Appendix: Supplementary Material

A Proof of Theorem and Lemmas

Lemma 1. *Let* $\mathbf{X}$ *be the* $(N+1)$*-mode matricization of* $\mathcal{X}$. *Denote* $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_I]$ *where each* $\mathbf{x}_i$ *is a column of* $\mathbf{X}$*, then*

$$\lambda_{\max} = 2/M \max\{|\mathbf{x}_i^{\mathrm{T}}\mathbf{y}|; i = 1, \cdots, I.\}.$$

Moreover, letting $i^* = \arg\max_i |\mathbf{x}_i^{\mathrm{T}}\mathbf{y}|$ *and* $(i_1^*, \cdots, i_N^*)$ *represents its corresponding indices in tensor space, then the initial non-zero solution of (11), denoted as* $(\sigma, \{\mathbf{w}^{(n)}\})$*, is given by*

$$\sigma = \epsilon, \mathbf{w}^{(1)} = sign(\mathbf{x}_{i^*}^{\mathrm{T}}\mathbf{y})\mathbf{1}_{i_1^*},\ \mathbf{w}^{(n)} = \mathbf{1}_{i_n^*}, \forall n = 2, \cdots, N.$$

where $\mathbf{1}_{i_n^*}$ *is a vector with all 0's except for a* 1 *in the* i_n^**-th coordinate.*

Proof. By using multilinear algebra, the problem (8) can be equivalently written as

$$\begin{aligned} \underset{\{\sigma, w^{(n)}\}}{min} \frac{1}{M}\|\mathbf{y} - \mathbf{X}(\sigma\mathbf{w}^{(N)} \otimes \cdots \otimes \mathbf{w}^{(1)}\|_2^2 + \lambda\sigma \prod\nolimits_{n=1}^{N} \|\mathbf{w}^{(n)}\|_1 + \alpha\sigma^2 \prod\nolimits_{n=1}^{N} \|\mathbf{w}^{(n)}\|_2^2 \\ \text{s.t.} \quad \sigma \geq 0,\ \|\mathbf{w}^{(n)}\|_1 = 1,\ n = 1, \cdots, N. \end{aligned} \tag{1}$$

where $\otimes$ denotes the Kronecker product operator.

This problem has the same $\lambda_{\max}$ as its corresponding elastic net problem by considering $(\sigma\mathbf{w}^{(N)} \otimes \cdots \otimes \mathbf{w}^{(1)})$ as a whole. Thus $\lambda_{\max}$ and the initial non-zero solution can be obtained as above by the Karush-Kuhn-Tucker (KKT) optimality conditions for the elastic net problem. □

Lemma 2. *If there exists* s *and* i_n *with* $|s| = \epsilon, n = 1, \cdots, N$ *such that*

$$\Gamma(s\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \cdots, \mathbf{1}_{i_N}; \lambda) \leq \Gamma(\{\mathbf{0}\}; \lambda), \tag{2}$$

it must be true that $\lambda \leq \lambda_0$.

Proof. By assumption, we can expand (2) as

$$J(s\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \cdots, \mathbf{1}_{i_N}) + \lambda\Omega(s\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \cdots, \mathbf{1}_{i_N}) \leq J(\{\mathbf{0}\}).$$

It follows that

$$\begin{aligned} \lambda &\leq \frac{1}{\epsilon}(J(\{\mathbf{0}\}) - J(s\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \cdots, \mathbf{1}_{i_N})) \\ &\leq \frac{1}{\epsilon}(J(\{\mathbf{0}\}) - \underset{\{i_1, \cdots, i_N\}, s = \pm\epsilon}{min} J(s\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \cdots, \mathbf{1}_{i_N})) \\ &= \lambda_0. \end{aligned}$$

□

Lemma 3. *For any* t *with* $\lambda_{t+1} = \lambda_t$*, we have* $\Gamma(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}; \lambda_{t+1}) \leq \Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_{t+1}) - \xi$.

