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

$$\begin{array}{ll} \quad\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \begin{pmatrix} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{pmatrix} & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Rank-1 model a[movie]b[user]

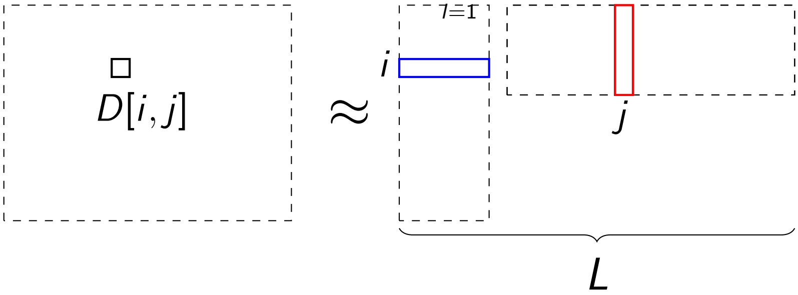
Ratings $\approx$ Mean rating +

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

Low-rank model

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

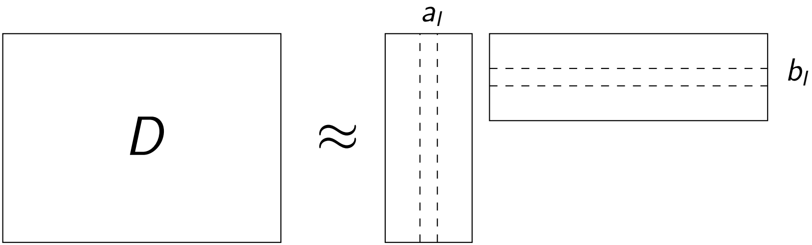
$$D[i,j] \approx L[i,j] := \sum_{l=1}^{r} a_l[i] b_l[j]$$



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

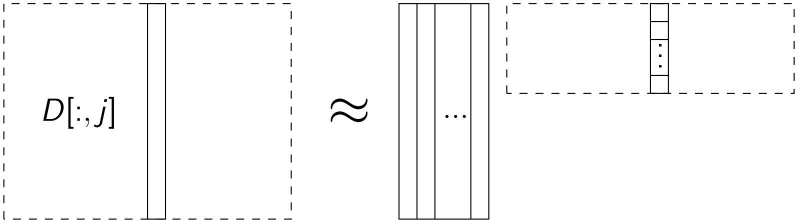
$b_l[j]$ is the contribution of factor l to user j

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

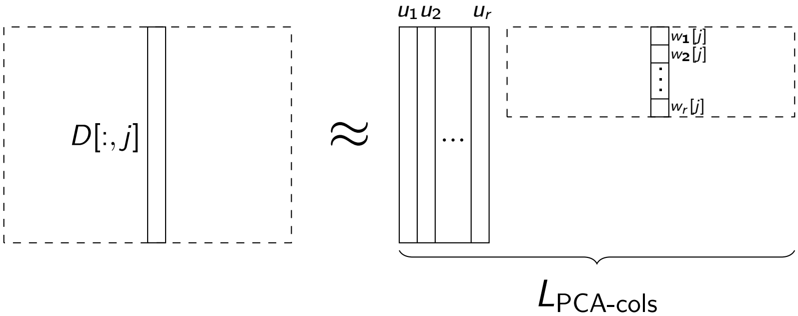
Principal directions are eigenvectors corresponding to r largest eigenvalues: u_1, u_2, ..., u_r

Assuming columns are centered:

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

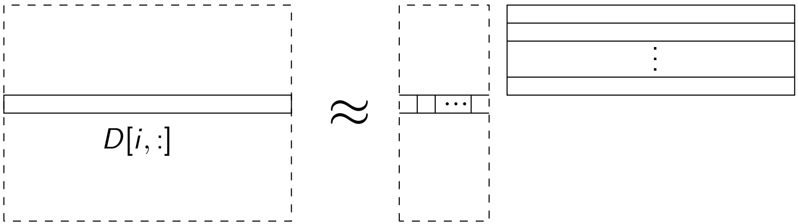
Approximation $D[:, j] \approx \sum_{l=1}^{r} u_l w_l[j]$

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



Wait a minute

Why rows and not columns?



