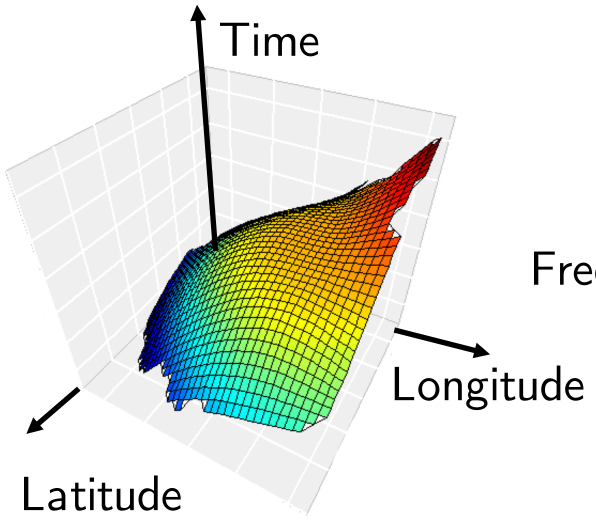
Learning from Multi-Way Data: Simple and Efficient Tensor Regression

Rose Yu, Yan Liu

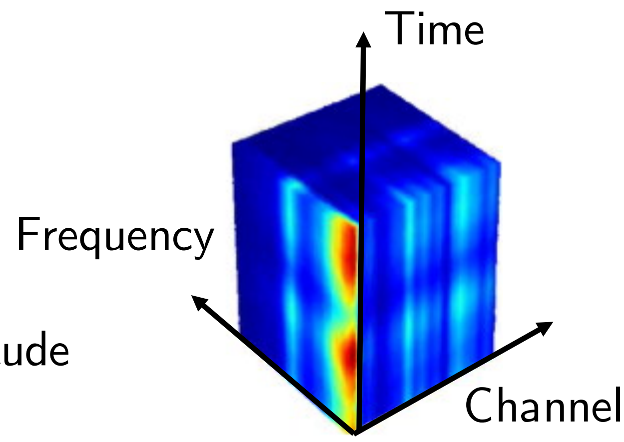
Poster #58, Tue 3-7 pm

Multi-Way Data

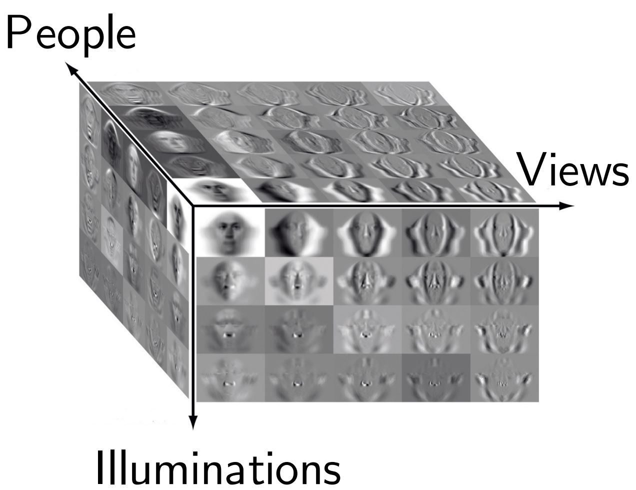
- Massive multi-way data emerges from many fields
  - Climate
  - Neural Science
  - ...
- Multi-way data contains multi-directional correlations



