A Coupled Manifold Optimization Framework to Jointly Model the Functional Connectomics and Behavioral Data Spaces

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 Medicine and 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. The problem of linking functional connectomics to behavior is extremely challenging due to the complex interactions between the two distinct, but related, data domains. We propose a coupled manifold optimization framework which projects fMRI data onto a low dimensional matrix manifold common to the cohort. The patient specific loadings simultaneously map onto a behavioral measure of interest via a second, non-linear, manifold. By leveraging the kernel trick, we can optimize over a potentially infinite dimensional space without explicitly computing the embeddings. As opposed to conventional manifold learning, which assumes a fixed input representation, our framework directly optimizes for embedding directions that predict behavior. Our optimization algorithm combines proximal gradient descent with the trust region method, which has good convergence guarantees. We validate our framework on resting state fMRI from fifty-eight patients with Autism Spectrum Disorder using three distinct measures of clinical severity. Our method outperforms traditional representation learning techniques in a cross validated setting, thus demonstrating the predictive power of our coupled objective.

1 Introduction

Steady state patterns of co-activity in resting state fMRI (rs-fMRI) are believed to reflect the intrinsic functional connectivity between brain regions [4]. Hence, there is increasing interest to use rs-fMRI as a diagnostic tool for studying neurological disorders such as autism, schizophrenia and ADHD. Unfortunately, the well reported confounds of rs-fMRI, coupled with patient heterogeneity makes the task of jointly analyzing rs-fMRI and behavior extremely challenging.

A. C. S. Chung et al. (Eds.): IPMI 2019, LNCS 11492, pp. 605–616, 2019.
https://doi.org/10.1007/978-3-030-20351-1_47

Behavioral Prediction from Neuroimaging Data. Joint analysis of rs-fMRI and behavioral data typically follows a two stage pipeline. Stage 1 is a feature selection or a representation learning step, while Stage 2 maps the learned features onto behavioral data through a statistical or machine learning model. Some notable examples of the Stage 1 feature extraction include graph theoretic measures which aggregate the associative relationships in the connectome, and dimensionality reduction techniques [5], which explain the variation in the data. From here, popular Stage 2 algorithms include Support Vector Machine (SVMs), kernel ridge regression [5]. This pipelined approach has been successful at classification for identifying disease subtypes and distinguishing between patients and healthy controls. However, there has been limited success in terms of predicting dimensional measures, such as behavioral severity from neuroimaging data.

The work of [3] develops a generative-discriminative basis learning framework, which decomposes the rs-fMRI correlation matrices into a group and patient level term. The authors use a linear regression to estimate clinical severity from the patient representation, and jointly optimize the group average, patient coefficients, and regression weights. In this work, we pose the problem of combining the neuroimaging and behavioral data spaces as a dual manifold optimization. Namely, we represent the each patient's fMRI data using a low rank matrix decomposition to project it onto a common vector space. The projection loadings are simultaneously used to construct a high dimensional non-linear embedding to predict a behavioral manifestation. We jointly optimize both representations in order to capture the complex relationship between the two domains.

Manifold Learning for Connectomics. Numerous manifold learning approaches have been employed to study complex brain topologies, especially in the context of disease classification. For example, the work of [11] used graph kernels on the spatio-temporal fMRI time series dynamics to distinguish between the autistic and healthy groups. Going one step further, [9] used higher order morphological kernels to classify ASD subpopulations.

While these methods are computationally efficient and simple in formulation, their generalization power is limited by the input data features. Often, subtle individual level changes are overwhelmed by group level confounds. We integrate the feature learning step directly into our framework by simultaneously optimizing both the embeddings and the projection onto the behavioral space. This optimization is also coupled to the brain basis, which helps us model the behavioral and neuroimaging data space jointly, and reliably capture individual variability. We leverage the kernel trick to provide both the representational flexibility and computational tractability to outperform a variety of baselines.

2 A Coupled Manifold Optimization (CMO) Framework

Figure 1 presents an overview of our Coupled Manifold Optimization (CMO) framework. The blue box represents our neuroimaging term. We group voxels into P ROIs, yielding the $P \times P$ input correlation matrices $\{\mathbf{\Gamma}_n\}_{n=1}^{N}$ for N patients. As seen, the correlation matrices are projected onto a low rank subspace

