007, pp. 82–96.

[30]. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Initiative ADN et al., “Random forest-based similarity measures for multi-modal classification of alzheimer’s disease”, NeuroImage, vol. 65, pp. 167–175, 2013. [PubMed: 23041336]

[31]. Singanamalli A, Wang H, Lee G, Shih N, Rosen M, Master S, Tomaszewski J, Feldman M, and Madabhushi A, “Supervised multiview canonical correlation analysis: Fused multimodal prediction of disease diagnosis and prognosis”, in Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 9038 International Society for Optics and Photonics, 2014, pp. 903–805.

[32]. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, Feng D, Fulham MJ et al., “Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer’s disease”, IEEE Transactions on Biomedical Engineering, vol. 62, no. 4, pp. 1132–1140, 2015. [PubMed: 25423647]

[33]. Liu J, Ji S, and Ye J, “Multi-task feature learning via efficient l 2, 1-norm minimization”, in UAI, 2009, pp. 339–348.

[34]. Chen J, Zhou J, and Ye J, “Integrating low-rank and group-sparse structures for robust multi-task learning”, in ACM KDD, 2011, pp. 42–50.

[35]. Adeli E, Meng Y, Li G, Lin W, and Shen D, “Multi-task prediction of infant cognitive scores from longitudinal incomplete neuroimaging data”, NeuroImage, 2018.

[36]. Ando RK and Zhang T, “A framework for learning predictive structures from multiple tasks and unlabeled data”, Journal of Machine Learning Research, vol. 6, no. Nov, pp. 1817–1853, 2005.

[37]. Bakker B and Heskes T, “Task clustering and gating for bayesian multitask learning”, Journal of Machine Learning Research, vol. 4, no. May, pp. 83–99, 2003.

[38]. Akshoomoff N, “Use of the mullen scales of early learning for the assessment of young children with autism spectrum disorders”, Child Neuropsychology, vol. 12, no. 4–5, pp. 269–277, 2006. [PubMed: 16911972]

[39]. Li G, Wang L, Shi F, Gilmore JH, Lin W, and Shen D, “Construction of 4d high-definition cortical surface atlases of infants: Methods and applications”, Medical image analysis, vol. 25, no. 1, pp. 22–36, 2015. [PubMed: 25980388]

[40]. Fischl B, “Freesurfer”, Neuroimage, vol. 62, no. 2, pp. 774–781, 2012. [PubMed: 22248573]

[41]. Fornito A, Wood SJ, Whittle S, Fuller J, Adamson C, Saling MM, Velakoulis D, Pantelis C, and Yiicel M, “Variability of the paracingulate sulcus and morphometry of the medial frontal cortex: associations with cortical thickness, surface area, volume, and sulcal depth”. Human brain mapping, vol. 29, no. 2, pp. 222–236, 2008. [PubMed: 17497626]

[42]. Rimol LM, Nesvåg R, Hagler DJ, Bergmann Ø, Fennema-Notestine C, Hartberg CB, Haukvik UK, Lange E, Pung CJ, Server A et al., “Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder”, Biological psychiatry, vol. 71, no. 6, pp. 552–560, 2012. [PubMed: 22281121]

[43]. Li G, Wang L, Shi F, Lin W, and Shen D, “Constructing 4d infant cortical surface atlases based on dynamic developmental trajectories of the cortex”, in MICCAI, 2014, pp. 89–96.

[44]. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, Elison JT, Swanson MR, Zhu H, Botteron KN et al., “Early brain development in infants at high risk for autism spectrum disorder”, Nature, vol. 542, no. 7641, pp. 348–351, 2017. [PubMed: 28202961]

[45]. Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT et al., “An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest”, NeuroImage, vol. 31, no. 3, pp. 968–980, 2006. [PubMed: 16530430]

[46]. Adeli-Mosabbeb E, Thung K-H, An L, Shi F, and Shen D, “Robust feature-sample linear discriminant analysis for brain disorders diagnosis”, in NIPS, 2015, pp. 658–666.

[47]. Wang M, Hua X-S, Yuan X, Song Y, and Dai L-R, “Optimizing multi-graph learning: towards a unified video annotation scheme” in ACM MM, 2007, pp. 862–871.

