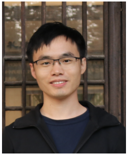
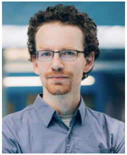
Low-Rank Matrix and Tensor Factorization for Speed Field Reconstruction

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

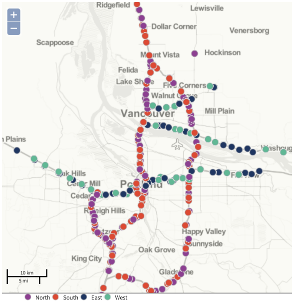
❶ **Slides**: https://xinychen.github.io/slides/MF_TF_SFR.pdf

❷ **Jupyter Notebook**: https://github.com/xinychen/transdim/blob/master/toy-examples/MF_TF_SFR.ipynb

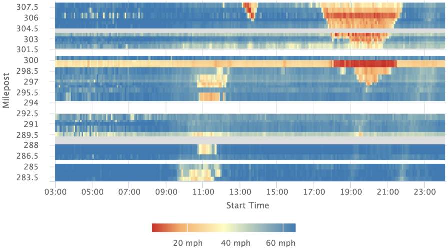
Outline

- **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
  Hankel Tensor and Its Factorization
- **Discussion**
  Which Model Is Better?
- **Conclusion**

- Portland highway traffic speed data[1]



Highway network & sensor locations

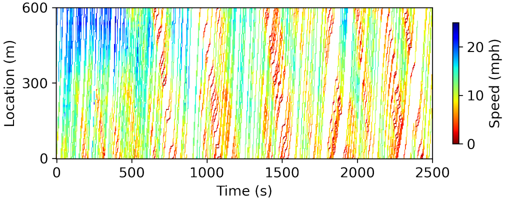


Traffic speed field

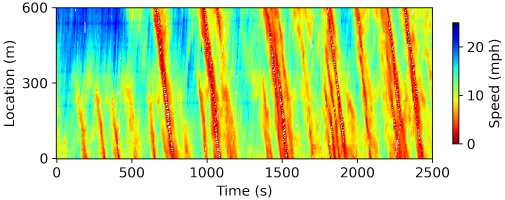
- Speed field $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ (N locations & T time steps)
- Speed field shows strong spatial/temporal dependencies

[1] https://portal.its.pdx.edu/home

Motivation



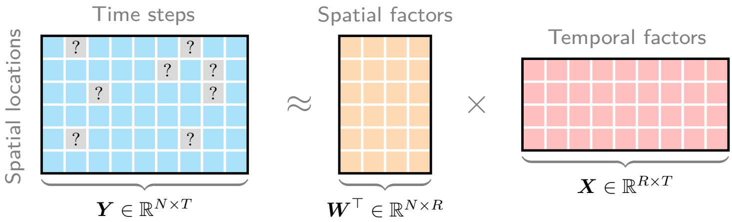
200-by-500 matrix (NGSIM) $\Downarrow$ Reconstruct speed field from sparse trajectories?



- 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!



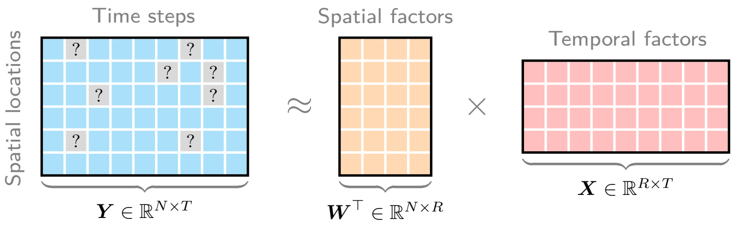
- MF optimization problem

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

with factor matrices $\boldsymbol{W}$ and $\boldsymbol{X}$. ($\|\cdot\|_F^2$ is the squared Frobenius norm.)

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

Matrix Factorization



- MF optimization problem

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

- Orthogonal projection $\mathcal{P}_\Omega : \mathbb{R}^{N \times T} \to \mathbb{R}^{N \times T}$?
    - Simple example: $\boldsymbol{Y} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ with $\Omega = \{(1,1),(2,2)\}$, we have

