

# Low-Rank Models

## Probability and Statistics for Data Science

Carlos Fernandez-Granda



These slides are based on the book [Probability and Statistics for Data Science](#) by Carlos Fernandez-Granda, available for purchase [here](#). A free preprint, videos, code, slides and solutions to exercises are available at <https://www.ps4ds.net>

# Motivation

Datasets where each data point is associated to 2 entities

1. *Recommender systems*: Rating given to movie  $i$  by user  $j$
2. *Computational genomics*: Expression level of gene  $i$  in cell  $j$
3. *Weather forecasting*: Temperature in location  $i$  at time  $j$

## Movie ratings

	Bob	Molly	Mary	Larry	
(	1	1	5	4	The Dark Knight
	2	1	4	5	Spiderman 3
	4	5	2	1	Love Actually
	5	4	2	1	Bridget Jones's Diary
	4	5	1	2	Pretty Woman
)	1	2	5	5	Superman 2

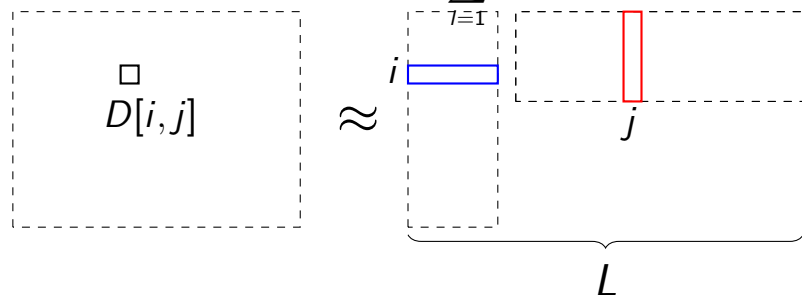
# Rank-1 model $a[\text{movie}]b[\text{user}]$

Ratings  $\approx$  Mean rating +

Dark Knight	$-0.45$	Bob	Molly	Mary	Larry
Spiderman 3	$-0.39$	(3.74	4.05	$-3.74$	$-4.05$ )
Love Actually	0.39				
BJ's Diary	0.38				
Pretty Woman	0.38				
Superman 2	$-0.45$				

# Low-rank model

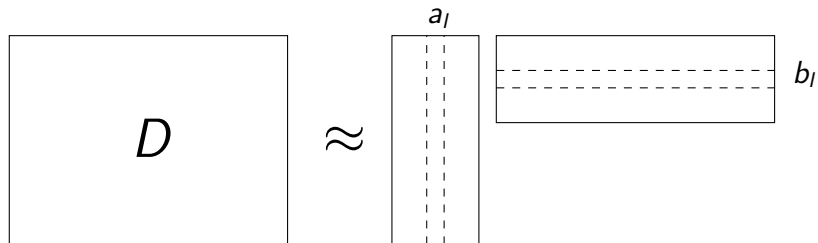
**Assumption:** Data depends on a small number of *factors*

$$D[i, j] \approx L[i, j] := \sum_{l=1}^r a_l[i] b_l[j]$$


$a_l[i]$  is the contribution of factor  $l$  to movie  $i$

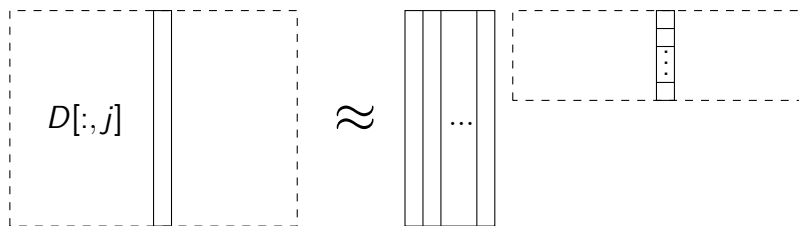
$b_l[j]$  is the contribution of factor  $l$  to user  $j$

Rank of  $L$ ?  $r$



## How do we fit the model?

Idea: Interpret *columns* as set of vectors



Best  $r$ -dimensional approximation of each column?



# Principal component analysis

Eigendecomposition of sample covariance matrix of columns

Principal directions are eigenvectors corresponding to  $r$  largest eigenvalues:

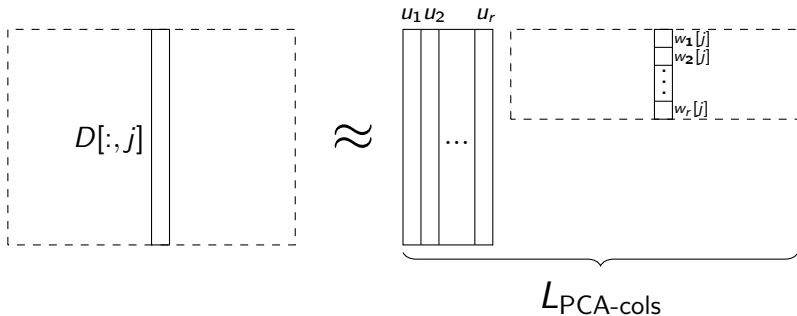
$u_1, u_2, \dots, u_r$

Assuming columns are centered:

Principal components:  $w_l[j] := u_l^T D[:, j]$

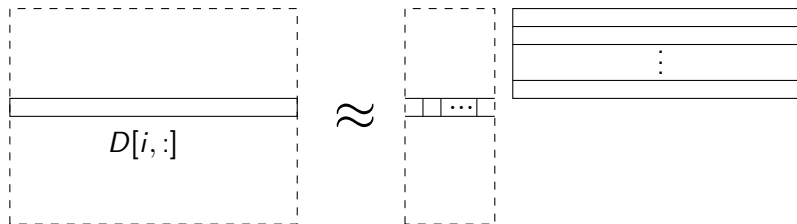
Approximation  $D[:, j] \approx \sum_{l=1}^r u_l w_l[j]$

$$L_{\text{PCA-cols}}[i, j] := \sum_{l=1}^r u_l[i] w_l[j]$$



Wait a minute

Why rows and not columns?



Best  $r$ -dimensional approximation of each row?

# PCA

Eigendecomposition of sample covariance matrix of rows

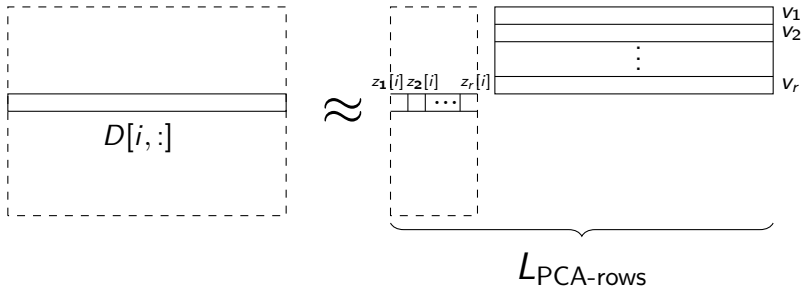
Principal directions are eigenvectors corresponding to  $r$  largest eigenvalues:  
 $v_1, v_2, \dots, v_r$

Assuming rows are centered:

Principal components:  $z_l[i] := D[i, :] v_l$

Approximation  $D[i, :] \approx \sum_{l=1}^r z_l[i] v_l^T$

$$L_{\text{PCA-rows}}[i, j] := \sum_{l=1}^r z_l[i] v_l[j]$$



Which one is better?

They are equivalent! (up to centering)

# Singular value decomposition

All matrices have an SVD ( $n_1 \leq n_2$ )

$$D = \underbrace{\begin{bmatrix} u_1 & u_2 & \cdots & u_{n_1} \end{bmatrix}}_U \underbrace{\begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & s_{n_1} \end{bmatrix}}_S \underbrace{\begin{bmatrix} v_1 & v_2 & \cdots & v_{n_1} \end{bmatrix}^T}_{V^T}$$

- ▶ Singular values  $s_1 \geq s_2 \geq \cdots \geq s_r \geq 0$
- ▶ Left singular vectors  $u_1, u_2, \dots, u_{n_1} \in \mathbb{R}^{n_1}$  are orthonormal
- ▶ Right singular vectors  $v_1, v_2, \dots, v_{n_1} \in \mathbb{R}^{n_2}$  are orthonormal

# Singular value decomposition

The diagram illustrates the Singular Value Decomposition (SVD) of a matrix  $D$ . On the left is a large rectangle labeled  $D$ . To its right is an equals sign followed by a summation from  $l=1$  to  $n_1$ . Each term in the summation is a rank-1 matrix  $K_l$ , which is represented by a bracketed group of two components: a vertical rectangle and a horizontal rectangle. The vertical rectangle is labeled  $u_l$  at the top, and the horizontal rectangle is labeled  $v_l$  at the top. A small square box labeled  $s_l$  is positioned to the left of the vertical rectangle, indicating the singular value. The entire group of  $u_l$  and  $v_l$  rectangles is bracketed at the bottom and labeled  $K_l$ .

$K_1, \dots, K_{n_1}$  are rank 1, orthogonal, unit norm

$$\text{Norm of } D = \sqrt{\sum_{l=1}^I s_l^2}$$

Rank- $r$  approximation?