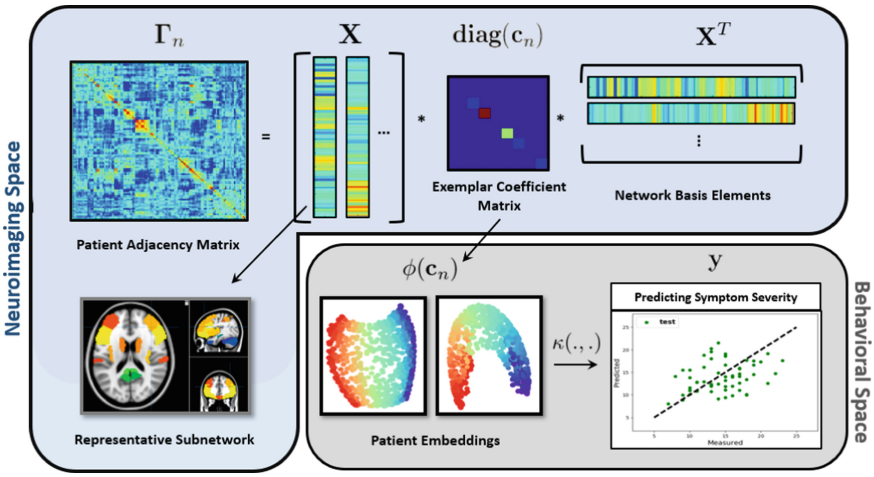

Fig. 1. Joint model for the functional connectomics and behavioral data. **Blue Box:** Matrix manifold representation **Gray Box:** Non-linear kernel ridge regression (Color figure online)

spanned by the group basis. The loadings are related to severity via a non-linear manifold and the associated kernel map, as indicated in the gray box.

Notice that $\mathbf{\Gamma}_n$ is positive semi-definite by construction. We employ a patient specific low rank decomposition $\mathbf{\Gamma}_n \approx \mathbf{Q}_n\mathbf{Q}_n^T$ to represent the correlation matrix. Each rank R factor $\{\mathbf{Q}_n \in \mathcal{R}^{P\times R}\}$, where $R \ll P$, projects onto a low dimensional subspace spanned by the columns of a group basis $\mathbf{X} \in \mathcal{R}^{P\times R}$. The vector $\mathbf{c}_n \in \mathcal{R}^{R\times 1}$ denotes the patient specific loading coefficients as follows:

$$\mathbf{\Gamma}_n \approx \mathbf{Q}_n\mathbf{Q}_n^T = \mathbf{X}\mathbf{diag}(\mathbf{c}_n)\mathbf{X}^T \tag{1}$$

where $\mathbf{diag}(\mathbf{c}_n)$ is a matrix with the entries of $\mathbf{c}_n$ on the leading diagonal, and the off-diagonal elements as 0. Equation (1) resembles a joint eigenvalue decomposition for the set $\{\mathbf{\Gamma}_n\}$ and was also used in [3]. The bases $\mathbf{X}_r \in \mathcal{R}^{P\times 1}$ capture co-activation patterns common to the group, while the coefficient loadings $\mathbf{c}_{nr}$ capture the strength of basis column r for patient n. Our key innovation is to use these coefficients to predict clinical severity via a non-linear manifold. We define an embedding map $\phi(\cdot) : \mathcal{R}^R \rightarrow \mathcal{R}^M$, which maps the native space representation of the coefficient vector $\mathbf{c}$ to an M dimensional embedding space, i.e. $\phi(\mathbf{c}) \in \mathcal{R}^{M\times 1}$. If $\mathbf{y}_n$ is the clinical score for patient n, we have the non-linear regression:

$$\mathbf{y}_n \approx \phi(\mathbf{c}_n)^T\mathbf{w} \tag{2}$$

with weight vector $\mathbf{w} \in \mathcal{R}^{M\times 1}$. Our joint objective combines Eqs. (1) and (2)

$$\mathcal{J}(\mathbf{X}, \{\mathbf{c}_n\}, \mathbf{w}) = \sum_n \left[||\mathbf{\Gamma}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)\mathbf{X}^T||_F^2 + \lambda||\mathbf{y}_n - \phi(\mathbf{c}_n)^T\mathbf{w}||_2^2 \right] \tag{3}$$

along with the constraint $\mathbf{c}_{nr} \geq 0$ to maintain positive semi-definiteness of $\{\mathbf{\Gamma}_n\}$. Here, λ controls the trade-off between the two representations. We include an ℓ_1 penalty on $\mathbf{X}$ to promote sparse solutions for the basis. We also regularize both the coefficients $\{\mathbf{c}_n\}$ and the regression weights $\mathbf{w}$ with ℓ_2 penalties to ensure that the objective is well posed. We add the terms $\gamma_1||\mathbf{X}||_1 + \gamma_2 \sum_n ||\mathbf{c}_n||_2^2 + \gamma_3||\mathbf{w}||_2^2$ to $\mathcal{J}(\cdot)$ in Eq. (3) with the penalties γ_1, γ_2 and γ_3 respectively.

2.1 Inferring the Latent Variables

We use alternating minimization to estimate the hidden variables $\{\mathbf{X}, \{\mathbf{c}_n\}, \mathbf{w}\}$. This procedure iteratively optimizes each unknown variable in Eq. (3) by holding the others constant until global convergence is reached.