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

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

Matrix Factorization

- MF optimization problem

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

- Partial derivatives

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

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

$$\begin{cases}\boldsymbol{W} := \boldsymbol{W}-\alpha\dfrac{\partial f}{\partial \boldsymbol{W}}\\ \boldsymbol{X} := \boldsymbol{X}-\alpha\dfrac{\partial f}{\partial \boldsymbol{X}}\end{cases} \quad \text{vs.} \quad \begin{cases}\alpha := \arg\min\limits_{\alpha} \ f(\boldsymbol{W}-\alpha\dfrac{\partial f}{\partial \boldsymbol{W}},\boldsymbol{X})\\ \boldsymbol{W} := \boldsymbol{W}-\alpha\dfrac{\partial f}{\partial \boldsymbol{W}}\\ \beta := \arg\min\limits_{\beta} \ f(\boldsymbol{W},\boldsymbol{X}-\beta\dfrac{\partial f}{\partial \boldsymbol{X}})\\ \boldsymbol{X} := \boldsymbol{X}-\beta\dfrac{\partial f}{\partial \boldsymbol{X}}\end{cases}$$

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

Matrix Factorization

- MF optimization problem

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

- Partial derivatives

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

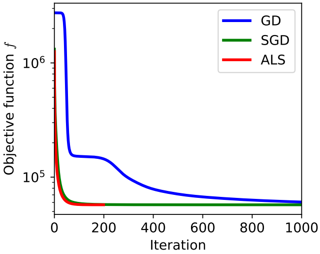
- Alternating least squares (**ALS**)

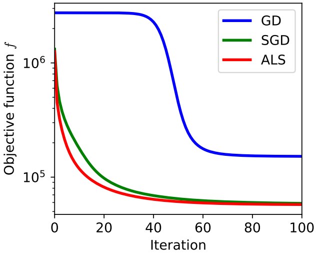
$$\begin{cases}\dfrac{\partial f}{\partial \boldsymbol{W}}=\boldsymbol{0}\\ \dfrac{\partial f}{\partial \boldsymbol{X}}=\boldsymbol{0}\end{cases} \quad\Longrightarrow\quad \begin{cases}\boldsymbol{w}_i:=\Big(\sum\limits_{t:(i,t)\in\Omega}\boldsymbol{x}_t\boldsymbol{x}_t^{\top}+\rho\boldsymbol{I}_R\Big)^{-1}\sum\limits_{t:(i,t)\in\Omega}\boldsymbol{x}_t y_{i,t}\\ \boldsymbol{x}_t:=\Big(\sum\limits_{i:(i,t)\in\Omega}\boldsymbol{w}_i\boldsymbol{w}_i^{\top}+\rho\boldsymbol{I}_R\Big)^{-1}\sum\limits_{i:(i,t)\in\Omega}\boldsymbol{w}_i y_{i,t}\end{cases}$$

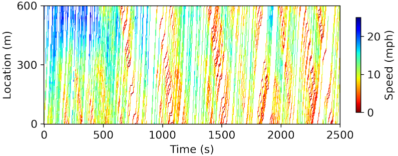
- Latent factors
  - $\boldsymbol{w}_i\in\mathbb{R}^R$, $i=1,2,\ldots,N$ are the columns of $\boldsymbol{W}$;
  - $\boldsymbol{x}_t\in\mathbb{R}^R$, $t=1,2,\ldots,T$ are the columns of $\boldsymbol{X}$.

Matrix Factorization

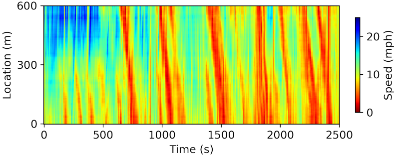
- Objective function f vs. iteration
  - Set rank $R = 10$, weight parameter $\rho = 10$;
  - Set GD step size $\alpha = 10^{-4}$.



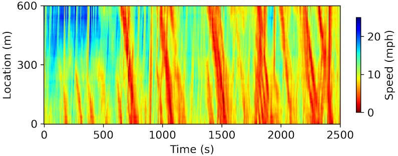




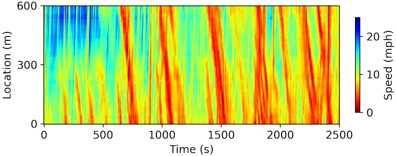
Sparse speed field



MF with GD



MF with SGD



MF with ALS

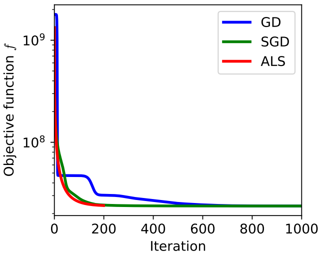
- Reconstruction errors

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

Matrix Factorization

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

- Dataset: 323 loop detectors & 8,064 time steps (288 per day)
- Objective function f vs. iteration
  - Set rank $R = 10$, weight parameter $\rho = 10^2$;
  - Set GD step size $\alpha = 2 \times 10^{-5}$.

