



**POLYTECHNIQUE
MONTRÉAL**

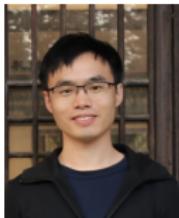
UNIVERSITÉ
D'INGÉNIERIE



Low-Rank Matrix and Tensor Factorization for Speed Field Reconstruction

Xinyu Chen

March 9, 2023



Ph.D. candidate
Xinyu Chen
Polytechnique Montréal



Supervisor
Prof. Nicolas Saunier
Polytechnique Montréal



Co-supervisor
Prof. Lijun Sun
McGill University

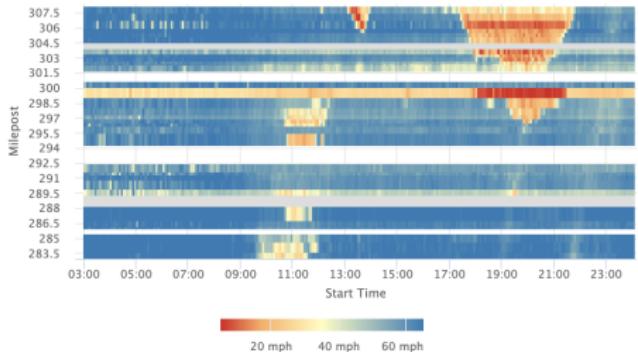
- ① **Slides:** https://xinychen.github.io/slides/MF_TF_SFR.pdf
- ② **Jupyter Notebook:** https://github.com/xinychen/transdim/blob/master/toy-examples/MF_TF_SFR.ipynb

Outline

- **Motivation**
- **Matrix Factorization**
 - Optimization Problem
 - GD vs. SGD vs. ALS
- **Smoothing Matrix Factorization**
 - Spatial/Temporal Smoothing
 - Alternating Minimization
- **Tensor Factorization**
 - Basic Idea
 - CP Tensor Factorization
 - Hankel Tensor and Its Factorization
- **Discussion**
 - Which Model Is Better?
- **Conclusion**

Motivation

- Portland highway traffic speed data¹



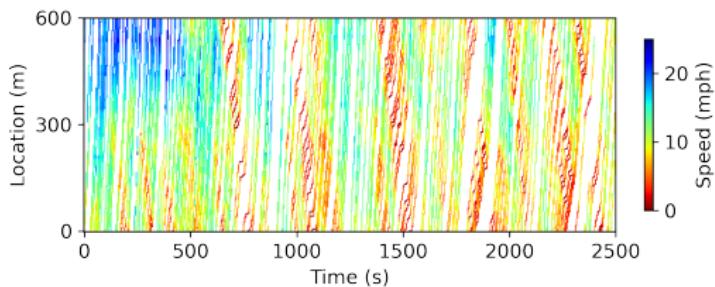
Traffic speed field

Highway network & sensor locations

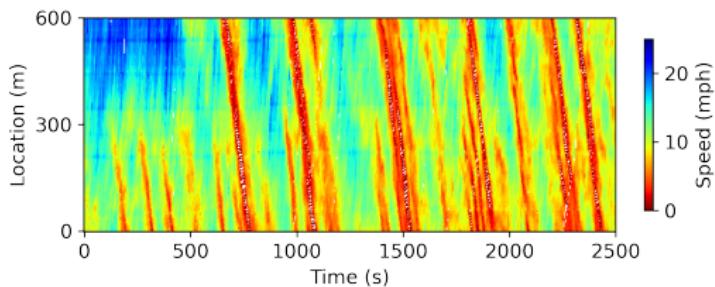
- Speed field $\mathbf{Y} \in \mathbb{R}^{N \times T}$ (N locations & T time steps)
- Speed field shows strong spatial/temporal dependencies

¹<https://portal.its.pdx.edu/home>

Motivation



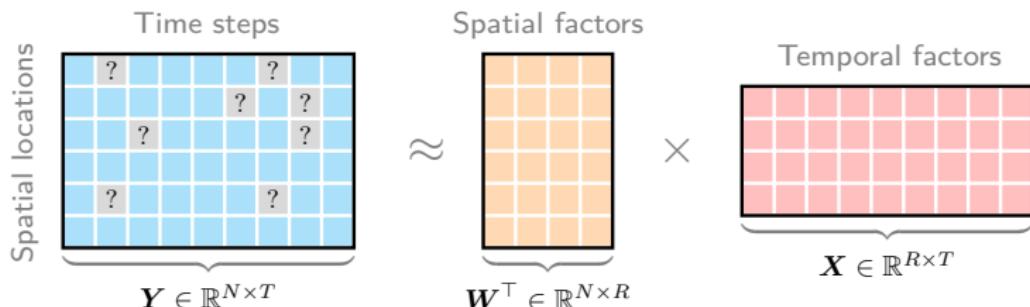
200-by-500 matrix
(NGSIM) \Downarrow Reconstruct speed field
from sparse trajectories?



