COLUMBIA | MAILMAN SCHOOL of PUBLIC HEALTH

Lawrence G. Chillrud
Environmental Health Sciences
Mailman School of Public Health
Columbia University

Principal Component Pursuit for Pattern Recognition from Incomplete Environmental Data

ENAR 2022 Spring Meeting

March 28, 2022

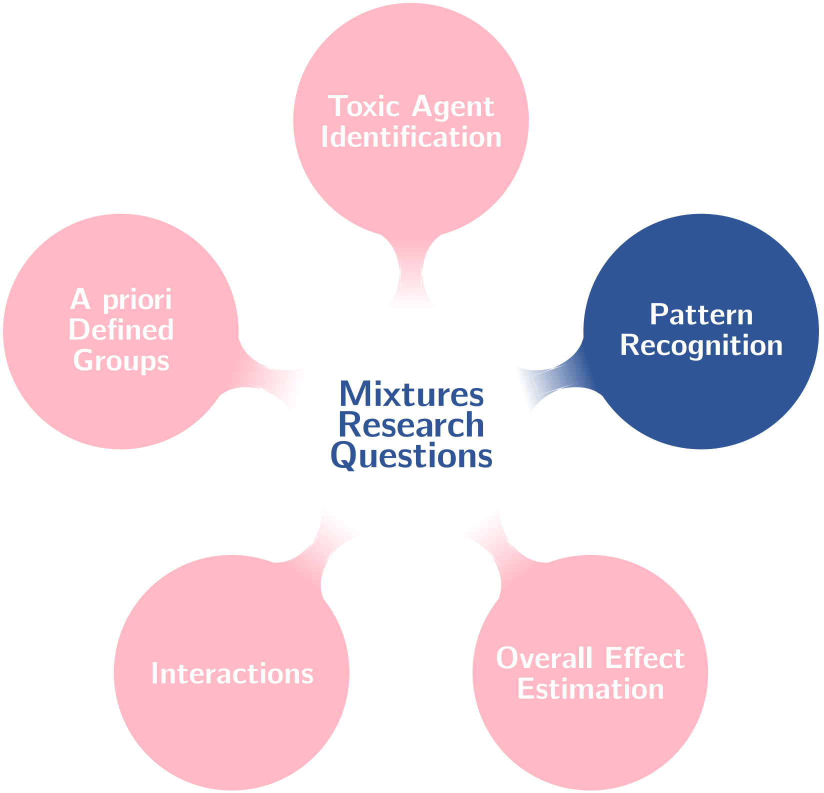
Outline

1. Background & motivation: mixtures, pattern recognition
2. Principal Component Pursuit (PCP) introduction
3. Block Missingness problem formulation
4. Extensions addressing block missingness
5. Results from simulated & applied analyses
6. Conclusion

Why study mixtures?

- Traditionally, health studies have focused on single-chemical analyses
  - E.g. lead exposure & brain development
  - This is unrealistic
- In reality, we are exposed to hundreds (thousands?) of chemicals
- The combination of exposures likely induces different responses

Why exposure pattern recognition?



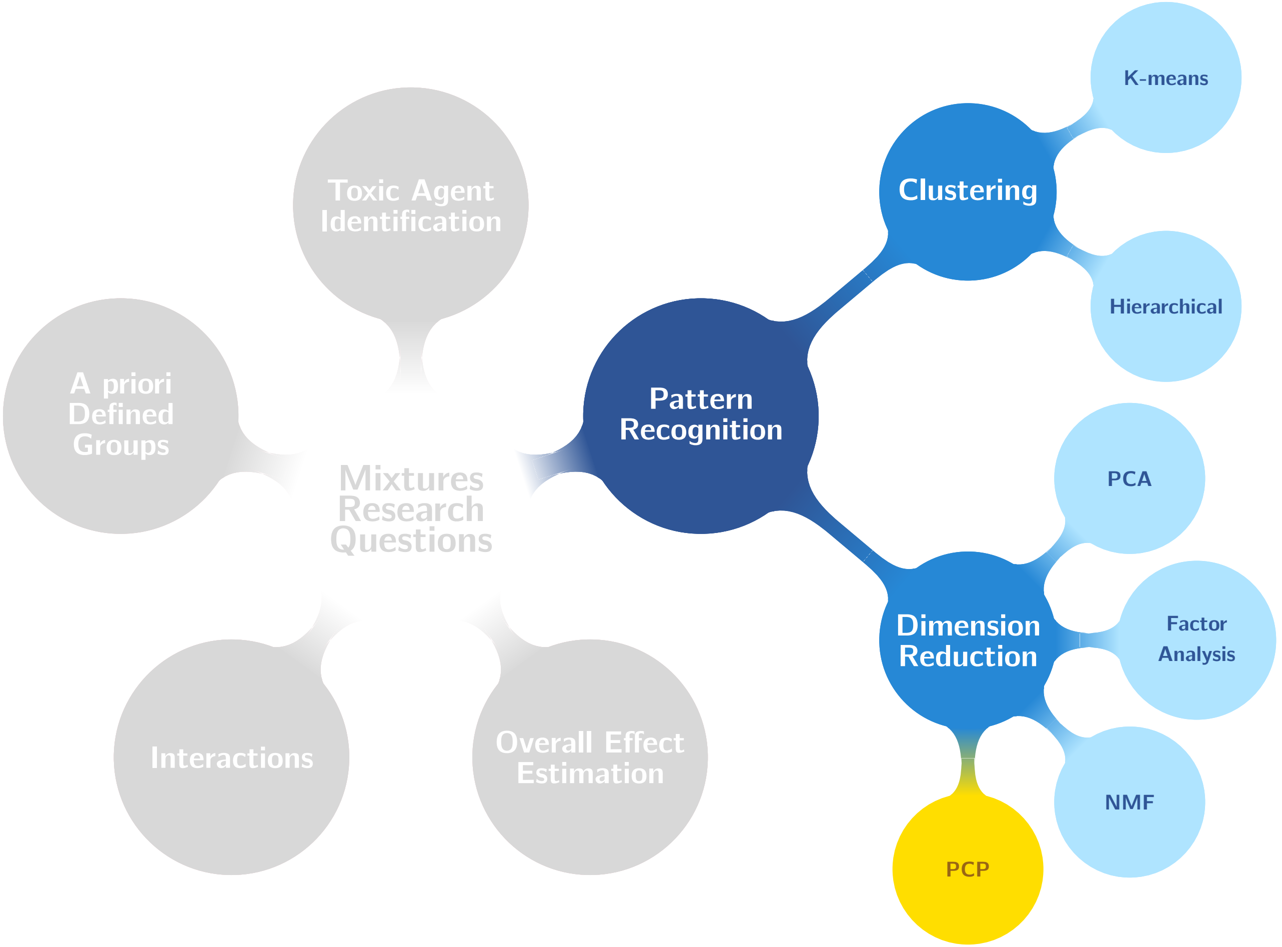
We'd like to identify:

