Boosted Sparse and Low-Rank Tensor Regression

Lifang He[1], Kun Chen[2], Wanwan Xu[2], Jiayu Zhou[3], Fei Wang[1]

[1]Department of Healthcare Policy and Research, Weill Cornell Medicine
[2]Department of Statistics, University of Connecticut
[3]Department of Computer Science and Engineering, Michigan State University

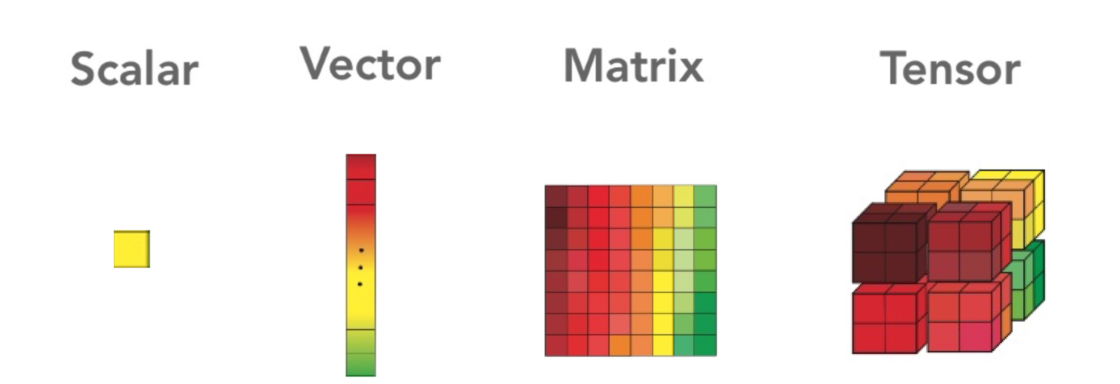
WangLab
Health Data Analytics
Weill Cornell Medicine

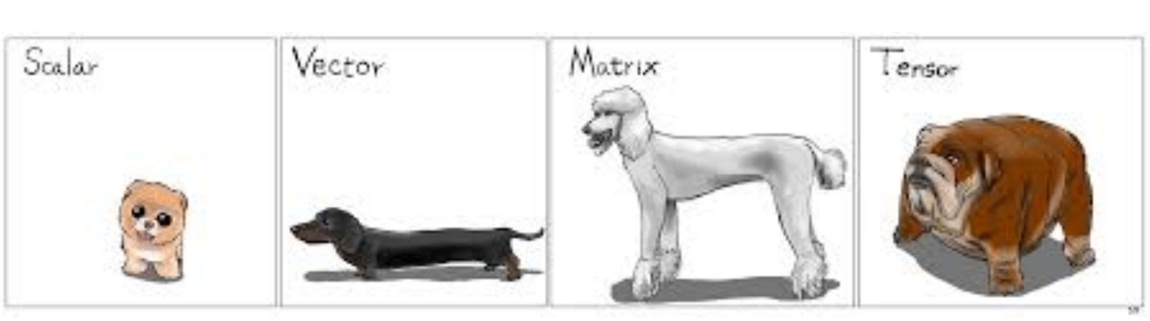
ABSTRACT

With the advanced capabilities for data acquisition, massive multiway data emerge from neuroscience: multi-channel EEG, volumetric or functional MRI, etc. Tensors and low-rank tensor decomposition are very powerful and versatile tools in machine learning for their ability to express and exploit multi-way data, where they are employed to approach a diverse number of tasks. However, tensor regression, which aims to learn a model with multilinear parameters, is especially suitable for applications with multi-directional relatedness, but has not been fully examined. In this paper, we propose a sparse and low-rank tensor regression model to relate a univariate outcome to a feature tensor, and take a divide-and-conquer strategy to simplify the task into a set of sparse unit-rank tensor factorization/regression problems (SURF). We then present a boosted estimation procedure to efficiently trace out the entire solution path. The superior performance of our approach is demonstrated on various real-world and synthetic examples.