[48]. Gabay D and Mercier B, A dual algorithm for the solution of non linear variational problems via finite element approximation. Institut de recherche d’informatique et d’automatique, 1975.

[49]. Lin Z, Liu R, and Su Z, “Linearized alternating direction method with adaptive penalty for low-rank representation”, in NIPS, 2011, pp. 612–620.

[50]. Bartels RH and Stewart G, “Solution of the matrix equation AX+ XB= C”, Communications of the ACM, vol. 15, no. 9, pp. 820–826, 1972.

[51]. Hu H, Lin Z, Feng J, and Zhou J, “Smooth representation clustering”, in CVPR, 2014, pp. 3834–3841.

[52]. Wen Z and Yin W, “A feasible method for optimization with orthogonality constraints”, Mathematical Programming, vol. 142, no. 1–2, pp. 397–434, 2013.

[53]. Liu G, Lin Z, Yan S, Sun J, Yu Y, and Ma Y, “Robust recovery of subspace structures by low-rank representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 171–184, 2013. [PubMed: 22487984]

[54]. Boyd S, Parikh N, Chu E, Peleato B, and Eckstein J, “Distributed optimization and statistical learning via the alternating direction method of multipliers”, Found. Trends. Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.

[55]. Argyriou A, Evgeniou T, and Pontil M, “Multi-task feature learning”, NIPS, vol. 19, pp. 41–48, 2007.

[56]. Ji S and Ye J, “An accelerated gradient method for trace norm minimization”, in ICML, 2009, pp. 457–464.

[57]. Hill J, Dierker D, Neil J, Inder T, Knutsen A, Harwell J, Coalson T, and Van Essen D, “A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants”, Journal of Neuroscience, vol. 30, no. 6, pp. 2268–2276, 2010. [PubMed: 20147553]

[58]. Li G, Wang L, Shi F, Lyall AE, Ahn M, Peng Z, Zhu H, Lin W, Gilmore JH, and Shen D, “Cortical thickness and surface area in neonates at high risk for schizophrenia”, Brain Structure and Function, vol. 221, no. 1, pp. 447–461, 2016. [PubMed: 25362539]

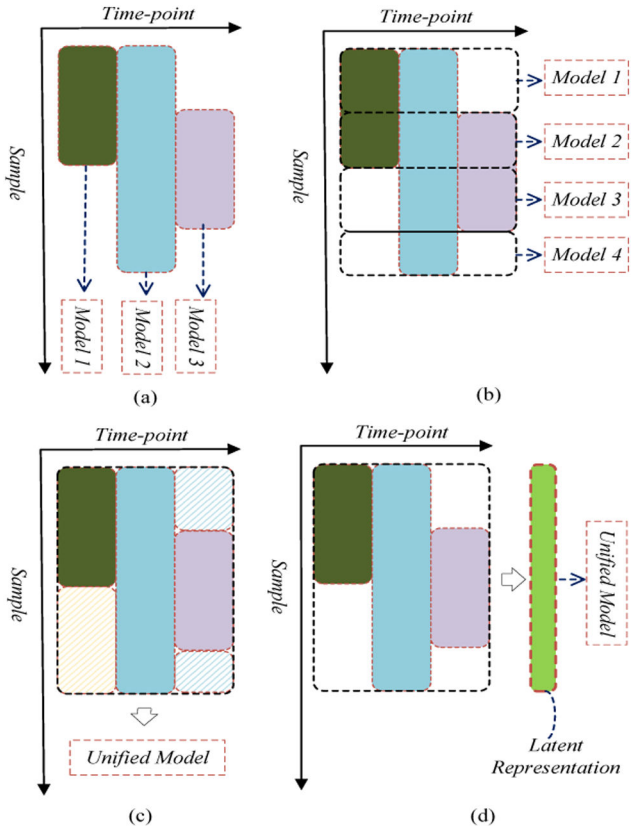


Fig. 1:

Strategies for handling the missing data: (a) learning one model for each data-point, (b) learning one model for each combination of multiple data-points, (c) learning a unified model based on the completed data, and (d) learning a unified model based on the latent representation for all data.