- How to learn from sparse spatiotemporal data?
- How to characterize spatial/temporal local dependencies?

Matrix Factorization

- Spatiotemporal data can be reconstructed by low-dimensional latent factors!



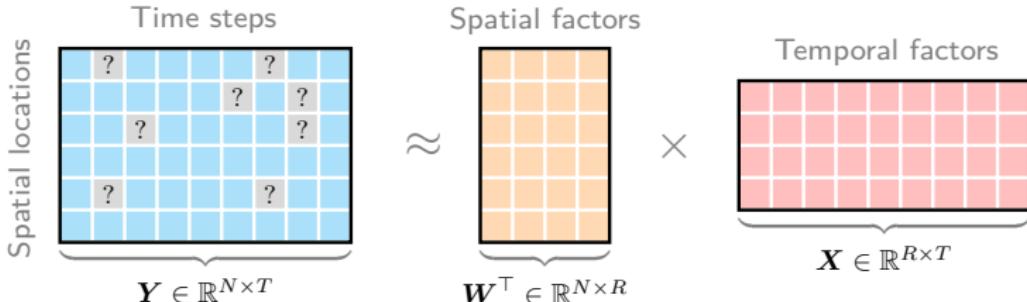
- MF optimization problem

$$\min_{\mathbf{W}, \mathbf{X}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2)$$

with factor matrices \mathbf{W} and \mathbf{X} .

- Objective function $f(\mathbf{W}, \mathbf{X})$ or f ;
- Rank $R \in \mathbb{N}^+$ ($R < \min\{N, T\}$);
- Orthogonal projection $\mathcal{P}_\Omega(\cdot)$.

Matrix Factorization



- MF optimization problem

$$\min_{\mathbf{W}, \mathbf{X}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2)$$

- Orthogonal projection $\mathcal{P}_\Omega : \mathbb{R}^{N \times T} \rightarrow \mathbb{R}^{N \times T}$?

- Simple example: $\mathbf{Y} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ with $\Omega = \{(1, 1), (2, 2)\}$, we have

$$\mathcal{P}_\Omega(\mathbf{Y}) = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad \mathcal{P}_\Omega^\perp(\mathbf{Y}) = \begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix} \quad (\text{On the complement})$$

- Role of regularization (with ρ): avoid overfitting.

Matrix Factorization

- MF optimization problem

$$\min_{\mathbf{W}, \mathbf{X}} \frac{1}{2} \left\| \mathcal{P}_{\Omega}(\mathbf{Y} - \mathbf{W}^{\top} \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2)$$

- Partial derivatives

$$\begin{cases} \frac{\partial f}{\partial \mathbf{W}} = -\mathbf{X} \mathcal{P}_{\Omega}^{\top} (\mathbf{Y} - \mathbf{W}^{\top} \mathbf{X}) + \rho \mathbf{W} \\ \frac{\partial f}{\partial \mathbf{X}} = -\mathbf{W} \mathcal{P}_{\Omega} (\mathbf{Y} - \mathbf{W}^{\top} \mathbf{X}) + \rho \mathbf{X} \end{cases}$$

- Gradient descent (**GD**) vs. Steepest gradient descent (**SGD**)

$$\begin{cases} \mathbf{W} := \mathbf{W} - \alpha \frac{\partial f}{\partial \mathbf{W}} \\ \mathbf{X} := \mathbf{X} - \alpha \frac{\partial f}{\partial \mathbf{X}} \end{cases} \quad \text{vs.} \quad \begin{cases} \alpha := \arg \min_{\alpha} f(\mathbf{W} - \alpha \frac{\partial f}{\partial \mathbf{W}}, \mathbf{X}) \\ \mathbf{W} := \mathbf{W} - \alpha \frac{\partial f}{\partial \mathbf{W}} \\ \beta := \arg \min_{\beta} f(\mathbf{W}, \mathbf{X} - \beta \frac{\partial f}{\partial \mathbf{X}}) \\ \mathbf{X} := \mathbf{X} - \beta \frac{\partial f}{\partial \mathbf{X}} \end{cases}$$