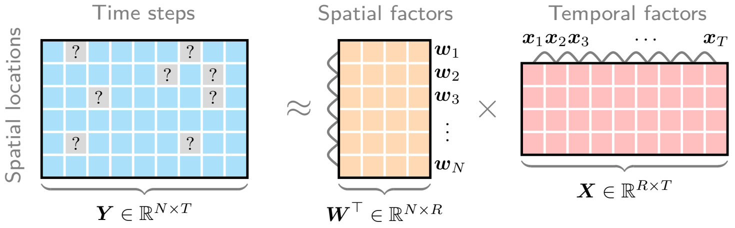


- Reconstruction errors

$$\text{MAPE} = \begin{cases} 9.14\% & \text{(GD)} \\ 9.12\% & \text{(SGD)} \\ 9.13\% & \text{(ALS)} \end{cases} \qquad \text{RMSE} = \begin{cases} 5.24 & \text{(GD)} \\ 5.24 & \text{(SGD)} \\ 5.24 & \text{(ALS)} \end{cases} \quad \text{(mph)}$$

Smoothing Matrix Factorization

- Spatial/temporal local dependencies are also important!



- Formulate spatial/temporal dependencies

$$\boldsymbol{W}\boldsymbol{\Psi}_1^\top = \begin{bmatrix} | & & | \\ \boldsymbol{w}_2 - \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_N - \boldsymbol{w}_{N-1} \\ | & & | \end{bmatrix}$$

$$\boldsymbol{X}\boldsymbol{\Psi}_2^\top = \begin{bmatrix} | & & | \\ \boldsymbol{x}_2 - \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_T - \boldsymbol{x}_{T-1} \\ | & & | \end{bmatrix}$$

Smoothing Matrix Factorization

- Formulate spatial/temporal dependencies

$$\boldsymbol{\Psi}=\begin{bmatrix}-1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1\end{bmatrix} \implies \begin{cases}\boldsymbol{W}\boldsymbol{\Psi}_1^\top & \text{with } \boldsymbol{\Psi}_1\in\mathbb{R}^{(N-1)\times N} \\ \boldsymbol{X}\boldsymbol{\Psi}_2^\top & \text{with } \boldsymbol{\Psi}_2\in\mathbb{R}^{(T-1)\times T}\end{cases}$$

- SMF optimization problem

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{X}}\ & \frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{Y}-\boldsymbol{W}^\top\boldsymbol{X})\right\|_F^2+\frac{\rho}{2}(\|\boldsymbol{W}\|_F^2+\|\boldsymbol{X}\|_F^2) \\ & +\frac{\lambda}{2}(\|\boldsymbol{W}\boldsymbol{\Psi}_1^\top\|_F^2+\|\boldsymbol{X}\boldsymbol{\Psi}_2^\top\|_F^2)\end{aligned}$$

- **Alternating minimization**

$$\boldsymbol{W}:=\{\boldsymbol{W}\mid\frac{\partial f}{\partial\boldsymbol{W}}=\boldsymbol{0}\}\qquad\boldsymbol{X}:=\{\boldsymbol{X}\mid\frac{\partial f}{\partial\boldsymbol{X}}=\boldsymbol{0}\}$$

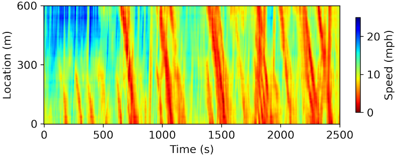
- Solving each matrix equation by the **conjugate gradient** method.

Smoothing Matrix Factorization

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

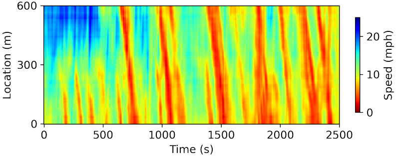
$$\text{MAPE} = \begin{cases} 50.66\% & \text{(GD)} \\ 45.13\% & \text{(SGD)} \\ 45.84\% & \text{(ALS)} \end{cases} \qquad \text{RMSE} = \begin{cases} 2.33 & \text{(GD)} \\ 2.79 & \text{(SGD)} \\ 2.80 & \text{(ALS)} \end{cases} \quad \text{(mph)}$$

SMF ($\lambda = 10$)



MAPE = **44.06%**, RMSE = 2.16mph

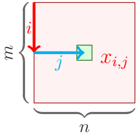
SMF ($\lambda = 10^2$)