Proximal gradient descent [7] is an efficient algorithm which provides good convergence guarantees for the non-differentiable ℓ_1 penalty on $\mathbf{X}$. However, it requires the objective to be convex in $\mathbf{X}$, which is not the case due to the bi-quadratic Frobenius norm expansion in Eq. (1). Hence, we introduce N constraints of the form $\mathbf{V}_n = \mathbf{X}\mathbf{diag}(\mathbf{c}_n)$, similar to the work of [3]. We enforce these constraints using the Augmented Lagrangians $\{\mathbf{\Lambda}_n\}$:

$$\mathcal{J}(\mathbf{X}, \{\mathbf{c}_n\}, \mathbf{w}, \{\mathbf{V}_n\}, \{\mathbf{\Lambda}_n\}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{V}_n\mathbf{X}^T||_F^2 + \lambda \sum_n ||\mathbf{y}_n - \phi(\mathbf{c}_n)^T\mathbf{w}||_2^2 + \sum_n \left[\mathrm{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n))\right] + \frac{1}{2}||\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)||_F^2 \right] \quad (4)$$

with $\mathbf{c}_{nr} \geq 0$ and $\mathrm{Tr}(\mathbf{M})$ denoting the trace operator. The additional terms $||\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)||_F^2$ regularize the trace constraints. Equation (4) is now convex in both $\mathbf{X}$ and the set $\{\mathbf{V}_n\}$, which allows us to optimize them via standard procedures. We iterate through the following four update steps till global convergence:

Proximal Gradient Descent on $\mathbf{X}$***:*** The gradient of $\mathcal{J}$ with respect to $\mathbf{X}$ is:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{X}} = \sum_n 2\left[\mathbf{X}\mathbf{V}_n^T - \mathbf{\Gamma}_n\right]\mathbf{V}_n - \mathbf{V}_n\mathbf{diag}(\mathbf{c}_n) + \mathbf{X}\mathbf{diag}(\mathbf{c}_n)^2 - \mathbf{\Lambda}_n\mathbf{diag}(\mathbf{c}_n)$$

With a learning rate of t, the proximal update with respect to $||\mathbf{X}||_1$ is given by:

$$\mathbf{X}^k = \mathbf{prox}_{||\cdot||_1}\left[\mathbf{X}^{k-1} - \left[\frac{t}{\gamma_1}\right]\frac{\partial \mathcal{J}}{\partial \mathbf{X}}\right] \; s.t. \; \mathbf{prox}_t(\mathbf{L}) = \mathbf{sgn}(\mathbf{L}) \circ (\mathbf{max}(|\mathbf{L}| - t, \mathbf{0}))$$

Where $\circ$ denotes the Hadamard product. Effectively, this update performs an iterative shrinkage thresholding on a locally smooth quadratic model of $||\mathbf{X}||_1$.

Kernel Ridge Regression for $\mathbf{w}$***:*** We denote $\mathbf{y}$ as the vector of the clinical severity scores and stack the patient embedding vectors i.e. $\phi(\mathbf{c_n}) \in \mathcal{R}^{M \times 1}$ into a matrix $\mathbf{\Phi}(\mathbf{C}) \in \mathcal{R}^{M \times N}$. The portion of $\mathcal{J}(\cdot)$ that depends on $\mathbf{w}$ is:

$$\mathcal{F}(\mathbf{w}) = \lambda||\mathbf{y} - \mathbf{\Phi}(\mathbf{C})^T\mathbf{w}||_2^2 + \gamma_3||\mathbf{w}||_2^2 \quad (5)$$

Setting the gradient of Eq. (5) to 0, and applying the matrix inversion lemma, the closed form solution for $\mathbf{w}$ is similar to kernel ridge regression:

$$\mathbf{w} = \mathbf{\Phi}(\mathbf{C})\Big[\mathbf{\Phi}(\mathbf{C})^T\mathbf{\Phi}(\mathbf{C}) + \frac{\gamma_3}{\lambda}\mathcal{I}_N\Big]^{-1}\mathbf{y} = \mathbf{\Phi}(\mathbf{C})\boldsymbol{\alpha} = \sum_j \boldsymbol{\alpha}_j\phi(\mathbf{c}_j) \tag{6}$$

where $\mathcal{I}_N$ is the identity matrix. Let $\kappa(\cdot,\cdot) : \mathcal{R}^M \times \mathcal{R}^M \rightarrow \mathcal{R}$ be the kernel map for ϕ, i.e. $\kappa(\mathbf{c}, \hat{\mathbf{c}}) = \phi(\mathbf{c})^\mathbf{T}\phi(\hat{\mathbf{c}})$. The dual variable $\boldsymbol{\alpha}$ can be expressed as $\boldsymbol{\alpha} = (\mathbf{K} + \frac{\gamma_3}{\lambda}\mathcal{I}_N)^{-1}\mathbf{y}$, where $\mathbf{K} = \mathbf{\Phi}(\mathbf{C})^T\mathbf{\Phi}(\mathbf{C})$ is the Gram matrix for the kernel $\kappa(\cdot,\cdot)$. Equation (6) implies that $\mathbf{w}$ lies in the span of the coefficient embeddings defining the manifold. We use the form of $\mathbf{w}$ in Eq. (6) to update the loading vectors in the following step, without explicitly parametrizing the vector $\phi(\mathbf{c}_n)$.