Best r-dimensional approximation of each row?

PCA

Eigendecomposition of sample covariance matrix of rows

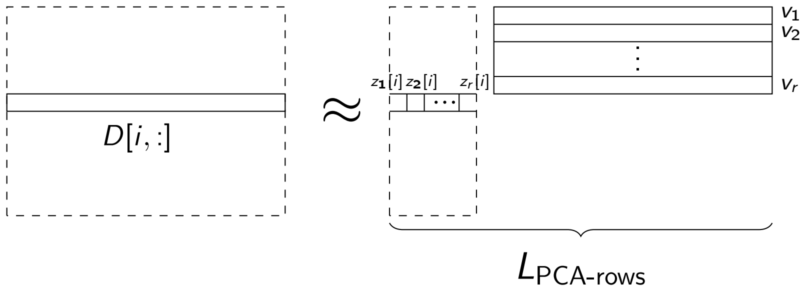
Principal directions are eigenvectors corresponding to r largest eigenvalues: v_1, v_2, ..., v_r

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)

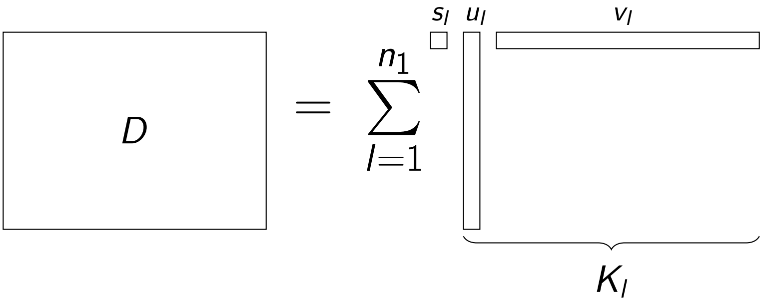
Singular value decomposition

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

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

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

Singular value decomposition



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

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

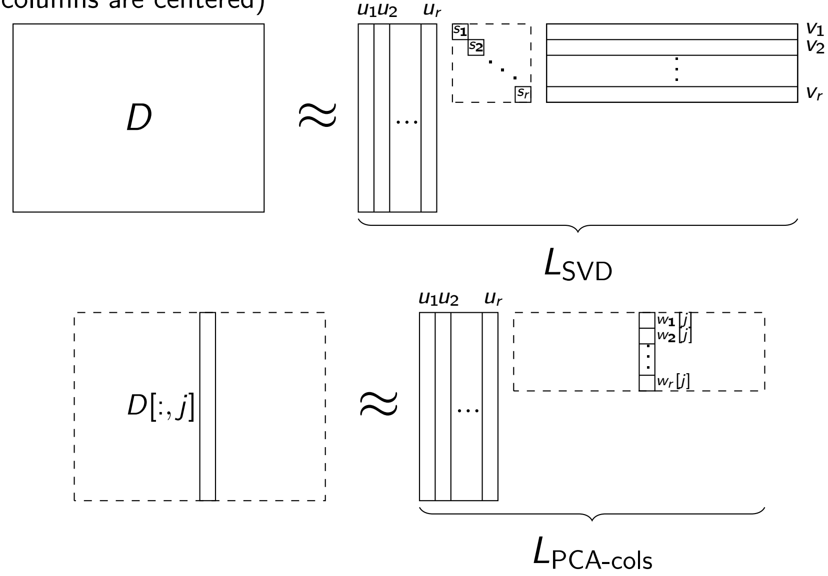
Rank-r approximation?

Truncated SVD

$$\boxed{L_{\mathrm{SVD}}} := \sum_{l=1}^{\boldsymbol{r}} \overset{s_l}{\square} \boxed{K_l}$$