Proof. This is obviously true if the backward step is taken since $\Gamma(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}; \lambda_t) \leq \Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_t) - \xi$ and $\lambda_{t+1} = \lambda_t$. So we only need to consider the forward step when $\lambda_{t+1} = \lambda_t$. If the claim is not true, then

$$J(\sigma_t, \{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) < \lambda_t\Omega(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \lambda_t\Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\}) + \xi = \lambda_t\epsilon + \xi.$$

That is,

$$\lambda_{t+1} = \lambda_t > \frac{1}{\epsilon}(J(\sigma_t, \{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \xi),$$

which contradicts with the fact that $\lambda_{t+1} = \min(\lambda_t, \frac{1}{\epsilon}(J(\sigma_t, \{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \xi))$. □

Lemma 4. *For any* t *with* $\lambda_{t+1} < \lambda_t$*, we have* $\Gamma(\hat{\mathbf{w}}_t^{(n)} + s_{i_n}\mathbf{1}_{i_n}; \lambda_t) > \Gamma(\hat{\mathbf{w}}_t^{(n)}; \lambda_t) - \xi$.

Proof. First of all, when $\lambda_{t+1} < \lambda_t$, it holds that $\Omega(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) = \Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\}) + \epsilon$. From $\lambda_{t+1} = \min(\lambda_t, \frac{1}{\epsilon}(J(\sigma_t, \{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \xi))$ and $\lambda_{t+1} < \lambda_t$, we know that

$$J(\sigma_t, \{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \xi = \lambda_{t+1}\epsilon = \lambda_{t+1}(\Omega(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\})),$$

that is, $\Gamma(\hat{\mathbf{w}}_t^{(n)}; \lambda_{t+1}) - \xi = \Gamma(\hat{\mathbf{w}}_{t+1}^{(n)}; \lambda_{t+1})$. Then we have

$$\begin{aligned}
\Gamma(\hat{\mathbf{w}}_t^{(n)}; \lambda_t) - \xi &= \Gamma(\hat{\mathbf{w}}_t^{(n)}; \lambda_{t+1}) - \xi + (\lambda_t - \lambda_{t+1})\Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\}) \\
&= \Gamma(\hat{\mathbf{w}}_{t+1}^{(n)}; \lambda_{t+1}) + (\lambda_t - \lambda_{t+1})\Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\}) \\
&= \Gamma(\hat{\mathbf{w}}_{t+1}^{(n)}; \lambda_t) + (\lambda_{t+1} - \lambda_t)(\Omega(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}) - \Omega(\sigma_t, \{\mathbf{w}_t^{(n)}\})) \\
&= \Gamma(\hat{\mathbf{w}}_{t+1}^{(n)}; \lambda_t) + (\lambda_{t+1} - \lambda_t)\epsilon < \Gamma(\hat{\mathbf{w}}_{t+1}^{(n)}; \lambda_t) = \min\{\Gamma(\hat{\mathbf{w}}_t^{(n)} + s_{i_n}\mathbf{1}_{i_n}; \lambda_t)\}. \square
\end{aligned}$$

Theorem 1. *For any t such that $\lambda_{t+1} < \lambda_t$, we have $(\sigma_t, \{\mathbf{w}_t^{(n)}\}) \to (\sigma(\lambda_t), \{\widetilde{\mathbf{w}}^{(n)}(\lambda_t)\})$ as $\epsilon, \xi \to 0$, where $(\sigma(\lambda_t), \{\widetilde{\mathbf{w}}^{(n)}(\lambda_t)\})$ denotes a coordinate-wise minimum point of Problem (7).*

Proof. First, by Lemma 3, we have $\Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_t) \leq \Gamma(\sigma_{t-1}, \{\mathbf{w}_{t-1}^{(n)}\}; \lambda_{t-1}) - \xi$ when $\lambda_t = \lambda_{t-1}$. Then it is easy to verify the series of inequalities