- Sources of exposure
- Behaviors leading to exposure

Linking patterns associated with adverse health outcomes could yield:

- Efficient policy & public health regulations
- Targeted interventions

Existing pattern recognition techniques:



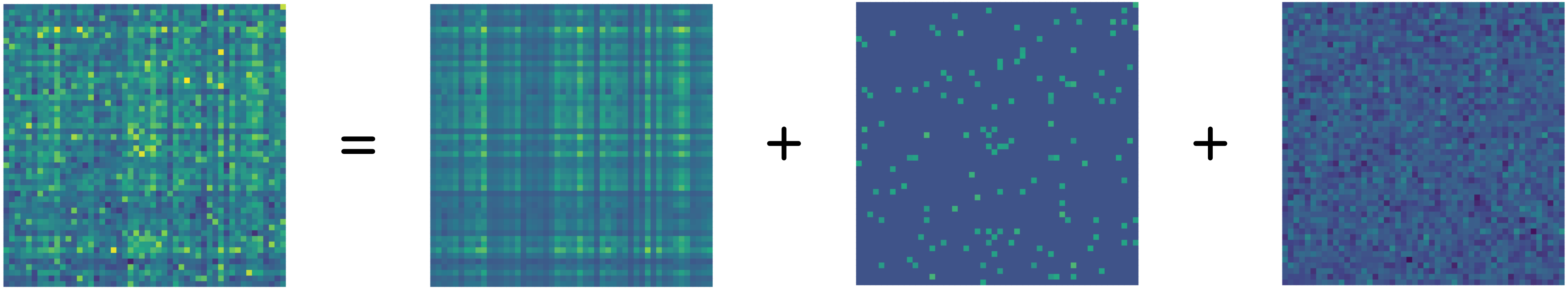
Limitations include:

- Choice of k patterns subjective
- Outliers may affect solution
- No standard for handling structured (block) missingness

Proposed solution:

- Principal Component Pursuit

Mixtures modeling



Data (D) = Low rank (L_0) + Sparse (S_0) + Noise (Z_0)

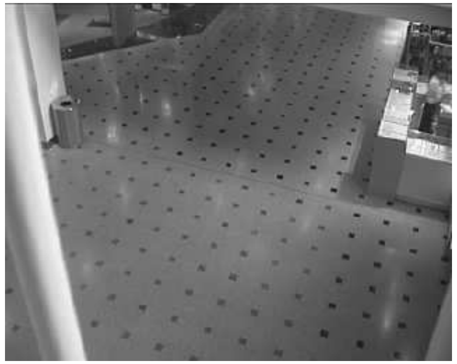
Principal Component Pursuit (PCP)

- Convex optimization algorithm from computer vision
- Performs dimension reduction by decomposing data matrix into:
  1. Low-rank *(L)* → consistent patterns
  2. Sparse *(S)* → unique or outlying events

$$\sqrt{PCP} := \min_{\boldsymbol{L},\boldsymbol{S}} ||\boldsymbol{L}||_* + \lambda||\boldsymbol{S}||_1 + \mu||\boldsymbol{L} + \boldsymbol{S} - \boldsymbol{D}||_F$$

(Zhang et al., 2021)

Original



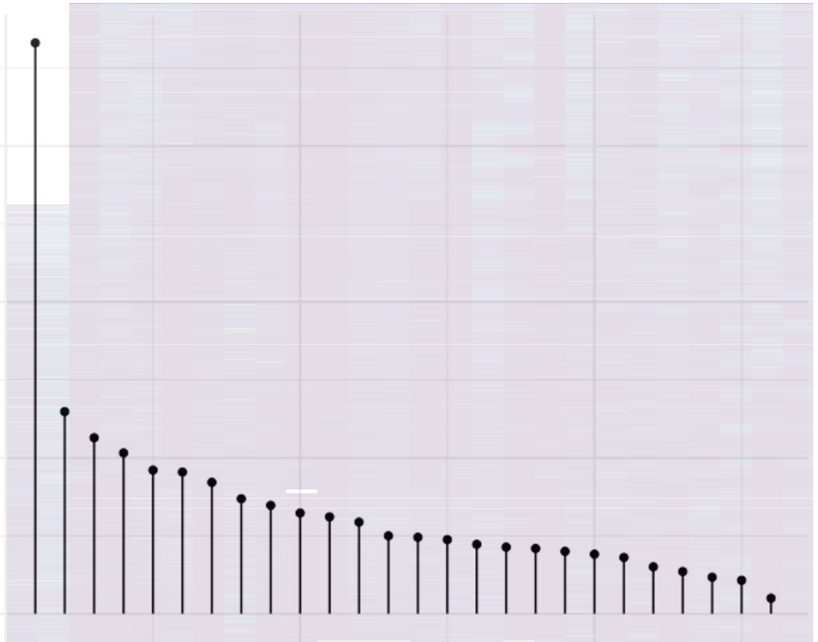
Low-rank *(L)*

Sparse *(S)*

Benefits of PCP

- Robust to noisy data
- Researcher does not need to choose k
- Extreme events not discarded & do not influence patterns
- Improved predictive accuracy over PCA