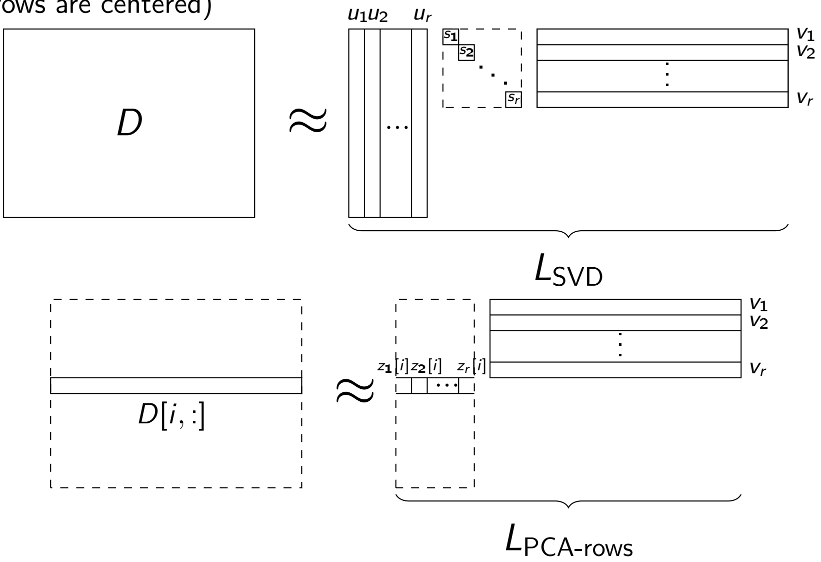
DD^T = sample covariance matrix of columns

(if columns are centered)



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

(if rows are centered)



SVD truncation and PCA are equivalent

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

$$L_{\mathsf{SVD}} = \arg \min_{\operatorname{rank}(L)=r} ||D - L||_{\mathsf{F}}$$

Optimal low-rank approximation!

Movie ratings

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

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

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) +

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

Estimated ratings

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

$$= \begin{array}{cc} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \begin{pmatrix} 1.34\,(1) & 1.19\,(1) & 4.66\,(5) & 4.81\,(4) \\ 1.55\,(2) & 1.42\,(1) & 4.45\,(4) & 4.58\,(5) \\ 4.45\,(4) & 4.58\,(5) & 1.55\,(2) & 1.42\,(1) \\ 4.43\,(5) & 4.56\,(4) & 1.57\,(2) & 1.44\,(1) \\ 4.43\,(4) & 4.56\,(5) & 1.57\,(1) & 1.44\,(2) \\ 1.34\,(1) & 1.19\,(2) & 4.66\,(5) & 4.81\,(5) \end{pmatrix} & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{B. Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