Trust Region Update for $\{\mathbf{c}_n\}$***:*** The objective function for each patient loading vector $\mathbf{c}_n$ decouples as follows when the other variables are fixed:

$$\mathcal{F}(\mathbf{c}_n) = \lambda||\mathbf{y}_n - \phi(\mathbf{c}_n)^T\mathbf{w}||_2^2 + \gamma_2||\mathbf{c}_n||_2^2 + \text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n))\right] + \frac{1}{2}||\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)||_F^2 \quad s.t. \quad \mathbf{c}_{nr} \geq 0 \tag{7}$$

We now substitute this form into Eq. (7) and use the kernel trick, to write:

$$||\mathbf{y}_n - \phi(\mathbf{c}_n)^T\mathbf{w}||_2^2 = ||\mathbf{y}_n - \sum_j \phi(\mathbf{c}_n)^T\phi(\hat{\mathbf{c}}_\mathbf{j})\boldsymbol{\alpha}_j||_2^2 = ||\mathbf{y}_n - \sum_j \kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)\boldsymbol{\alpha}_j||_2^2$$

where $\{\hat{\mathbf{c}}_\mathbf{n}\}$ denotes the coefficient vector estimates from the previous step to compute $\mathbf{w}$. Notice that the kernel trick buys a second advantage, in that we only need to optimize over the first argument of $\kappa(\cdot,\cdot)$. Since kernel functions typically have a nice analytic form, we can easily compute the gradient $\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ and hessian $\nabla^2\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ of $\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ with respect to $\mathbf{c}_n$.

Given this, the gradient of $\mathcal{F}(\cdot)$ with respect to $\mathbf{c}_n$ takes the following form:

$$\mathbf{g}_n = \frac{\partial\mathcal{F}}{\partial\mathbf{c}_n} = \mathbf{c}_n \circ \left[\left[\mathcal{I}_R \circ (\mathbf{X}^T\mathbf{X})\right]\mathbf{1}\right] - \left[\mathcal{I}_R \circ (\mathbf{\Lambda}_n^T\mathbf{X} + \mathbf{V}_n^T\mathbf{X})\right]\mathbf{1} + 2\gamma_2\mathbf{c}_n$$
$$-\lambda\sum_i \boldsymbol{\alpha}_i\left[2\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i)\mathbf{y}_i - \sum_k \boldsymbol{\alpha}_k\left[\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i)\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_k) + \kappa(\mathbf{c}_n, \hat{\mathbf{c}}_k)\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i)\right]\right]$$

where $\mathbf{1}$ is the vector of all ones. Notice that the top line of the gradient term is from the matrix decomposition and regularization terms, and the bottom line corresponds to the kernel regression. The Hessian $\mathbf{H}_n = \partial^2\mathcal{F}/\partial\mathbf{c}_n^2$ can be similarly computed. Due to space limitations, we have omitted its explicit form.

Given the low dimensionality of $\mathbf{c}_n$, we derive a trust region optimizer for this variable. The trust region algorithm provides guaranteed convergence, like the popular gradient descent method, with the speedup of second-order procedures. The algorithm iteratively updates $\mathbf{c}_n$ according to the descent direction $\mathbf{p}_k$, i.e. $\mathbf{c}_n^{(k+1)} = \mathbf{c}_n^{(k)} + \mathbf{p}_k$. The vector $\mathbf{p}_k$ is computed via the following quadratic objective, which is a second order Taylor expansion of $\mathcal{F}$ around $\mathbf{c}_n^k$:

$$\mathbf{p} = \arg\min_{\mathbf{p}} \mathcal{F}(\mathbf{c}_n^k) + \mathbf{g}_n^k(\mathbf{c}_n^k)^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\mathbf{H}_n^k(\mathbf{c}_n^k)\mathbf{p} \;\; s.t. \; ||\mathbf{p}||_2 \leq \delta_k, \; \mathbf{c}_{nr}^k + \mathbf{p}_r \geq 0$$

where $\mathbf{g}_n(\cdot)$ and $\mathbf{H}_n(\cdot)$ are the gradient and Hessian referenced above evaluated at the current iterate $\mathbf{c}_n^k$. We recursively search for a suitable trust region radius δ_k such that we are guaranteed sufficient decrease in the objective at each iteration. This algorithm has a lower bound on the function decrease per update, and with an appropriate choice of the δ_k, converges to a local minimum of $\mathcal{F}$ [12].

Augmented Lagrangian Update for $\mathbf{V}_n$ ***and*** $\mathbf{\Lambda}_n$***:*** Each $\{\mathbf{V}_n\}$ has a closed form solution, while the dual variables $\{\mathbf{\Lambda}_n\}$ are updated via gradient ascent:

$$\mathbf{V}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{X}^T + 2\mathbf{\Gamma}_n\mathbf{X} - \mathbf{\Lambda}_n)(\mathcal{I}_R + 2\mathbf{X}^T\mathbf{X})^{-1} \tag{8}$$

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k(\mathbf{V}_n - \mathbf{X}\mathbf{diag}(\mathbf{c}_n)) \tag{9}$$

We cycle through the updates in Eqs. (8–9) to ensure that the proximal constraints are satisfied with increasing certainty at each step. We choose the learning rate parameter η_k for the gradient ascent step of the Augmented Lagrangian to guarantee sufficient decrease for every iteration of alternating minimization.

