Joint Sparse and Low-Rank Regularized Multi-Task Multi-Linear Regression for Prediction of Infant Brain Development with Incomplete Data

Ehsan Adeli, Yu Meng, Gang Li, Weili Lin, and Dinggang Shen(✉)

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA
dgshen@med.unc.edu

Abstract. Studies involving dynamic infant brain development has received increasing attention in the past few years. For such studies, a complete longitudinal dataset is often required to precisely chart the early brain developmental trajectories. Whereas, in practice, we often face missing data at different time point(s) for different subjects. In this paper, we propose a new method for prediction of infant brain development scores at future time points based on longitudinal imaging measures at early time points with possible missing data. We treat this as a multi-dimensional regression problem, for predicting multiple brain development scores (*multi-task*) from multiple previous time points (*multi-linear*). To solve this problem, we propose an objective function with a joint ℓ_1 and low-rank regularization on the mapping weight tensor, to enforce feature selection, while preserving the structural information from multiple dimensions. Also, based on the bag-of-words model, we propose to extract features from longitudinal imaging data. The experimental results reveal that we can effectively predict the brain development scores assessed at the age of four years, using the imaging data as early as two years of age.

1 Introduction

The early postnatal period witnesses dynamic brain development, which has not been sufficiently explored. Such assessments can be essential steps in identifying and treating the early neurodevelopmental disorders, as well as understanding how brain develops. Longitudinal neuroimaging analysis of the early postnatal brain development, especially for scoring of an individual's brain development, is a very interesting and important problem. However, this is quite challenging, due to rapid brain changes during this stage. In this paper, we present a novel method to extract informative brain MRI features and propose a multi-task multi-linear regression model for predicting brain development scores in future time points.

To conduct this study, we use longitudinal MRI data from healthy infant subjects, with each subject scanned at every 3 months in the first year, every 6

Supported in part by NIH grants MH100217, MH108914, MH107815 and MH110274.

© Springer International Publishing AG 2017
M. Descoteaux et al. (Eds.): MICCAI 2017, Part I, LNCS 10433, pp. 40–48, 2017.
DOI: 10.1007/978-3-319-66182-7_5

months in the second year, and every 12 months from the third year. At the age of 48 months, five brain development scores are assessed for each subject, which characterize how an individual's brain has developed. We seek to predict these five scores purely from the neuroimaging data in multiple previous time points. However, we face quite a number of challenges. (1) In certain time points, there are missing neuroimaging data for some subjects, due to subject's no show-up or dropout. This poses a major challenge for the task of prediction. (2) We have multiple brain development scores to predict. As these scores are acquired from same subjects, they are essentially inter-related and can benefit each other for the prediction tasks. Hence, we have a *multi-task* problem at hand. (3) Each subject is scanned at multiple time points in the first 48 months; therefore, we need to build multiple models (*multi-linear*), which are also inter-related. (4) The neuroimaging data at each time point are extremely high-dimensional, and therefore we need an intuitive feature extraction and dimensionality reduction technique to avoid the so-called Small-Sample-Size (SSS) problem, in which the number of subjects is way much less than the number of features. (5) Often all features acquired from neuroimaging data are not necessarily relevant and useful for the prediction tasks. Specially, the features from the very earlier time points can be less effective in predicting the future scores. Hence, we need to enforce selecting the most important features for a reliable and accurate prediction model.

Accordingly, we design a novel framework to address all the above challenges. Specifically, first, we propose a model based on Bag-of-Words (BoW) [11] to extract meaningful low-dimensional features from the high-dimensional neuroimaging data, denoted as brain fingerprints. Then, we propose a novel Multi-Task Multi-Linear Regression (MTMLR) framework to take advantage of the existing inherent structure and inter-relation between the tasks and between the time points, by using low-rank tensor regularization as a natural underpinning for preserving this underlying structural information. We also include a ℓ_1 regularization on the same tensor to enforce selection of the most relevant features. Furthermore, our MTMLR formulations can deal with incomplete data by neglecting the time points with no data for any specific subject. The obtained prediction results indicate that our framework can accurately predict the brain development scores as early as at 24 months of age.