u_1

Rating $\approx$ Mean rating (3) +

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

$s_1 v_1$

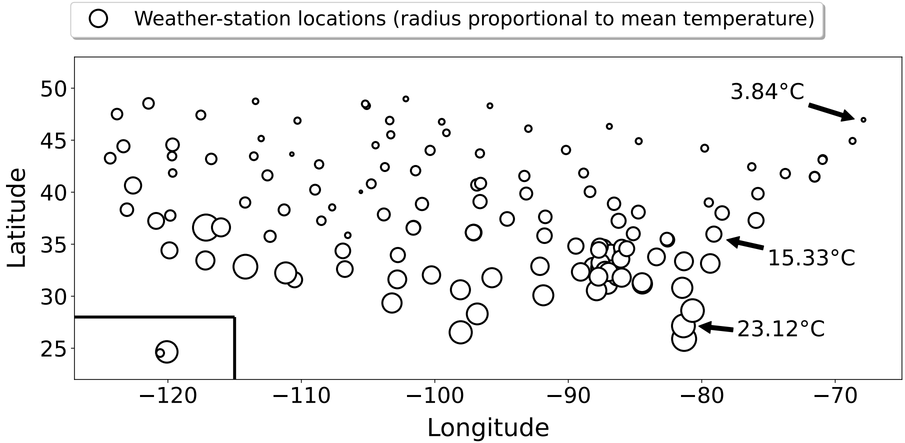
Rating $\approx$ Mean rating (3) +

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

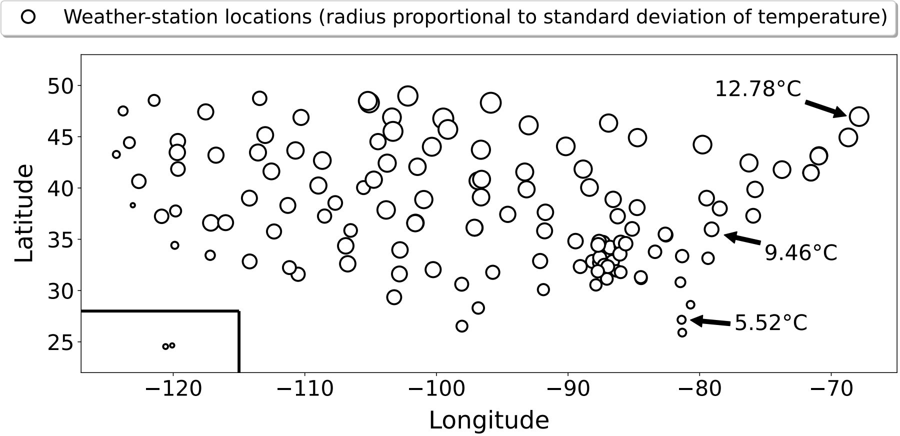
Temperatures

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

Mean at each station

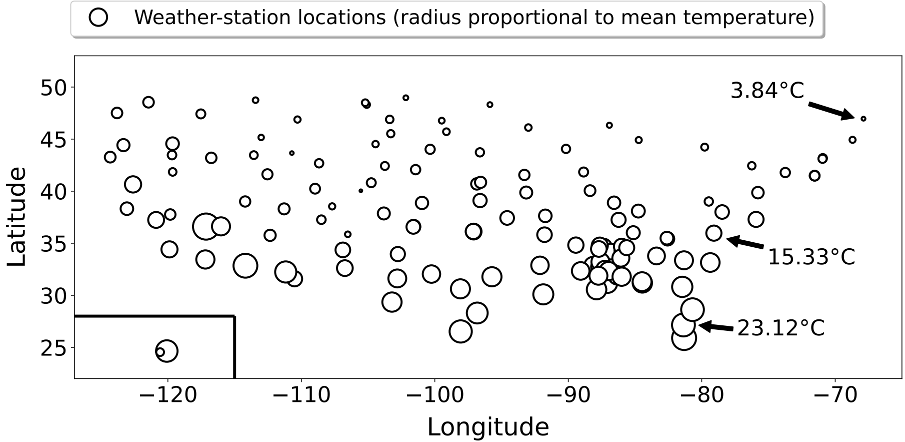


Standard deviation

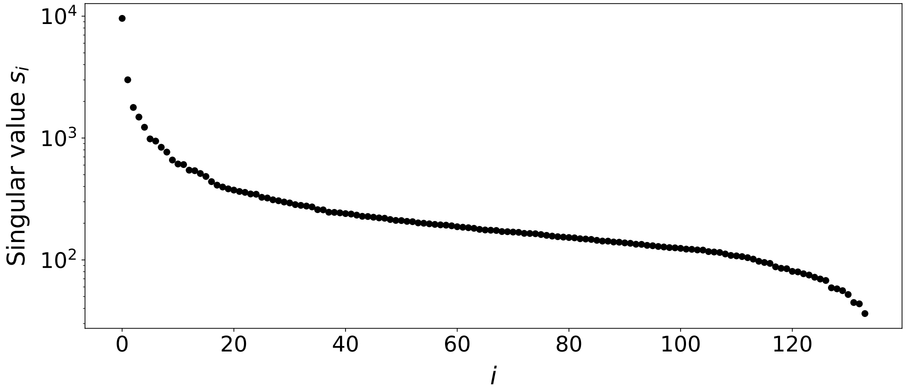


Goal: Analyze temporal patterns

We center temperatures at each station (removing sample mean)



