Multidimensional analysis and detection of informative features in human brain white matter

Adam Richie-Halford[1], Jason Yeatman[2], Noah Simon[3] & Ariel Rokem[4]
1. eScience Institute, 3. Dept. of Biostatistics, 4. Dept. of Psychology, University of Washington, 2. Stanford University

ISMRM 2021
Contact: richford@uw.edu

Introduction

White matter structure is important to normal brain function. Diffusion tensor imaging (DTI) measures the white matter *in vivo* by fitting a diffusion model in every voxel.

Challenge: Make sense of the massive dimensionality of DTI data with comparatively few subjects in any given study.

One approach is to reduce the dimensionality of each tensor to scalars, e.g. mean diffusivity (MD) or fractional anisotropy (FA).
Tractometry reduces the dimensionality of this data by quantifying diffusion metrics along tract profiles.

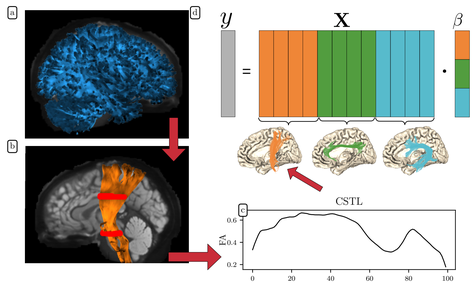

Figure 1: Tractometry data flow (a) Whole brain tractography generates streamlines approximating the trajectories of white matter connections. **(b)** Tractometry classifies these streamlines into anatomical bundles. In this case, we show the left corticospinal tract (CSTL) over a mid-saggital anatomical slice. **(c)** Tractometry further extracts bundle profiles, quantifications of various diffusion metrics along the length of the fiber bundle. Here, we show one subject's fractional anisotropy (FA) profile for CSTL. **(d)** the phenotypical target data and tractometric features can be organized into a linear model, $\hat{y} = \mathbf{X}\hat{\beta}$. The feature matrix $\mathbf{X}$ is color-coded to reveal a natural group structure: the left group contains k features from the CSTL, the middle group contains k features from the left cingulum cingulate, and the right group contains k features from the left arcuate. The coefficients in $\hat{\beta}$ follow the same natural grouping.

Conclusion

- Novel method for analysis of dMRI tractometry data
  - Accurate prediction of phenotypic information
  - Interpretable results and identification of important features
- Applicable to both localized and global phenomena
- Packaged as open-source software called AFQ-Insight:
  `https://github.com/richford/AFQ-Insight`
- Integrates into broader AFQ software ecosystem:
  - pyAFQ: creates tractometry data from raw dMRI
  - AFQ-Browser: visualization, analysis, and sharing

Results: Classifying Patients with ALS

- Previous study measured dMRI in patients with amyotrophic lateral sclerosis (ALS) (Sarica et al. 2017).
  - 24 ALS patients and 24 demographically matched controls
  - Previous state of the art achieved 80% accuracy using random forests and *a priori* feature selection.
- Our method outperforms previous results, with a cross-validated accuracy of 83% and an AUC-ROC of 0.88.
- Moreover, it automatically identifies the corticospinal tract (CST) as the critical feature for differentiating patients with ALS from controls
- Thus, well-known features of the disease can be recovered through an automated, data-driven approach.