Climate Measurements



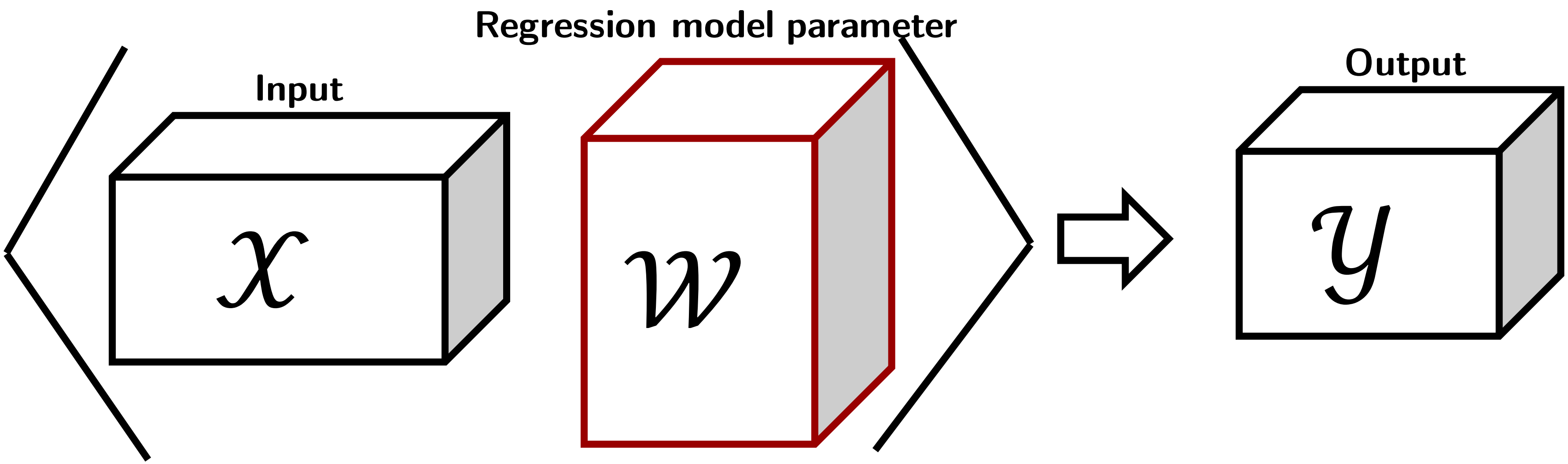
Multi-channel EEG



Facial Recognition

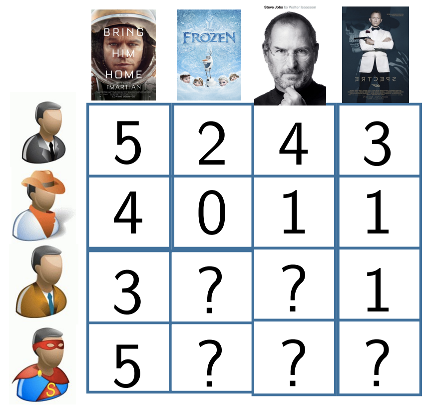
Tensor Regression

- Multi-way data can be naturally represented as **tensors**
- Tensor Regression: large-scale supervised learning from multi-way data
- Goal: learn a regression model with multi-linear parameters



