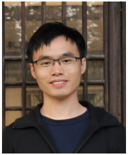
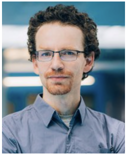
Low-Rank Matrix and Tensor Factorization for Speed Field Reconstruction

Xinyu Chen

March 9, 2023





Ph.D. candidate
Xinyu Chen
Polytechnique Montréal

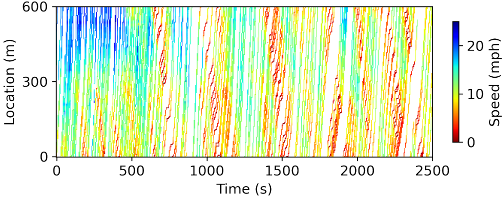
Supervisor
Prof. Nicolas Saunier
Polytechnique Montréal

Co-supervisor
Prof. Lijun Sun
McGill University

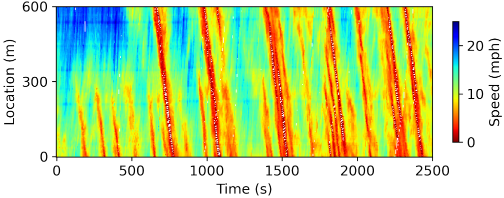
Outline

- **Motivation**
- **Matrix Factorization**
  - Optimization Problem
  - GD vs. SGD vs. ALS
- **Smoothing Matrix Factorization**
  - Spatial/Temporal Smoothing
  - Alternating Minimization
- **Temporal Matrix Factorization**
- **Tensor Factorization**
  - Basic Idea
  - CP Tensor Factorization
  - Hankel Tensor and Its Factorization
- **More Material**

Motivation

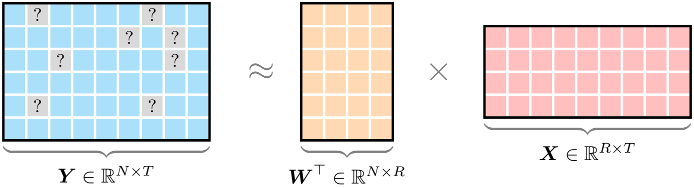


⇓ Reconstruct speed field from sparse trajectories?



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 $\boldsymbol{X}$.

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**)

$$\begin{cases} \boldsymbol{W} := \boldsymbol{W}-\alpha\dfrac{\partial f}{\partial \boldsymbol{W}} \\ \boldsymbol{X} := \boldsymbol{X}-\alpha\dfrac{\partial f}{\partial \boldsymbol{X}} \end{cases}$$

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**)

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

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

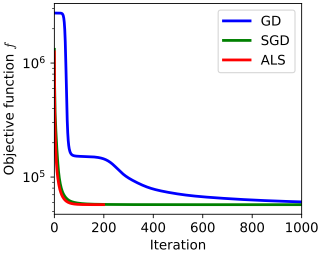
$$\begin{cases}\dfrac{\partial f}{\partial \boldsymbol{W}}=-\boldsymbol{X}\mathcal{P}_{\Omega}^{\top}(\boldsymbol{Y}-\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{W}\\ \dfrac{\partial f}{\partial \boldsymbol{X}}=-\boldsymbol{W}\mathcal{P}_{\Omega}(\boldsymbol{Y}-\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{X}\end{cases}$$

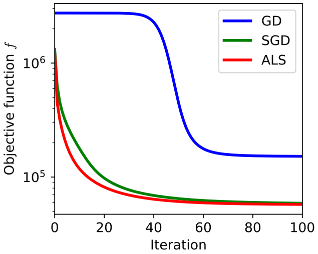
- Alternating least squares (**ALS**)