2 Materials and Feature Extraction

To conduct this study, we use the longitudinal MRI data from 24 healthy infant subjects. For each subject, T1-, T2-, and diffusion-weighted MR images are acquired at nine different time points (*i.e.*, 0, 3, 6, 9, 12, 18, 24, 36 and 48 months), and five brain development scores are acquired for each subject at 48 months, including Visual Reception Scale (VRS), Fine Motor Scale (FMS), Receptive Language Scale (RLS), Expressive Language Scale (ELS), and Early Learning Composite (ELC). Note that the fifth score (*i.e.*, ELC) can be interpreted as the composite of the other four. As discussed earlier, we have missing imaging data for some of the subjects at certain time points. Figure 1 illustrates the formation of our dataset, in which black blocks indicate missing data.

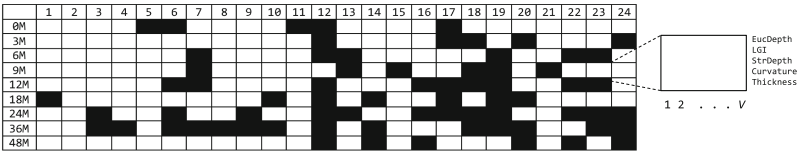

Fig. 1. Longitudinal infant dataset, containing 24 subjects (columns), each scanned at 9 different time points (rows). Each block contains the cortical morphological attributes of all vertices on the cortical surface for a specific subject at a specific time point. Black blocks show the missing data at the respective time points.

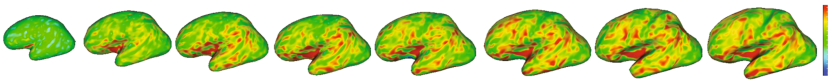

Fig. 2. Longitudinal cortical thickness maps on the inflated cortical surface for a representative subject.

All images are processed using an infant-specific computational pipeline for cortical surface reconstruction and registration, similar to [8]. Then, five attributes are extracted for each vertex on the cortical surfaces. These attributes are: the sulcal depth as Euclidian distance from the surface hull (EucDepth) [6], local gyrification index (LGI) [6], curve sulcal depth along the streamlines (StrDepth) [6], mean curvature [7], and cortical thickness [7] (Fig. 2).

The attributes for all vertices on the cortical surface of each subject lead to an extremely high-dimensional set of data. To slash the dimensionality of the feature vector, we consider each vertex as a 5D vector, containing its 5 attributes. Using a model similar to BoW [11], we group the similar 5D vectors to create a high-level profile for each cortical surface. Specifically, we create a pool from these vectors from all subjects in the dataset, and cluster them into $d = 100$ different clusters, based on *weighted* Euclidean distance. Then, a d-dimensional vector can be simply used to represent each subject, corresponding to the frequencies of its vertices lying in each of these d clusters. But it is important to note that not all of the 5 attributes on the surface are equally important. That is why we employ a *weighted* Euclidean distance to conduct the clustering. To calculate the weights for each attribute, corresponding to the relevance of that attribute with the brain development score, we employ a paired t-test between the attribute values and the score to be predicted (*e.g.*, ELC). The percentages of the vertices having a p-value of less than 0.05 are calculated for each attribute. These percentage values show the importance of the attributes. We normalize these values to have a sum equal to 1, and use them to weight the attributes in the distance function.

After the above procedure, we have a d-dimensional feature vector for each time point of each subject. This vector encodes the structural characteristics of the cortical surface, and is denoted as the fingerprint of the subject's brain. It intuitively encodes the formation of attributes on the cortical morphology, and hence can be used to predict the brain development scores (see the next Section).

3 Joint Sparse and Low-Rank Regularized MTMLR

With the problem description discussed earlier, we have N subjects, scanned at T different time points, with S different brain development scores assessed from each subject. We extract d different features from the subjects at each time point (Sect. 2). Figure 3 illustrates different settings for a regression problem, in which the loss function $L(.)$, and the regularization of the mapping coefficients $\mathcal{R}(.)$ are defined on vectors, matrices or tensors depending on the problem nature (See the notations[1]). We seek to find the best mapping for the prediction of the scores, knowing that joint learning of multiple relevant tasks can outperform learning each task separately [3,10].