- Fixed step size α (**GD**) vs. optimal step sizes $\{\alpha, \beta\}$ (**SGD**)

Matrix Factorization

- MF optimization problem

$$\min_{\mathbf{W}, \mathbf{X}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2)$$

- Partial derivatives

$$\begin{cases} \frac{\partial f}{\partial \mathbf{W}} = -\mathbf{X} \mathcal{P}_\Omega^\top (\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) + \rho \mathbf{W} \\ \frac{\partial f}{\partial \mathbf{X}} = -\mathbf{W} \mathcal{P}_\Omega (\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) + \rho \mathbf{X} \end{cases}$$

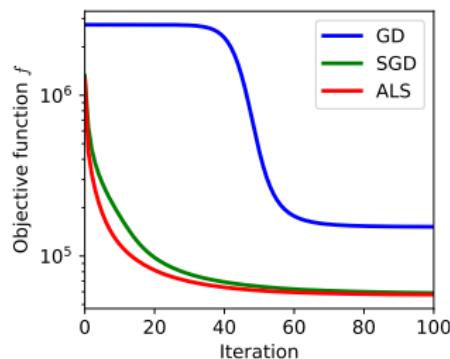
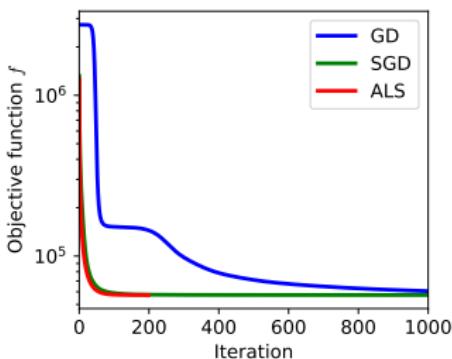
- Alternating least squares (**ALS**)

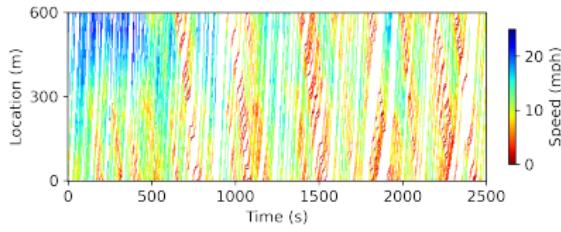
$$\begin{cases} \mathbf{w}_i := \left(\sum_{t:(i,t) \in \Omega} \mathbf{x}_t \mathbf{x}_t^\top + \rho \mathbf{I}_R \right)^{-1} \sum_{t:(i,t) \in \Omega} \mathbf{x}_t y_{i,t} \\ \mathbf{x}_t := \left(\sum_{i:(i,t) \in \Omega} \mathbf{w}_i \mathbf{w}_i^\top + \rho \mathbf{I}_R \right)^{-1} \sum_{i:(i,t) \in \Omega} \mathbf{w}_i y_{i,t} \end{cases}$$

- $\mathbf{w}_i \in \mathbb{R}^R$, $i = 1, 2, \dots, N$ are the columns of \mathbf{W} ;
- $\mathbf{x}_t \in \mathbb{R}^R$, $t = 1, 2, \dots, T$ are the columns of \mathbf{X} .

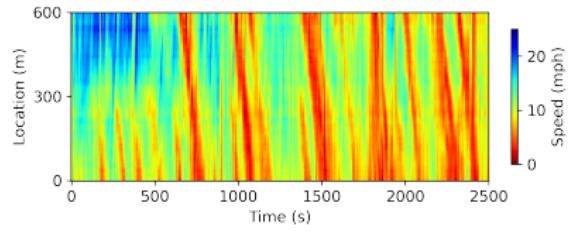
Matrix Factorization

- Objective function f vs. iteration
 - Set rank $R = 10$, weight parameter $\rho = 10$;
 - Set GD step size $\alpha = 10^{-4}$.

