Integrating Neural Networks and Dictionary Learning for Multidimensional Clinical Characterizations from Functional Connectomics Data

Niharika Shimona D'Souza[1(✉)], Mary Beth Nebel[2,3], Nicholas Wymbs[2,3], Stewart Mostofsky[2,3,4], and Archana Venkataraman[1]

[1] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA
Shimona.Niharika.Dsouza@jhu.edu
[2] Center for Neurodevelopmental and Imaging Research, Kennedy Krieger Institute, Baltimore, USA
[3] Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA
[4] Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, USA

Abstract. We propose a unified optimization framework that combines neural networks with dictionary learning to model complex interactions between resting state functional MRI and behavioral data. The dictionary learning objective decomposes patient correlation matrices into a collection of shared basis networks and subject-specific loadings. These subject-specific features are simultaneously input into a neural network that predicts multidimensional clinical information. Our novel optimization framework combines the gradient information from the neural network with that of a conventional matrix factorization objective. This procedure collectively estimates the basis networks, subject loadings, and neural network weights most informative of clinical severity. We evaluate our combined model on a multi-score prediction task using 52 patients diagnosed with Autism Spectrum Disorder (ASD). Our integrated framework outperforms state-of-the-art methods in a ten-fold cross validated setting to predict three different measures of clinical severity.

1 Introduction

Resting state fMRI (rs-fMRI) tracks steady-state co-activation patterns i.e. functional connectivity, in the brain in the absence of a task paradigm. Recently, there has been an increasing interest in using rs-fMRI to study neurodevelopmental disorders, such as autism and schizophrenia. These disorders are characterized by immense patient heterogeneity in terms of their clinical manifestations. However, the high data dimensionality coupled with external noise greatly confound a unified characterization of behavior and connectivity data.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11766, pp. 709–717, 2019.
https://doi.org/10.1007/978-3-030-32248-9_79

Most techniques relating rs-fMRI to behavior focus on discriminating patients from controls. In the simplest case, statistical tests are used to identify group differences in rs-fMRI features [8]. From the machine learning side, neural network architectures have become popular for investigating neuroimaging correlates of developmental disorders [8]. However, very few works handle continuous valued severity prediction from connectomics data. One recent example is the work of [3], which develops a convolutional neural network (CNN) to predict two cognitive measures directly from brain connectomes. A more traditional example is the work of [2], which combines a dictionary learning on patient correlation matrices with a linear regression on the patient loadings to predict clinical severity. Their joint optimization procedure helps the authors extract representations that generalize to unseen data. A more pipelined approach is presented in [6]. They decouple feature selection from prediction to estimate multiple severity measures jointly from voxel-ROI correlations. In contrast, our method *jointly optimizes* for an *anatomically interpretable basis* and a complex *non-linear behavioral encoding* that explain connectivity and clinical severity simultaneously.

We propose one of the first end-to-end frameworks that embeds a traditional model-based representation (dictionary learning) with deep networks into a single optimization. We borrow from the work of [2] to project the patient correlation matrices onto a shared basis. However, in a notable departure from prior work, we couple the patient projection onto the dictionary with a neural network for multi-score behavioral prediction. We *jointly optimize for the basis, patient representation, and neural network weights* by combining gradient information from the two objectives. We demonstrate that our unified framework provides us with the necessary representational flexibility to model complex interactions in the brain, and to learn effectively from limited training data. Our unique optimization strategy outperforms state-of-the-art baseline methods at estimating a generalizable multi-dimensional patient characterization.

2 Multidimensional Clinical Characterization from Functional Connectomics

Figure 1 illustrates our framework. The blue box denotes our dictionary learning representation, while the gray box is the neural network architecture. Let N be the number of patients and P be the number of regions in our brain parcellation. We decompose the correlation matrix $\mathbf{\Gamma}_n \in \mathcal{R}^{P\times P}$ for each patient n, via K dictionary elements of a shared basis $\mathbf{X} \in \mathcal{R}^{P\times K}$, and a subject-specific loading vector $\mathbf{c}_n \in \mathcal{R}^{K\times 1}$. Thus, our dictionary learning objective $\mathcal{D}$ is as follows:

$$\mathcal{D}(\mathbf{X}, \{\mathbf{c}_n\}; \{\mathbf{\Gamma}_n\}) = \sum_n \left[||\mathbf{\Gamma}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)\mathbf{X}^T||_F^2 + \gamma_2 ||\mathbf{c}_n||_2^2 \right] + \gamma_1 ||\mathbf{X}||_1 \quad (1)$$

where $\mathbf{diag}(\mathbf{c}_n)$ denotes a matrix with the entries of $\mathbf{c}_n$ on the leading diagonal and the non-diagonal entries as 0. Since $\mathbf{\Gamma}_n$ is positive semi-definite, we add the constraint $\mathbf{c}_{nk} \geq 0$. The columns of $\mathbf{X}$ capture representative patterns of co-activation common to the cohort. The loadings $\mathbf{c}_{nk}$ capture the network strength