As can be seen in Fig. 3(c), a MTMLR task is defined by aggregating the predictions from each time point t from the t^{th} fiber of the data tensor, $\mathbf{X}^t$, using the respective mapping coefficients, $\mathbf{W}^t$. All these mapping coefficients $\mathbf{W}^t, \forall 1 \leq t \leq T$, are stacked together to form a tensor of order three, $\mathcal{W}$. As it is apparent, tensor $\mathcal{W}$ has ample intertwined dependencies along its different dimensions, since each of its fibers hold mapping coefficients from different time points of same subjects predicting the same set of scores. Hence, it is a quite feasible assumption that this tensor should be rank deficient. But the rank function is not a well-defined function and is often approximated by the nuclear norm. As a result, to include this in the optimization objective, we can define the regularization term as $\mathcal{R}(\mathcal{W}) = \lambda\|\mathcal{W}\|_*$. However, all features from all time points might not be beneficial in building the prediction model, we propose to include a joint sparse and low-rank regularization. As discussed in the literature [5,12], a mixture of ℓ_1 and nuclear norms often makes the model less sensitive to the feature size and variations. Hence, the regularization term would be:

$$\mathcal{R}(\mathcal{W}) = \lambda_1\|\mathcal{W}\|_* + \lambda_2\|\mathcal{W}\|_1. \tag{1}$$

The loss function, evaluating the level of misprediction, would require to aggregate over all combinations of scores and subjects across different time

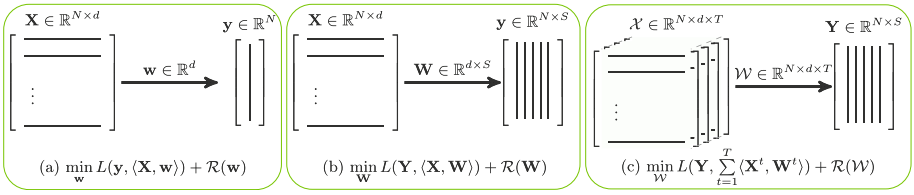

Fig. 3. Illustration of different regression models: (a) Linear Regression, (b) Multi-Task Regression, (c) Multi-Task Multi-Linear Regression.

[1] Bold capital letters denote matrices (*e.g.*, $\mathbf{A}$), small bold letters are vectors (*e.g.*, $\mathbf{a}$), and non-bold letters denote scalars (*e.g.*, a). Tensors are represented by calligraphic typeface letters (*e.g.*, $\mathcal{W}$). $\|.\|_*$ and $\|.\|_1$ designate the nuclear and ℓ_1 norms, respectively, while $\langle ., . \rangle$ denotes the inner product. $\mathbf{W}_{(n)}$ denotes the mode-n matricization of the tensor $\mathcal{W}$, *i.e.*, unfolding $\mathcal{W}$ from its n^{th} dimension to form a matrix.

points. As stated before, in our longitudinal study, we have missing data in several time points. To deal with this incomplete data, we define a mask matrix, $\mathbf{A}$, analogous to the block in Fig. 1. Each element of this matrix (a_i^t) would indicate if there exists the neuroimaging data for subject i at time point t. As a result we have:

$$L(\mathbf{Y}, \mathcal{X}, \mathcal{W}) = \sum_{s=1}^{S} \sum_{i=1}^{N} \sum_{t=1}^{T} a_i^t . \left(y_i^s - \langle \mathbf{x}_i^t, \mathbf{w}^{s,t} \rangle\right)^2. \tag{2}$$

Optimization: In order to optimize the objective function with the loss function (2) and regularization (1), we use the Alternating Direction Method of Multipliers (ADMM) [1]. To do this, we utilize a convex surrogate for the rank of a tensor, which is approximated using the nuclear norm. Similar to previous works [10,12], a good convex proxy for that is defined as the average of the nuclear norms of each matricization of $\mathcal{W}$:

$$\|\mathcal{W}\|_* = \frac{1}{O} \sum_{n=1}^{O} \|\mathbf{W}_{(n)}\|_*, \tag{3}$$

where O is the tensor order ($O = 3$ in our case). This reduces the problem to minimizing the matrix nuclear norms (sum of eigenvalues of the matrix), which is widely studied in the literature [9,10]. Therefore, the objective function would become:

$$\min_{\mathcal{W}} L(\mathbf{Y}, \mathcal{X}, \mathcal{W}) + \frac{\lambda_1}{O} \sum_{n=1}^{O} \|\mathbf{W}_{(n)}\|_* + \lambda_2 \|\mathcal{W}\|_1. \tag{4}$$