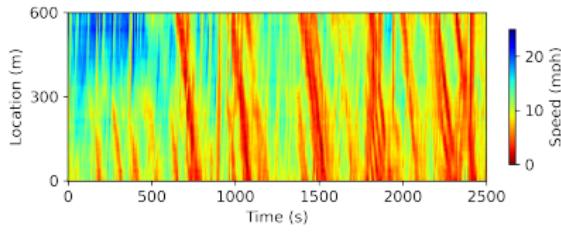




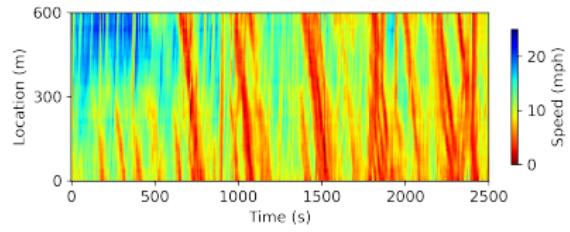
Sparse speed field



MF with GD



MF with SGD



MF with ALS

- Reconstruction errors

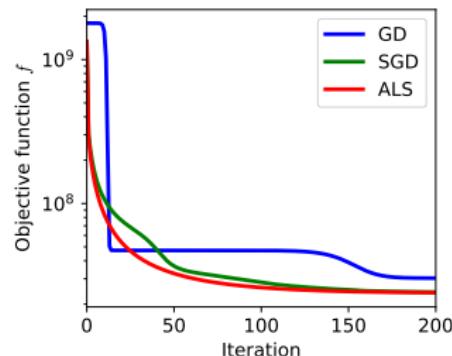
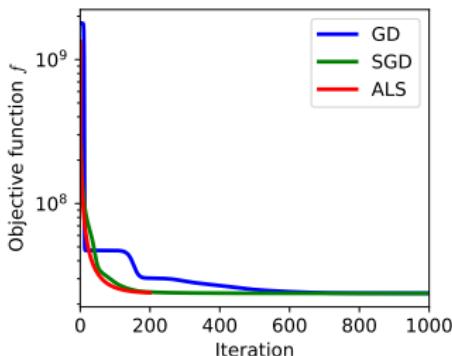
$$\text{MAPE} = \begin{cases} 50.66\% & (\text{GD}) \\ 45.13\% & (\text{SGD}) \\ 45.84\% & (\text{ALS}) \end{cases}$$

$$\text{RMSE} = \begin{cases} 2.33 & (\text{GD}) \\ 2.79 & (\text{SGD}) \quad (\text{mph}) \\ 2.80 & (\text{ALS}) \end{cases}$$

Matrix Factorization

Seattle freeway traffic speed dataset (randomly mask 60% entries)

- Dataset: 323 loop detectors & 8064 time steps (288 per day)
- Objective function f vs. iteration
 - Set rank $R = 10$, weight parameter $\rho = 10^2$;
 - Set GD step size $\alpha = 2 \times 10^{-5}$.



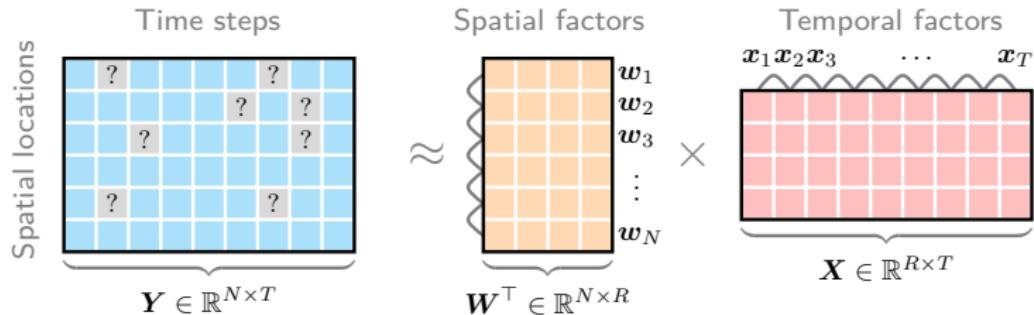
- Reconstruction errors

$$\text{MAPE} = \begin{cases} 9.14\% & (\text{GD}) \\ 9.12\% & (\text{SGD}) \\ 9.13\% & (\text{ALS}) \end{cases}$$

$$\text{RMSE} = \begin{cases} 5.24 & (\text{GD}) \\ 5.24 & (\text{SGD}) \quad (\text{mph}) \\ 5.24 & (\text{ALS}) \end{cases}$$

Smoothing Matrix Factorization

- Spatial/temporal local dependencies are also important!