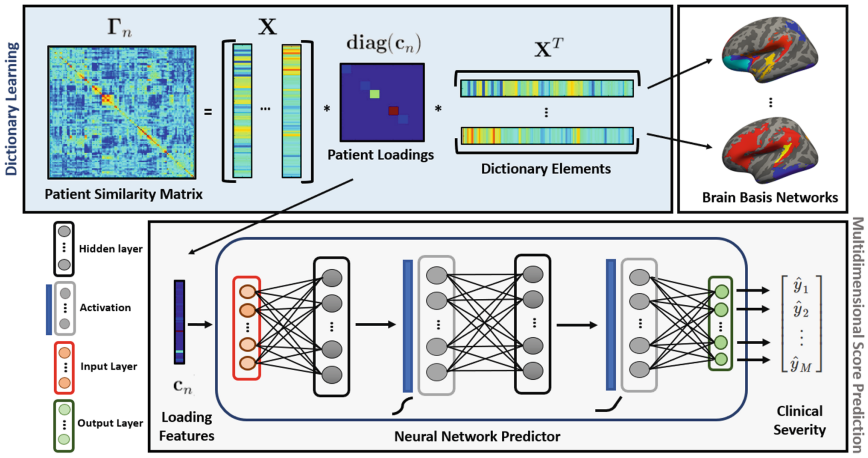

Fig. 1. A unified framework for integrating neural networks and dictionary learning. **Blue Box:** Dictionary Learning from correlation matrices **Gray Box:** Neural Network architecture for multidimensional score prediction (Color figure online)

of basis k in patient n. We add an ℓ_1 penalty to $\mathbf{X}$ to encourage sparsity, and an ℓ_2 penalty to $\{\mathbf{c}_n\}$ to ensure that the objective is well posed. γ_1 and γ_2 are the corresponding regularization weights. The loadings $\mathbf{c}_n$ are also the input features to a neural network. The network parameters $\mathbf{\Theta}$ encode a series of non-linear transformations that map the input features to behavior. $\mathbf{Y}_n \in \mathcal{R}^{M\times 1}$ is a vector of M concatenated clinical measures, which describe the location of patient n on the behavioral spectrum. $\hat{\mathbf{Y}}_n$ is estimated using the latent representation $\mathbf{c}_n$. We employ the Mean Square Error (MSE) to define the network loss $\mathcal{L}$:

$$\mathcal{L}(\{\mathbf{c}_n\}, \mathbf{\Theta}; \{\mathbf{Y}_n\}) = \sum_n \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) = \lambda \sum_n ||\hat{\mathbf{Y}}_n - \mathbf{Y}_n||_F^2 \tag{2}$$

Since $\mathcal{L}$ is added to $\mathcal{D}$ defined in Eq. (1), λ balances the contribution of the dictionary learning and neural network terms to the objective.

Our proposed network architecture is highlighted in the gray box. Our modeling choices require us to carefully control for two key network design aspects: representational capacity, and convergence of the optimization. Given the low dimensionality of the input $\mathbf{c}_n$, we opt for a simple fully connected Artificial Neural Network (ANN) with two hidden layers and a width of 40. We use a tanh function $\left(\tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}\right)$ as the activation to the first hidden layer. We then use a SoftPlus $(SP(x) = \log(1 + \exp(x)))$, a smooth approximation to the Rectified Linear Unit, as the activation for the second layer. Experimentally, we found that these modeling choices are robust to issues with saturation and vanishing gradients, which commonly confound neural network training.