To optimize the above objective, we require a set of auxiliary variables, leading to:

$$\begin{aligned} \min_{\mathcal{W},\mathcal{U},\{\mathbf{V}_n\}_{n=1}^{O}} \quad & L(\mathbf{Y}, \mathcal{X}, \mathcal{U}) + \frac{\lambda_1}{O} \sum_{n=1}^{O} \|\mathbf{V}_n\|_* + \lambda_2 \|\mathcal{W}\|_1 \\ \text{s.t.} \quad & \mathcal{U} = \mathcal{W} \wedge \mathbf{V}_n = \mathbf{W}_{(n)}, \forall n \in \{1, \ldots, O\}. \end{aligned} \tag{5}$$

Using ADMM [1], we write the augmented Lagrangian function. Then, we iteratively optimize for each of the optimization variables, $\mathcal{W}, \mathcal{U}, \{\mathbf{V}_n\}_{n=1}^{O}$, while fixing the others. Solving for $\mathcal{U}$, we would have a linear-quadratic function, which is convex and can be optimized efficiently. Solving for $\mathcal{W}$ would require minimization of the ℓ_1 norm, which can be done using the soft thresholding operator as a proximal operator for ℓ_1 norm [1]. Solving for each of the $\{\mathbf{V}_n\}_{n=1}^{O}$ variables requires separate minimization of the matrix nuclear norms. This can also be done using the Singular Value Thresholding (SVT) algorithm [2].

Lemma 1. *Minimizing the optimization objective in Eq. (5) using ADMM would converge to the optimal value.*

Proof. The objective in (5) is convex, since all its associated terms are convex functions. It is previously proven [1,4] that the alternative optimization in

ADMM converges to the optimal value, under this condition, if there are two variables associated with the alternative optimization. Considering our objective function, one can figure out that $\mathcal{W}$ is the only variable that is contingent on the others (through the constraints). Hence, the other variables are optimized independent from each other at each given iteration. So, if we hypothetically stack all matrices $\{\mathbf{V}_n\}_{n=1}^{O}$ into a tensor $\mathcal{V}$, then concatenate this tensor with $\mathcal{U}$ and name it $\mathcal{Z} = [\mathcal{U}, \mathcal{V}]$, the optimization procedure using ADMM is analogues to an alternating optimization between two variables $\mathcal{Z}$ and $\mathcal{W}$. Accordingly, ADMM would converge to the optimal solution for the objective in Eq. (5). □

4 Experiments

First, to evaluate the attributes that we have used to describe the cortical surfaces, we examine their weights obtained in Sect. 2. Figure 4 shows the percentage of the vertices with $p < 0.05$ for predicting ELC at different time points. It is obvious that, at earlier ages, the curvature appears to be more relevant, while, at the later time points, the cortical thickness shows quite important. As discussed earlier, we used these weights (normalized to sum to 1) to extract our BoW features for each subject at each time point, denoted as brain fingerprint.

To conduct the prediction experiments, we performed 10-fold cross-validation and calculated the root mean square error (RMSE) and the absolute correlation coefficient (R) between the predicted and the actual values for all five scores. The obtained results of using the neuroimaging data up to a specific time point are listed in Table 1, with the tuning hyerparameters fixed, as $\lambda_1 = \lambda_2 = 1/\sqrt{\min(N,d,T)}$. Note that the scores are all normalized with the min and max of possible values for each score separately, such that all scores range in $[0, 1]$. The mean $\pm$ standard deviation of the scores after normalization are $0.54 \pm 0.26, 0.60 \pm 0.29, 0.52 \pm 0.25, 0.58 \pm 0.21$ and 0.62 ± 0.27, respectively. As