- Formulate spatial/temporal dependencies

$$\mathbf{W}\Psi_1^\top = \begin{bmatrix} & & & \\ \mathbf{w}_2 - \mathbf{w}_1 & \cdots & \mathbf{w}_N - \mathbf{w}_{N-1} & \\ & & & \end{bmatrix}$$
$$\mathbf{X}\Psi_2^\top = \begin{bmatrix} & & & \\ \mathbf{x}_2 - \mathbf{x}_1 & \cdots & \mathbf{x}_T - \mathbf{x}_{T-1} & \\ & & & \end{bmatrix}$$

Smoothing Matrix Factorization

- Formulate spatial/temporal dependencies

$$\Psi = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \implies \begin{cases} \mathbf{W}\Psi_1^\top & \text{with } \Psi_1 \in \mathbb{R}^{(N-1) \times N} \\ \mathbf{X}\Psi_2^\top & \text{with } \Psi_2 \in \mathbb{R}^{(T-1) \times T} \end{cases}$$

- SMF optimization problem

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{X}} \quad & \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{Y} - \mathbf{W}^\top \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2) \\ & + \frac{\lambda}{2} (\|\mathbf{W}\Psi_1^\top\|_F^2 + \|\mathbf{X}\Psi_2^\top\|_F^2) \end{aligned}$$

- **Alternating minimization**

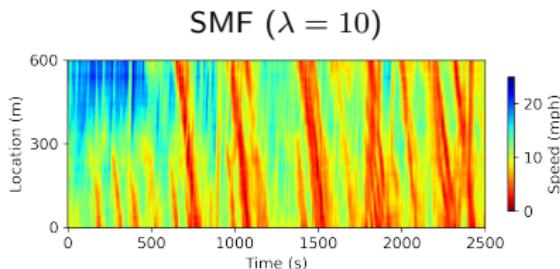
$$\mathbf{W} := \{ \mathbf{W} \mid \frac{\partial f}{\partial \mathbf{W}} = \mathbf{0} \} \quad \mathbf{X} := \{ \mathbf{X} \mid \frac{\partial f}{\partial \mathbf{X}} = \mathbf{0} \}$$

- Solving each matrix equation by the **conjugate gradient** method.

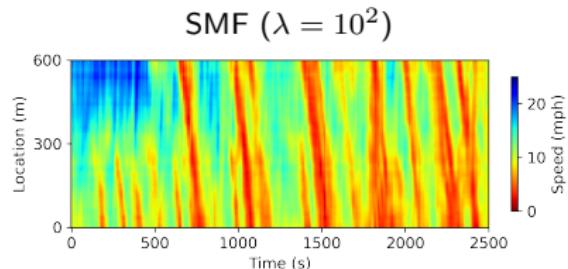
Smoothing Matrix Factorization

- Speed field reconstruction
 - Set rank $R = 10$, weight parameter $\rho = 10$.
 - Recall that the reconstruction errors of MF:

$$\text{MAPE} = \begin{cases} 50.66\% & (\text{GD}) \\ 45.13\% & (\text{SGD}) \\ 45.84\% & (\text{ALS}) \end{cases} \quad \text{RMSE} = \begin{cases} 2.33 & (\text{GD}) \\ 2.79 & (\text{SGD}) \quad (\text{mph}) \\ 2.80 & (\text{ALS}) \end{cases}$$



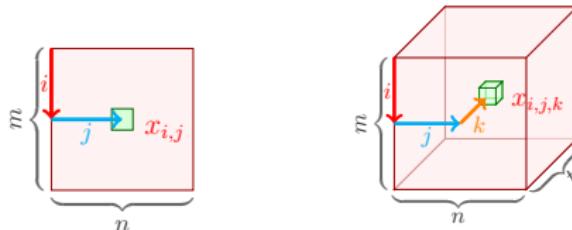
MAPE = 44.06%, RMSE = 2.16mph



MAPE = 48.00%, RMSE = 1.60mph

Tensor Factorization

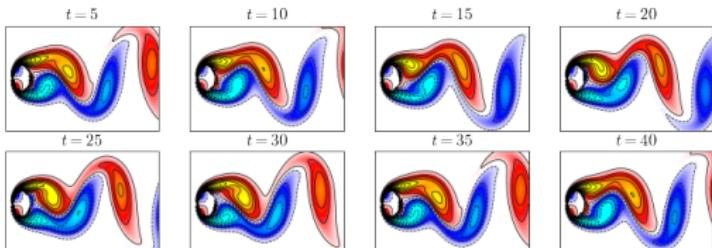
- What is tensor? $\mathbf{X} \in \mathbb{R}^{m \times n}$ vs. $\mathcal{X} \in \mathbb{R}^{m \times n \times t}$