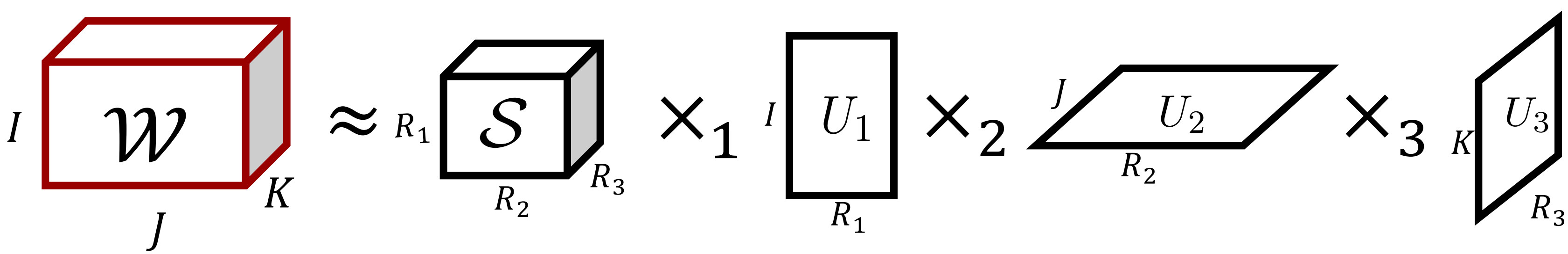
Low-Rank Structure

- Low-rank structures can capture multi-linear correlations



Collaborative Filtering

- Tucker rank: high-order singular value decomposition



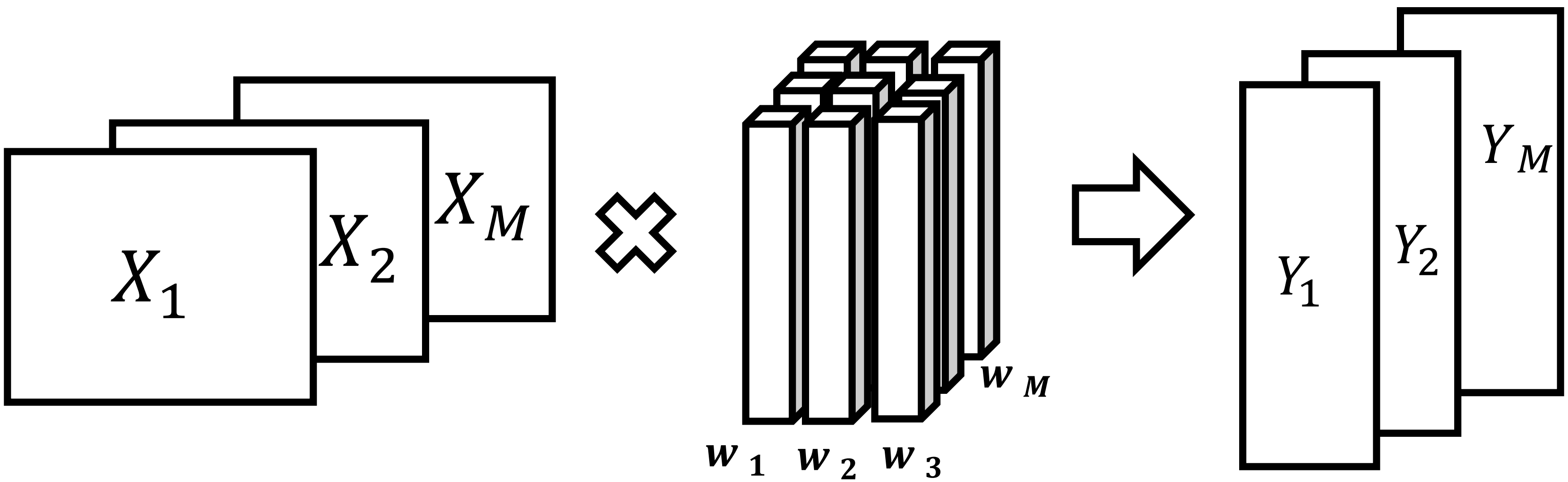
Low-Rank Tensor Regression

- Predictor tensor $\mathcal{X}$; response tensor $\mathcal{Y}$
- Regression model $\langle \mathcal{X}, \mathcal{W} \rangle$: e.g. $\sum_{m=1}^{M} \mathcal{X}_{:,:,m} \mathcal{W}_{:,:,m}$
- Loss function $\mathcal{L}(\hat{\mathcal{Y}}; \mathcal{Y})$: e.g. $\|\hat{\mathcal{Y}} - \mathcal{Y}\|_F^2$
- Goal: Learn a parameter tensor $\mathcal{W}$ with low-rank constraint

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \{ \mathcal{L} (\langle \mathcal{X}, \mathcal{W} \rangle ; \mathcal{Y}) \}$$
$$\text{subject to} \quad \operatorname{rank}(\mathcal{W}) \leq R$$

Examples

- $\mathcal{Y} = cov\langle\mathcal{X}, \mathcal{W}\rangle + \mathcal{E}$ [Zhao et al. 2011]
- $\mathcal{Y} = vec(\mathcal{X})^T vec(\mathcal{W}) + vec(\mathcal{E})$ [Zhou et al. 2013]
- $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$ [Romera-Paredes et al., 2013]



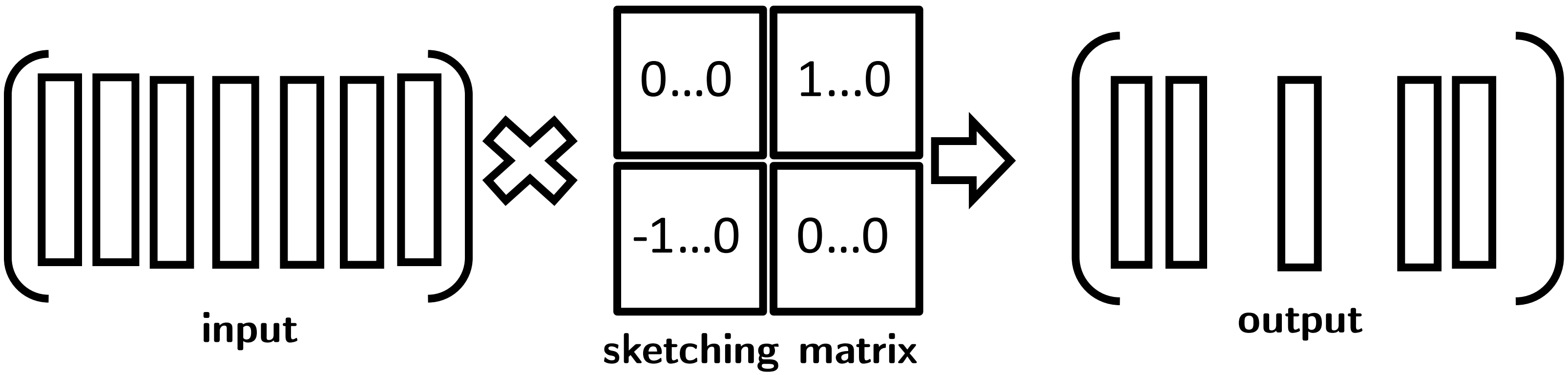
Multi-linear Multi-task Learning [Romera-Paredes et al., 2013]