Prediction on Unseen Data: We use the estimates $\{\mathbf{X}^*, \mathbf{w}^*, \{\mathbf{c}_n^*\}\}$ obtained from the training data to compute the loading vector $\bar{\mathbf{c}}$ for an unseen patient. We must remove the data term in Eq. (4), as the corresponding value of $\bar{\mathbf{y}}$ is unknown for the new patient. Hence, the kernel terms in the gradient and hessian disappear. We also assume that the conditions for the proximal operator hold with equality; this eliminates the Augmented Lagrangians in the computation. The objective in $\bar{\mathbf{c}}$ reduces to the following quadratic form:

$$\frac{1}{2}\bar{\mathbf{c}}^T\bar{\mathbf{H}}\bar{\mathbf{c}} + \bar{\mathbf{f}}^T\bar{\mathbf{c}} \;\; s.t. \;\; \bar{\mathbf{A}}\bar{\mathbf{c}} \leq \bar{\mathbf{b}} \tag{10}$$

Note that the formulation is similar to the trust region update we used previously. For an unseen patient, the parameters from Eq. (10) are:

$$\bar{\mathbf{H}} = 2(\mathbf{X}^T\mathbf{X}) \circ (\mathbf{X}^T\mathbf{X}) + 2\gamma_2\mathcal{I}_R$$
$$\bar{\mathbf{f}} = -2\mathcal{I}_R \circ (\mathbf{X}^T\mathbf{\Gamma}_n\mathbf{X})\mathbf{1}; \;\; \bar{\mathbf{A}} = -\mathcal{I}_R \;\; \bar{\mathbf{b}} = \mathbf{0}$$

The Hessian $\bar{\mathbf{H}}$ is positive definite, which leads to an efficient quadratic programming solution to Eq. (10). The severity score for the test patient is estimated by $\bar{\mathbf{y}} = \phi(\bar{\mathbf{c}})^T\mathbf{w}^* = \sum_j \kappa(\bar{\mathbf{c}}, \mathbf{c}_j^*)\boldsymbol{\alpha}_j^*$, where $\boldsymbol{\alpha}^* = \left[\mathbf{\Phi}(\mathbf{C}^*)^T\mathbf{\Phi}(\mathbf{C}^*) + \frac{\gamma_3}{\lambda}\mathcal{I}_N\right]^{-1}\mathbf{y}$.

2.2 Baseline Comparison Methods

We compare our algorithm with the standard manifold learning pipeline to predict the target severity score. We consider two classes of representation learning techniques motivated from the machine learning and graph theoretic literature. From here, we construct a non-linear regression model similar to our manifold learning term in Eq. (3). Our five baseline comparisons are as follows:

1. Principal Component Analysis (PCA) on the stacked $\frac{P\times(P-1)}{2}$ correlation coefficients followed by a kernel ridge regression (kRR) on the projections
2. Kernel Principal Component Analysis (kPCA) on the correlation coefficients followed by a kRR on the embeddings
3. Node Degree computation (D_N) based on the thresholded correlation matrices followed by a kRR on the P node features
4. Betweenness Centrality (C_B) on the thresholded correlation matrices followed by a kRR on the P node features
5. Decoupled Matrix Decomposition (Eq.(3)) and kRR on the loadings $\{\mathbf{c}_n\}$.

Baseline 5 helps us evaluate and quantify the advantage provided by our joint optimization approach as opposed to a pipelined prediction of clinical severity.

3 Experimental Results:

rs-fMRI Dataset and Preprocessing. We validate our method on a cohort of 58 children with high-functioning ASD (Age: 10.06 ± 1.26, IQ: 110 ± 14.03). rs-fMRI scans were acquired on a Phillips 3T Achieva scanner using a single-shot, partially parallel gradient-recalled EPI sequence 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 a standard pre-processing pipeline, consisting of slice time correction, rigid body realignment, normalization to the EPI version of the MNI template, Comp Corr [1], nuisance regression, spatial smoothing by a 6 mm FWHM Gaussian kernel, and bandpass filtering between 0.01–0.1 Hz. We use the Automatic Anatomical Labeling (AAL) atlas to define 116 cortical, subcortical and cerebellar regions. We subtract the contribution of the first eigenvector 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 the methods.

We consider three separate measures of clinical severity quantifying different impairments associated with ASD. The Autism Diagnostic Observation Schedule (ADOS) [8] captures social and communicative deficits of the patient along with repetitive behaviors (dynamic range: 0–30). The Social Responsiveness Scale (SRS) [8] characterizes impaired social functioning (dynamic range: 70–200). Finally, the Praxis score [2] quantifies motor control, tool usage and gesture imitation skills in ASD patients (dynamic range: 0–100).