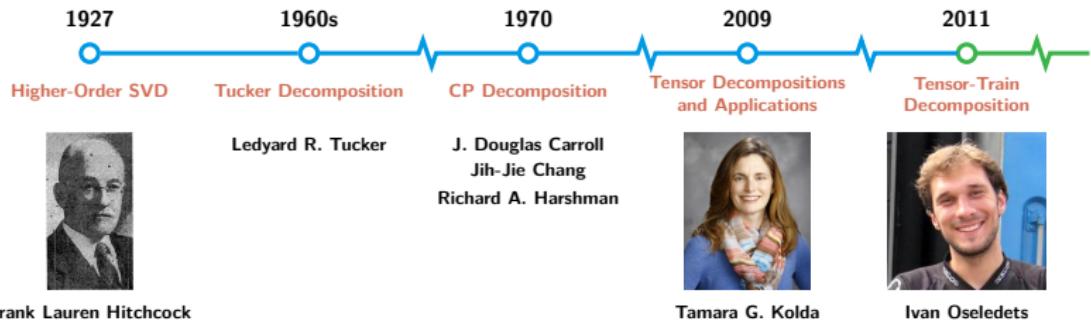
- Tensors are everywhere!



Color image with
RGB channels

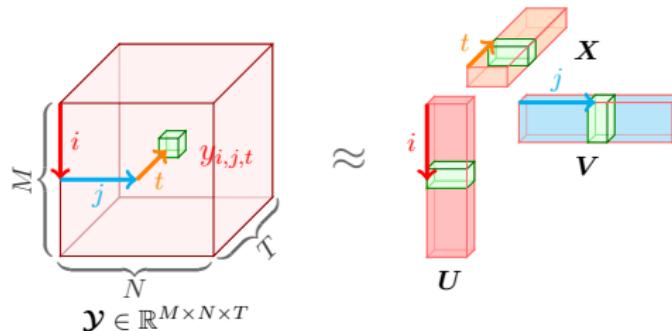


Dynamical system (fluid flow)



CP Tensor Factorization

- Factorize \mathcal{Y} into the combination of three rank- R factor matrices (i.e., low-dimensional latent factors).



- Understanding CP factorization^{2,3}:

$$\left\{ \begin{array}{l} y_{i,j,t} \approx \sum_{r=1}^R u_{i,r} v_{j,r} x_{t,r} \quad (\text{sum of latent factors}) \\ \mathcal{Y} \approx \sum_{r=1}^R \mathbf{u}_r \otimes \mathbf{v}_r \otimes \mathbf{x}_r \quad (\text{sum of rank-one tensors}) \end{array} \right.$$

²CANDECOMP/PARAFAC (CP) decomposition.

³The symbol \otimes denotes the outer product.

Hankel Tensor and Its Factorization

- Hankel matrix

- Given $\mathbf{y} = (1, 2, 3, 4, 5, 6)^\top$ and window length $\tau = 3$, we have

$$\mathcal{H}_\tau(\mathbf{y}) = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix} \in \mathbb{R}^{4 \times 3}$$

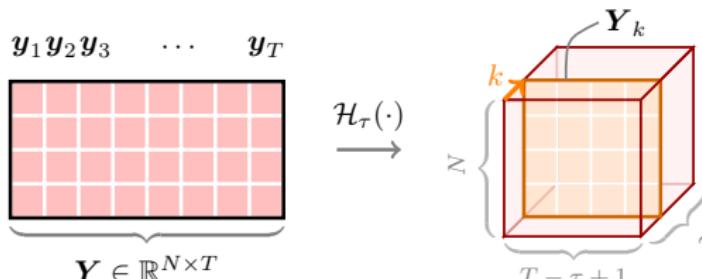
- On time series?

- (Hankelization) Hankel tensor $\mathcal{H}_\tau(\mathbf{Y})$

- Tensor size: $N \times (T - \tau + 1) \times \tau$;

- Slices: $\mathbf{Y}_k = \begin{bmatrix} | & | & | & | \\ \mathbf{y}_k & \mathbf{y}_{k+1} & \cdots & \mathbf{y}_{T-\tau+k} \\ | & | & | & | \end{bmatrix}, k = 1, 2, \dots, \tau$;

- Slice size: $N \times (T - \tau + 1)$.