2.1 Joint Optimization Strategy

We use alternating minimization to iteratively optimize for the dictionary elements $\mathbf{X}$, the patient projections $\{\mathbf{c}_n\}$, and ANN parameters $\mathbf{\Theta}$. Here, we sequentially optimize for each hidden variable in the objective by fixing the rest, until global convergence. We use Proximal Gradient Descent to handle the non-differentiable ℓ_1 penalty in Eq. (1), which requires the rest of the objective to be convex in $\mathbf{X}$. We circumvent this issue by the strategy in [2]. Namely, we introduce N constraints of the form $\mathbf{V}_n = \mathbf{X}\mathbf{diag}(\mathbf{c}_n)$, and substitute them into the Frobenius norm terms in Eq. (1). These constraints are enforced using the Augmented Lagrangians $\{\mathbf{\Lambda}_n\}$. If Tr[$\mathbf{Q}$] denotes the trace operation, we add N terms of the form $\text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n))\right] + 0.5\ ||\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)||_F^2$ to Eq. (1). We then iterate through the following four steps until convergence.

***Proximal Gradient Descent on* $\mathbf{X}$.** Each step of the proximal algorithm constructs a locally smooth quadratic model of $||\mathbf{X}||_1$ based on the gradient of $\mathcal{D}$ with respect to $\mathbf{X}$. Using this model, the algorithm iteratively updates $\mathbf{X}$ through shrinkage thresholding. We fix the learning rate for this step at 10^{-4}.

***Updating the Neural Network Weights* $\mathbf{\Theta}$.** We optimize the weights $\mathbf{\Theta}$ according to the loss function $\mathcal{L}$ using backpropagation to estimate gradients. There are several obstacles in training a neural network to generalize and few available theoretical guarantees to guide design considerations. We pay careful attention to this, since the global optimization procedure couples the updates between $\mathbf{\Theta}$ and $\{\mathbf{c}_n\}$. We employ the ADAM optimizer, which is robust to small datasets. We randomly initialize at the first main update. We found a learning rate of 10^{-4}, scaled by 0.9 every 5 epochs to be sufficient for encoding the training data, while avoiding bad local minima and over-fitting. We train for 50 epochs with a batch-size of 12. Finally, we fix the obtained weights to update $\{\mathbf{c}_n\}$.

***L-BFGS Update for* $\{\mathbf{c}_n\}$.** The objective for each $\mathbf{c}_n$ decouples as follows:

$$\mathcal{J}(\mathbf{c}_n) = \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) + \gamma_2||\mathbf{c}_n||_2^2 + \text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n))\right] + 0.5||\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)||_F^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \tag{3}$$

Notice that we can use a standard backpropagation algorithm to compute the gradient of $\ell_{\mathbf{\Theta}}(.)$ with respect to $\mathbf{c}_n$, denoted by $\nabla\ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n)$. The gradient of $\mathcal{J}$ with respect to $\mathbf{c}_n$, denoted $\mathbf{g}(\mathbf{c}_n)$, can then be computed as follows:

$$\mathbf{g}(\mathbf{c}_n) = \nabla\ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) + \mathbf{c}_n \circ \left[\left[\mathcal{I}_K \circ (\mathbf{X}^T\mathbf{X})\right]\mathbf{1}\right] - \left[\mathcal{I}_K \circ (\mathbf{\Lambda}_n^T\mathbf{X} + \mathbf{V}_n^T\mathbf{X})\right]\mathbf{1} + 2\gamma_2\mathbf{c}_n$$

where $\mathbf{1}$ is the vector of all ones, and $\circ$ represents the Hadamard product. The first term is from the ANN, while the rest are from the modified dictionary learning objective. The gradient combines information from the ANN function landscape with that from the correlation matrix estimation. For each iteration r, the BFGS [9] algorithm recursively constructs a positive-definite Hessian approximation $\mathbf{B}(\mathbf{c}_n^r)$ based on the gradients estimated. Next, we iteratively compute

a descent direction $\mathbf{p}$ for $\mathbf{c}_n^r$ using the following bound-constrained objective:

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{J}(\mathbf{c}_n^r) + \mathbf{g}(\mathbf{c}_n^r)^T\mathbf{p} + 0.5\mathbf{p}^T\mathbf{B}(\mathbf{c}_n^r)\mathbf{p} \quad s.t. \quad \mathbf{c}_{nk}^r + \mathbf{p}_k \geq 0 \tag{4}$$

We then update $\mathbf{c}_n$ as: $\mathbf{c}_n^{r+1} = \mathbf{c}_n^r + \delta\mathbf{p}^*$, repeating this procedure until convergence. Effectively, the BFGS update leverages second-order curvature information through each $\mathbf{B}(\mathbf{c}_n)$ estimation. In practice, δ is set to 0.9.