Characterizing the Non-linear Patient Manifold: Based on simulated data, we observed that the standard exponential kernel provides a good recovery performance in the lower part of the dynamic range, while polynomial kernels are more suited for modeling the larger behavioral scores, as shown in Fig. 2. Thus, we use a mixture of both kernels to capture the complete behavioral characteristics:

$$\kappa(\mathbf{c}_i, \mathbf{c}_j) = \exp\left[-\frac{||\mathbf{c}_i - \mathbf{c}_j||_2^2}{\sigma^2}\right] + \frac{\rho}{l}\left(\mathbf{c}_j^T \mathbf{c}_i + 1\right)^l$$

We vary the kernel parameters across 2 orders of magnitude and select the settings: ADOS $\{\sigma^2 = 1, \rho = 0.8, l = 2.5\}$, SRS $\{\sigma^2 = 1, \rho = 2, l = 1.5\}$ and

Praxis $\{\sigma^2 = 1, \rho = 0.5, l = 1.5\}$. The varying polynomial orders reflect the differences in the dynamic ranges of the scores.

Predicting ASD Clinical Severity. We evaluate every algorithm in a ten fold cross validation setting, i.e. we train the model on a 90% split of our data, and report the performance on the unseen 10%. The number of components was fixed at 15 for PCA and at 10 for k-PCA. For k-PCA, we use an RBF kernel with the coefficient parameter 0.1. There are two free parameters for the kRR, namely, the kernel parameter C and ℓ_2 parameter β. We obtain the best performance for the following settings: ADOS $\{C = 0.1, \beta = 0.2\}$, SRS $\{C = 0.1, \beta = 0.8\}$, and Praxis $\{C = 0.01, \beta = 0.2\}$. For the graph theoretic baselines, we obtained the best performance by thresholding the entries of $\{\mathbf{\Gamma}_n\}$ at 0.2. We fixed the parameters in our CMO framework using a grid search for $\{\lambda, \gamma_1, \gamma_2, \gamma_3\}$. The values were varied between $(10^{-3} - 10)$. The performance is insensitive to λ and γ_3, which are fixed at 1. The remaining parameters were set at $\{\gamma_1 = 10, \gamma_2 = 0.7, \gamma_3 = 1\}$ for all the scores. We fix the number of networks, R, at the knee point of the eigenspectrum of $\{\mathbf{\Gamma}_n\}$, i.e. $(R = 8)$.

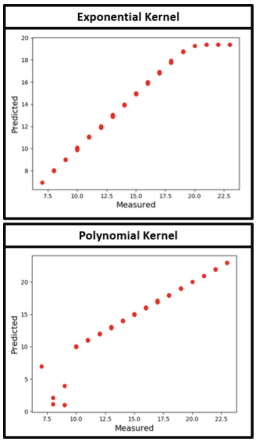

Fig. 2. Recovery **Top:** Exponential **Bottom:** Polynomial kernel

Performance Comparison. Figures 3, 4, and 5 illustrate the regression performance for ADOS, SRS, and Praxis respectively. The bold $\mathbf{x} = \mathbf{y}$ line indicates ideal performance. The red points denote the training fit, while the blue points indicate testing performance. Note that baseline testing performance tracks the mean value of the data (indicated by the horizontal black line). In comparison, our method not only consistently fits the training set more faithfully, but also generalizes much better to unseen data. We emphasize that even the pipelined treatment using the matrix decomposition in Eq. (3), followed by a kernel ridge regression on the learnt projections fails to generalize. This finding makes a strong case for coupling the two representation terms in our CMO strategy. We conjecture that the baselines fail to capture representative connectivity patterns that explain both the functional neuroimaging data space and the patient behavioral heterogeneity. On the other hand, our CMO framework leverages the underlying structure of the correlation matrices through the basis manifold representation. At the same time, it seeks those embedding directions that are predictive of behavior. As reported in Table 1, our method quantitatively outperforms the baselines approaches, in terms of both the Median Absolute Error (MAE) and the Mutual Information (MI) metrics.

Clinical Interpretation. Figure 6 illustrates the subnetworks $\{\mathbf{X}_r\}$ trained on ADOS. The colorbar indicates subnetwork contributions to the AAL regions.

Table 1. Performance evaluation using **Median Absolute Error (MAE)** & **Mutual Information (MI)**. Lower MAE & higher MI indicate better performance.

Score	Method	MAE train	MAE test	MI train	MI test
ADOS	PCA & kRR	1.29	3.05	1.46	0.87
	k-PCA & kRR	1.00	2.94	1.48	0.38
	C_B & kRR	2.10	2.93	1.03	0.95
	D_N & kRR	2.09	3.03	0.97	0.96
	Decoupled	2.11	3.11	0.82	1.24
	CMO Framework	**0.035**	**2.73**	**3.79**	**2.10**
SRS	PCA & kRR	7.39	19.70	2.78	3.30
	k-PCA & kRR	5.68	18.92	2.85	1.74
	C_B & kRR	11.00	17.72	2.32	3.66
	D_N & kRR	11.46	17.79	2.24	3.60
	Decoupled	15.9	18.61	2.04	3.71
	CMO Framework	**0.09**	**13.28**	**5.28**	**4.36**
Praxis	PCA & kRR	5.33	12.5	2.50	2.68
	k-PCA & kRR	4.56	11.15	2.56	1.51
	C_B & kRR	8.17	12.61	1.99	3.05
	D_N & kRR	8.18	13.14	2.00	3.20
	Decoupled	10.11	13.33	3.28	1.53
	CMO Framework	**0.13**	**9.07**	**4.67**	**3.87**