Adapting PCP for use in environmental health

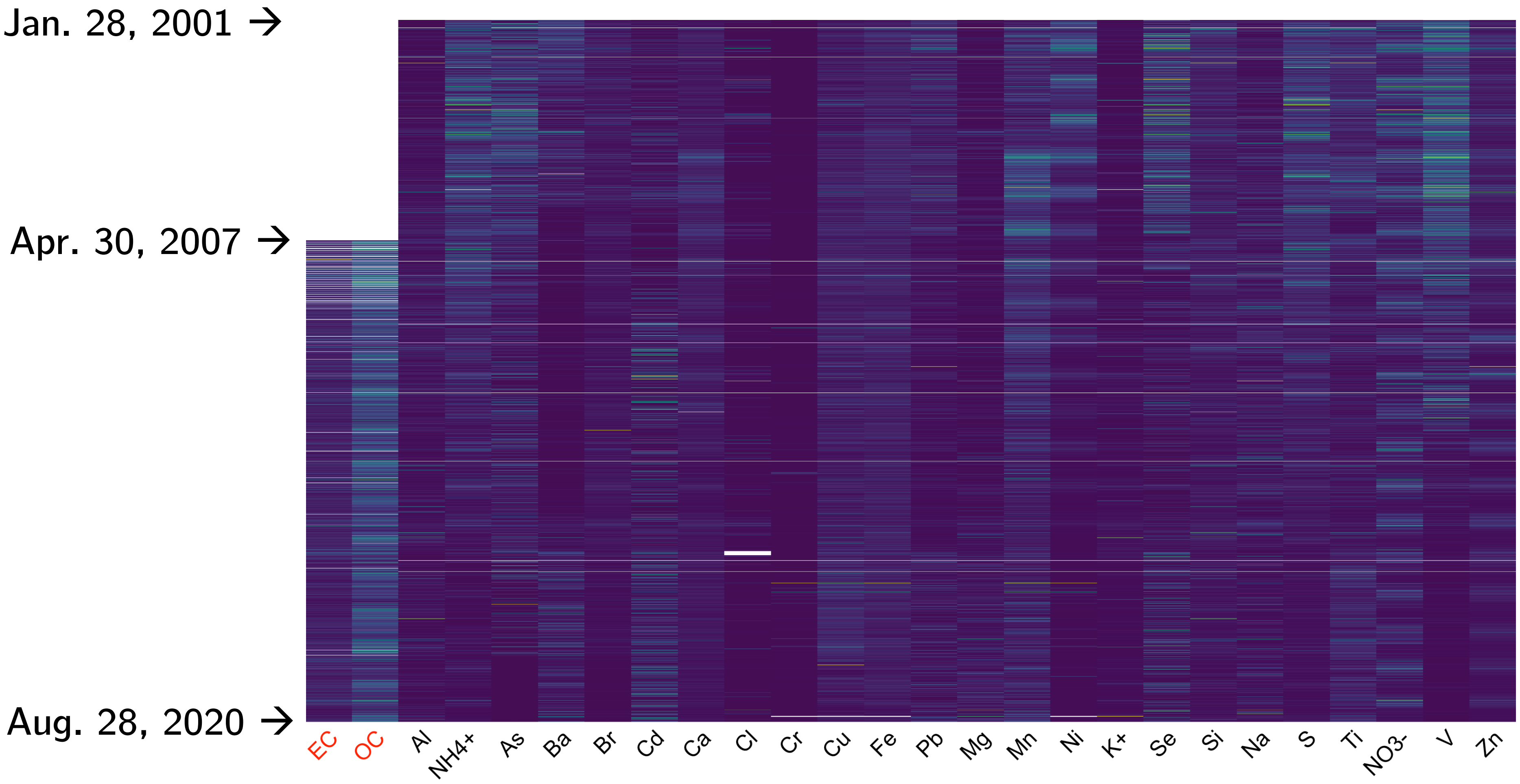


$$ncPCP = \min_{L,S} \mathbb{1}_{rank(L) \le r} + \lambda||S||_1 + \mu||L + S - D||_F$$

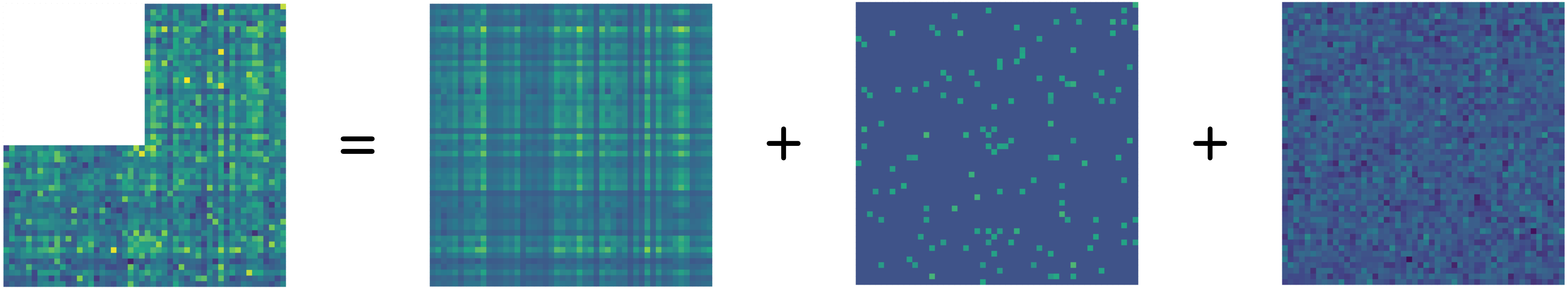
$$\sqrt{PCP} = \min_{L,S} ||L||_* + \lambda||S||_1 + \mu||L + S - D||_F$$

(Gibson et al., 2021)

EPA AQS $PM_{2.5}$ data: NYC, 2001 - 2020



The block missingness problem



Data $(\tilde{D})$ = Low rank (L_0) + Sparse (S_0) + Noise (Z_0)

The problem: can we recover L_0 from $\tilde{D}$?

How does PCP handle block missingness?