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

Matrix Factorization

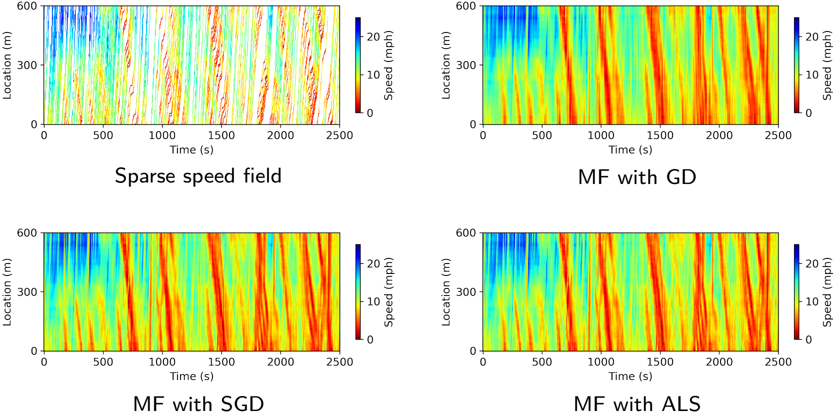
- Objective function f vs. iteration
  - Rank $R = 10$, weight parameter $\rho = 10$;
  - GD step size $\alpha = 10^{-4}$.





- Reconstruction errors

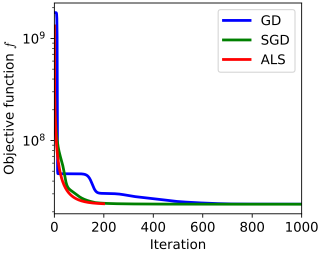
$$\text{MAPE} = \begin{cases} 50.66\% & \text{(GD)} \\ 45.13\% & \text{(SGD)} \\ 45.84\% & \text{(ALS)} \end{cases} \qquad \text{RMSE} = \begin{cases} 2.33 & \text{(GD)} \\ 2.79 & \text{(SGD)} \\ 2.80 & \text{(ALS)} \end{cases} \quad \text{(mph)}$$

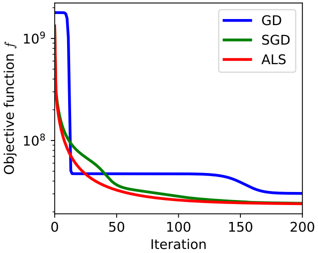

Sparse speed field
MF with GD
MF with SGD
MF with ALS
Location (m)
Time (s)
Speed (mph)
0
300
600
500
1000
1500
2000
2500
10
20

Matrix Factorization

Seattle freeway speed dataset (randomly generate 60% missing values)

- Objective function f vs. iteration
  - Rank $R = 10$, weight parameter $\rho = 10^2$;
  - GD step size $\alpha = 2 \times 10^{-5}$.



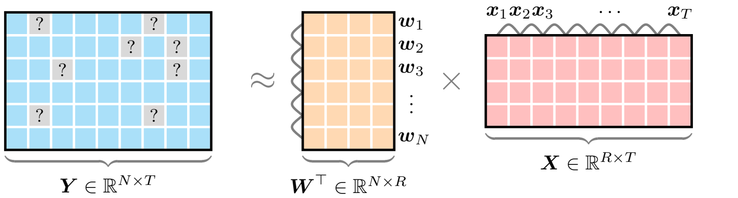


- Reconstruction errors

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

Smoothing Matrix Factorization

- Spatial/temporal local dependencies are also important!



- Spatial/temporal smoothing

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

Smoothing Matrix Factorization

- SMF optimization problem

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

- Partial derivatives

$$\begin{cases}\dfrac{\partial f}{\partial \boldsymbol{W}}=-\boldsymbol{X}\mathcal{P}_{\Omega}^{\top}(\boldsymbol{Y}-\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{W}+\lambda\boldsymbol{W}\boldsymbol{\Psi}_1^{\top}\boldsymbol{\Psi}_1\\ \dfrac{\partial f}{\partial \boldsymbol{X}}=-\boldsymbol{W}\mathcal{P}_{\Omega}(\boldsymbol{Y}-\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{X}+\lambda\boldsymbol{X}\boldsymbol{\Psi}_2^{\top}\boldsymbol{\Psi}_2\end{cases}$$