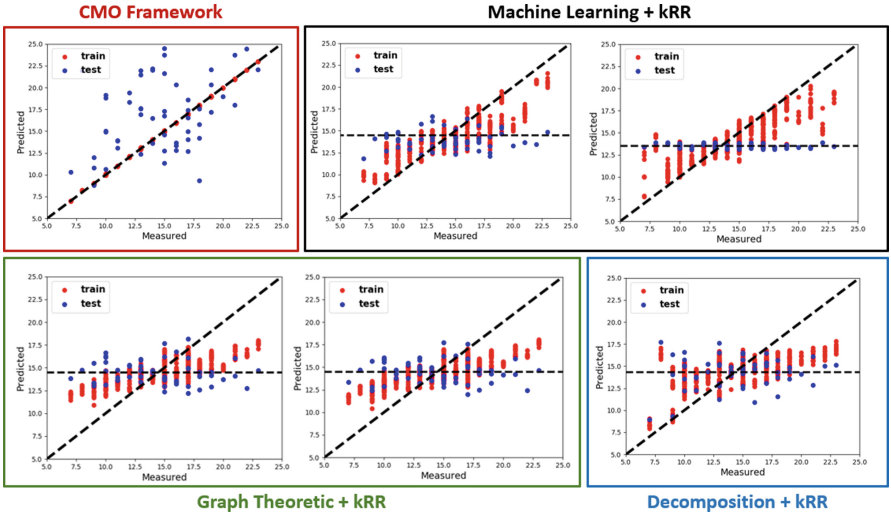

Fig. 3. Prediction performance for the ADOS score for **Red Box:** CMO framework. **Black Box: (L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box: (L)** Node degree centrality and kRR **(R)** Betweenness centrality and kRR **Blue Box:** Matrix decomposition from Eq. (3) followed by kRR (Color figure online)

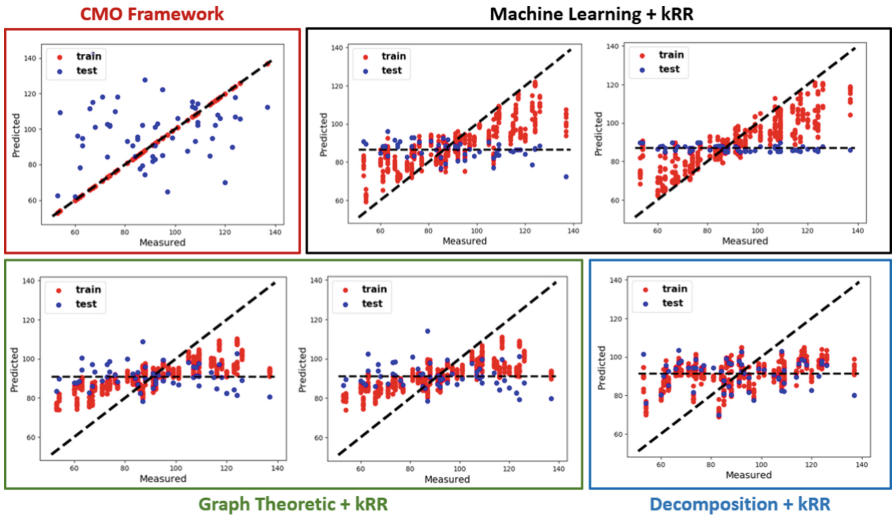

Fig. 4. Prediction performance for the SRS score for **Red Box:** CMO framework. **Black Box: (L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box: (L)** Node degree centrality and kRR **(R)** Betweenness centrality and kRR **Blue Box:** Matrix decomposition from Eq. (3) followed by kRR (Color figure online)

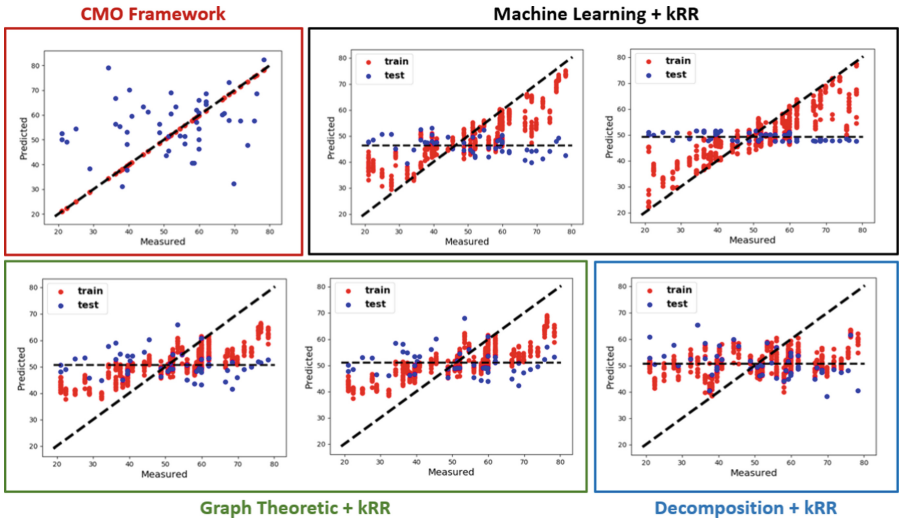

Fig. 5. Prediction performance for the Praxis score for **Red Box:** CMO framework. **Black Box: (L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box: (L)** Node degree centrality and kRR **(R)** Betweenness centrality and kRR **Blue Box:** Matrix decomposition from Eq. (3) followed by kRR (Color figure online)