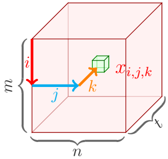


MAPE = 48.00%, RMSE = **1.60mph**

Tensor Factorization

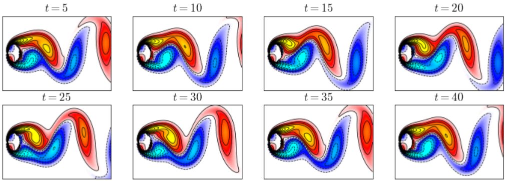
- What is tensor? $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ vs. $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m \times n \times t}$



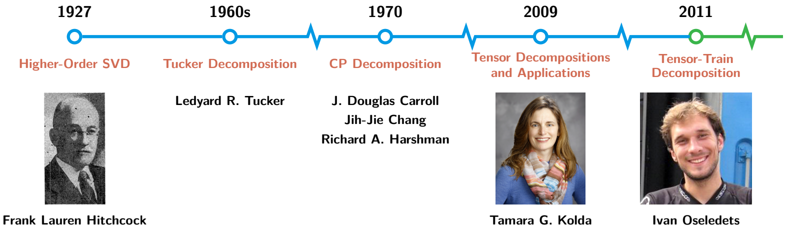


- Tensors are everywhere!

Color image with RGB channels



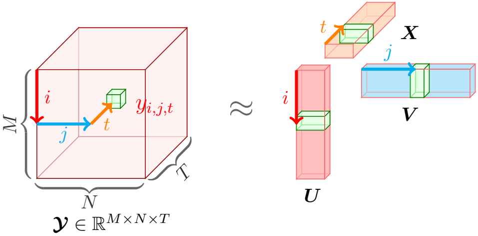
Dynamical system (fluid flow)



1927
Higher-Order SVD
Frank Lauren Hitchcock
1960s
Tucker Decomposition
Ledyard R. Tucker
1970
CP Decomposition
J. Douglas Carroll
Jih-Jie Chang
Richard A. Harshman
2009
Tensor Decompositions and Applications
Tamara G. Kolda
2011
Tensor-Train Decomposition
Ivan Oseledets

CP Tensor Factorization

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



- Understanding CP factorization[2,3]:

$$\begin{cases} y_{i,j,t} \approx \sum_{r=1}^{R} u_{i,r} v_{j,r} x_{t,r} & \text{(sum of latent factors)} \\ \boldsymbol{\mathcal{Y}} \approx \sum_{r=1}^{R} \boldsymbol{u}_r \otimes \boldsymbol{v}_r \otimes \boldsymbol{x}_r & \text{(sum of rank-one tensors)} \end{cases}$$

[2] CANDECOMP/PARAFAC (CP) decomposition.
[3] The symbol $\otimes$ denotes the outer product.

Hankel Tensor and Its Factorization

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

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

  - On time series $\boldsymbol{y} = (y_1, y_2, \ldots, y_5)^\top$ with $\tau = 2$:

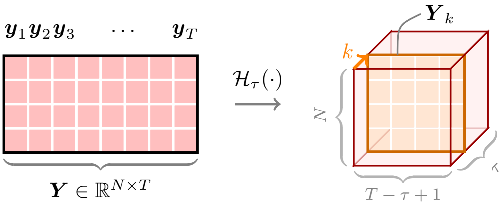
$$\mathcal{H}_\tau(\boldsymbol{y}) = \begin{bmatrix} y_1 & y_2 \\ y_2 & y_3 \\ y_3 & y_4 \\ y_4 & y_5 \end{bmatrix} \approx \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \otimes \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\Longrightarrow \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = \mathcal{H}_\tau^{-1}\left(\begin{bmatrix} v_1x_1 & v_1x_2 \\ v_2x_1 & v_2x_2 \\ v_3x_1 & v_3x_2 \\ v_4x_1 & v_4x_2 \end{bmatrix}\right) = \begin{bmatrix} v_1x_1 \\ (v_1x_2 + v_2x_1)/2 \\ (v_2x_2 + v_3x_1)/2 \\ (v_3x_2 + v_4x_1)/2 \\ v_4x_2 \end{bmatrix}$$

  - Automatic temporal modeling.

Hankel Tensor and Its Factorization

- (Hankelization) Hankel tensor $\mathcal{H}_\tau(\boldsymbol{Y})$
  - Tensor size: $N \times (T-\tau+1) \times \tau$;
  - Slices: $\boldsymbol{Y}_k = \begin{bmatrix} | & | & & | \\ \boldsymbol{y}_k & \boldsymbol{y}_{k+1} & \cdots & \boldsymbol{y}_{T-\tau+k} \\ | & | & & | \end{bmatrix}$, $k = 1, 2, \ldots, \tau$;
  - Slice size: $N \times (T-\tau+1)$.