Singular values of centered data

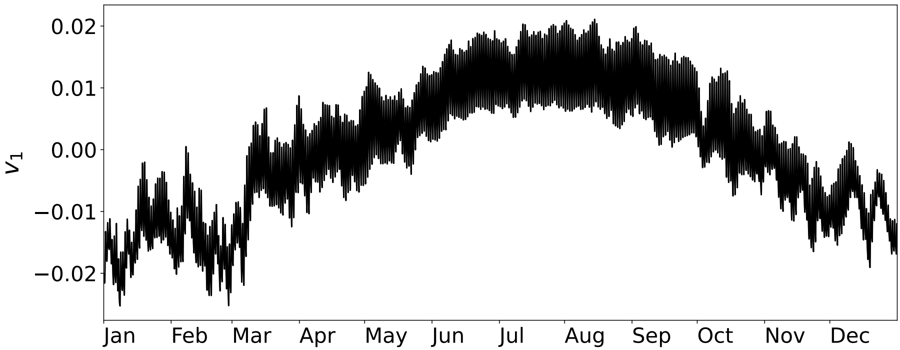


Rank 1 model

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

v_1

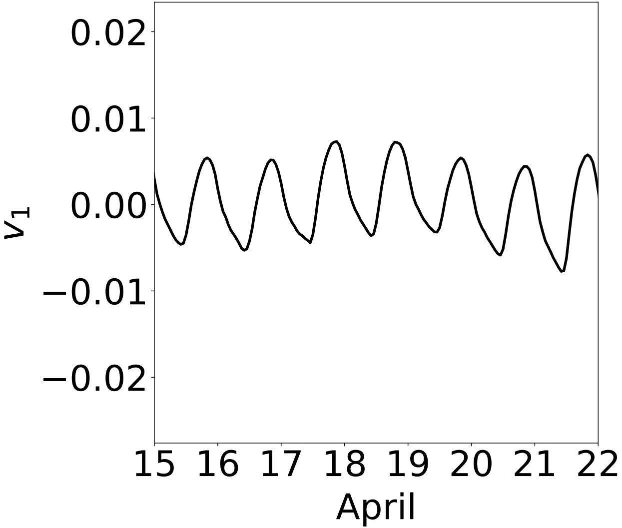
$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] v_1[\text{time}]$$