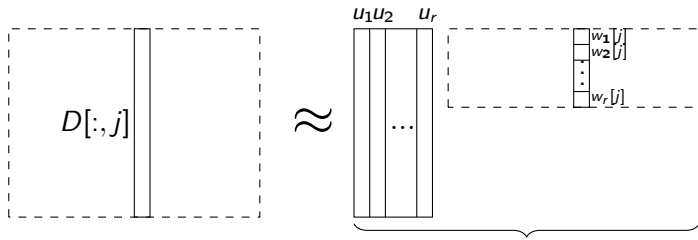
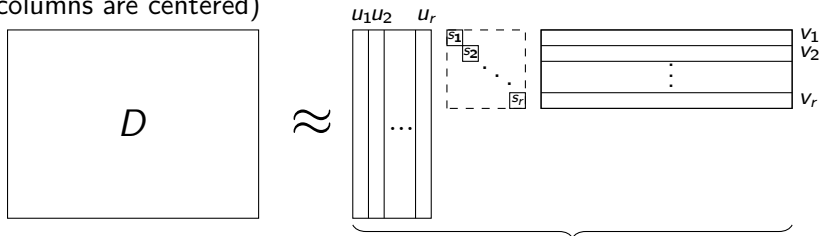


## Truncated SVD

$$\boxed{L_{\text{SVD}}} \quad \coloneqq \quad \sum_{l=1}^r s_l \boxed{K_l}$$

$DD^T$  = sample covariance matrix of columns

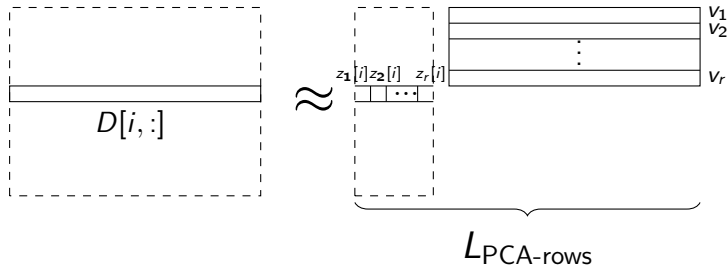
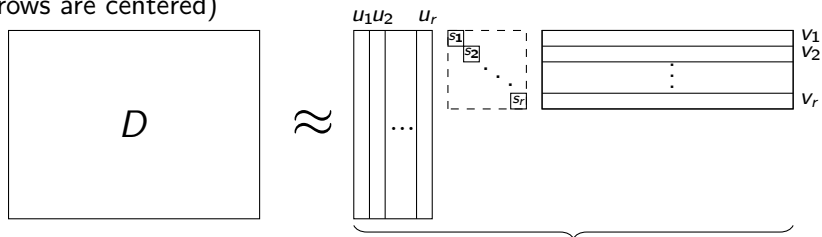
(if columns are centered)



$L_{PCA-cols}$

$D^T D =$  sample covariance matrix of rows

(if rows are centered)



## SVD truncation and PCA are equivalent

$$L_{\text{SVD}} = L_{\text{PCA-rows}} = L_{\text{PCA-cols}} \quad (\text{up to centering!})$$

$$L_{\text{SVD}} = \arg \min_{\text{rank}(L)=r} \|D - L\|_F$$

Optimal low-rank approximation!

## Movie ratings

$$D := \begin{array}{ccccc} & \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} & \\ \left( \begin{array}{cccc} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{array} \right) & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

$$m(D) := \frac{1}{24} \sum_{i=1}^6 \sum_{j=1}^4 D[i,j] = 3$$

## Movie ratings

$$\begin{aligned} D_{\text{ct}} &:= D - m(D) = USV^T \\ &= U \begin{bmatrix} 7.79 & 0 & 0 & 0 \\ 0 & 1.62 & 0 & 0 \\ 0 & 0 & 1.55 & 0 \\ 0 & 0 & 0 & 0.62 \end{bmatrix} V^T \end{aligned}$$

## Rank-1 model

Rating  $\approx$  Mean rating (3) +

Dark Knight	$-0.45$				
Spiderman 3	$-0.39$				
Love Actually	$0.39$	Bob	Molly	Mary	Larry
BJ's Diary	$0.38$	$(3.74$	$4.05$	$-3.74$	$-4.05)$
Pretty Woman	$0.38$				
Superman 2	$-0.45$				

## Estimated ratings

$$L[\text{movie}, \text{user}] := m(D) + s_1 u_1[\text{movie}] v_1[\text{user}]$$

	Bob	Molly	Mary	Larry	
=	1.34 (1)	1.19 (1)	4.66 (5)	4.81 (4)	The Dark Knight
	1.55 (2)	1.42 (1)	4.45 (4)	4.58 (5)	Spiderman 3
	4.45 (4)	4.58 (5)	1.55 (2)	1.42 (1)	Love Actually
	4.43 (5)	4.56 (4)	1.57 (2)	1.44 (1)	B. Jones's Diary
	4.43 (4)	4.56 (5)	1.57 (1)	1.44 (2)	Pretty Woman
	1.34 (1)	1.19 (2)	4.66 (5)	4.81 (5)	Superman 2