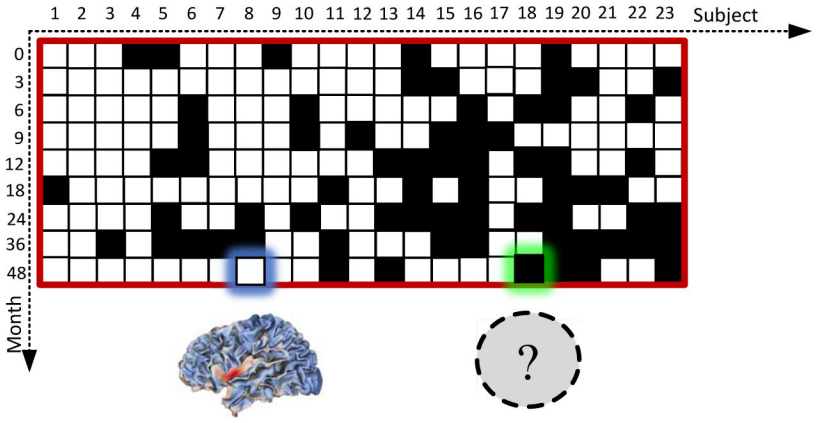


Fig. 2:

Illustration of our dataset. White blocks indicate the availability of imaging data for subjects at specific time-points, while black blocks indicate the missing data. For instance, the white block highlighted in blue represents the availability of MRI data for the 8^{th} subject at its 48^{th} month (e.g., the brain attribute maps underneath it), while the black block highlighted in green indicates a missing MRI scan (the question mark). We have 23 subjects with measured development scores, indicated by the red rectangle. The missing rates from the 1^{st} time-point to last time-point are 21.7%, 21.7%, 26.1%, 26.1%, 39.1%, 30.4%, 47.8%, 56.5%, and 26.1%, respectively.

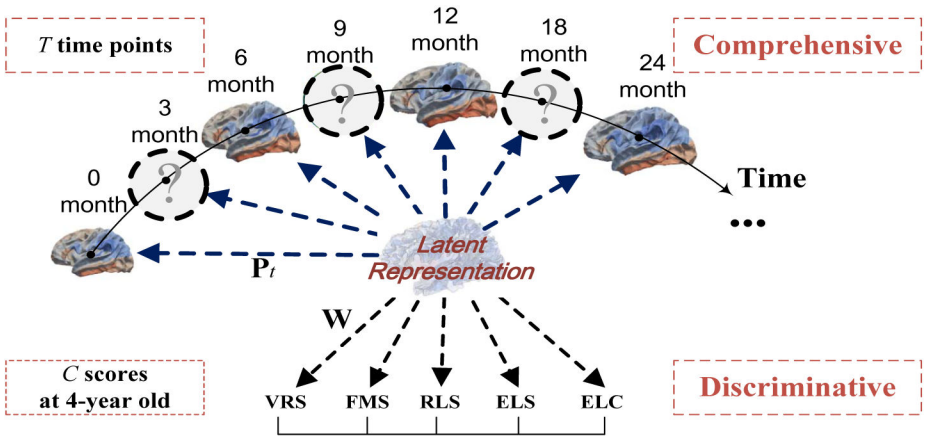
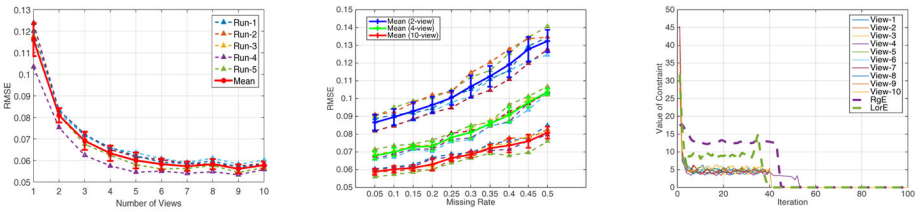


Fig. 3:

Illustration of Latent Partial Multi-view Representation Learning. Our model uncovers the comprehensive and discriminative latent representation (termed as latent atlas from medical image field) jointly from incomplete observations, based on which the multi-task (C scores) multi-view (T time-points) prediction model is learned.