Annual seasonal component

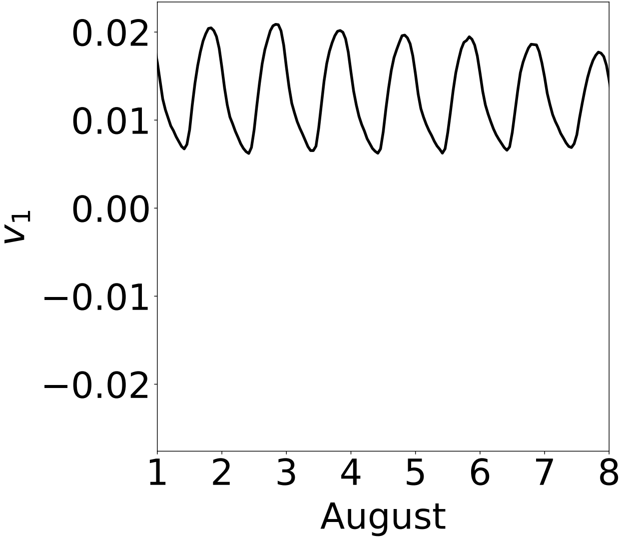
Zooming in... Daily pattern

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



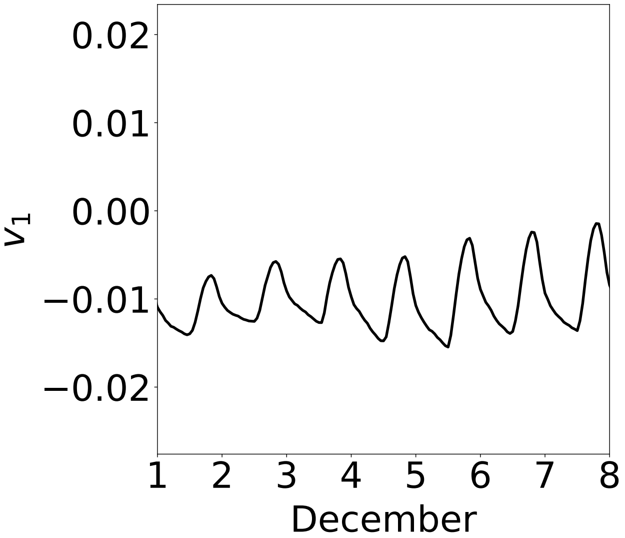
Daily pattern

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



Daily pattern

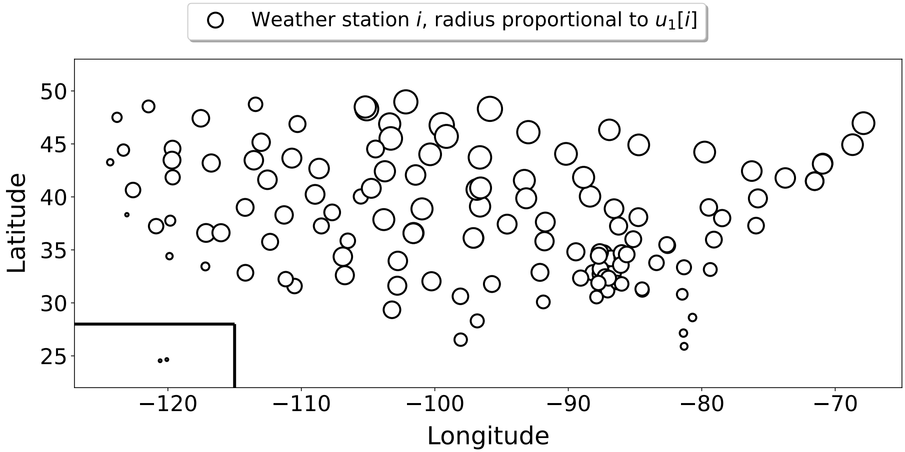
$$D[\text{station}, \text{time}] \approx m_{\text{station}} + s_1 u_1[\text{station}] {\color{red} v_1[\text{time}]}$$



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

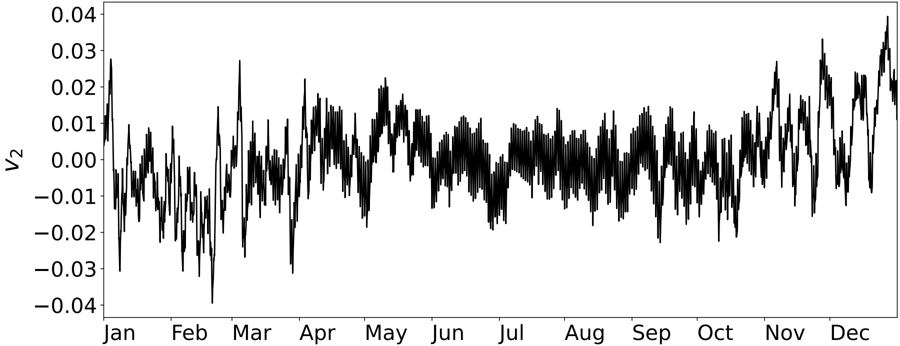


Rank 2 model

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

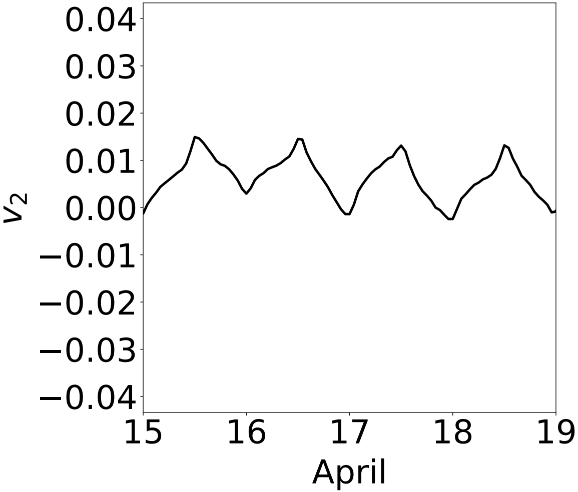
v_2

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



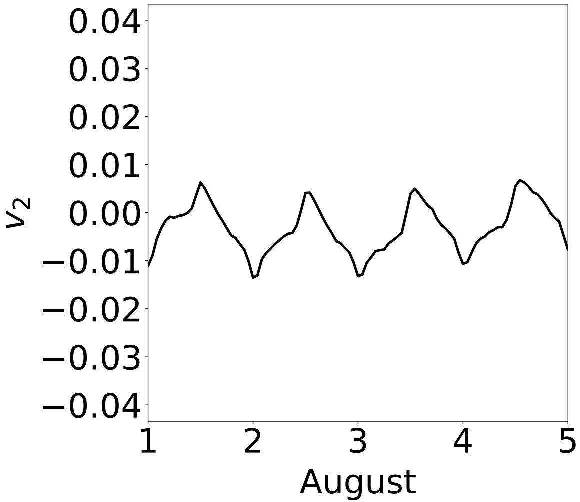
If we zoom in?