Augmented Lagrangian Update for the Constraint Variables. We have a closed form solution for computing the constraint argument $\{\mathbf{V}_n\}$. The dual Lagrangians, i.e. $\{\mathbf{\Lambda}_n\}$ are updated via gradient ascent. We cycle through the collective updates for these two variables until convergence. We use a learning rate of 10^{-4}, scaled by 0.75 at each iteration of gradient ascent.

Prediction on Unseen Data. We use cross validation to assess our framework. For a new patient, we compute the loading vector $\bar{\mathbf{c}}$ using the estimates $\{\mathbf{X}^*, \mathbf{\Theta}^*\}$ obtained during training. We remove the contribution of the ANN term from the joint objective, as we do not know the corresponding value of $\bar{\mathbf{Y}}$ for a new patient. The proximal operator conditions are assumed to hold with equality, removing the Lagrangian terms. The optimization in $\bar{\mathbf{c}}$ takes the following form:

$$\begin{gathered} 0.5\ \bar{\mathbf{c}}^T\bar{\mathbf{H}}\bar{\mathbf{c}} + \bar{\mathbf{f}}^T\bar{\mathbf{c}} \quad s.t. \quad \bar{\mathbf{A}}\bar{\mathbf{c}} \leq \bar{\mathbf{b}} \\ \bar{\mathbf{H}} = 2(\mathbf{X}^T\mathbf{X}) \circ (\mathbf{X}^T\mathbf{X}) + 2\gamma_2\mathcal{I}_K \\ \bar{\mathbf{f}} = -2\mathcal{I}_K \circ (\mathbf{X}^T\mathbf{\Gamma}_n\mathbf{X})\mathbf{1}; \quad \bar{\mathbf{A}} = -\mathcal{I}_K \quad \bar{\mathbf{b}} = \mathbf{0} \end{gathered} \tag{5}$$

This formulation is similar to Eq. (4) from the BFGS update for $\{\mathbf{c}_n\}$. $\bar{\mathbf{H}}$ is also positive definite, thus giving an efficient quadratic programming solution to Eq. (5). We estimate the score vector $\bar{\mathbf{Y}}$ by a forward pass through the ANN.

2.2 Baseline Comparisons

We compare against two baselines that predict severity scores from correlation matrices $\mathbf{\Gamma}_n \in \mathcal{R}^{P\times P}$. The first has a joint optimization flavor similar to our work, while the second uses a CNN to exploit the structure in $\{\mathbf{\Gamma}_n\}$:

1. The Generative-Discriminative Basis Learning framework in [2]
2. BrainNet Convolutional Neural Network (CNN) from [3]

Implementation Details. The model in [2] adds a linear predictive term $\gamma||\mathbf{C}^T\mathbf{w} - \mathbf{y}||_2^2 + \lambda_3||\mathbf{w}||_2^2$ to the dictionary learning objective in Eq. (1). They estimate a single regression vector $\mathbf{w}$ to compute a scalar measure $\mathbf{y}_n$ from the loading matrix $\mathbf{C} \in \mathcal{R}^{K\times N}$. To provide a fair comparison, we modify this discriminative term to $\gamma||\mathbf{C}^T\mathbf{W} - \mathbf{Y}||_2^2 + \lambda_3||\mathbf{W}||_2^2$, to predict the vectors $\{\mathbf{Y}_n \in \mathcal{R}^{M\times 1}\}_{n=1}^N$ using the weight matrix $\mathbf{W} \in \mathcal{R}^{K\times M}$. Using the guidelines in [2], we fixed λ_3 and γ at 1, and swept the other parameters over a suitable range. We set number of networks to $K = 8$, which is the knee point of the eigenspectrum for $\{\mathbf{\Gamma}_n\}$.

The network architecture in [3] predicts two cognitive measures from correlation matrices. In our case, $\{\mathbf{\Gamma}_n\}$ are of size $P \times P$. For our comparison, we modify the output layer to be of size M. We use the recommended guidelines in [3] for setting the learning rate, batch-size and momentum during training.

For our framework, the trade-off λ from Eq. (2) balances the dictionary learning and neural network losses in the joint optimization. The generalization is also governed by γ_1 and γ_2 from the dictionary learning. Using a grid search, we fix $\{\gamma_1 = 10, \gamma_2 = 0.1, \lambda = 0.1\}$. The number of networks K is fixed to 8.