$u_1$

Rating  $\approx$  Mean rating (3) +

Dark Knight	$-0.45$	Bob	Molly	Mary	Larry
Spiderman 3	$-0.39$	(3.74	4.05	$-3.74$	$-4.05$ )
Love Actually	$0.39$				
BJ's Diary	$0.38$				
Pretty Woman	$0.38$				
Superman 2	$-0.45$				

$S_1 V_1$

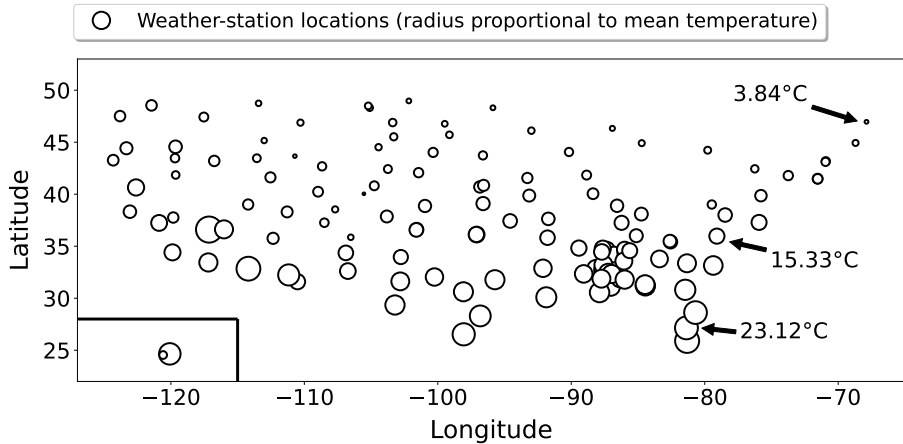
Rating  $\approx$  Mean rating (3) +

Dark Knight	$\begin{pmatrix} -0.45 \\ -0.39 \\ 0.39 \\ 0.38 \\ 0.38 \\ -0.45 \end{pmatrix}$	Bob	Molly	Mary	Larry
Spiderman 3		(3.74	4.05	-3.74	-4.05)
Love Actually					
BJ's Diary					
Pretty Woman					
Superman 2					

# Temperatures

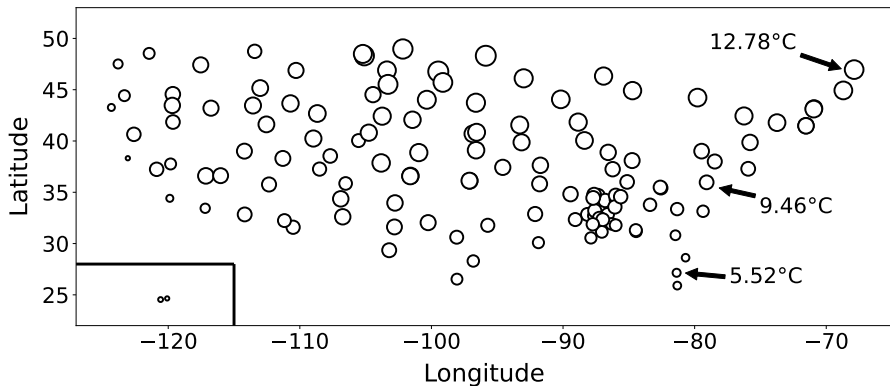
Dataset of hourly temperatures at 134 weather stations in the United States over one year

## Mean at each station



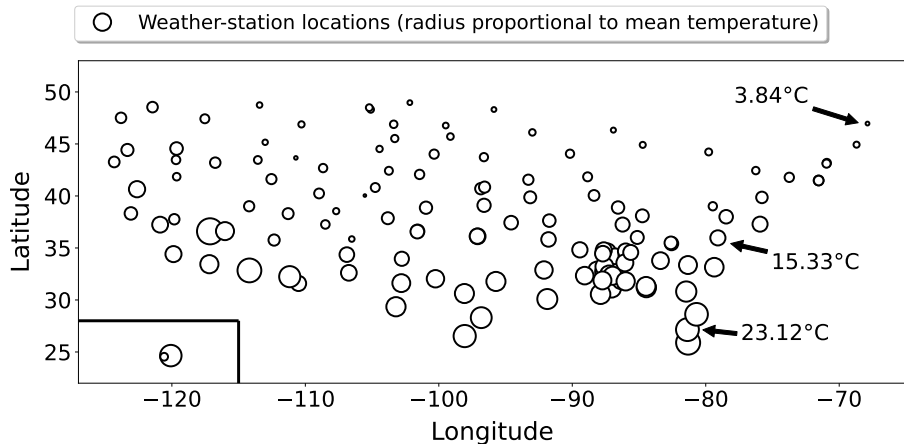
# Standard deviation

○ Weather-station locations (radius proportional to standard deviation of temperature)