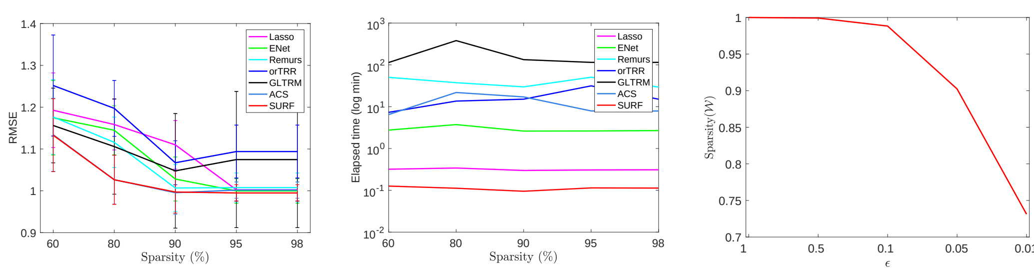
$$\Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_t) \leq \Gamma(\sigma_{t-1}, \{\mathbf{w}_{t-1}^{(n)}\}; \lambda_{t-1}) - \xi \leq \cdots \leq \Gamma(\sigma_{t-p}, \{\mathbf{w}_{t-p}^{(n)}\}; \lambda_{t-p}) - p\xi \tag{3}$$

holds when $\lambda_t = \lambda_{t-1} = \cdots = \lambda_{t-p}$ and p is the value such that $\lambda_{t-p} < \lambda_{t-p-1}$. As $\epsilon, \xi \to 0$, a straightforward consequence of (3) is that the sequence of the objective function values is monotonically decreasing at λ_t, that is,

$$\Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_t) \leq \Gamma(\sigma_{t-1}, \{\mathbf{w}_{t-1}^{(n)}\}; \lambda_t) \leq \cdots \leq \Gamma(\sigma_{t-p}, \{\mathbf{w}_{t-p}^{(n)}\}; \lambda_t). \tag{4}$$

Using Lemma 4, we know that λ_t gets reduced such that $\lambda_{t+1} < \lambda_t$ only occurs in the forward step when $\Gamma(\sigma_{t+1}, \{\mathbf{w}_{t+1}^{(n)}\}; \lambda_t) > \Gamma(\sigma_t, \{\mathbf{w}_t^{(n)}\}; \lambda_t) - \xi$. This means that even by searching over all possible coordinate descent directions in each subproblem (with the size of update fixed at ϵ), the objective function at λ_t can not be further reduced. Since each subproblem is strongly convex w.r.t $(\sigma, \mathbf{w}^{(n)})$, it has a unique solution. Therefore, when $\epsilon, \xi \to 0$ and at the time λ_t gets reduced to λ_{t+1}, we can say a coordinate-wise minimum point of $\Gamma(\cdot)$ is reached for λ_t, which completes the proof. $\square$

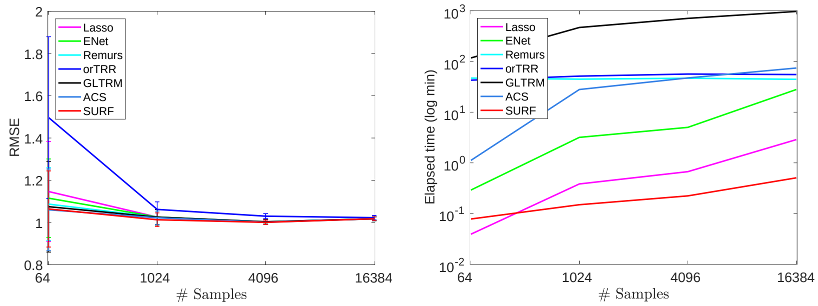
B Some Experimental Results



(a) Test error (b) Running time (c) Sparsity vs. ϵ

Figure 1: Results with increasing sparsity level ($S\%$) of true $\mathbf{W}$ on synthetic 2D data (a)-(b), and (c) sparsity results of $\mathcal{W}$ versus step size for SURF.