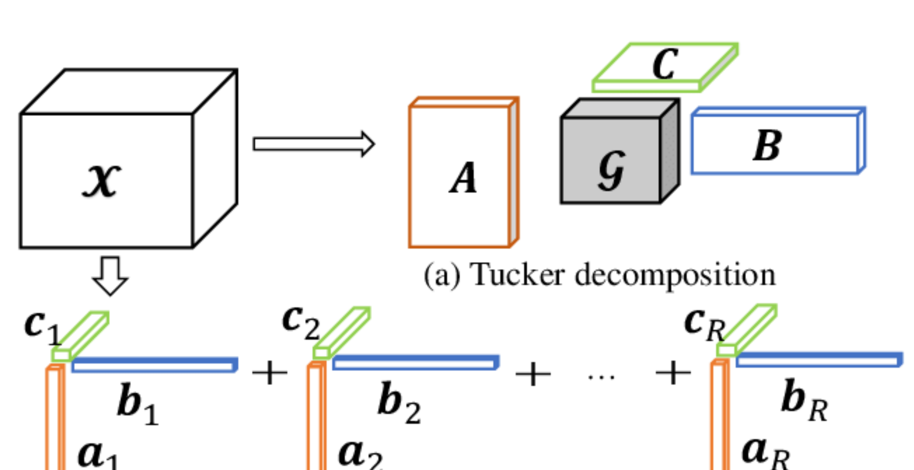
BACKGROUND

Vector to Tensor





Low-Rank Tensor Decomposition



(a) Tucker decomposition

(b) CP decomposition

Sparse and Low-rank Tensor Regression

Tensor Regression Problem

$$\min_{\mathcal{W}} \frac{1}{M}\sum_{m=1}^{M} L(\langle \mathcal{X}^m, \mathcal{W}\rangle, y^m) + \lambda\Omega(\mathcal{W}).$$

$$y^m = \langle \mathcal{X}^m, \mathcal{W}\rangle + \varepsilon^m$$

$\mathcal{X}^m$: input tensor predictor; y^m: scalar response;
W: weight tensor parameter.
L(.) and Ω(.): *loss function and regularizer*.

Formulation

$$\min_{\sigma_r, w_r^{(n)}} \frac{1}{M}\sum_{m=1}^{M}\Big(y^m - \big\langle \mathcal{X}^m, \sum_{r=1}^{R}\sigma_r \mathbf{w}_r^{(1)}\circ\cdots\circ\mathbf{w}_r^{(N)}\big\rangle\Big)^2 + \sum_{r=1}^{R}\sum_{n=1}^{N}\lambda_{r,n}\|\mathbf{w}_r^{(n)}\|_1,$$
$$\text{s.t.} \quad \|\mathbf{w}_r^{(n)}\|_1 = 1,\ n = 1,\cdots,N,\ r = 1,\cdots,R.$$

Challenge:
1. CP rank R needs to be pre-specified;
2. Parameter identifiability issues;
3. Time-consuming for many parameters need to be adjusted.

Solution: Divide-and-Conquer: Sequential pursue for sparse tensor regression.

Sparse Unit-Rank Tensor Factorization (SURF)

$$\widehat{\mathcal{W}}_r = \min_{\mathcal{W}_r} \frac{1}{M}\sum_{m=1}^{M}(y_r^m - \langle \mathcal{X}^m, \mathcal{W}_r\rangle)^2 + \lambda_r\|\mathcal{W}_r\|_1,$$
$$\text{s.t.} \quad \text{rank}(\mathcal{W}_r) \le 1.$$

Where y_r^m is the current residue of response with

$$y_r^m := \begin{cases} y^m, & \text{if } r = 1 \\ y_{r-1}^m - \langle \mathcal{X}^m, \widehat{\mathcal{W}}_{r-1}\rangle, & \text{otherwise.} \end{cases}$$

Formulation

$$\widehat{\mathcal{W}} = \min_{W} \frac{1}{M}\sum_{m=1}^{M}(y^m - \langle \mathcal{X}^m, \mathcal{W}\rangle)^2 + \lambda\|\mathcal{W}\|_1 + \alpha\|\mathcal{W}\|_F^2,$$
$$\text{s.t.} \quad \text{rank}(\mathcal{W}) \le 1.$$