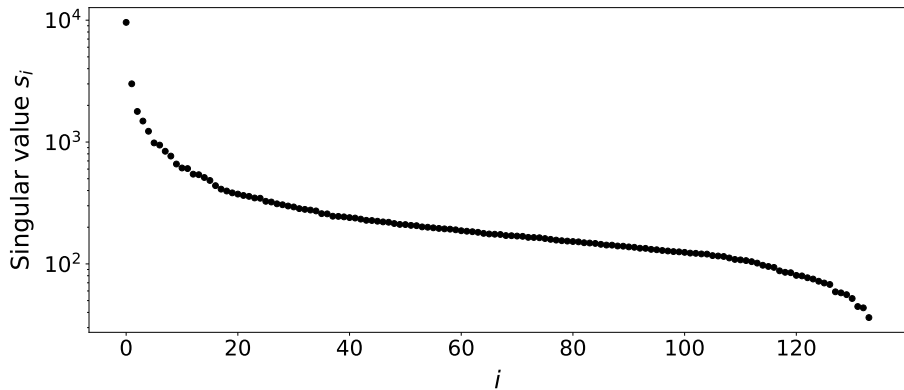


## Goal: Analyze temporal patterns

We center temperatures at each station (removing sample mean)



## Singular values of centered data



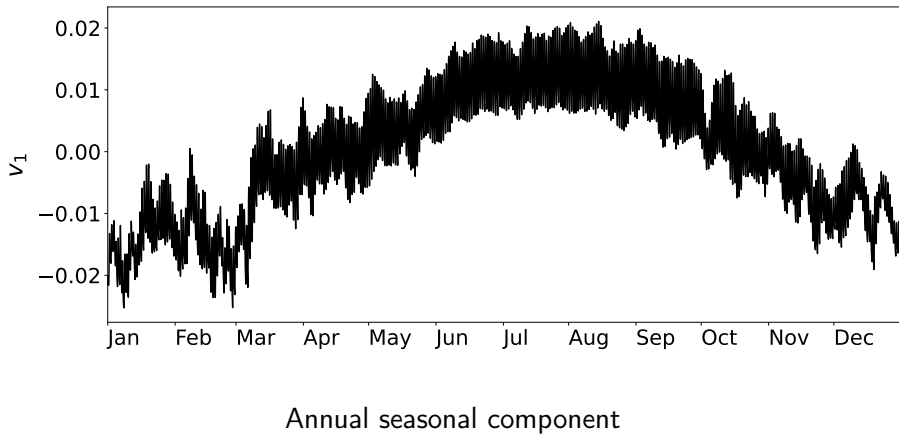
## Rank 1 model

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$



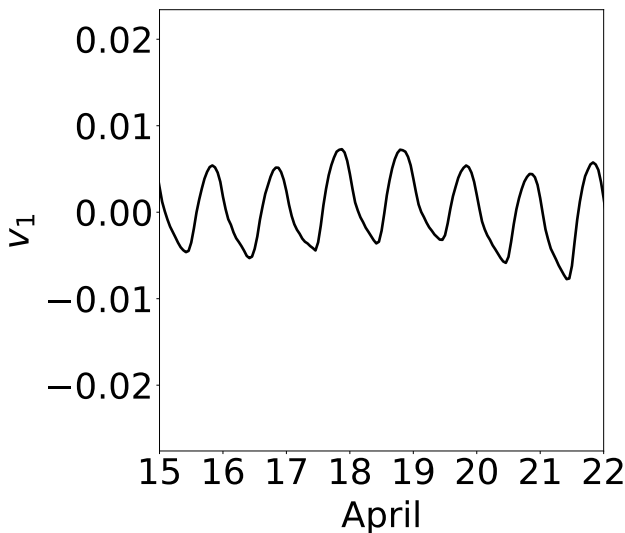
$v_1$

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$



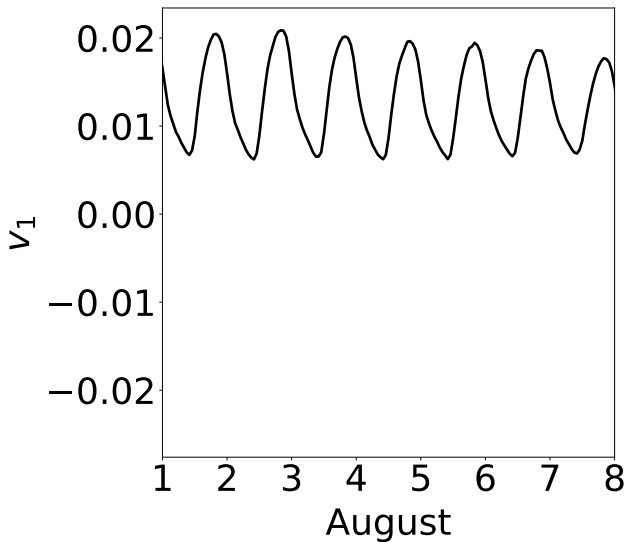
## Zooming in... Daily pattern

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$



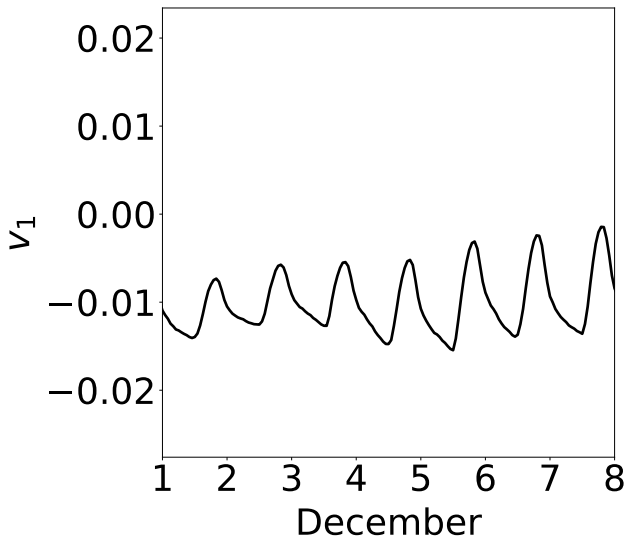