Related Work

- Alternating least square (ALS) [Romera-Paredes et al. 2013]
  - Empirically effective
  - Sub-optimal solution
- Spectral regularization [Tomiyoka et al. 2014]
  - Nice convex behavior
  - Slow convergence rate
- Greedy matching pursuit [Yu et al. 2014]
  - Fast convergence
  - Memory bottleneck

Subsampled Tensor Projected Gradient (TPG)

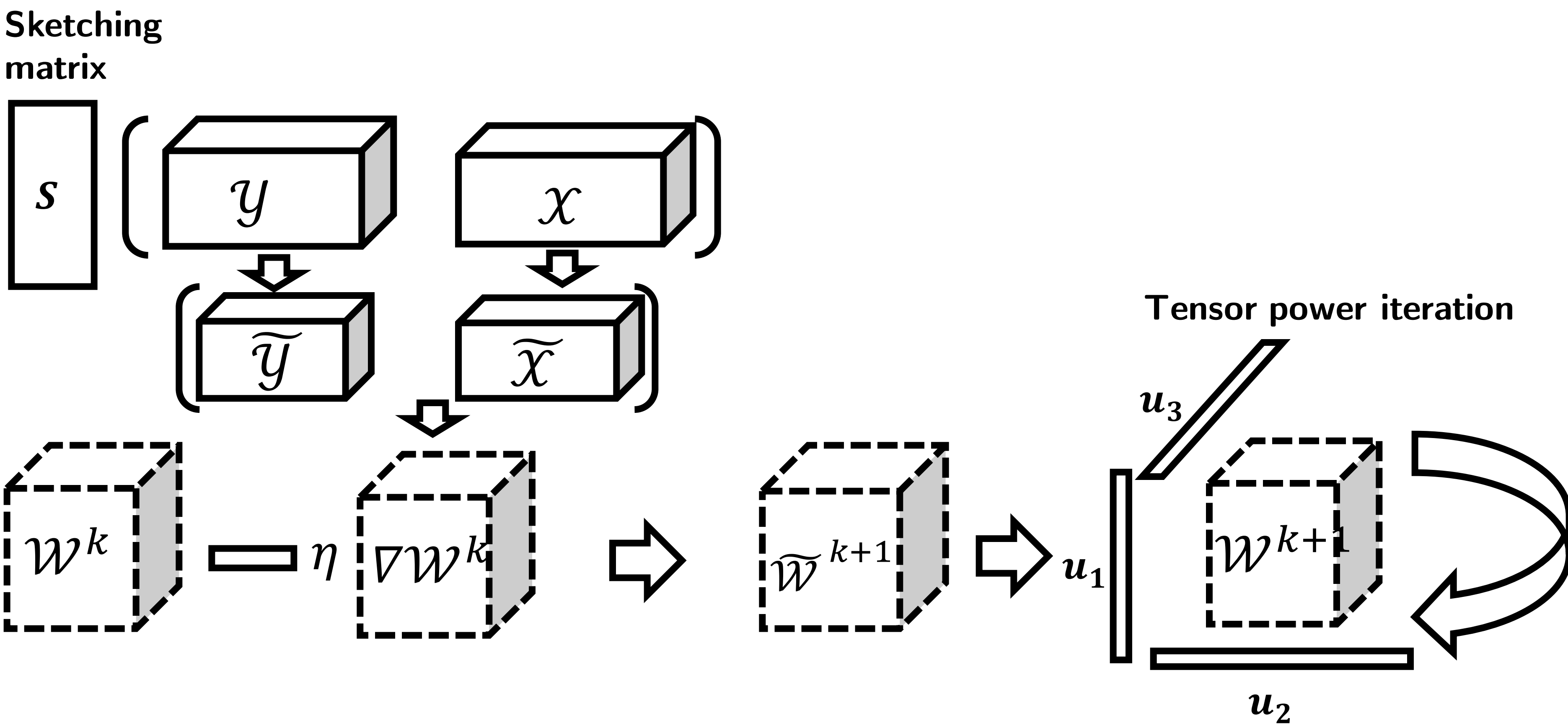
- Data ⇨ Random sketching [Woodruff 2014]
- Model ⇨ Iterative hard thresholding [Thomas and Davies 2009]



