Principal component pursuit (PCP) for exposure pattern recognition

Lawrence Gardner Chillrud[1]*, Elizabeth Atkeson Gibson[1], Yanelli Nunez[1], Robert Edward Colgan[2], Rachel Hannah Tao[3], Junhui Zhang[4], Jingkai Yan[5], John Wright[5], Jeff Goldsmith[6], Marianthi-Anna Kioumourtzoglou[1]

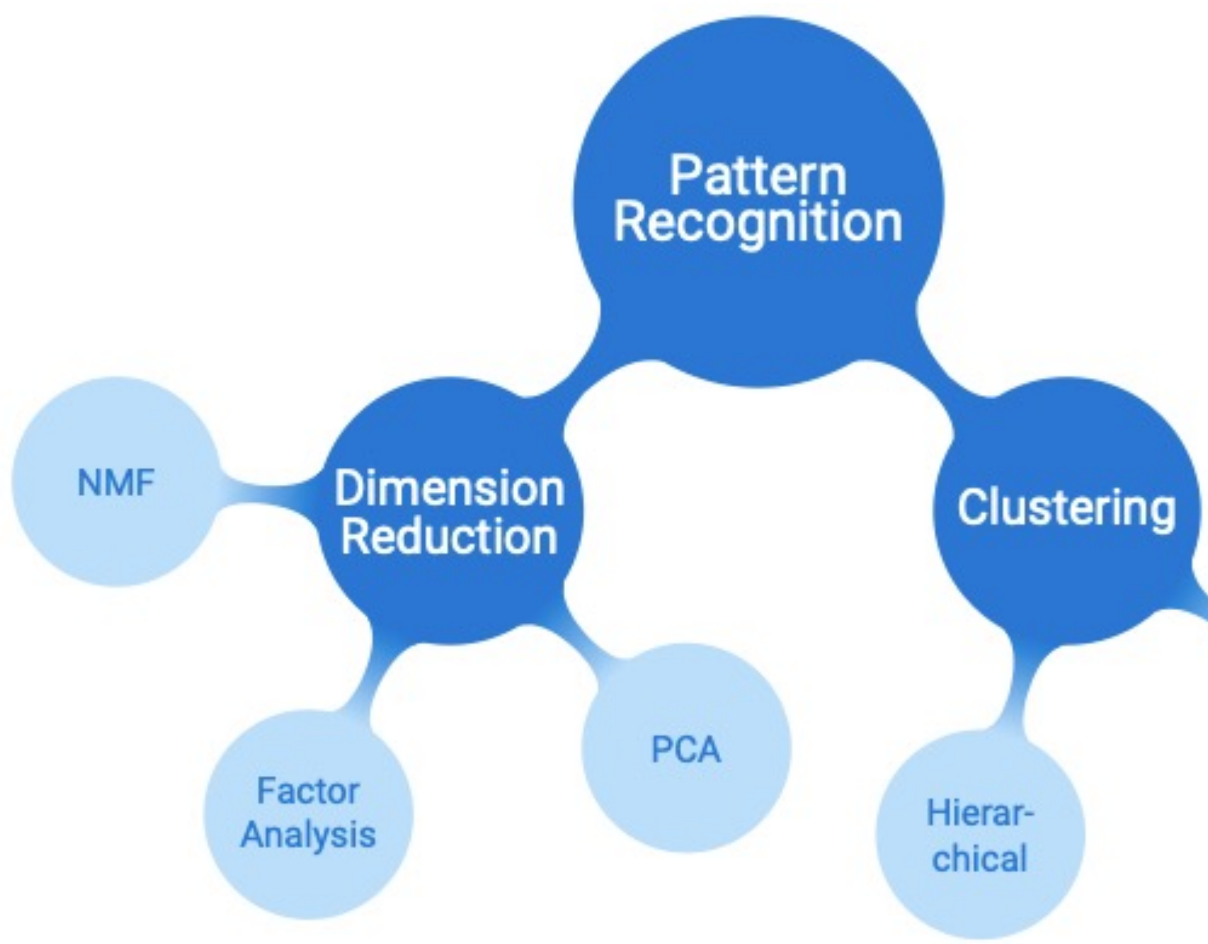
Departments of [1]Environmental Health Sciences, [2]Computer Science, [3]Epidemiology, [4]Applied Physics and Applied Mathematics, [5]Electrical Engineering, [6]Biostatistics, Columbia University, New York, NY, USA

*Correspondence: lgc2139@cumc.columbia.edu

ISEE '21 NEW YORK CITY

Key findings: PCP can serve as a useful and robust technique to identify exposure patterns that are amenable to public health messaging and research.