3 Experimental Evaluation and Results

Data and Preprocessing. We validate our method on a cohort of 52 children with high-functioning ASD. Rs-fMRI data is acquired on a Phillips 3T Achieva scanner (EPI, with TR/TE = 2500/30 ms, flip angle = 70°, res = $3.05 \times 3.15 \times 3$ mm, having 128 or 156 time samples). We use the pre-processing pipeline in [2], which consists of slice time correction, rigid body realignment, normalization to the EPI version of the MNI template, Comp Corr, nuisance regression, spatial smoothing by a 6 mm FWHM Gaussian kernel, and bandpass filtering (0.01−0.1 Hz). We defined 116 regions using the Automatic Anatomical Labeling (AAL) atlas. The contribution of the first eigenvector is subtracted from the regionwise correlation matrices because it is roughly constant and biases the predictions. The residual correlation matrices, $\{\mathbf{\Gamma}_n\}$, are used as inputs for all three methods.

We use three clinical measures quantifying various impairments associated with ASD. The Autism Diagnostic Observation Schedule (ADOS) [5] is scored by a clinician and captures socio-communicative deficits along with repetitive behaviors (dynamic range: 0−30). The Social Responsiveness Scale (SRS) [5] is scored through a parent/teacher questionnaire, and quantifies impaired social functioning (dynamic range: 70−200). On the other hand, Praxis measures the ability to perform skilled motor gestures on command, by imitation, and for tool usage. Two trained research-reliable raters score a videotaped performance based on total percent correct gestures (dynamic range: 0−100).

Performance Characterization. Figure 3 illustrates the *multi-score regression* performance of each method based on ten fold cross validation. Our quantitative metrics are median absolute error (MAE) and mutual information (MI) between the actual and computed scores. Lower MAE and higher MI indicate better performance. The orange points indicate training fit, while the blue points denote performance on held out samples. The $\mathbf{x} = \mathbf{y}$ line indicates ideal performance (Fig. 2).

Observe that both the Generative-Discriminative model and the BrainNet CNN perform comparably to our model for predicting ADOS. However, our model outperforms the baselines in terms of MAE and MI for SRS and Praxis, with the blue points following the $\mathbf{x} = \mathbf{y}$ line more closely. Generally, we find that as we vary the free parameters, the baselines predict *one of the three scores well (in Fig. 3, ADOS), but fit the rest poorly.* In contrast, only our framework learns a

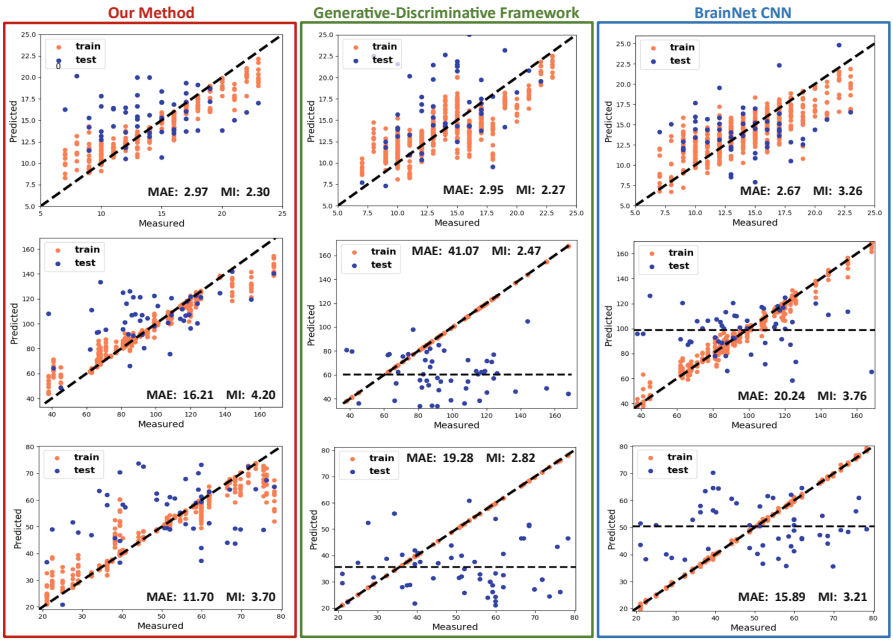

Fig. 2. Multi-Score Prediction performance for **Top:** ADOS **Middle:** SRS **Bottom:** Praxis by **Red Box:** Our Framework. **Green Box:** Generative-Discriminative Framework from [2]. **Blue Box:** BrainNet CNN from [3] (Color figure online)