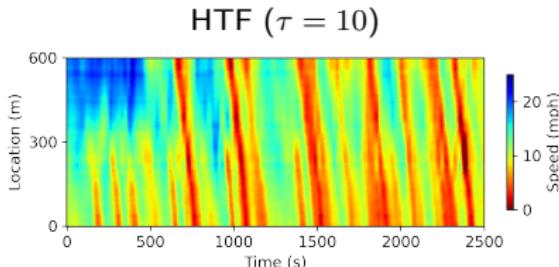


Hankel Tensor and Its Factorization

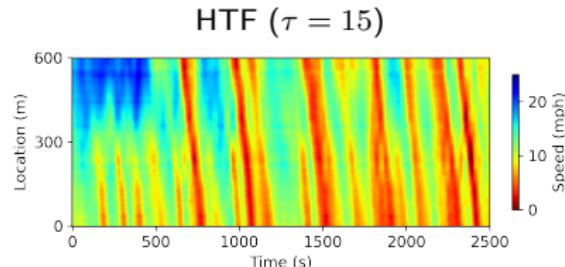
- HTF optimization problem

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V}, \mathbf{X}} \quad & \frac{1}{2} \left\| \mathcal{P}_{\tilde{\Omega}} \left(\mathcal{H}_\tau(\mathbf{Y}) - \sum_{r=1}^R \mathbf{u}_r \otimes \mathbf{v}_r \otimes \mathbf{x}_r \right) \right\|_F^2 \\ & + \frac{\rho}{2} (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{X}\|_F^2) \end{aligned}$$

- HTF's advantage/disadvantage over MF:
 - ✓ Automatic temporal modeling ✗ High memory consumption
- Speed field reconstruction
 - Set rank $R = 10$, weight parameter $\rho = 10$;
 - Recall that SMF: MAPE = 48.00% & RMSE = 1.60mph.

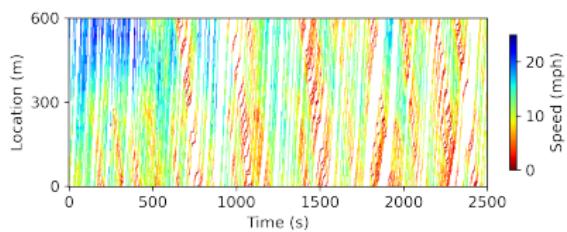


MAPE = 41.40%, RMSE = 1.42mph

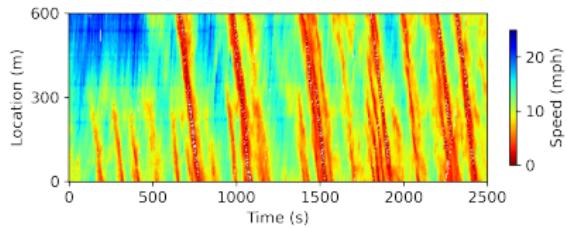


MAPE = 43.97%, RMSE = 1.42mph

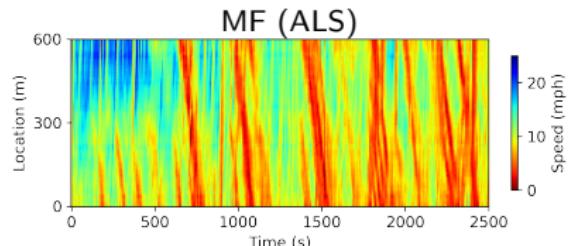
Which Model Is Better?



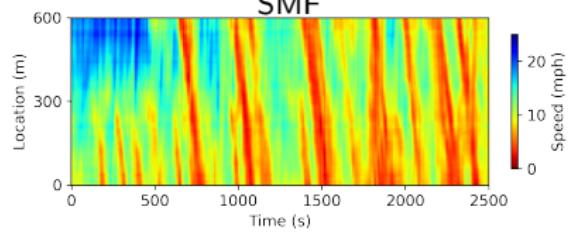
Sparse speed field



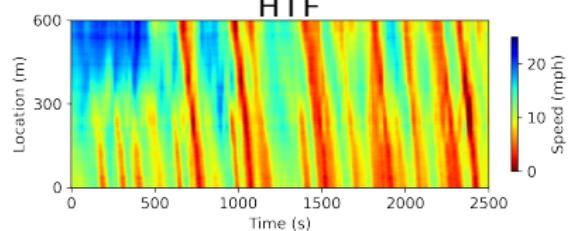
Ground truth speed field



MF (ALS)