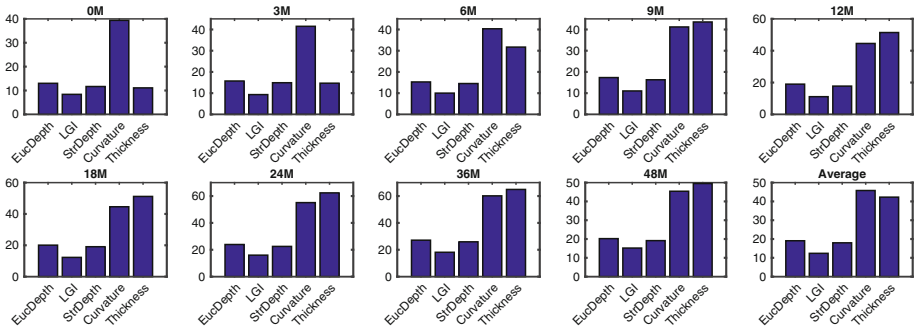

Fig. 4. Percentage of the vertices that are not rejected at the 5% significance level for predicting the Early Learning Composite (ELC) score from each of the five features, at different time points. The last one in the second row shows the average value across all time points for the features.

Table 1. The RMSE and correlation coefficient, R, performance metrics for the prediction results, through 10-fold cross-validation.

		0–3M	0–6M	0–9M	0–12M	0–18M	0–24M	0–36M	0–48M
VRS	RMSE	0.21 ± 0.16	0.20 ± 0.11	0.20 ± 0.09	0.18 ± 0.12	0.18 ± 0.12	0.18 ± 0.10	0.17 ± 0.12	0.17 ± 0.10
	R	0.60	0.68	0.66	0.67	0.69	0.71	0.72	0.72
FMS	RMSE	0.20 ± 0.15	0.19 ± 0.17	0.19 ± 0.13	0.21 ± 0.11	0.18 ± 0.17	0.18 ± 0.16	0.18 ± 0.12	0.18 ± 0.11
	R	0.58	0.61	0.66	0.66	0.69	0.70	0.70	0.71
RLS	RMSE	0.22 ± 0.13	0.21 ± 0.12	0.21 ± 0.15	0.21 ± 0.17	0.21 ± 0.13	0.20 ± 0.15	0.20 ± 0.12	0.20 ± 0.09
	R	0.59	0.60	0.62	0.65	0.65	0.66	0.66	0.67
ELS	RMSE	0.20 ± 0.13	0.19 ± 0.10	0.20 ± 0.09	0.19 ± 0.12	0.18 ± 0.12	0.17 ± 0.13	0.18 ± 0.10	0.17 ± 0.12
	R	0.61	0.65	0.67	0.68	0.68	0.70	0.71	0.71
ELC	RMSE	0.21 ± 0.11	0.20 ± 0.11	0.18 ± 0.10	0.17 ± 0.09	0.19 ± 0.10	0.18 ± 0.10	0.19 ± 0.12	0.17 ± 0.09
	R	0.63	0.66	0.68	0.70	0.72	0.73	0.73	0.74

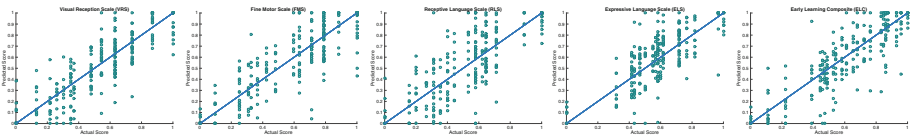

Fig. 5. Scatter plots of the actual (horizontal axis) and the predicted (vertical axis) values of the five scores (From left to right: VRS, FMS, RLS, ELS and ELC), at the 24M time point, for 10 different runs.

can be seen in the table, after the age of 24 months, the results are consistently predicted with a relatively good approximation (for both the RMSE and R). One of the main reasons why the results have not been improved much after that might be due to the fact that we have too much missing data in the later time points. Additionally, the scatter plots for 10 different runs of 10-fold cross-validation for predicting the scores at the age of 24M are depicted in Fig. 5. This Figure demonstrates that, in general, the scores are predicted reasonably good.