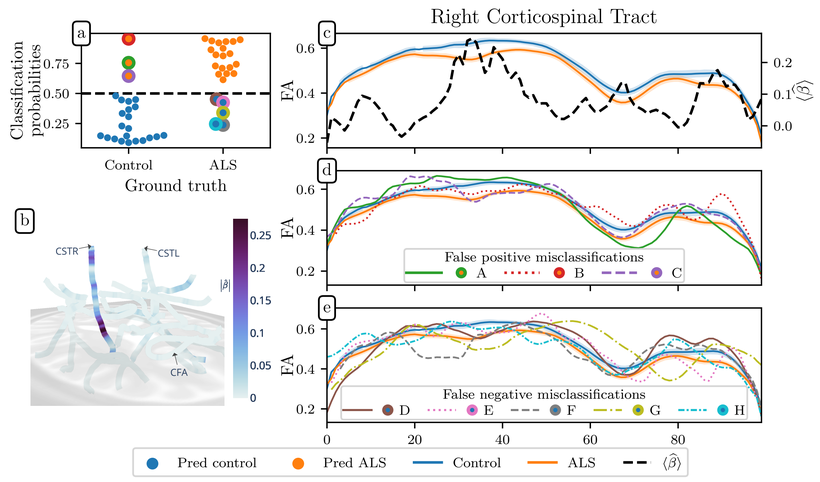

Figure 2: SGL accurately and interpretably predicts ALS diagnosis. (a) Classification probabilities for ALS diagnosis, with controls on the left, patients on the right, predicted controls in blue, and predicted patients in orange. That is, orange dots on the left represent false positives, while blue dots on the right represent false negatives. We achieve 83% accuracy with an ROC AUC of 0.88. **(b)** SGL coefficients are presented on the core fibers of major fiber bundles. They exhibit high group sparsity and are concentrated in the FA of the CST. The brain is oriented with the right hemisphere in the foreground and anterior to the right of the page. The CSTL, CSTR, and CFA bundles are indicated for orientation. **(c)** SGL identifies three portions of the CST as important, where $\hat{\beta}$ (dashed line, right axis) has large values. These are centered around nodes 30, 65, and 90, corresponding to locations of substantial differences in FA between the ALS and control groups (shaded areas indicates standard error of the mean). **(d)** Bundle profiles for false positive classifications. Line colors correspond to the marker edge color in the top left plot. These individuals have reduced FA in the CST portions which SGL identified as important. Their misclassification is coherent with the feature importance and the group differences in FA. **(e)** Individual bundle profiles for false negative classifications. These individuals have bundle profiles which oscillate between the group means.

Results: Predicting "Brain Age"

- In a regression setting, SGL can predict "brain age" in three previous studies.
  - Weston-Havens (WH): 76 subjects with ages 6-50 (Yeatman et al. 2014)
  - Healthy Brain Network (HBN): 978 subjects with ages 5-21 (Alexander et al. 2017)
  - Cam-CAN: 640 subjects with ages 18-88 (Shafto et al. 2014)
- We predict subjects' chronological age with competitive performance (see Figure 3).
- Older subjects have higher residual variance, reflecting the automatically-chosen log-transformation and implying that brain age becomes more difficult to predict as we age chronologically.
- In contrast to the ALS classification case, brain age feature importance is dense and non-localized.