- Projected gradient descent: $\mathcal{W}^{k+1} = P_R\left(\mathcal{W}^{k} - \eta \nabla \mathcal{W}^{k}\right)$
  1. Gradient descent step
  2. Low-rank projection step

Subsampled Tensor Projected Gradient (TPG)

- Random sketching as data subsampling
- Iterative hard thresholding as dimensional reduction



Theoretical Analysis

Definition: [*Restricted Isometry Property (RIP)] The isometry constant of* $\mathcal{X}$ *is the smallest number* δ_R *such as the following holds for all* $\mathcal{W}$ *with Tucker rank at most* R.

$$(1-\delta_R)\|\mathcal{W}\|_F^2 \leq \|\langle \mathcal{X}, \mathcal{W}\rangle\|_F^2 \leq (1+\delta_R)\|\mathcal{W}\|_F^2$$

- RIP Characterizes matrices which are nearly orthonormal
- Regression model imposes the RIP assumption w.r.t. matrix rank instead of tensor rank

Theoretical Analysis

Theorem: *For tensor regression model $\mathcal{Y} = \langle \mathcal{X}, \mathcal{W} \rangle + \mathcal{E}$, suppose the predictor tensor $\mathcal{X}$ satisfies RIP condition with isometry constant δ_R <1/3. With step-size $\eta = \frac{1}{1+\delta_R}$, TPG computes a feasible solution $\mathcal{W}^*$ such that the estimation error $\|\mathcal{W} - \mathcal{W}^*\|_F^2 < \frac{1}{1-\delta_{2R}} \|\mathcal{E}\|_F^2$ in at most $\left\lceil \frac{1}{\log 1/\alpha} \log \frac{\|\mathcal{Y}\|_F^2}{\|\mathcal{E}\|_F^2} \right\rceil$ iterations for an universal constant α.*

- Weak assumption on RIP constant
- Converge in a fixed number of iterations
- Memory requirement linear in the problem size

I : Multi-Linear Multi-Task Learning

- Multi-task learning where tasks have multi-directional relatedness
- E.g. predict ratings for restaurants on three aspects: food, service, and overall quality

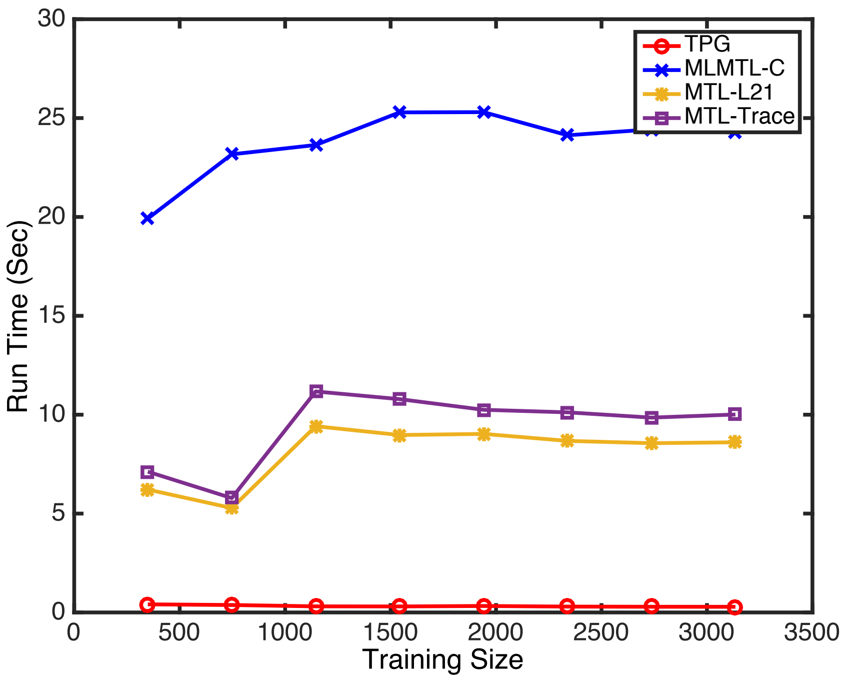
$$\widehat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} \|Y^t - X^T w^t\|_F^2 \right\}$$
$$\text{subject to} \quad \operatorname{rank}(\mathcal{W}) \leq R$$