## Daily pattern

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$



## Daily pattern

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$

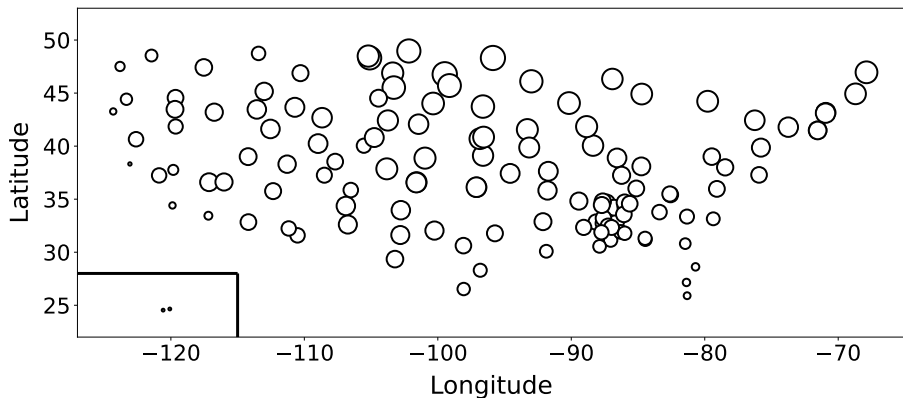


$u_1$

$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$

Seasonal / daily pattern component for each station

○ Weather station  $i$ , radius proportional to  $u_1[i]$

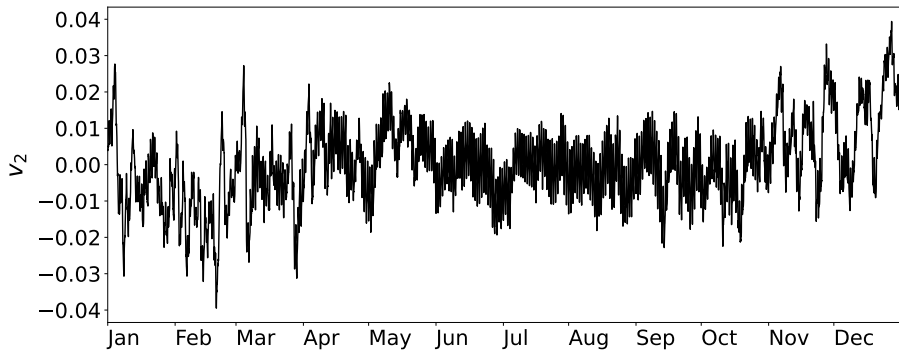


## Rank 2 model

$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$

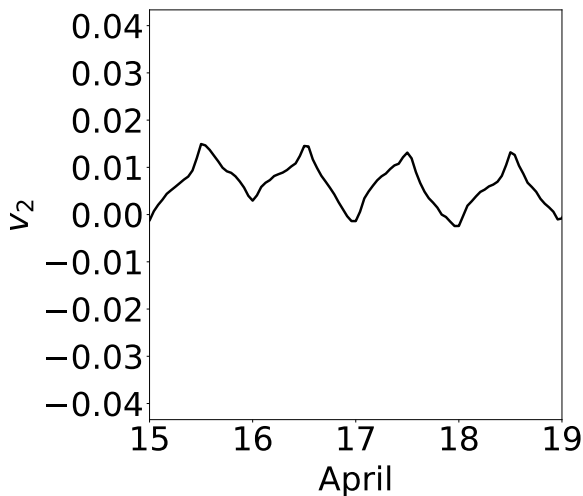
$v_2$

$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$



If we zoom in?