To compare the proposed method with other baseline techniques on our application, we adopt several methods with the same 10-fold cross-validation experimental settings on the 0–24M experiment (as in 8^{th} column of Table 1). The methods in comparison are the same formulation as ours but only with the nuclear norm regularization (denoted as MTMLR$_*$), only with the ℓ_1 norm regularization (denoted as MTMLR$_1$), concatenating all the features from all time points and conducting a sparse feature selection followed by only a multi-task regression (denoted as SFS+MTR), support vector regression (denoted as SFS+SVR), or simple ridge regression (SFS+RR). The R measure results, showing the correlation of the predicted and the original values, are provided in Table 2. As it is apparent from the results, the proposed method yields the best results for almost all of the five

Table 2. Comparison results from different methods with the R measure.

	VRS	FMS	RLS	ELS	ELC
Proposed	**0.71**	**0.70**	0.66	**0.70**	**0.73**
MTMLR$_*$	0.65	0.62	**0.68**	0.61	0.66
MTMLR$_1$	0.48	0.56	0.39	0.51	0.53
SFS+MTR	0.39	0.43	0.35	0.40	0.46
SFS+SVR	0.31	0.35	0.23	0.26	0.31
SFS+RR	0.19	0.25	0.25	0.21	0.28

brain development scores. This is attributed to the fact that, using our joint regularization technique, we can preserve the underlying structural information hidden in the multi-dimensional data, while enforcing feature selection to use the most beneficial features. The three latter methods concatenate the features from different time points and hence they are losing a great deal of structural information. On the other hand, since the dimensionality of the feature vector will become large, the SFS technique might not necessarily capture the best features. The last two methods further lose the dependency between the tasks, as they predict each task separately, and hence achieve lower prediction performances.

5 Conclusions

In this paper, we proposed a multi-task multi-linear regression model with a joint sparse and nuclear norm tensor regularization for predicting postnatal brain development scores from multiple previous time points. Our proposed tensor regularization helps better leveraging structure information in multi-dimensional set of data, while enforcing feature selection to ensure that most beneficial features are used in building the model. We also discussed the convergence properties of the proposed optimization algorithm. Furthermore, we presented a method to extract meaningful low-dimensional features from the cortical surfaces of infant brains, denoted as brain fingerprints. As shown by the results, the combination of our brain fingerprinting and regression model can lead to reasonable predictions, while outperforming all baseline models.

References

1. Boyd, S., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
2. Cai, J.F., Candès, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
3. Caruana, R.: Multitask learning. In: Thrun, S., Pratt, L. (eds.) Learning to Learn, pp. 95–133. Springer, New York (1998)
4. Eckstein, J., Yao, W.: Understanding the convergence of the alternating direction method of multipliers: theoretical and computational perspectives. Pac. J. Optim. **11**(4), 619–644 (2015)
5. Gaiffas, S., Lecué, G.: Sharp oracle inequalities for high-dimensional matrix prediction. IEEE Trans. Inf. Theor. **57**(10), 6942–6957 (2011)
6. Li, G., et al.: Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age. J. Neurosci. **34**(12), 4228–4238 (2014)
7. Li, G., et al.: Construction of 4D high-definition cortical surface atlases of infants: methods and applications. Med. Image Anal. **25**(1), 22–36 (2015)
8. Meng, Y., et al.: Learning-based subject-specific estimation of dynamic maps of cortical morphology at missing time points in longitudinal infant studies. Hum. Brain Mapp. **37**(11), 4129–4147 (2016)
9. Mosabbeb, E.A., et al.: Robust feature-sample linear discriminant analysis for brain disorders diagnosis. In: NIPS, pp. 658–666 (2015)

10. Romera-Paredes, B., Aung, H., Bianchi-Berthouze, N., Pontil, M.: Multilinear multitask learning. In: ICML, pp. 1444–1452 (2013)
11. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE TPAMI **31**(4), 591–606 (2009)
12. Song, X., Lu, H.: Multilinear regression for embedded feature selection with application to FMRI analysis. In: AAAI (2016)