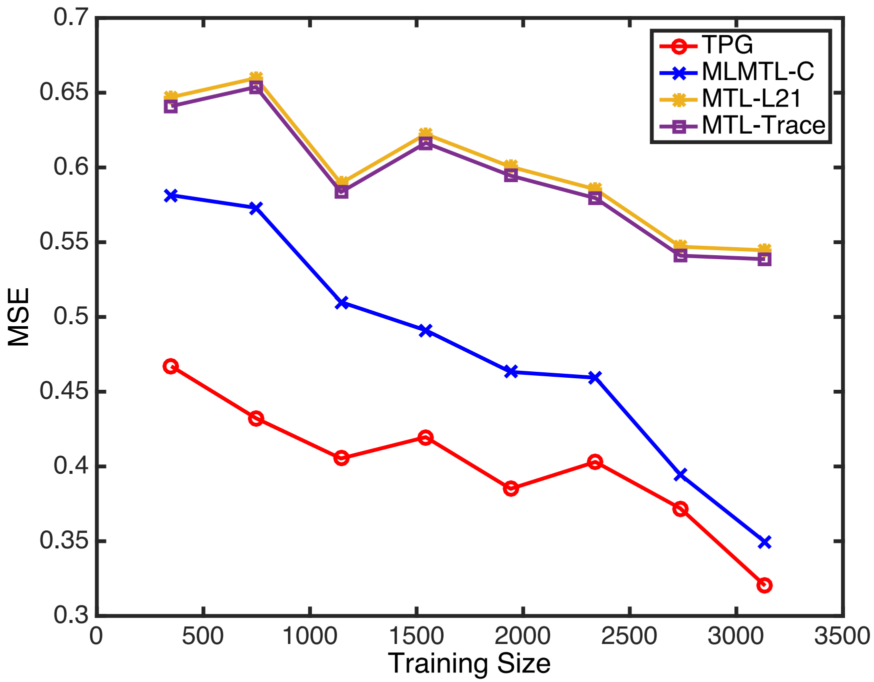
Baselines

- OLS: OLS estimator without low-rank constraint
- THOSVD (De Lathauwer et al., 2000b): a two-step heuristic approach that first solves the least square and then performs truncated singular value decomposition
- Greedy (Yu et al., 2014): a fast tensor regression solution that sequentially estimates rank one sub-space based on Orthogonal Matching Pursuit
- ADMM (Tomiyoka et al., 2014): alternating direction method of multipliers for nuclear norm regularized optimization

Exp: Multi-linear Multi-task Learning

- 45 restaurant features: geographical position, cuisine type, price band, and etc.
- 138 customers with 15,362 rating records





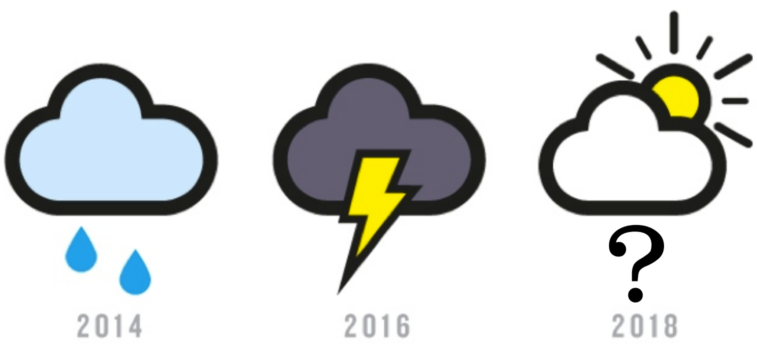
II : Spatio-Temporal Forecasting

- Uses multivariate historical observations to predict future values
- Bayesian spatio-temporal models [Cressie 2008] are not scalable