It is also interesting to examine the performance of our method with varying sparsity level of $\mathbf{W}$ and sample size. For this purpose, we compare the prediction error and running time (log min) of all methods on the synthetic data. When studying one of factors, other factors are fixed to $M = 500$, $I = 16$, $R = 50$, $S = 80$ (similar as in the main panel). Figure 1(a)-(b) shows the results for the case of $S = \{60, 80, 90, 95, 98\}$ on synthetic 2D data, where $S\%$ indicates the sparsity level of true $\mathbf{W}$. As can be seen from the plots, SURF generates slightly better predictions than other existing methods when the true $\mathbf{W}$ is sparser. Moreover, as shown in Figure 1(c), it is also interesting to note that smaller stepsizes give much less sparsity for SURF. Figure 2 shows the results with increasing number of samples. Overall, SURF gives better predictions at a lower computational cost. Particularly, Remurs and orTRR do not change too much as increasing number of samples, this may due to early stop in the iteration process when searching for optimized solution.



(a) Test error (b) Running time

Figure 2: Results with increasing number of samples on synthetic 2D data.

C Description of Data Preprocessing

We preprocessed the DTI and MRI acquisitions on 656 subjects as follows. T1-weighted MRI data was acquired using the ADNI-2 sequence, and processed using the FreeSurfer[1], followed by [1]. For DTI data, each subject's raw data were aligned to the b0 image using the FSL[2] eddy-correct tool to correct for head motion and eddy current distortions. The gradint table is also corrected accordingly. Non-brain tissue is removed from the diffusion MRI using the Brain Extraction Tool (BET) from FSL [2]. To correct for echo-planar induced (EPI) susceptibility artifacts, which can cause distortions at tissue-fluid interfaces, skull-stripped b0 images are linearly aligned and then elastically registered to their respective preprocessed structural MRI using Advanced Normalization Tools (ANTs[3]) with SyN nonlinear registration algorithm [3]. The resulting 3D deformation fields are then applied to the remaining diffusion-weighted volumes to generate full preprocessed diffusion MRI dataset for the brain network reconstruction. In the meantime, 84 ROIs is parcellated from T1-weighted MRI using Freesufer.

Based on these 84 ROIs, we reconstruct four types of brain connectivity matrices for each subject, using the following four tensor-based deterministic tractography algorithms: Fiber Assignment by Continuous Tracking (FACT) [4], the 2nd-order Runge-Kutta (RK2) [5], interpolated streamline (SL) [6], and the tensorline (TL) [7]. Each resulted connectivity matrix for each subject is 84×84. To avoid computation bias, we normalize each connectivity matrix by dividing by its maximum value, as matrices derived from different tractography methods have different scales and ranges.

References

[1] Liang Zhan, Jiayu Zhou, Yalin Wang, Yan Jin, Neda Jahanshad, Gautam Prasad, Talia M Nir, Cassandra D Leonardo, Jieping Ye, Paul M Thompson, et al. Comparison of nine tractography algorithms for detecting abnormal structural brain networks in alzheimer's disease. *Frontiers in aging neuroscience*, 7:48, 2015.

[2] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

[3] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[4] Susumu Mori, Barbara J Crain, Vadappuram P Chacko, and Peter Van Zijl. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of neurology*, 45(2):265–269, 1999.

[1] `https://surfer.nmr.mgh.harvard.edu`
[2] `http://www.fmrib.ox.ac.uk/fsl`
[3] `http://stnava.github.io/ANTs/`

[5] Peter J Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. In vivo fiber tractography using dt-mri data. *Magnetic resonance in medicine*, 44(4):625–632, 2000.

[6] Thomas E Conturo, Nicolas F Lori, Thomas S Cull, Erbil Akbudak, Abraham Z Snyder, Joshua S Shimony, Robert C McKinstry, Harold Burton, and Marcus E Raichle. Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences*, 96(18):10422–10427, 1999.

[7] Mariana Lazar, David M Weinstein, Jay S Tsuruda, Khader M Hasan, Konstantinos Arfanakis, M Elizabeth Meyerand, Benham Badie, Howard A Rowley, Victor Haughton, Aaron Field, et al. White matter tractography using diffusion tensor deflection. *Human brain mapping*, 18(4):306–321, 2003.