General Version

Let $\mathcal{W} = \sigma\mathbf{w}^{(1)}\circ\cdots\circ\mathbf{w}^{(N)}$, $\mathbf{y} = [y^1,\cdots,y^M]$, and $\mathcal{X} = [\mathcal{X}^1,\cdots,\mathcal{X}^M]$.

$$\min_{\widehat{w}^{(n)}} \frac{1}{M}\|\mathbf{y}^{\mathrm{T}} - \mathbf{Z}^{(-n)\mathrm{T}}\widehat{\mathbf{w}}^{(n)}\|_2^2 + \alpha\beta^{(-n)}\|\widehat{\mathbf{w}}^{(n)}\|_2^2 + \lambda\|\widehat{\mathbf{w}}^{(n)}\|_1,$$

where $\mathbf{Z}^{(-n)} = \mathcal{X}\times_1 \mathbf{w}^{(1)} \times_2 \cdots \times_{n-1} \mathbf{w}^{(n-1)} \times_{n+1} \cdots \times_N \mathbf{w}^{(N)}$, $\widehat{\mathbf{w}}^{(n)} = \sigma\mathbf{w}^{(n)}$ and $\beta^{(-n)} = \prod_{l\neq n}\|\mathbf{w}^{(l)}\|_2^2$.

Subproblem

$$\min_{\widehat{\mathbf{w}}^{(n)}} \underbrace{\frac{1}{M}\|\widehat{\mathbf{y}} - \widehat{\mathbf{Z}}^{(-n)}\widehat{\mathbf{w}}^{(n)}\|_2^2}_{J(\widehat{\mathbf{w}}^{(n)})} + \lambda\underbrace{\|\widehat{\mathbf{w}}^{(n)}\|_1}_{\Omega(\widehat{\mathbf{w}}^{(n)})}$$

where $\widehat{\mathbf{y}} = (\mathbf{y}, \mathbf{0})^{\mathrm{T}}$ and $\widehat{\mathbf{Z}}^{(-n)} = (\mathbf{Z}^{(-n)}, \sqrt{\alpha\beta^{(-n)}M}\mathbf{I})^{\mathrm{T}}$.

Fast Stagewise/Boosted Optimization

The n-th Objective Function

$$\Gamma(\widehat{\mathbf{w}}^{(n)};\lambda) = J(\widehat{\mathbf{w}}^{(n)}) + \lambda\Omega(\widehat{\mathbf{w}}^{(n)}).$$

Algorithm

Algorithm 1 Fast Stagewise Unit-Rank Tensor Factorization (SURF)

Input: Training data $\mathcal{D}$, a small stepsize $\epsilon > 0$ and a small tolerance parameter ξ [2]

Output: Solution paths of $(\sigma, \{\mathbf{w}^{(n)}\})$.

1: *Initialization*: take a forward step with $(\{\hat{i}_1,\cdots,\hat{i}_N\},\hat{s}) = \underset{\{i_1,\cdots,i_N\},s=\pm\epsilon}{arg\,min} J(s\mathbf{1}_{i_1},\mathbf{1}_{i_2},\cdots,\mathbf{1}_{i_N})$, and

$$\sigma_0 = \epsilon, \mathbf{w}_0^{(1)} = sign(\widehat{s})\mathbf{1}_{\widehat{i}_n},\ \mathbf{w}_0^{(n)} = \mathbf{1}_{\widehat{i}_n}(n\neq 1),\ \lambda_0 = (J(\{\mathbf{0}\}) - J(\sigma_0,\{\mathbf{w}_0^{(n)}\}))/\epsilon. \tag{11}$$

Set the active index sets $I_0^{(n)} = \{\hat{i}_n\}$ for $n = 1,\cdots,N$; $t = 0$.

