Missing Value Interpolation

Machine Learning for Earth System Sciences

2nd February 2021

Anomaly

- X(s,t) : measurement of quantity ‘X’ at location ‘s’ and time ‘t’
- μ(s): local mean, or “climatology”,

  eg. Mean daily temperature in Kharagpur
- μ(t): temporal mean, or “climatology”,

  eg. Mean aggregate rainfall in August
- μ(s,t): local mean with respect to time

  eg. Mean daily temperature for February in Kharagpur
- Anomaly = observation - climatology

Anomaly

- Anomaly $Y(s,t) = X(s,t) - \mu$ (positive, negative, zero)
- μ can be $\mu(s)$, $\mu(t)$ or $\mu(s,t)$ depending on context
- Anomaly Y likely to be spatially spatially coherent
- Many neighboring locations simultaneously have anomaly with large magnitude: anomaly event
- Anomaly event can be positive or negative!
- Eg. Heat wave: positive anomaly event w.r.t. temperature

  Drought: negaitive anomaly event w.r.t. rainfall

Any random data matrix

	T1	T2	T3	T4	T5
S1	32	30	19	23	32
S2	33	16	28	35	26
S3	25	33	13	15	33
S4	27	14	34	29	15
S5	38	26	32	35	18
S6	15	18	24	26	28

Can you predict missing values?

	T1	T2	T3	T4	T5
S1	32	30	19	x	32
S2	33	x	28	35	26
S3	25	33	x	15	33
S4	27	x	34	x	15
S5	x	26	32	35	18
S6	15	18	24	26	x

Can you predict missing values?

	T1	T2	T3	T4	T5
S1	32	30	32	x	32
S2	33	x	34	35	33
S3	25	23	x	26	26
S4	27	x	28	x	28
S5	x	18	19	23	18
S6	15	14	13	15	x

Geophysical Data Matrix

	T1	T2	T3	T4	T5
S1	32	30	32	35	32
S2	33	32	34	35	33
S3	25	23	24	26	26
S4	27	26	28	29	28
S5	18	18	19	23	18
S6	15	14	13	15	15

Location-wise interpolation

	T1	T2	T3	T4	T5	Local Mean
S1	32	30	32	x	32	31
S2	33	x	34	35	33	34
S3	25	23	x	26	26	25
S4	27	x	28	x	28	28
S5	x	18	19	23	18	19
S6	15	14	13	15	x	14

Time-wise Interpolation (makes no sense here)

	T1	T2	T3	T4	T5
S1	32	30	32	x	32
S2	33	x	34	35	33
S3	25	23	x	26	26
S4	27	x	28	x	28
S5	x	18	19	23	18
S6	15	14	13	15	x
Daily Mean	26.4	21.25	25.2	24.75	27.4

Spatio-temporal Interpolation

- $X(s,t) = \mu(s) + \eta(s,t) + e(s,t)$
- $\mu(s)$ = local mean/climatology
- $\eta(s,t) + e(s,t)$ = anomaly
- $\eta(s,t)$: predictable part of anomaly, $e(s,t)$: unpredictable
- Missing value prediction = climatology + anomaly prediction
- Climatology: estimate from past data or recent values!

Location-wise interpolation

	T1	T2	T3	T4	T5	Estimated Climatology
S1	32	30	32	x	32	31.5
S2	33	x	34	35	33	33.75
S3	25	23	x	26	26	25
S4	27	x	28	x	28	27.67
S5	x	18	19	23	18	19.5
S6	15	14	13	15	x	14.25

Anomaly Prediction

$$X_{s,t_1} = a \cdot X_{s,t_2} + b\, X_{s,t_3} \quad \forall$$

$$X_{s_1,t} = a \cdot X_{s_2,t} + b\, X_{s_3,t}$$

- Calculate anomaly at neighboring locations!
- Estimate of η(s,t) = Mean of neighboring anomalies!

- Basically a simplified model of autoregression!
- Only “neighboring” locations used as predictors with equal coefficients
- Climatology: bias term of regression!

Anomaly Calculation

	T1	T2	T3	T4	T5	Estimated Climatology
S1	0.5	-1.5	0.5	x	0.5	31.5
S2	-0.75	x	0.25	1.25	-0.75	33.75
S3	0	-2	x	1	1	25
S4	-0.67	x	0.33	x	0.33	27.67
S5	x	-1.5	-0.5	3.5	-1.5	19.5
S6	0.75	-0.25	-1.25	0.75	x	14.25

$X(s,t) - \mu(s) \neq \eta(s,t) + \cancel{\epsilon(s,t)}$

ANOMALY

GLOBAL

SPATIO-TEMP CORR.