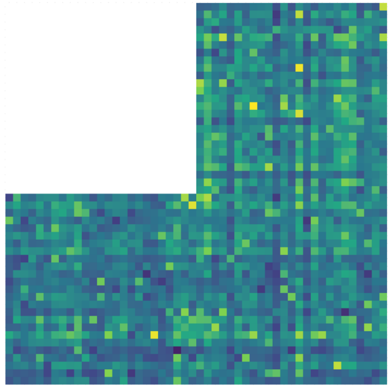
A key operation in PCP is taking a projected gradient step on the rank-r matrix $\boldsymbol{L}$. Formally,

$$\boldsymbol{L}_{k+1} = \mathcal{P}_r\left[\boldsymbol{L}_k - t\mathcal{P}_\Omega(\boldsymbol{L}_k + \boldsymbol{S}_k - \boldsymbol{D})\right]$$

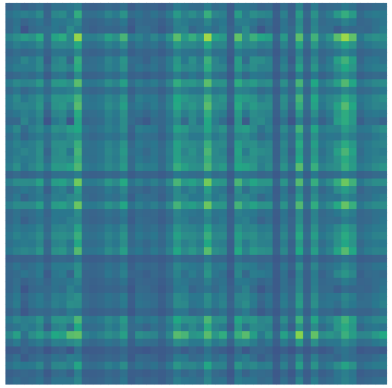
where $\mathcal{P}_r$ finds the closest rank-r approximation to a given input matrix.

→ This is just a truncated SVD / PCA

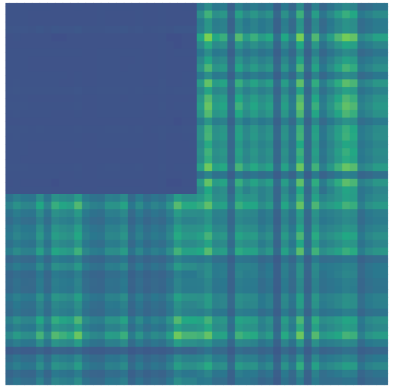
$\mathcal{P}_r$ struggles with block missingness...



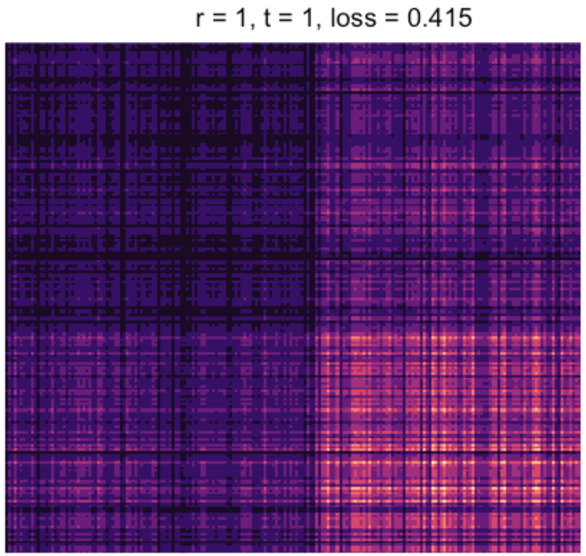
Data $(\tilde{D})$



Target L_0



$\mathcal{P}_r(\tilde{D}) = \hat{L}$



PCP

...so PCP will struggle with it as well

$$L_{k+1} = \mathcal{P}_r\left[L_k - t\mathcal{P}_\Omega(L_k + S_k - D)\right]$$

$$L_{k+1} = \mathcal{P}_{nystrom}\left[L_k - t\mathcal{P}_\Omega(L_k + S_k - D)\right]$$

Structure-aware Nystrom extension to $\mathcal{P}_r$

$\mathcal{P}_{nystrom}(\boldsymbol{W})$:

$$\widehat{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{W}_{12}[\mathcal{P}_r(\boldsymbol{W}_{22})]^{\dagger}\boldsymbol{W}_{21} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix}$$

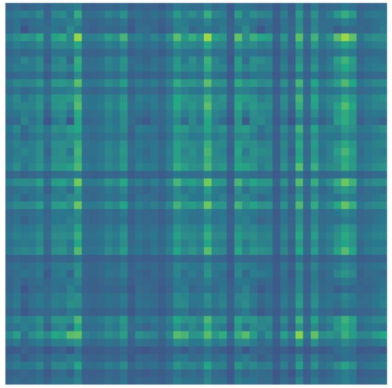
$$\mathcal{P}_r\left(\widehat{\boldsymbol{W}}\right)$$

Main idea: $\boldsymbol{W}_{11} = \boldsymbol{W}_{12}[\mathcal{P}_r(\boldsymbol{W}_{22})]^{\dagger}\boldsymbol{W}_{21}$

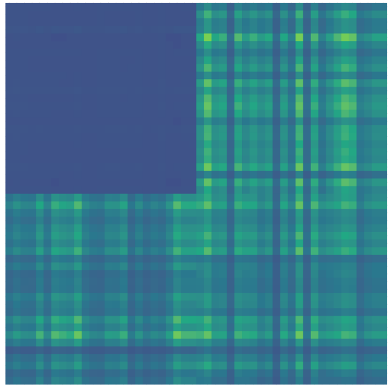
Key takeaways from $\mathcal{P}_{nystrom}$

- We are reconstructing the missing block from observed data
- This formulation is exact in no-noise conditions
  - As noise levels increase it becomes harder to recover missing block
- Main assumption: The missing block is characterized by the same patterns governing the observed blocks