(a) Effect of the number of view. (b) Effect of missing rate. (c) Convergence experiment.

Fig. 4:

Experiments on synthetic data. The solid lines in (a) and (b) are plotted by averaging five different runs shown as the dash lines.

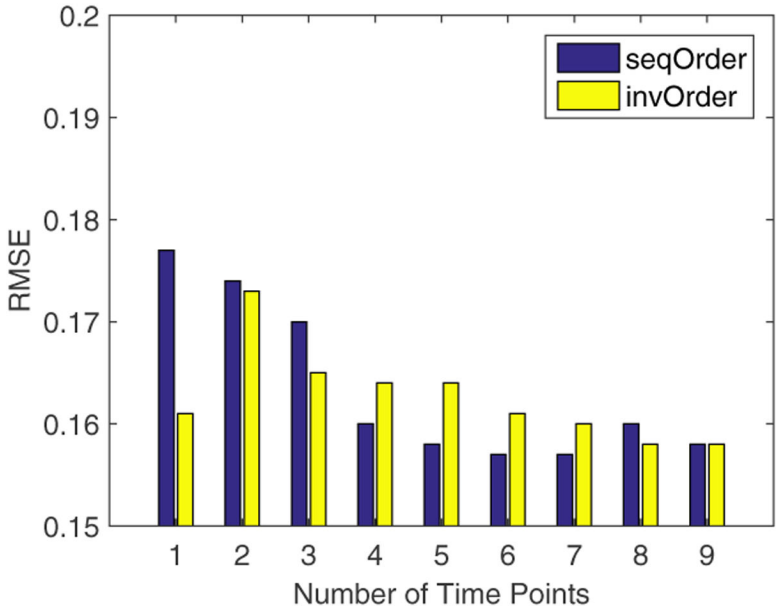


Fig. 5:
The performance trend with adding data at more time-points from two directions.

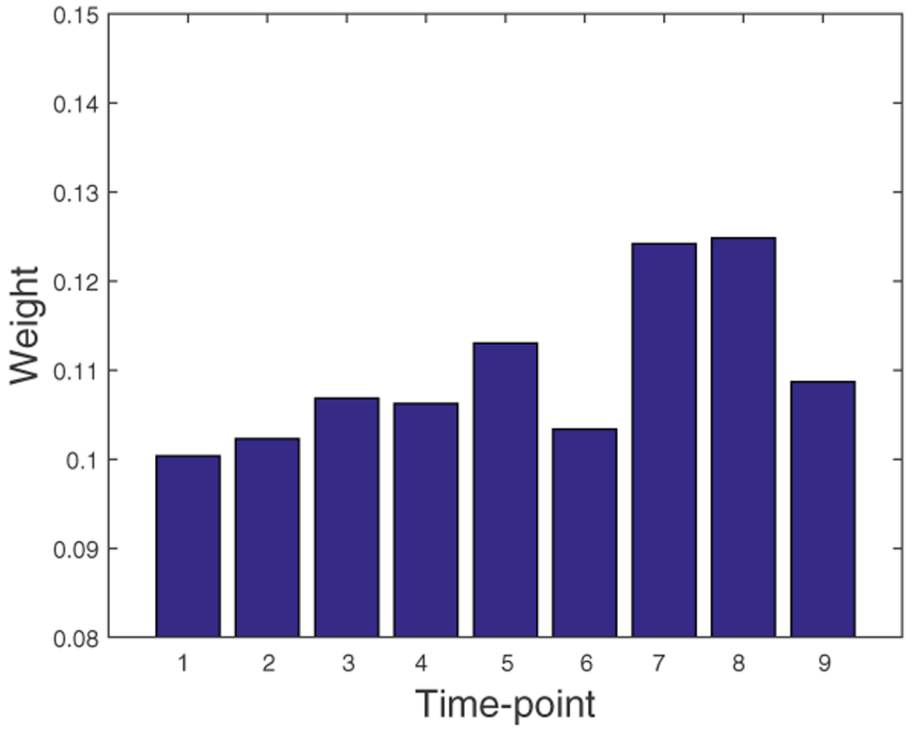
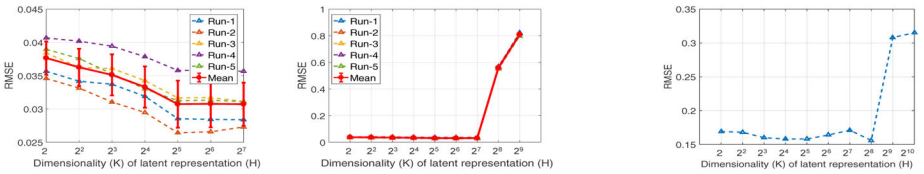