2: **repeat**

3: *Backward step*:

$$(\widehat{n},\widehat{i}_{\widehat{n}}) := \arg\min_{n,i_n\in I_t^{(n)}} J(\widehat{\mathbf{w}}_t^{(n)} + s_{i_n}\mathbf{1}_{i_n}),\ \text{where } s_{i_n} = -sign(\widehat{w}_{ti_n}^{(n)})\epsilon. \tag{12}$$

if $\Gamma(\widehat{\mathbf{w}}_t^{(\widehat{n})} + s_{\widehat{i}_{\widehat{n}}}\mathbf{1}_{\widehat{i}_{\widehat{n}}};\lambda_t) - \Gamma(\widehat{\mathbf{w}}_t^{(\widehat{n})};\lambda_t) \le -\xi$, **then**

$$\sigma_{t+1} = \|\widehat{\mathbf{w}}_t^{(\widehat{n})} + s_{\widehat{i}_{\widehat{n}}}\mathbf{1}_{\widehat{i}_{\widehat{n}}}\|_1,\ \mathbf{w}_{t+1}^{(\widehat{n})} = (\widehat{\mathbf{w}}_t^{(\widehat{n})} + s_{\widehat{i}_{\widehat{n}}})/\sigma_{t+1},\ \mathbf{w}_{t+1}^{(-\widehat{n})} = \mathbf{w}_t^{(-\widehat{n})},$$

$$\lambda_{t+1} = \lambda_t,\quad I_{t+1}^{(n)} := \begin{cases} I_t^{(\widehat{n})}\setminus\{\widehat{i}_{\widehat{n}}\}, & \text{if } w_{(t+1)\widehat{i}_{\widehat{n}}}^{(\widehat{n})} = 0 \\ I_t^{(n)}, & \text{otherwise.}\end{cases}$$

4: **else** *Forward step*:

$$(\widehat{n},\widehat{i}_{\widehat{n}},\widehat{s}_{\widehat{i}_{\widehat{n}}}) := \arg\min_{n,i_n,s=\pm\epsilon} J(\widehat{\mathbf{w}}_t^{(n)} + s_{i_n}\mathbf{1}_{i_n}), \tag{13}$$

$$\sigma_{t+1} = \|\widehat{\mathbf{w}}_t^{(\widehat{n})} + \widehat{s}_{\widehat{i}_{\widehat{n}}}\mathbf{1}_{\widehat{i}_{\widehat{n}}}\|_1,\ \mathbf{w}_{t+1}^{(\widehat{n})} = (\widehat{\mathbf{w}}_t^{(\widehat{n})} + \widehat{s}_{\widehat{i}_{\widehat{n}}})/\sigma_{t+1},\ \mathbf{w}_{t+1}^{(-\widehat{n})} = \mathbf{w}_t^{(-\widehat{n})},$$

$$\lambda_{t+1} = \min[\lambda_t, \frac{J(\sigma_t,\{\mathbf{w}_t^{(n)}\}) - J(\sigma_{t+1},\{\mathbf{w}_{t+1}^{(n)}\}) - \xi}{\Omega(\sigma_{t+1},\{\mathbf{w}_{t+1}^{(n)}\}) - \Omega(\sigma_t,\{\mathbf{w}_t^{(n)}\})}],\quad I_{t+1}^{(n)} := \begin{cases} I_t^{(n)}\cup\{\widehat{i}_n\}, & \text{if } n = \widehat{n} \\ I_t^{(n)}, & \text{otherwise.}\end{cases}$$

5: Set $t = t + 1$.

6: **until** $\lambda_t \le 0$

Forward Step

The solution at each forward step is

$$(\widehat{n},\widehat{i}_{\widehat{n}}) := \underset{n,i_n}{arg\,max}\ 2|\widehat{\mathbf{e}}^{(n)\mathrm{T}}\widehat{\mathbf{Z}}^{(-n)}\mathbf{1}_{i_n}| - \epsilon Diag(\widehat{\mathbf{Z}}^{(-n)\mathrm{T}}\widehat{\mathbf{Z}}^{(-n)})^{\mathrm{T}}\mathbf{1}_{i_n},$$

$$\widehat{s} = sign(\widehat{\mathbf{e}}^{(n)\mathrm{T}}\widehat{\mathbf{Z}}^{(-\widehat{n})}\mathbf{1}_{\widehat{i}_{\widehat{n}}})\epsilon,$$

where $\widehat{\mathbf{e}}^{(n)} = \widehat{\mathbf{y}} - \widehat{\mathbf{Z}}^{(-n)}\widehat{\mathbf{w}}^{(n)}$ is a constant at each iteration.