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left\{ \|\widehat{\mathcal{X}} - \mathcal{Y}\|_F^2 + \mu \sum_{m=1}^{M} \operatorname{trace}\left(\widehat{\mathcal{X}}_{:,:,m} L \widehat{\mathcal{X}}_{:,:,m}^T\right) \right\}$$

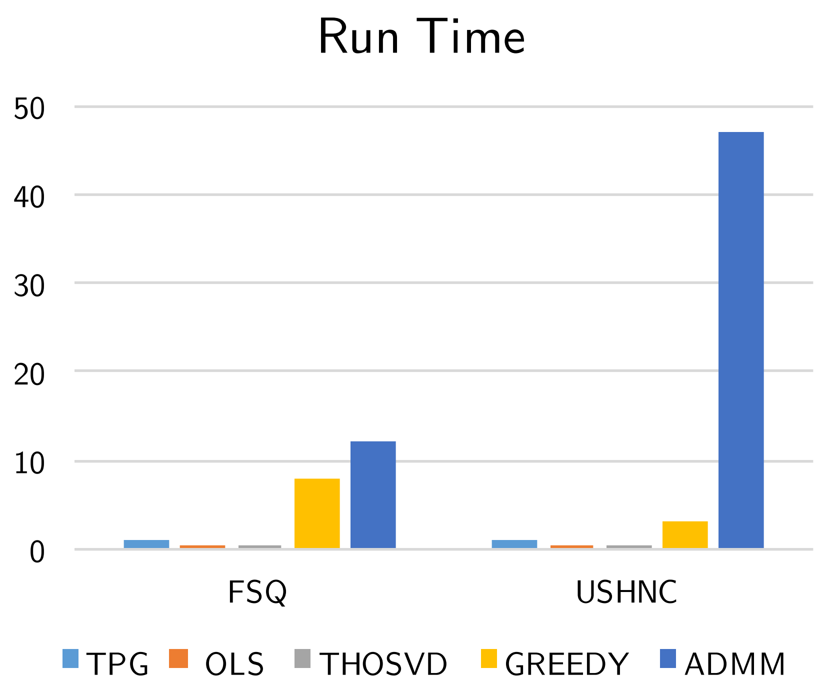
$$\text{subject to } \operatorname{rank}(\mathcal{W}) \leq R$$

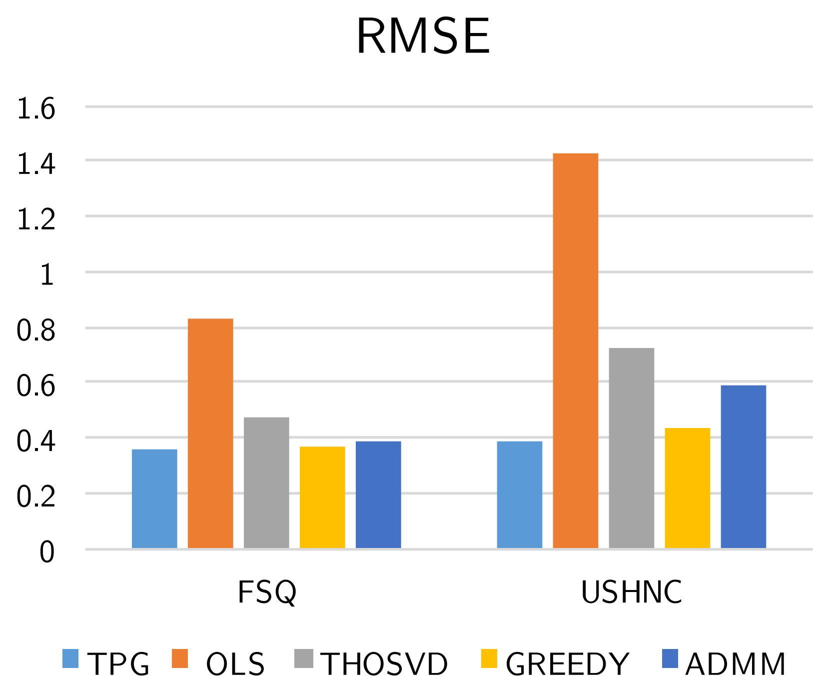
$$\widehat{\mathcal{X}}_{t,p,m} = \left[\mathcal{X}_{t-1,:,m}, \mathcal{X}_{t-2,:,m} \ldots \mathcal{X}_{t-K,:,m}\right] \cdot \mathcal{W}_{:,p,m}$$