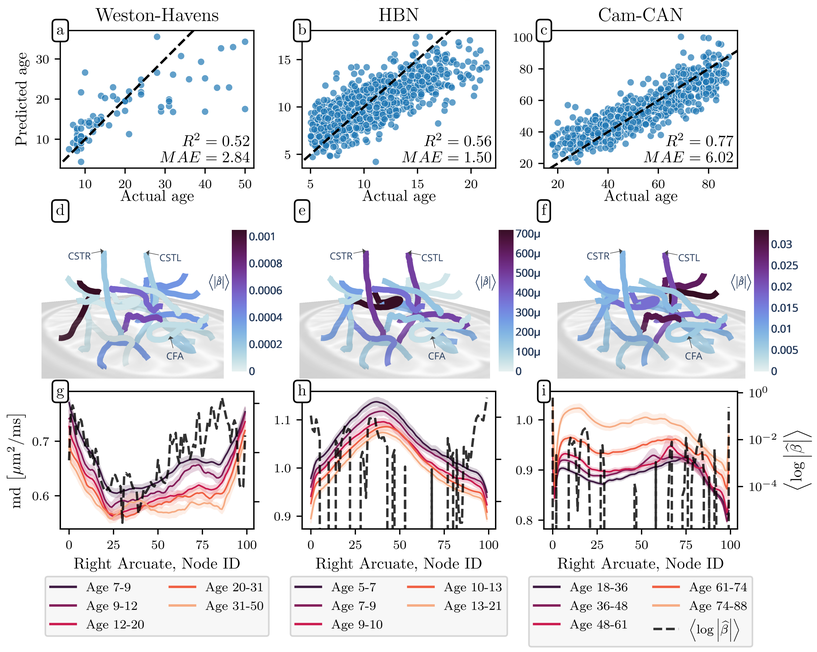

Figure 3: Predicting age with tractometry and SGL. (top) The predicted age vs. true age of each individual from the test splits (i.e., when each subject's data was held out in fitting the model) for the **(a)** WH, **(b)** HBN, and **(c)** Cam-CAN datasets; an accurate prediction falls close to the $y = x$ line (dashed). The mean absolute error (MAE) and coefficient of determination R^2 are presented in the lower right of each scatter plot. **(middle)** Feature importance for predicting age from tractometry in the **(d)** WH, **(e)** HBN, and **(f)** Cam-CAN datasets. The orientation of the brain is that same as in Figure 2b, however because the coefficients exhibit high global sparsity (as opposed to group sparsity), we plot the mean of the absolute value of $\hat{\beta}$ for each bundle on the core fiber. The global distrubution of the $\hat{\beta}$ coefficients reflects the fact that aging is not confined to a single white matter bundle. **(bottom)** Age quintile bundle profiles for the **(g)** WH, **(h)** HBN, and **(i)** Cam-CAN datasets.

Methods

We use AFQ to generate tractometry data as input to our model and then fit a linear model to the data,

$$y = \mathbf{X} \cdot \beta,$$

where

$$y := \text{phenotype } (n_{\text{subjects}} \times 1),$$
$$\mathbf{X} := \text{tractometry data } (n_{\text{subjects}} \times n_{\text{features}}),$$
$$\beta := \text{regression coefficients } (n_{\text{features}} \times 1).$$

The feature matrix, $\mathbf{X}$, has inherent group structure, containing groups for each diffusion metric and each fiber bundle. The high dimensionality of the data ($n_{\text{features}} \sim \mathcal{O}(10^4)$) requires regularization to avoid overfitting, via the Sparse Group Lasso (SGL),

$$\widehat{\beta} = \min_{\beta} \left\{ \underbrace{\|\widehat{y} - \mathbf{X} \cdot \beta\|_2^2}_{\text{Linear regression}} + \underbrace{(1-\alpha)\lambda \sum_{\ell} \sqrt{p_\ell} \|\beta^{(\ell)}\|_2}_{\text{Group Lasso Penalty}} + \underbrace{\alpha\lambda \|\beta\|_1}_{\text{Lasso Penalty}} \right\}$$

SGL enforces sparsity at both the **inter**-group level (using $\alpha = 0$) and **intra**-group level (using $\alpha = 1$). The hyperparameters λ, α are optimized through nested k-fold cross-validation.

References

[1] Richie-Halford et al. Multidimensional analysis and detection of informative features in diffusion mri measurements of human white matter. *bioRxiv*, 2019.

[2] Simon et al. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[3] Yeatman et al. Tract profiles of white matter properties: automating fiber-tract quantification. *PloS one*, 7(11):e49790, 2012.

[4] Sarica et al. The corticospinal tract profile in amyotrophic lateral sclerosis. *Human brain mapping*, 38(2):727–739, 2017.

[5] Yeatman et al. Lifespan maturation and degeneration of human brain white matter. *Nature communications*, 5:4932, 2014.

[6] Alexander et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4:170181, December 2017.

[7] Shafto et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14:204, 2014.

Acknowledgments

Alfred P. Sloan Foundation
Gordon and Betty Moore Foundation
eScience Institute
Advancing Data-Intensive Discovery in All Fields
Google Cloud Platform