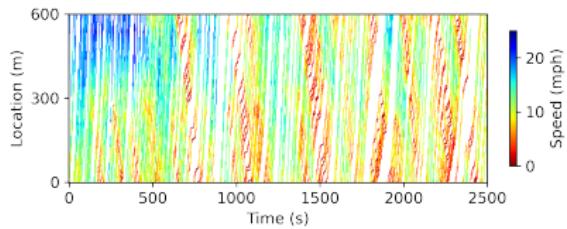


SMF

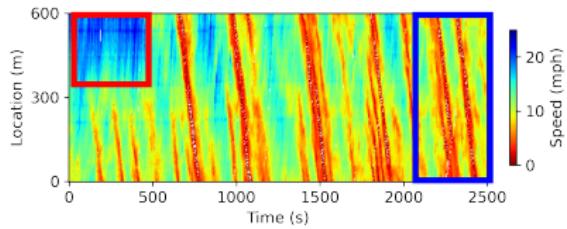


HTF

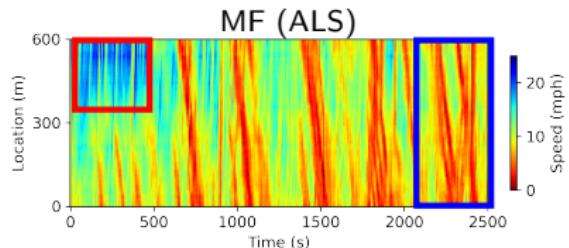
Which Model Is Better?



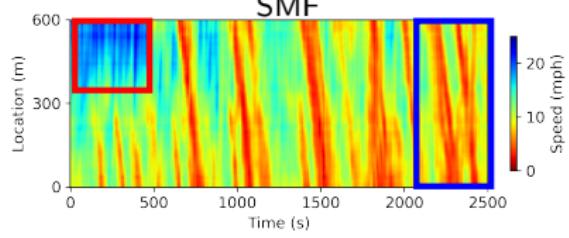
Sparse speed field



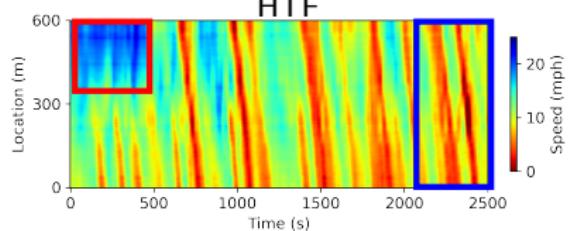
Ground truth speed field



MF (ALS)



SMF



HTF

Which Model Is Better?

- Seattle freeway traffic speed data
 - Randomly mask 60% entries;
 - SMF: set $R = 10$, $\rho = 10^2$, $\lambda = 2 \times 10^2$;
 - HTF: set $\tau = 6$, $R = 10$, $\rho = 10^2$;
 - Reconstruction errors

$$\text{MAPE} = \begin{cases} 9.13\% & (\text{ALS}) \\ 9.01\% & (\text{SMF}) \\ \mathbf{8.67\%} & (\text{HTF}) \end{cases} \quad \text{RMSE} = \begin{cases} 5.24 & (\text{ALS}) \\ 5.14 & (\text{SMF}) \\ \mathbf{5.02} & (\text{HTF}) \end{cases} \text{ (mph)}$$

Which Model Is Better?

- Gray image inpainting
 - Randomly mask 90% pixels;
 - MF: set $R = 50$, $\rho = 10^{-1}$;
 - SMF: set $R = 50$, $\rho = 10^{-1}$, $\lambda = 10$.



Incomplete image



MF



SMF



Ground truth

Conclusion

- How to reconstruct sparse speed field?
 - ✓ Matrix factorization (**MF**) ✓ Tensor factorization (**TF**)
- The importance of spatiotemporal modeling in low-rank methods?
 - Spatial/temporal **smoothing** regularization:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{X}} \quad & \frac{1}{2} \left\| \mathcal{P}_{\Omega}(\mathbf{Y} - \mathbf{W}^{\top} \mathbf{X}) \right\|_F^2 + \frac{\rho}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{X}\|_F^2) \\ & + \frac{\lambda}{2} (\|\mathbf{W} \Psi_1^{\top}\|_F^2 + \|\mathbf{X} \Psi_2^{\top}\|_F^2) \end{aligned}$$

- Automatic temporal modeling via **Hankelization**:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V}, \mathbf{X}} \quad & \frac{1}{2} \left\| \mathcal{P}_{\tilde{\Omega}} \left(\mathcal{H}_{\tau}(\mathbf{Y}) - \sum_{r=1}^R \mathbf{u}_r \otimes \mathbf{v}_r \otimes \mathbf{x}_r \right) \right\|_F^2 \\ & + \frac{\rho}{2} (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{X}\|_F^2) \end{aligned}$$



POLYTECHNIQUE
MONTRÉAL

UNIVERSITÉ
D'INGÉNIERIE



Thanks for your attention!

Any Questions?

About me:

-  Homepage: <https://xinychen.github.io>
-  GitHub: <https://github.com/xinychen> (3K+ stars)
-  Blog: <https://medium.com/@xinyu.chen> (60K+ views)
-  How to reach me: chenxy346@gmail.com