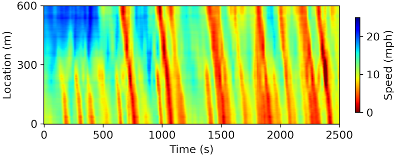
Hankel Tensor and Its Factorization

- HTF optimization problem

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

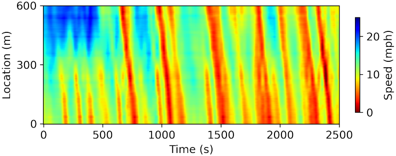
- HTF's advantage/disadvantage over MF:
  - ✓ Automatic temporal modeling ✗ High memory consumption
- Speed field reconstruction
  - ○ Set rank $R=10$, weight parameter $\rho=10$;
  - ○ Recall that SMF: MAPE = 48.00% & RMSE = 1.60mph.

HTF ($\tau=10$)



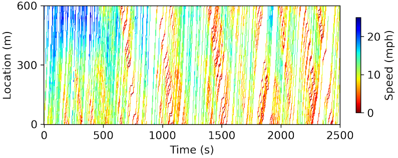
MAPE = **41.40%**, RMSE = **1.42mph**

HTF ($\tau=15$)

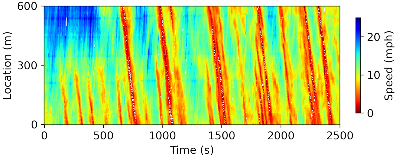


MAPE = 43.97%, RMSE = **1.42mph**

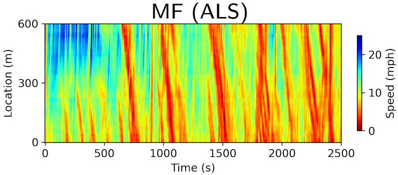
Which Model Is Better?

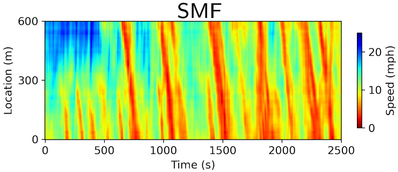


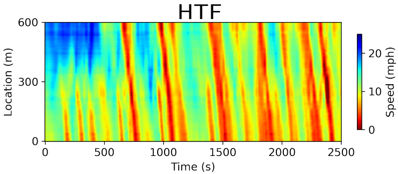
Sparse speed field



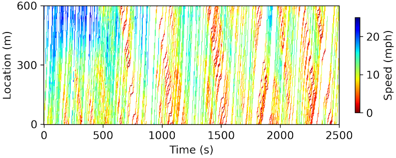
Ground truth speed field



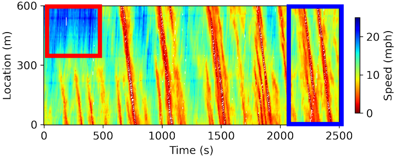




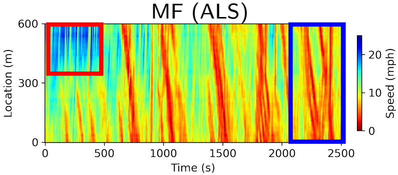
Which Model Is Better?

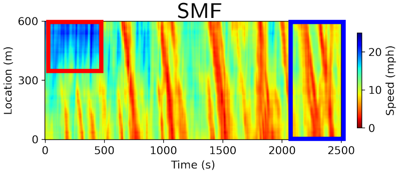


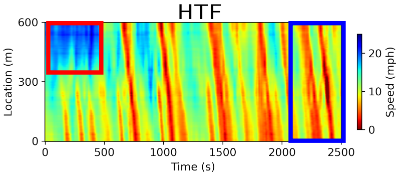
Sparse speed field



Ground truth speed field







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

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

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:

$$\min_{\boldsymbol{W},\boldsymbol{X}} \frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{Y}-\boldsymbol{W}^\top\boldsymbol{X})\right\|_F^2+\frac{\rho}{2}(\|\boldsymbol{W}\|_F^2+\|\boldsymbol{X}\|_F^2)$$
$$+\frac{\lambda}{2}(\|\boldsymbol{W}\boldsymbol{\Psi}_1^\top\|_F^2+\|\boldsymbol{X}\boldsymbol{\Psi}_2^\top\|_F^2)$$

  - Automatic temporal modeling via **Hankelization**:

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

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