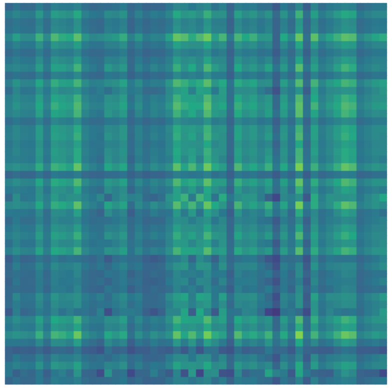
Simulation results



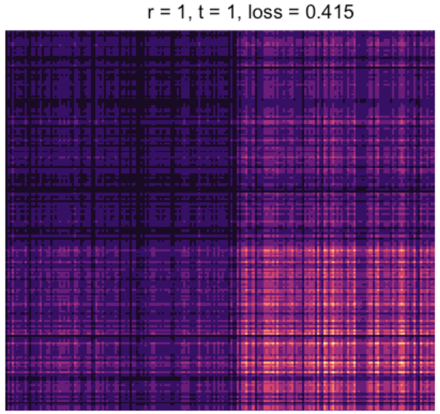
Target $\boldsymbol{L_0}$



$\mathcal{P}_r(\tilde{\boldsymbol{D}})$



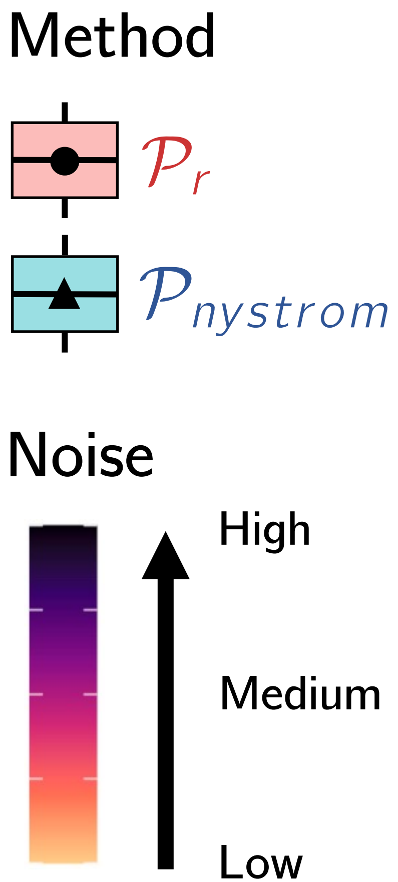
$\mathcal{P}_{nystrom}(\tilde{\boldsymbol{D}})$