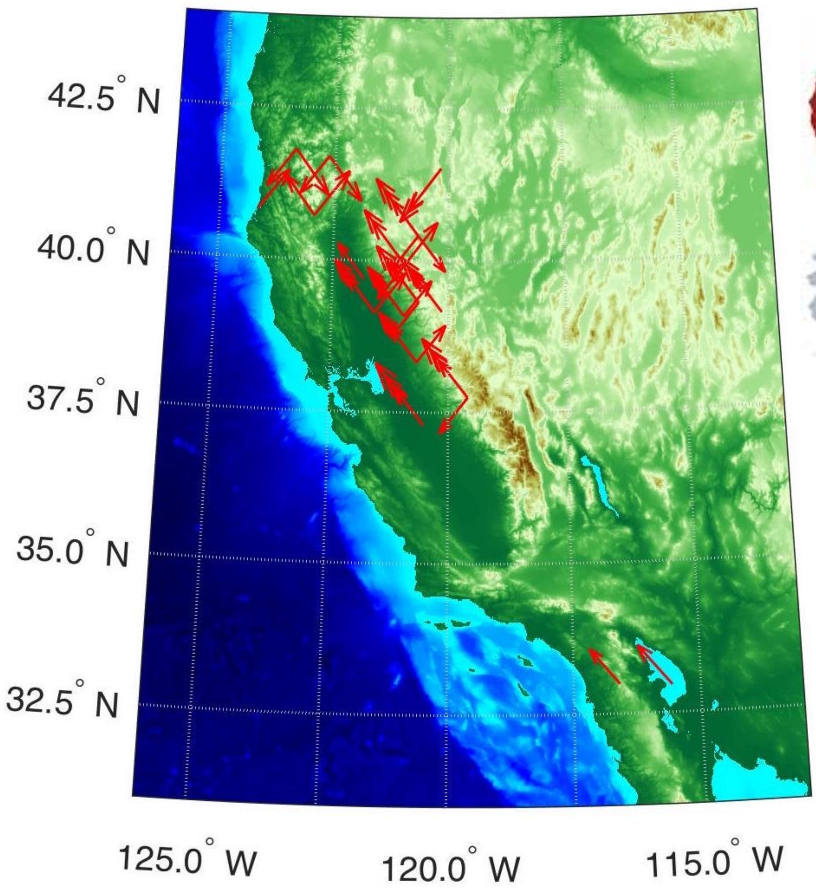
Exp: Spatio-Temporal Forecasting

- **Foursquare**: Hourly check-in records of 739 users in 34 different venue categories over a period of 3,474 hours
- **USHCN**: Five variables collected across more than 1,200 locations and spans over 45,384 time stamps





Exp: Spatio-Temporal Forecasting



Velocity vector plot of learned atmosphere circulation

Discussion & Conclusion

- TPG: Random sketching + iterative hard thresholding
- Fixed number of iterations and linear memory requirement
- Further acceleration with second-order Newton's method

Thank You!

Data available on http://www-bcf.usc.edu/~liu32/data.html

Details about tensor regression: http://roseyu.com/

References

- [1] Rose Yu, Dehua Cheng, Yan Liu. Accelerated Online Low-Rank Tensor Learning for Multivariate Spatio-Temporal Streams *International Conference on Machine Learning* (ICML), 2015
- [2] Rose Yu*, Mohammad Taha Bahadori*, Yan Liu. Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning *Advances in Neural Information Processing Systems* (NIPS), 2014, Spotlight
- [3] Romera-Paredes, Bernardino, et al. "Multilinear multitask learning." *Proceedings of the 30th International Conference on Machine Learning*. 2013.
- [4] Wimalawarne, Kishan, Masashi Sugiyama, and Ryota Tomioka. "Multitask learning meets tensor factorization: task imputation via convex optimization." *Advances in neural information processing systems*. 2014
- [5] Woodruff, David P. "Sketching as a Tool for Numerical Linear Algebra." *Foundations and Trends® in Theoretical Computer Science* 10.1–2 (2014): 1-157.
- [6] Blumensath, Thomas, and Mike E. Davies. "Iterative hard thresholding for compressed sensing." *Applied and Computational Harmonic Analysis* 27.3 (2009): 265-274.
- [7] Cressie, Noel, and Christopher K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.