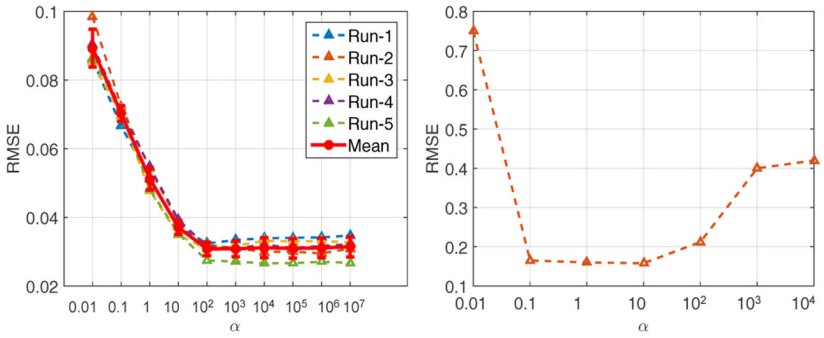
Fig. 6:
The estimated weights $\boldsymbol{\omega}$ for multiple time-points.



(a) Synthetic data. (b) Real-world data.

Fig. 7:

Impact of the dimensionality for the latent representation **H** on the prediction performance.



(a) Synthetic data. (b) Real-world data.

Fig. 8:
Impact of the hyperparameter α on the prediction performance.

TABLE I:

Results of using different numbers of time-points (in terms of RMSE).

Time-Points	VRS	FMS	RLS	ELS	ELC	Average
0 Month (1 time-point)	0.167	0.208	0.160	0.187	0.161	0.177
01–03 Month (2 time-points)	0.164	0.206	0.159	0.183	0.159	0.174
01–06 Month (3 time-points)	0.162	0.205	0.151	0.176	0.155	0.170
01–09 Month (4 time-points)	0.162	0.189	0.143	0.163	0.143	0.160
01–12 Month (5 time-points)	0.158	0.190	0.137	0.164	0.138	0.158
01–18 Month (6 time-points)	0.157	0.190	0.137	0.164	0.136	0.157
01–24 Month (7 time-points)	0.159	0.191	0.138	0.162	0.137	0.157
01–36 Month (8 time-points)	0.162	0.194	0.140	0.165	0.141	0.160
01–48 Month (9 time-points)	0.162	0.189	0.139	0.165	0.138	0.158

TABLE II:

Performance comparison (in terms of RMSE). For the compared methods, the first row depicts the results with the missing values substituted with zeros, and the second row depicts the results with the missing values substituted with the averaged values of the observed ones.

Method	VRS	FMS	RLS	ELS	ELC	Average
NN	0.200	0.220	0.259	0.291	0.209	0.236
	0.219	0.259	0.165	0.196	0.182	0.204
MtJFS	0.284	0.296	0.279	0.278	0.286	0.285
	0.276	0.273	0.189	0.214	**0.134**	0.217
RMTL	0.313	0.321	0.242	0.229	0.273	0.276
	0.146	0.200	0.178	0.188	0.137	0.170
TrMTL	0.349	0.373	0.280	0.256	0.319	0.315
	0.279	0.276	0.192	0.217	0.136	0.220
Proposed	0.162	**0.189**	**0.139**	**0.165**	0.138	**0.158**

IEEE Trans Med Imaging. Author manuscript; available in PMC 2019 April 05.