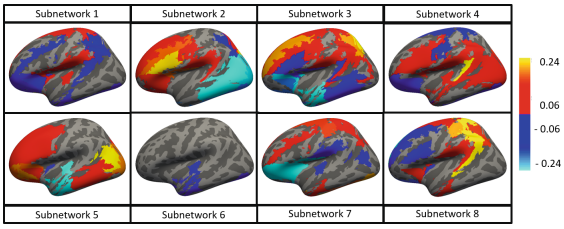

Fig. 3. Eight subnetworks identified by our model from multi-score prediction. The blue and green regions are anticorrelated with the red and orange regions. (Color figure online)

representation that predicts all three clinical measures *simultaneously*, and hence overall outperforms the baselines. We believe that the representational flexibility of neural networks along with our joint optimization is key to generalization.

Figure 3 illustrates the subnetworks in $\{\mathbf{X}_k\}$. Regions storing positive values are anticorrelated with negative regions. From a clinical standpoint, Subnetwork 8 includes the somatomotor network (SMN) and competing, i.e. anticorrelated, contributions from the default mode network (DMN). Subnetwork 3 also has contributions from the DMN and SMN, both of which have been widely

reported in ASD [4]. Along with the DMN, Subnetworks 5 and 2 contain positive and competing contributions from the higher order visual processing areas (i.e. occipital and temporal lobes) respectively. These findings concur with behavioral reports of reduced visual-motor integration in ASD [4]. Finally, Subnetworks 2, 3, and 8 exhibit central executive control network and insula contributions, believed to be critical for switching between self-referential and goal-directed behavior [7].

4 Conclusion

We have introduced the first unified framework to combine classical optimization with the modern-day representational power of neural networks. This integrated strategy allows us to characterize and predict multidimensional behavioral severity from rs-fMRI connectomics data. Namely, our dictionary learning term provides us with interpretability in the brain basis for clinical impairments. Our predictive term cleverly exploits the ability of neural networks to learn rich representations from data. The joint optimization procedure helps learn informative connectivity patterns from limited training data. Our framework makes very few assumptions about the data and can be adapted to work with complex clinical score prediction scenarios. For example, we are developing an extension our method to handle case/control severity prediction using a mixture density network (MDN) [1] in lieu of a regression network. The MDN models a mixture of Gaussians to fit the target bimodal distribution. Accordingly, the network loss function is a negative log-likelihood, which differs from conventional formulations. This is another scenario that may advance our understanding of neuropsychiatric disorders. In the future, we will also explore extensions that simultaneously integrate functional, structural and dynamic information.

Acknowledgements. This work was supported by the National Science Foundation CRCNS award 1822575, National Science Foundation CAREER award 1845430, National Institute of Mental Health (R01 MH085328-09, R01 MH078160-07, K01 MH109766 and R01 MH106564), National Institute of Neurological Disorders and Stroke (R01NS048527-08), and the Autism Speaks foundation.

References

1. Bishop, C.M.: Mixture density networks. Technical report. Citeseer (1994)
2. D'Souza, N.S., Nebel, M.B., Wymbs, N., Mostofsky, S., Venkataraman, A.: A generative-discriminative basis learning framework to predict clinical severity from resting state functional MRI data. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 163–171. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_19
3. Kawahara, J., et al.: BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
4. Nebel, M.B., et al.: Intrinsic visual-motor synchrony correlates with social deficits in autism. Bio. Psych. **79**(8), 633–641 (2016)

5. Payakachat, N., et al.: Autism spectrum disorders: a review of measures for clinical, health services and cost-effectiveness applications. Expert Rev. Pharmacoeconomics Outcomes Res. **12**(4), 485–503 (2012)
6. Rahim, M., et al.: Joint prediction of multiple scores captures better individual traits from brain images. NeuroImage **158**, 145–154 (2017)
7. Sridharan, D., et al.: A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Nat. Acad. Sci. **105**(34), 12569–12574 (2008)
8. Vieira, S., et al.: Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. **74**, 58–75 (2017)
9. Wright, S., et al.: Numerical optimization. Springer Sci. **35**(67–68), 7 (1999)