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



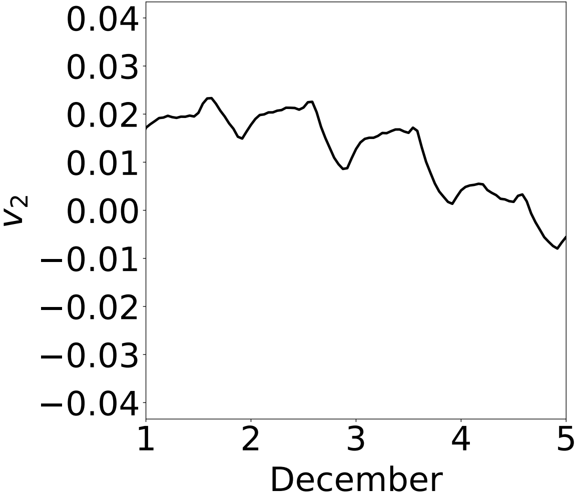
If we zoom in?

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



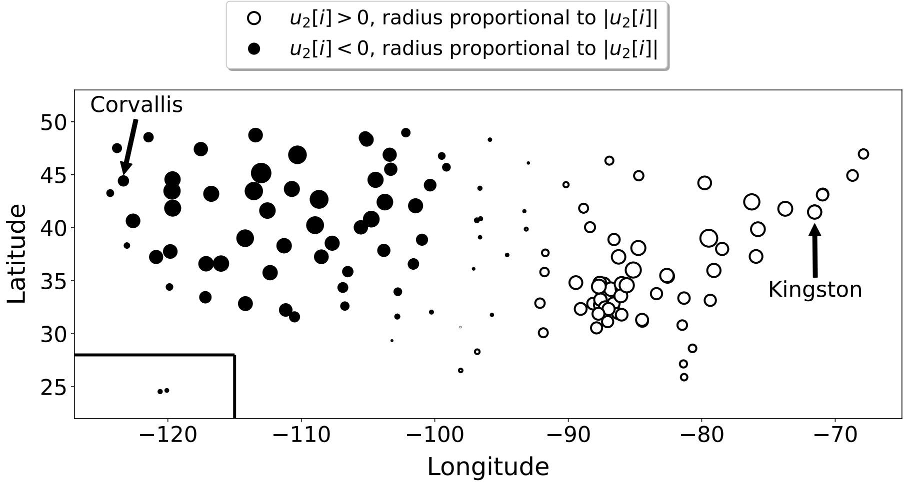
If we zoom in?

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



u_2

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

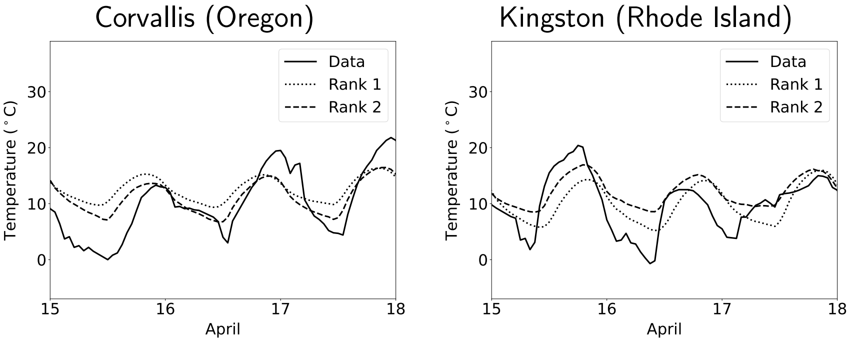


Rank-2 model

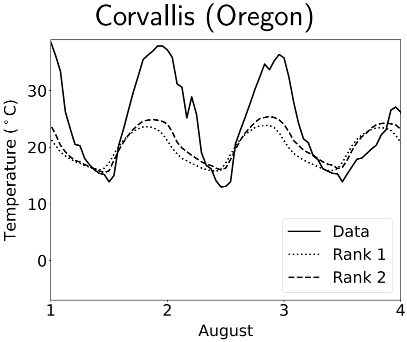
$$m_{\text{Corvallis}} + s_1 u_1[\text{Corvallis}] v_1[\text{time}] + s_2 u_2[\text{Corvallis}] v_2[\text{time}]$$
$$= 12.3 + 554\, v_1[\text{time}] - 218\, v_2[\text{time}]$$