Backward Step

The solution at each backward step is

$$(\widehat{n},\widehat{i}_{\widehat{n}}) := \underset{n,i_n\in I^{(n)}}{arg\,min}\ 2sign(\widehat{w}_{i_n}^{(n)})\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{Z}}^{(-n)}\mathbf{1}_{i_n} + \epsilon Diag(\widehat{\mathbf{Z}}^{(-n)\mathrm{T}}\widehat{\mathbf{Z}}^{(-n)})^{\mathrm{T}}\mathbf{1}_{i_n}.$$

Computation Analysis

At each iteration, $\mathbf{Z}^{(-n)}$ $(n\neq\hat{n})$ can be updated by

$$\mathbf{Z}_{t+1}^{(-n)} = \frac{1}{\sigma_{t+1}}(\sigma_t\mathbf{Z}_t^{(-n)} + \mathbf{Z}_t^{(-n,-\hat{n})}\times_{\hat{n}}\widehat{s}_{\widehat{i}_{\widehat{n}}}\mathbf{1}_{\widehat{i}_{\widehat{n}}}),$$

where $(-n,-\hat{n})$ denotes every mode except n and $\hat{n}$.

The computational complexity of our approach per iteration is

$$O(M\sum_{n\neq\hat{n}}^{N}(\prod_{s\neq n,\hat{n}} I_s + 5I_n) + 2MI_{\hat{n}})$$

In contrast, the ACS algorithm has to be run for each fixed λ, and within each of suchproblems, each iteration requires

$$O(M\prod_{n=1}^{N} I_n)$$

Convergence Analysis

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

Lemma 2: For any t with $\lambda_{t+1} < \lambda_t$, we have $\Gamma(\hat{\mathbf{w}}_t^{(n)} + s_{i_n}\mathbf{1}_{i_n};\lambda_t) > \Gamma(\hat{\mathbf{w}}_t^{(n)};\lambda_t) - \xi$.

Lemma 1 and Lemma 2 proves the following convergence theorem.

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 the SURF problem.

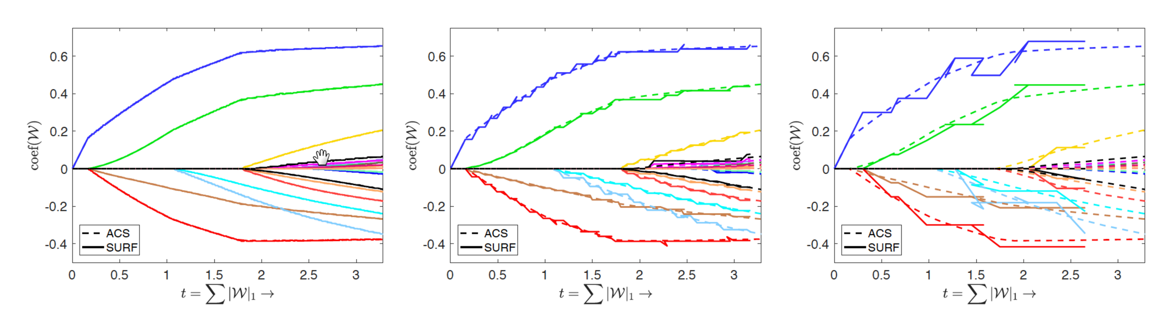
Experiments

Summarization of Compared Methods

Table 1: Compared methods. α and λ are regularized parameters; R is the CP rank.