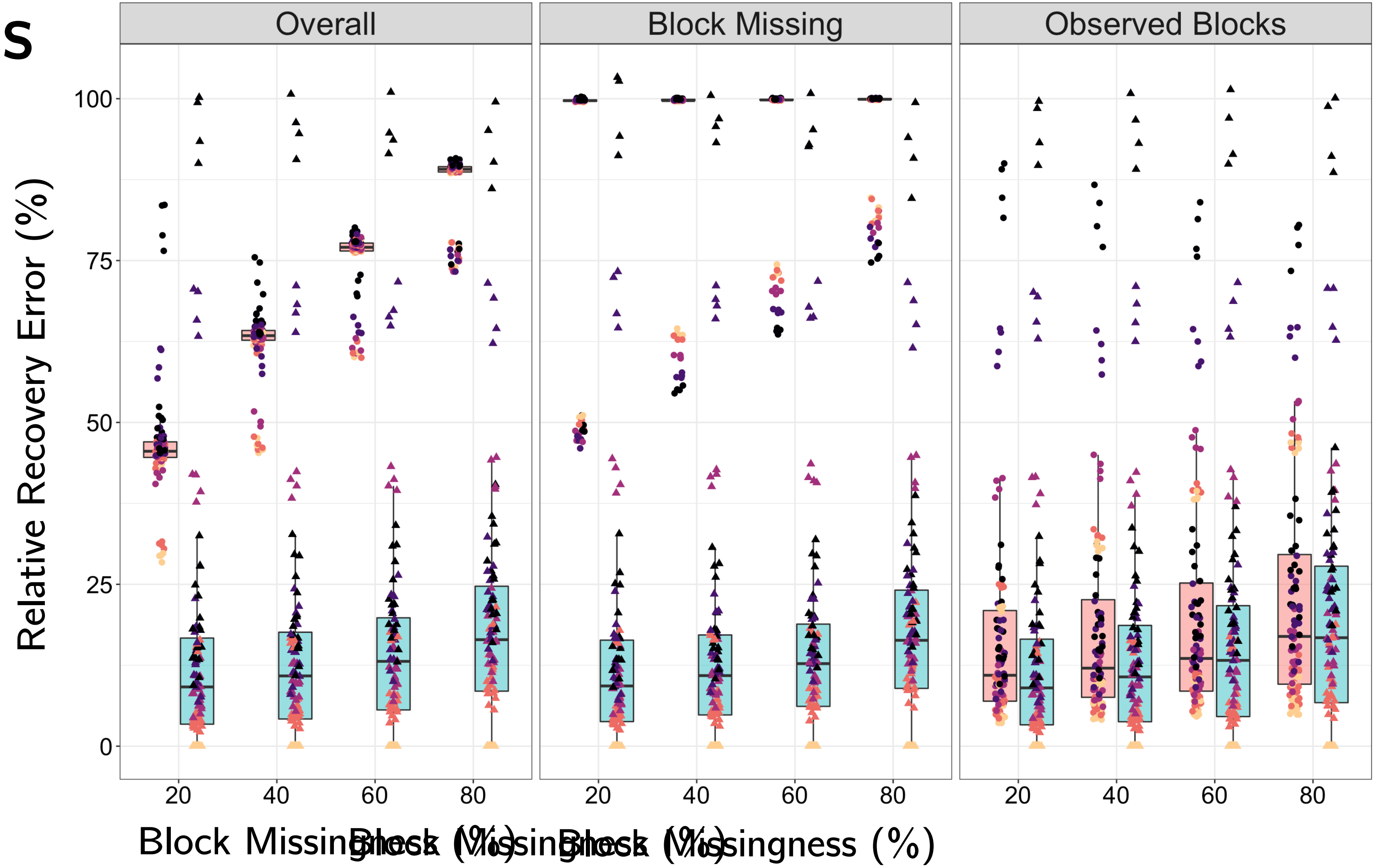


← PCP w/ $\mathcal{P}_r$

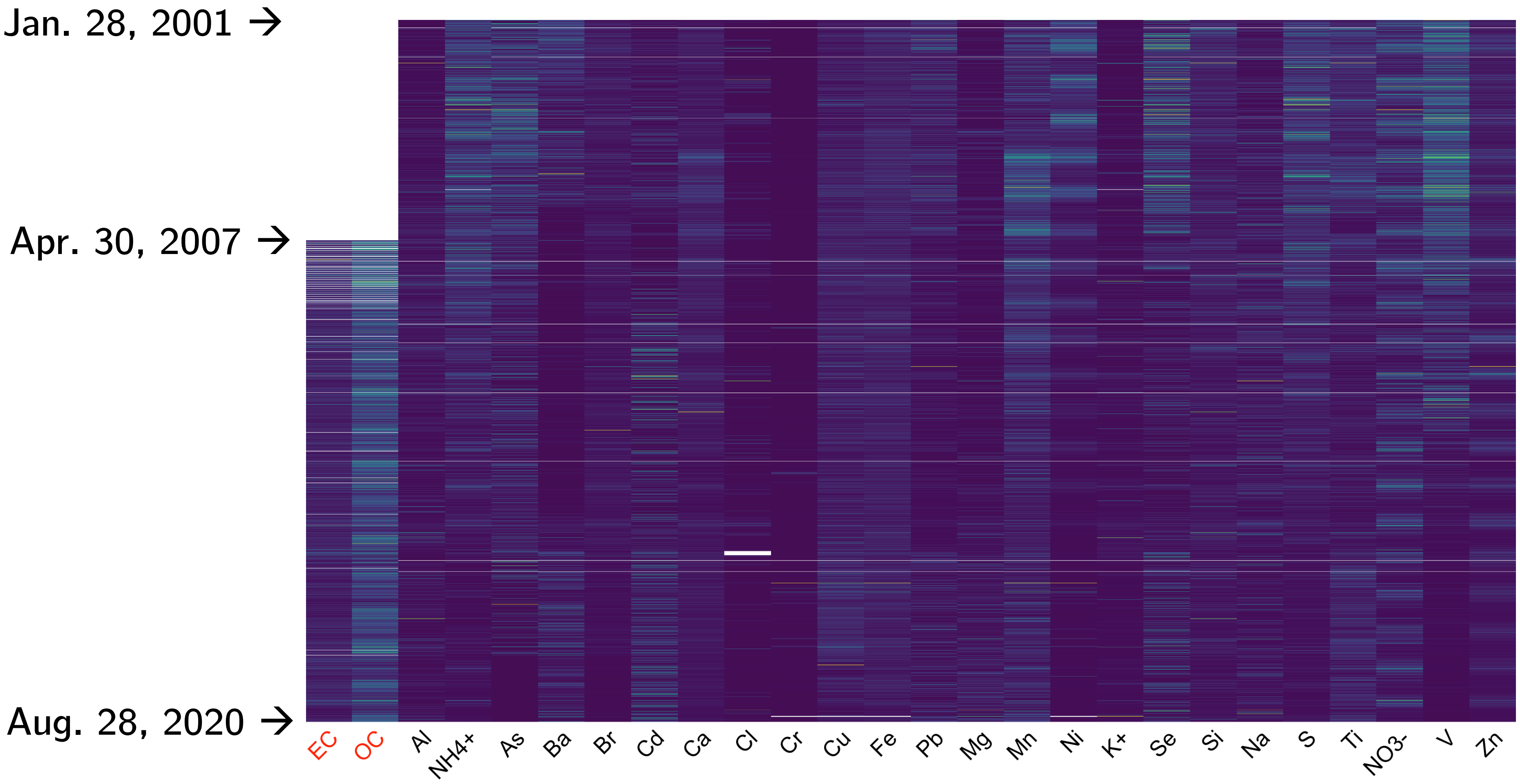
← PCP w/ $\mathcal{P}_{nystrom}$

Simulation results

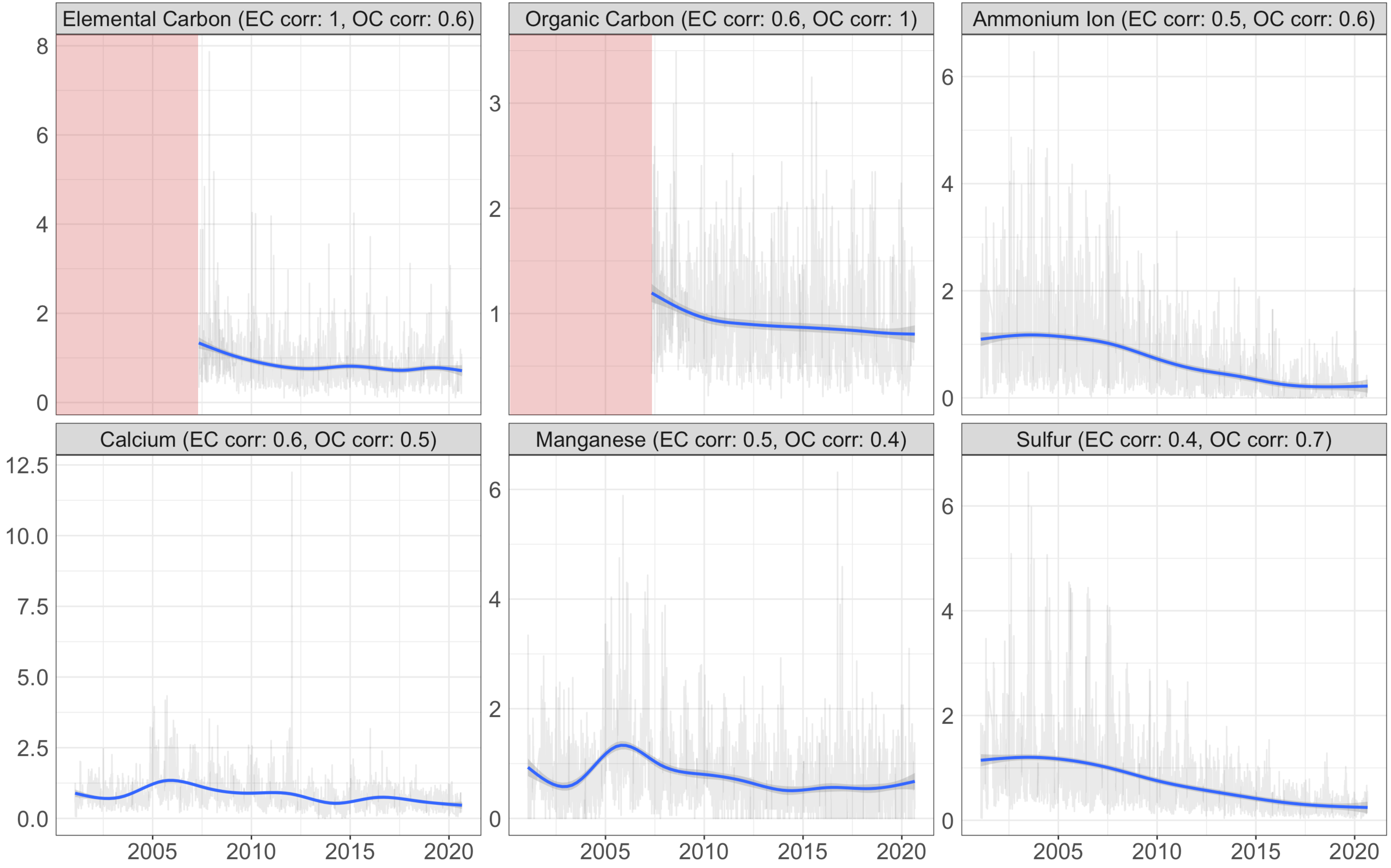




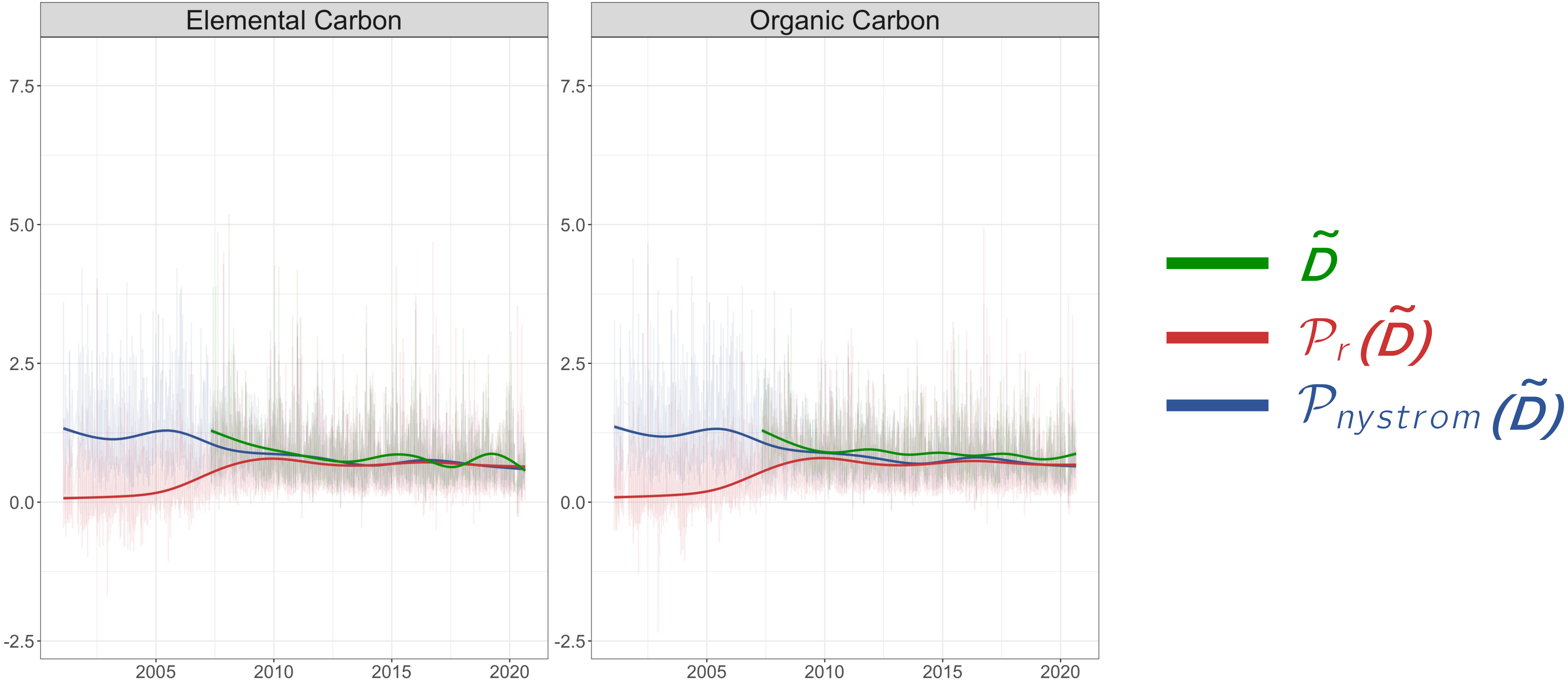
EPA AQS $PM_{2.5}$ data: NYC, 2001 - 2020



NYC Data


Elemental Carbon (EC corr: 1, OC corr: 0.6)
Organic Carbon (EC corr: 0.6, OC corr: 1)
Ammonium Ion (EC corr: 0.5, OC corr: 0.6)
Calcium (EC corr: 0.6, OC corr: 0.5)
Manganese (EC corr: 0.5, OC corr: 0.4)
Sulfur (EC corr: 0.4, OC corr: 0.7)
0 2 4 6 8
1 2 3
0 2 4 6
0.0 2.5 5.0 7.5 10.0 12.5
0 2 4 6
0 2 4 6
2005 2010 2015 2020

Results – EPA AQS $PM_{2.5}$ data: NYC, 2001 - 2020



Conclusion

- The Nystrom extension improves recovery of missing block
- This formulation is exact in no-noise conditions
  - As noise levels increase it becomes harder to recover the missing block
- Main assumption: The missing block is characterized by the same patterns governing the observed blocks
- PCP equipped w/Nystrom extension serves as a useful pattern recognition tool

Future Work github.com/Columbia-PRIME/pcpr

- Tackle overlapping block missingness
- Explore extensions for high-noise situations
- Investigate uncertainty characterization

Acknowledgements

Columbia University PRIME Team:


Marianthi-Anna Kioumourtzoglou
Environmental Health


John Wright
Electrical Engineering

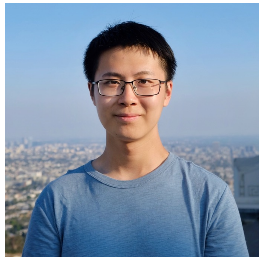

Jeff Goldsmith
Biostatistics


Elizabeth Gibson
Environmental Health


Yanelli Núñez
Environmental Health


Jaime Benavides
Environmental Health


Robert Colgan
Computer Science


Jingkai Yan
Electrical Engineering


Rachel Tao
Epidemiology


Junhui Zhang
Applied Physics &
Applied Mathematics

makLab

Arin Balalian, Tanya Butt, Ilan Cerna-Turoff, Gali Cohen, Vivian Do, Maggie Li, Robbie Parks, Sebastian Rowland, Roheeni Saxena, Jenni Shearston, Sabah Usmani

Supported by:

NIEHS
R01 ES028805
P30 ES009089
F31 ES030263

Contact:

lgc2139@cumc.columbia.edu
lawrence-chillrud.github.io

Mathematical intuition behind why $W_{11} = W_{12}[\mathcal{P}_r(W_{22})]^{\dagger} W_{21}$

$W_{22}^{\dagger}$

Simple scenario:

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} U_1V_1 & U_1V_2 \\ U_2V_1 & U_2V_2 \end{bmatrix} \quad rank(W) = r$$

$$W_{11} = U_1 V_1$$

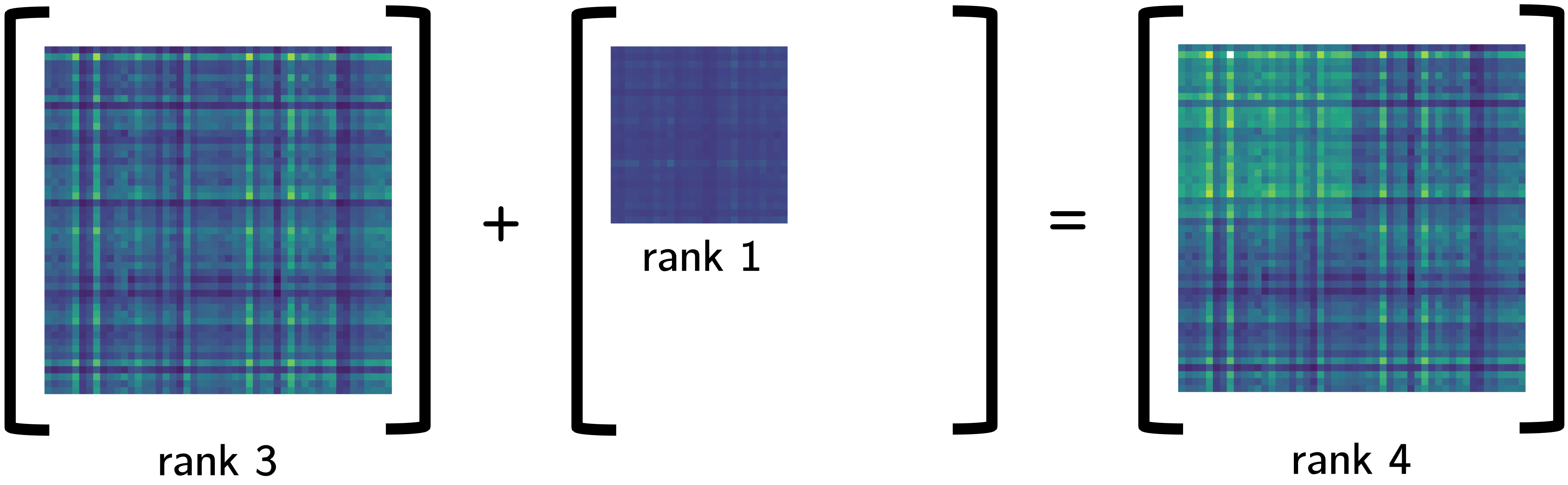
$$W_{12} W_{22}^{\dagger} W_{21} = U_1 \underbrace{V_2 (U_2 V_2)^{\dagger} U_2}_{I_{r \times r}} V_1 = U_1 V_1 = W_{11}$$

~~U_2, V_2 are invertible~~

~~$rank(U_2) = rank(V_2) \leqslant r$~~

$$V_2 (U_2 V_2)^{\dagger} U_2 = V_2 (U_2 V_2)^{-1} U_2 = V_2 V_2^{-1} U_2^{-1} U_2 = I_{r \times r}$$

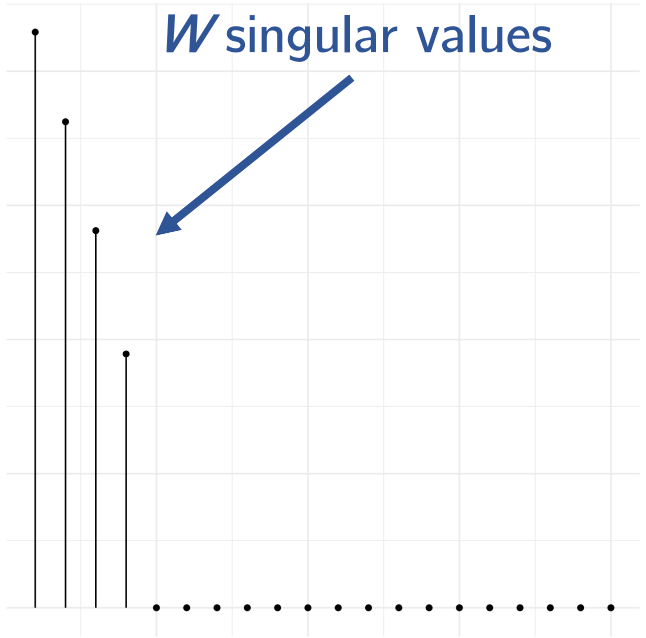
When $rank(\boldsymbol{U}_2) = rank(\boldsymbol{V}_2) < r$



Mathematical intuition behind why $W_{11} = W_{12}[\mathcal{P}_r(W_{22})]^\dagger W_{21}$

Simple scenario was noise free!

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \quad rank(W) = r$$



Real world scenario is not

$$\widetilde{W} = \begin{bmatrix} \widetilde{W}_{11} & \widetilde{W}_{12} \\ \widetilde{W}_{21} & \widetilde{W}_{22} \end{bmatrix} \quad rank(\widetilde{W}) > r$$