Anomaly Calculation

	T1	T2	T3	T4	T5	Estimated Climatology
S1	0.5	-1.5	0.5	x	0.5	31.5
S2	-0.75	x	0.25	1.25	-0.75	33.75
S3	0	-2	x	1	1	25
S4	-0.67	x	0.33	x	0.33	27.67
S5	x	-1.5	-0.5	3.5	-1.5	19.5
S6	0.75	-0.25	-1.25	0.75	x	14.25

$|x - 1.25| \approx 0.9$ $|x - 0.5| \approx 2.5$ $|x - 0.5| \approx 2$

Anomaly Interpolation from Neighbors

	T1	T2	T3	T4	T5	Estimated Climatology
S1	0.5	-1.5	0.5	1.25	0.5	31.5
S2	-0.75	-1.5	0.25	1.25	-0.75	33.75
S3	0	-2	0.33	1	1	25
S4	-0.67	-2	0.33	1	0.33	27.67
S5	0.75	-1.5	-0.5	3.5	-1.5	19.5
S6	0.75	-0.25	-1.25	0.75	-1.5	14.25

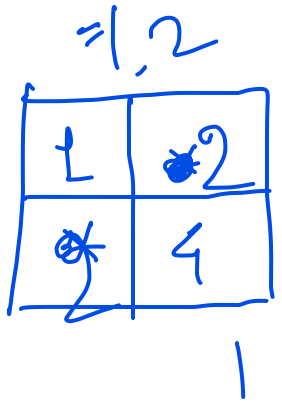
Missing value prediction

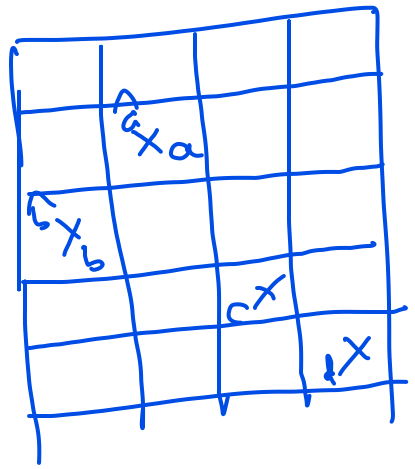
	T1	T2	T3	T4	T5	Estimated Climatology
S1	32	30	32	33	0.5	31.5
S2	33	-1.5	0.25	1.25	-0.75	33.75
S3	0	-2	0.33	1	1	25
S4	-0.67	-2	0.33	1	0.33	27.67
S5	0.75	-1.5	-0.5	3.5	-1.5	19.5
S6	0.75	-0.25	-1.25	0.75	-1.5	14.25

Missing value prediction

	T1	T2	T3	T4	T5	Estimated Climatology
S1	32	30	32	33	0.5	31.5
S2	33	-1.5	0.25	1.25	-0.75	33.75
S3	0	-2	0.33	1	1	25
S4	-0.67	-2	0.33	1	0.33	27.67
S5	0.75	-1.5	-0.5	3.5	-1.5	19.5
S6	0.75	-0.25	-1.25	0.75	-1.5	14.25

Alternative Approaches





- Low-rank matrix completion!
- Theoretical Result: If a limited fraction of entries in a matrix are missing, they can be estimated, provided the matrix rank is low!
- Rank of matrix = number of linearly independent rows/columns
- Spatio-temporal data matrices: “approximately” low-rank!

Not a low-rank matrix

	T1	T2	T3	T4	T5
S1	32	30	19	23	32
S2	33	16	28	35	26
S3	25	33	13	15	33
S4	27	14	34	29	15
S5	38	26	32	35	18
S6	15	18	24	26	28

“Approximately” low-rank matrix

	T1	T2	T3	T4	T5
S1	32	30	32	35	32
S2	33	32	34	35	33
S3	25	23	24	26	26
S4	27	26	28	29	28
S5	18	18	19	23	18
S6	15	14	13	15	15

Low-rank matrix Completion

- Frame it as an optimization problem
- Partially observed matrix M
- Set of observed entries: Ω

$$\textit{minimize} \quad \textit{rank}(X), \textit{ s.t. } X(\Omega) = M(\Omega)$$

- rank(X) is a non-convex function, difficult to optimize
- Relaxation: Nuclear Norm (sum of singular values)!
- Can be solved by specialized optimization algorithms (Candes and Recht, 2008)

Another Approach: Matrix Factorization

- Low rank matrix X = A.B
- A: (S x K) matrix, B: (K x T) matrix, K much lower than S,T
- Can we find suitable A and B matrices?
- Number of unknown values: SK + KT
- Each entry in X: X(s,t) = A(s,1)B(1,t) + A(s,2)B(2,t) + …… + A(s,k)B(k,t)
- Number of observed values in X >> SK + KT!
- Number of equations >> Number of variables (overdetermined system)
- (A*,B*) = least square solutions of A,B -> estimate X* = A*.B*!