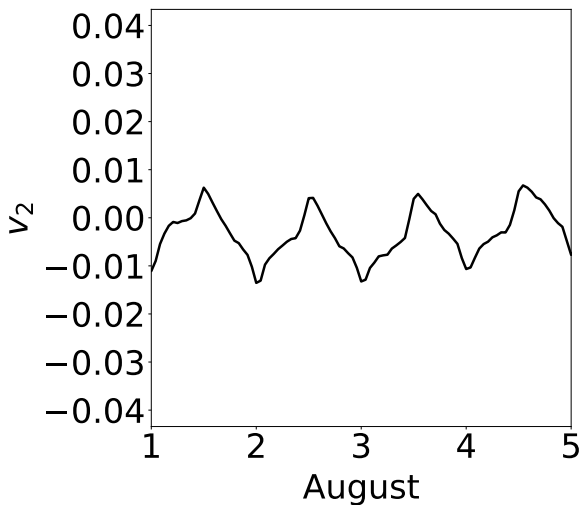
$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$





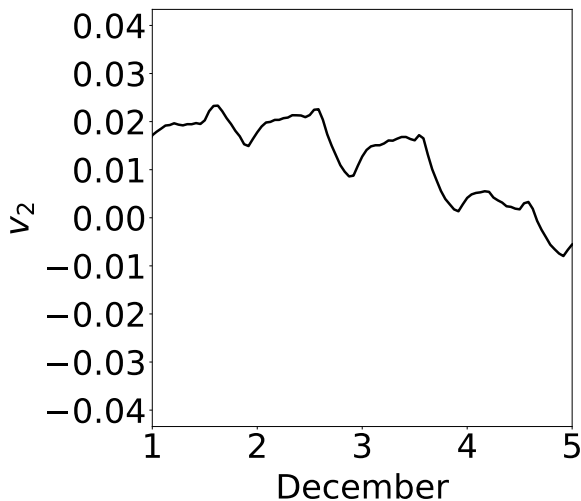
If we zoom in?

$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$



If we zoom in?

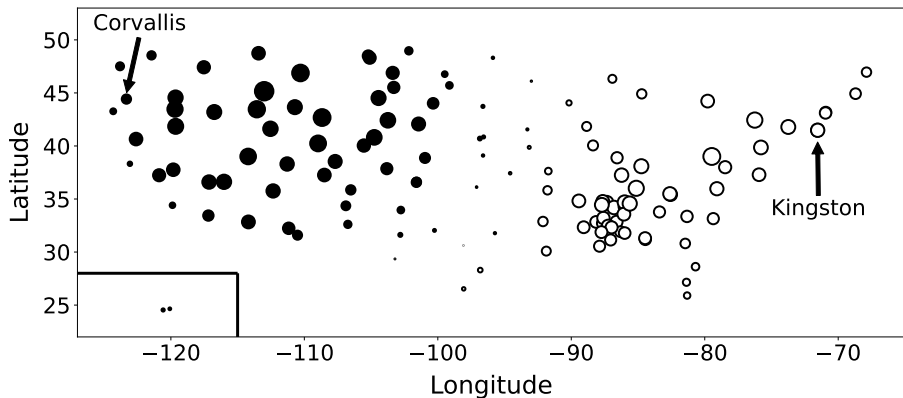
$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$



$u_2$

$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$

- $u_2[i] > 0$ , radius proportional to  $|u_2[i]|$
- $u_2[i] < 0$ , radius proportional to  $|u_2[i]|$



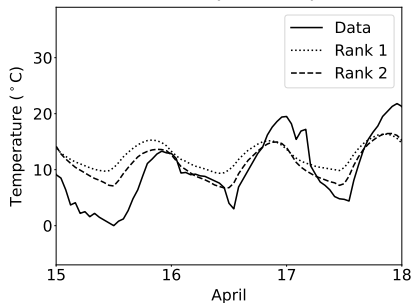
## Rank-2 model

$$\begin{aligned} m_{\text{Corvallis}} + s_1 u_1[\text{Corvallis}] v_1[\text{time}] + s_2 u_2[\text{Corvallis}] v_2[\text{time}] \\ = 12.3 + 554 v_1[\text{time}] - 218 v_2[\text{time}] \end{aligned}$$