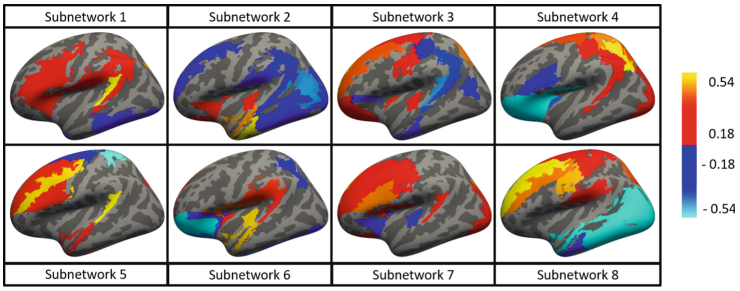

Fig. 6. Eight subnetworks identified by our model from the prediction of ADOS. The blue & green regions are anticorrelated with the red & orange regions. (Color figure online)

Regions storing negative values are anticorrelated with positive regions. From a clinical standpoint, Subnetwork 4 includes the somatomotor network (SMN) and competing i.e. anticorrelated contributions from the default mode network (DMN), previously reported in ASD [6]. Subnetwork 8 comprises of the SMN and competing contributions from the higher order visual processing areas in the occipital and temporal lobes. These findings are in line with behavioral reports of reduced visual-motor integration in ASD [6]. Though not evident from the surface plots, Subnetwork 5 includes anticorrelated contributions from subcortical regions, mainly, the amygdala and hippocampus, believed to be important for socio-emotional regulation in ASD. Finally, Subnetwork 6 has competing contributions from the central executive control network and insula, which are critical for switching between self-referential and goal-directed behavior [10].

Figure 7 compares Subnetwork 2 obtained from ADOS, SRS and Praxis prediction. There is a significant overlap in the bases subnetworks obtained by training across the different scores. This strengthens the hypothesis that our method is able to identify representative, as well as predictive connectivity patterns.

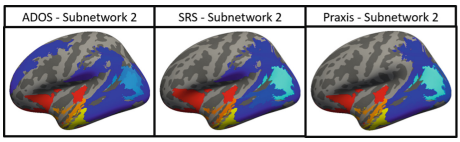

Fig. 7. Subnetwork 2 obtained from **L:** ADOS **M:** SRS and **R:** Praxis prediction

4 Conclusion

We have introduced a Coupled Manifold Optimization strategy that jointly analyzes data from two distinct, but related, domains through its shared projection. In contrast to conventional manifold learning, we optimize for the relevant embedding directions that are predictive of clinical severity. Consequently,

our method captures representative connectivity patterns that are important for quantifying and understanding the spectrum of clinical severity among ASD patients. We would like to point out that our framework makes very few assumptions about the data and can be adapted to work with different similarity matrices and clinical scores. We believe that our method could potentially be an important diagnostic tool for the cognitive assessment of various neuropsychiatric disorders. We are working on a multi-score extension which jointly analyses different behavioral domains. We will explore extensions of our representation that simultaneously integrate functional, structural and behavioral information.

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

References

1. Behzadi, Y., et al.: A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage **37**(1), 90–101 (2007)
2. Dowell, L.R., et al.: Associations of postural knowledge and basic motor skill with dyspraxia in autism: implication for abnormalities in distributed connectivity and motor learning. Neuropsychology **23**(5), 563 (2009)
3. 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
4. Fox, M.D., et al.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. **8**(9), 700 (2007)
5. Murphy, K.P.: Machine learning: a probabilistic perspective (2012)
6. Nebel, M.B., et al.: Intrinsic visual-motor synchrony correlates with social deficits in autism. Biol. Psych. **79**(8), 633–641 (2016)
7. Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends® Opt. **1**(3), 127–239 (2014)
8. Payakachat, N., et al.: Autism spectrum disorders: a review of measures for clinical, health services and cost-effectiveness applications. Exp. Rev. Pharmacoeconomics Outcomes Res. **12**(4), 485–503 (2012)
9. Soussia, M., Rekik, I.: High-order connectomic manifold learning for autistic brain state identification. In: Wu, G., Laurienti, P., Bonilha, L., Munsell, B.C. (eds.) CNI 2017. LNCS, vol. 10511, pp. 51–59. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67159-8_7
10. Sridharan, D., et al.: A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Natl. Acad. Sci. **105**(34), 12569–12574 (2008)
11. Thiagarajan, J.J., et al.: Multiple kernel sparse representations for supervised and unsupervised learning. IEEE Trans. Image Process. **23**(7), 2905–2915 (2014)
12. Wright, S., et al.: Numerical optimization. Springer Sci. **35**(67–68), 7 (1999)