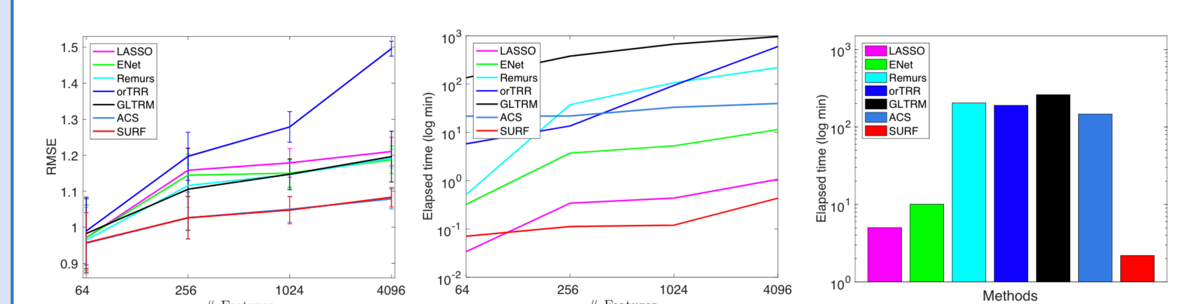
Methods	LASSO	ENet	Remurs	orTRR	GLTRM	ACS	SURF
Input Data Type	Vector	Vector	Tensor	Tensor	Tensor	Tensor	Tensor
Regularization	ℓ_1 (**w**)	ℓ_1/ℓ_2 (**w**)	Nuclear/ℓ_1 ($\mathcal{W}$)	ℓ_2 ($\mathbf{W}^{(n)}$)	ℓ_1/ℓ_2 ($\mathbf{W}^{(n)}$)	ℓ_1/ℓ_2 ($\mathcal{W}_r$)	ℓ_1/ℓ_2 ($\mathcal{W}_r$)
Rank Explored	—	—	—	Optimized	Fixed	Increased	Increased
Hyperparameters	λ	α,λ	λ_1,λ_2	α, R	α,λ,R	α,λ,R	α,λ,R

Empirical Analysis on Synthetic Data



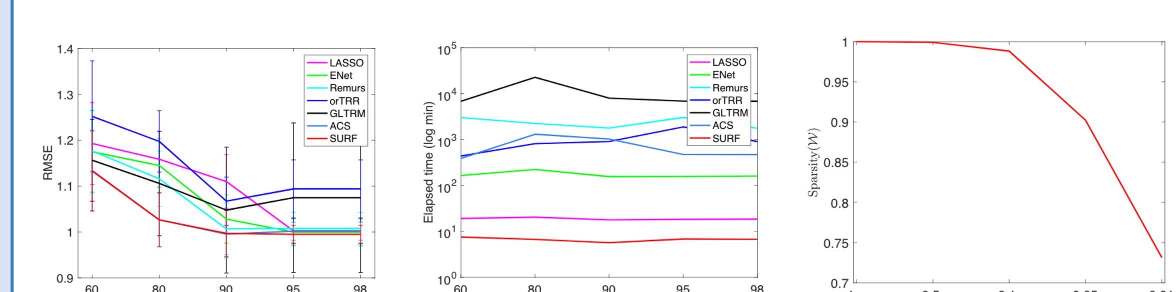
(a) $\epsilon = 0.01$ (b) $\epsilon = 0.1$ (c) $\epsilon = 0.5$

Figure 1: Comparison of solution paths of SURF (solid line) and ACS (dashed line) with different step sizes on synthetic data. The path of estimates $\mathcal{W}$ for each λ is treated as a function of $t = \|\mathcal{W}\|_1$.