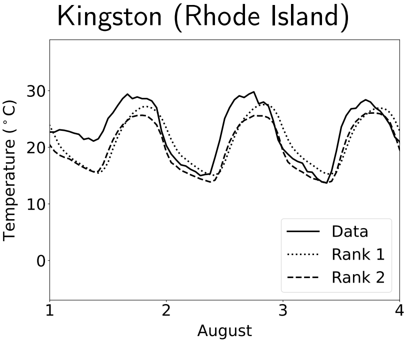
$$m_{\text{Kingston}} + s_1 u_1[\text{Kingston}] v_1[\text{time}] + s_2 u_2[\text{Kingston}] v_2[\text{time}]$$
$$= 9.7 + 853\, v_1[\text{time}] + 311\, v_2[\text{time}]$$

Comparison of rank-1 and rank-2 models

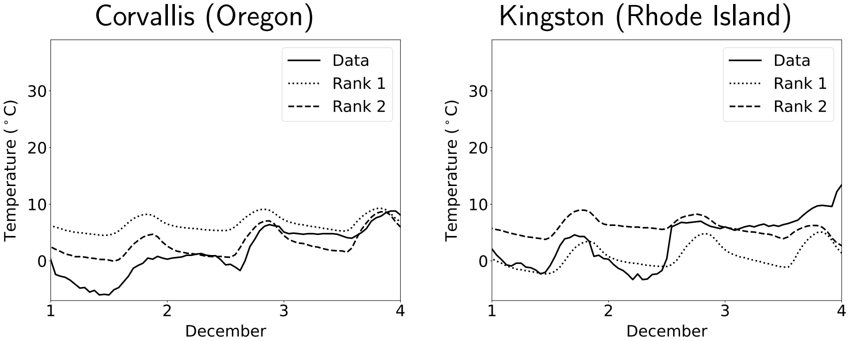


Comparison of rank-1 and rank-2 models



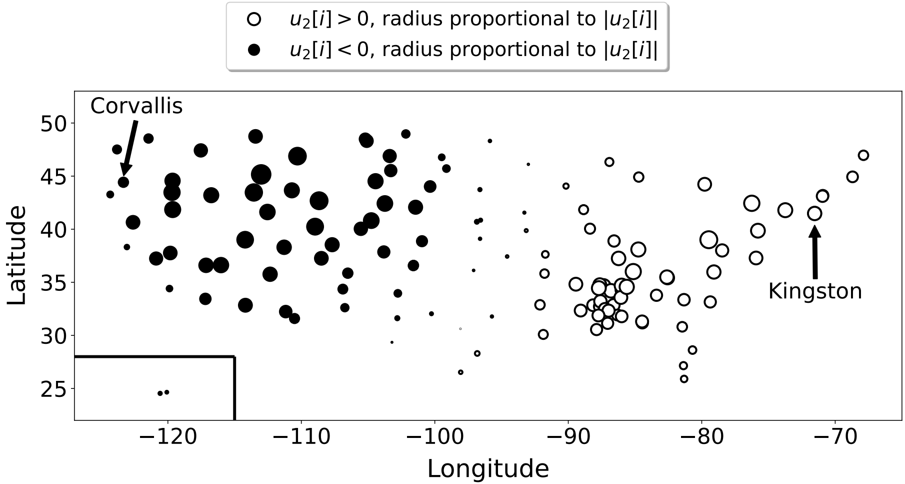


Comparison of rank-1 and rank-2 models



u_2

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



What have we learned?

How to interpret low-rank models

Connection to PCA

How to fit them by truncating the SVD