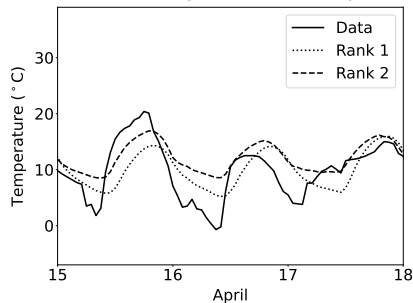
$$\begin{aligned} m_{\text{Kingston}} + s_1 u_1[\text{Kingston}] v_1[\text{time}] + s_2 u_2[\text{Kingston}] v_2[\text{time}] \\ = 9.7 + 853 v_1[\text{time}] + 311 v_2[\text{time}] \end{aligned}$$

# Comparison of rank-1 and rank-2 models

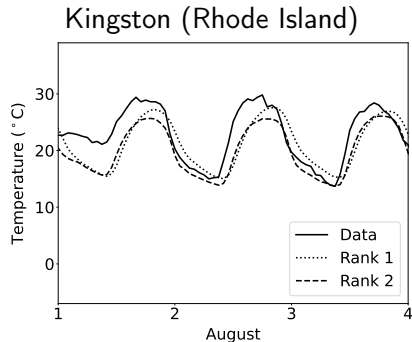
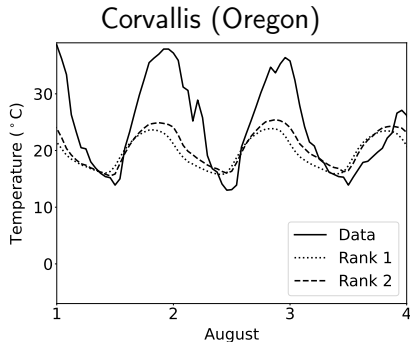
Corvallis (Oregon)



Kingston (Rhode Island)

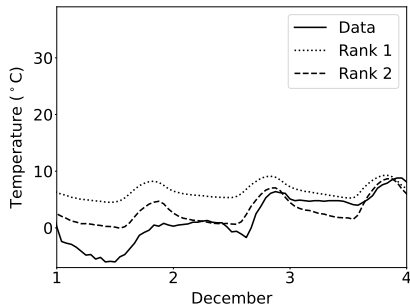


# Comparison of rank-1 and rank-2 models

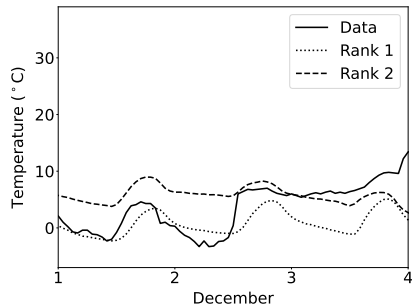


# Comparison of rank-1 and rank-2 models

Corvallis (Oregon)



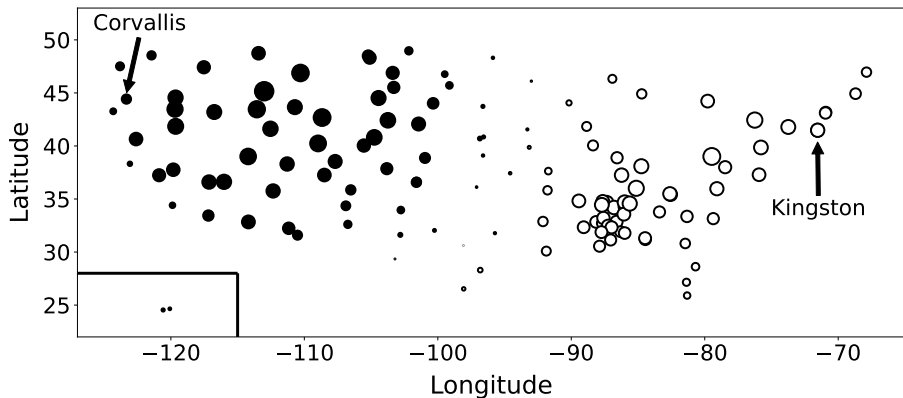
Kingston (Rhode Island)



$u_2$

$$D[s, t] \approx m_{\text{station}} + s_1 u_1[s] v_1[t] + s_2 u_2[\text{station}] v_2[\text{time}]$$

- $u_2[i] > 0$ , radius proportional to  $|u_2[i]|$
- $u_2[i] < 0$ , radius proportional to  $|u_2[i]|$





# What have we learned?

How to interpret low-rank models

Connection to PCA

How to fit them by truncating the SVD