(a) Test error (b) Running time (c) Running time on MRI data

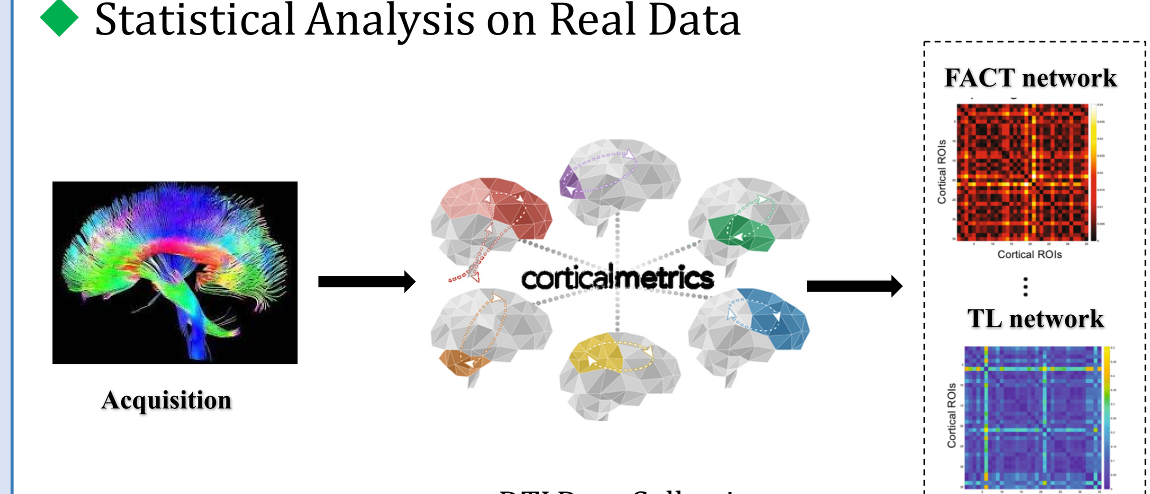
Figure 2: Results with increasing number of features on synthetic 2D data (a)-(b), and (c) real 3D MRI data of features $240\times175\times176$ with fixed hyperparameters (without cross validation).



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

Figure 3: 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.

Statistical Analysis on Real Data



DTI Data Collection

Table 2: Results on different DTI datasets (mean ± std.). Column 2 indicates the used metrics RMSE, Sparsity of Coefficients (SC) and CPU execution time (in mins).

Datasets	Metrics	Comparative Methods						
		LASSO	ENet	Remurs	orTRR	GLTRM	ACS	SURF
DTI_{fact}	RMSE	2.94±0.34	2.92±0.32	2.91±0.32	3.48±0.21	3.09±0.35	2.81±0.24	2.81±0.23
	Sparsity	0.99±0.01	0.97±0.01	0.66±0.13	0.00±0.00	0.90±0.10	0.92±0.02	0.95±0.01
	Time	6.4±0.3	46.6±4.6	161.3±9.3	27.9±5.6	874.8±29.6	60.8±24.4	1.7±0.2
DTI_{rk2}	RMSE	3.18±0.36	3.16±0.42	2.97±0.30	3.76±0.44	3.26±0.46	2.90±0.31	2.91±0.32
	Sparsity	0.99±0.01	0.95±0.03	0.37±0.09	0.00±0.00	0.91±0.06	0.93±0.02	0.94±0.01
	Time	5.7±0.3	42.4±2.9	155.0±10.7	10.2±0.1	857.4±22.5	63.0±21.6	5.2±0.8
DTI_{sl}	RMSE	3.06±0.34	2.99±0.34	2.93±0.27	3.56±0.41	3.14±0.39	2.89±0.38	2.87±0.35
	Sparsity	0.98±0.01	0.95±0.01	0.43±0.17	0.00±0.00	0.87±0.03	0.90±0.03	0.93±0.02
	Time	5.8±0.3	45.0±1.0	163.6±9.0	7.5±0.9	815.4±6.5	66.3±44.9	1.5±0.1
DTI_{tl}	RMSE	3.20±0.40	3.21±0.59	2.84±0.35	3.66±0.35	3.12±0.32	2.82±0.33	2.83±0.32
	Sparsity	0.99±0.01	0.96±0.03	0.44±0.13	0.00±0.00	0.86±0.03	0.90±0.02	0.91±0.02
	Time	5.5±0.2	42.3±1.4	159.6±7.6	26.6±3.1	835.8±9.9	96.7±43.2	3.8±0.5
Combined	RMSE	3.02±0.37	2.89±0.41	2.81±0.31	3.33±0.27	3.26±0.45	2.79±0.31	2.78±0.29
	Sparsity	0.99±0.00	0.97±0.01	0.34±0.22	0.00±0.00	0.91±0.19	0.97±0.01	0.99±0.00
	Time	8.8±0.6	71.9±2.3	443.5±235.7	48.4±8.0	1093.6±49.7	463.2±268.1	6.2±0.5

NOTE: The full paper and code are available at https://github.com/LifangHe/SURF.
Email: lifanghescut@gmail.com