- **Alternating minimization**

$$\begin{cases}\boldsymbol{W}:=\{\boldsymbol{W}\mid \boldsymbol{X}\mathcal{P}_{\Omega}^{\top}(\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{W}+\lambda\boldsymbol{W}\boldsymbol{\Psi}_1^{\top}\boldsymbol{\Psi}_1=\boldsymbol{X}\mathcal{P}_{\Omega}^{\top}(\boldsymbol{Y})\}\\ \boldsymbol{X}:=\{\boldsymbol{X}\mid \boldsymbol{W}\mathcal{P}_{\Omega}(\boldsymbol{W}^{\top}\boldsymbol{X})+\rho\boldsymbol{X}+\lambda\boldsymbol{X}\boldsymbol{\Psi}_2^{\top}\boldsymbol{\Psi}_2=\boldsymbol{W}\mathcal{P}_{\Omega}(\boldsymbol{Y})\}\end{cases}$$

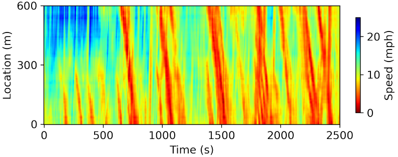
- Solving each matrix equation by the **conjugate gradient** method.

Smoothing Matrix Factorization

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

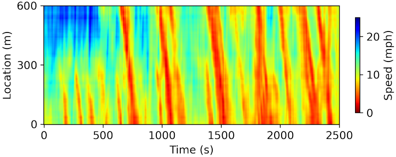
$$\text{MAPE} = \begin{cases} 50.66\% & \text{(GD)} \\ 45.13\% & \text{(SGD)} \\ 45.84\% & \text{(ALS)} \end{cases} \qquad \text{RMSE} = \begin{cases} 2.33 & \text{(GD)} \\ 2.79 & \text{(SGD)} \\ 2.80 & \text{(ALS)} \end{cases} \quad \text{(mph)}$$

$\lambda = 10$



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

$\lambda = 10^2$

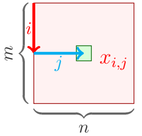


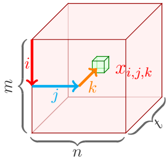
MAPE = 48.00%, RMSE = **1.60mph**

Temporal Matrix Factorization

Tensor Factorization

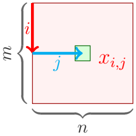
- What is tensor? $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ vs. $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m \times n \times t}$

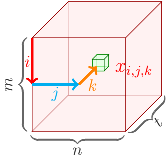




Tensor Factorization

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



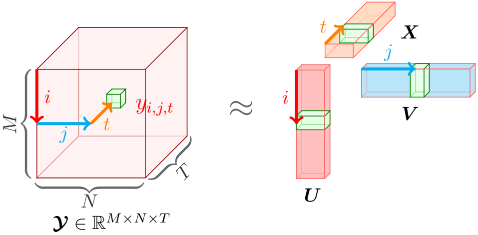


- Revisit tensor factorization

1927	1960s	1970	2009	2011
Higher-Order SVD	Tucker Decomposition	CP Decomposition	Tensor Decompositions and Applications	Tensor-Train Decomposition
 Frank Lauren Hitchcock	Ledyard R. Tucker	J. Douglas Carroll Jih-Jie Chang Richard A. Harshman	 Tamara G. Kolda	 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[1]:

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

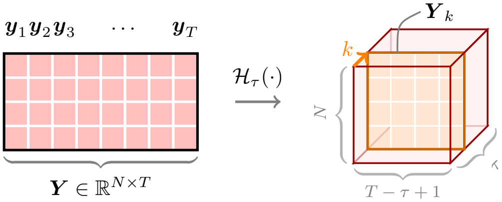
[1] The symbol $\otimes$ denotes the outer product.

Hankel Tensor and Its Factorization

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

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

- (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

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

- HTF's advantage/disadvantage over MF
  - Advantage: automatic temporal modeling, domain knowledge
  - Disadvantage: high memory consumption
- Speed field reconstruction
  - Rank $R=10$, weight parameter $\rho=10$.

More Material

- Slides:
- NGSIM speed field data:
- Seattle freeway traffic speed data:
- Jupyter Notebook:
- LaTeX drawing:
- ...