Background



Why pattern recognition?
- Identify sources of exposure
- Identify behaviors that lead to exposure
- Can help drive targeted interventions, inform policy, and shape public health regulations

Limitations of many existing pattern recognition techniques in environmental health (EH) include:
- Results often uninterpretable
- Outliers may affect solution
- Chemical concentrations may be below the limit of detection (LOD)

Methods

We have adapted PCP, a robust dimensionality reduction algorithm in computer vision, to pattern recognition in EH.

PCP is a convex optimization program that decomposes a data matrix into:
- Low-rank matrix ($\hat{L}$) → consistent patterns of exposure
- Sparse matrix ($\hat{S}$) → unique or outlying exposure events

Original

$\hat{L}$

$\hat{S}$

Adaptations for environmental health:

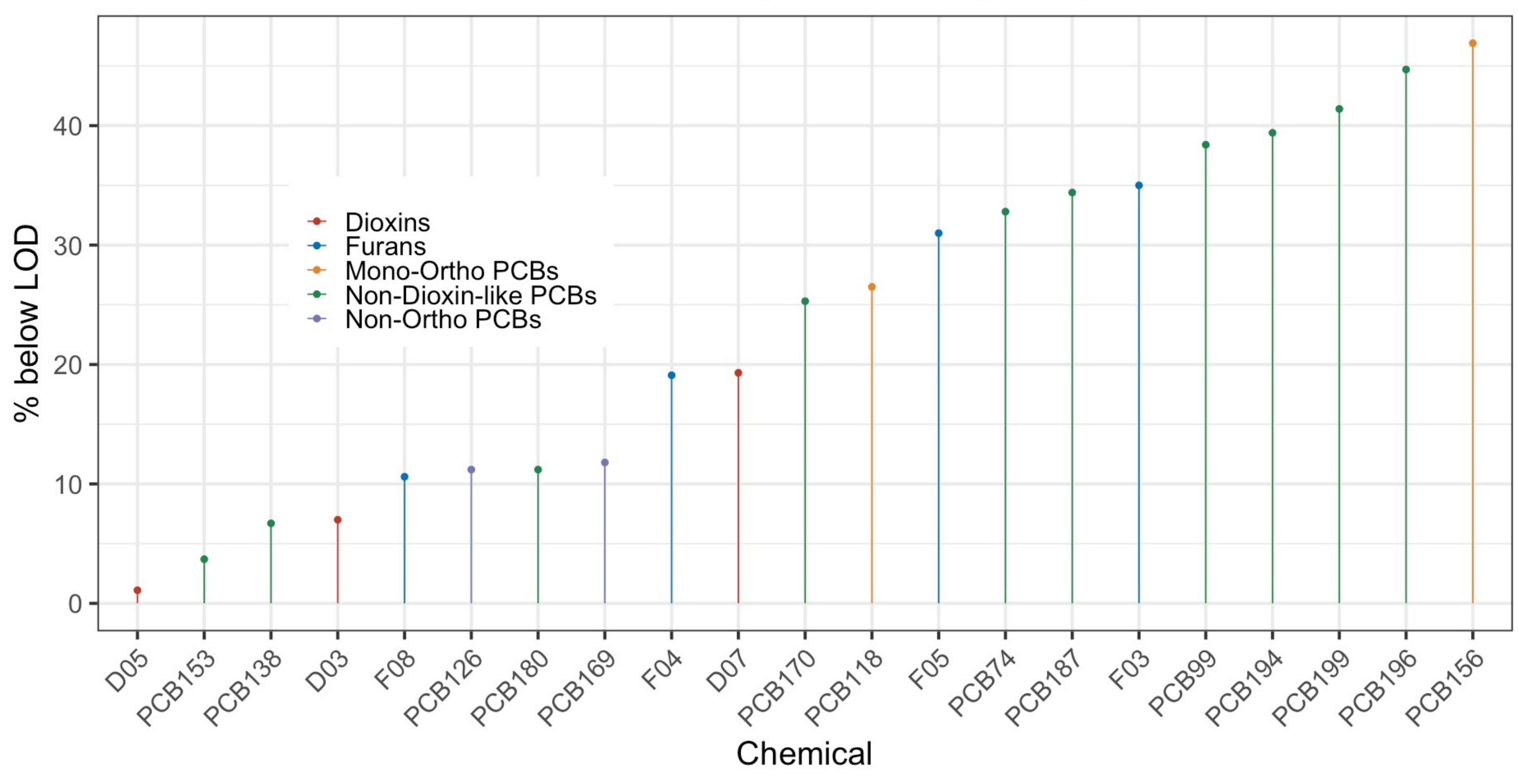
1. Non-negativity constraint on $\hat{L}$
2. Can accommodate missingness
3. Novel penalties for observations < LOD
4. Non-convex approach that better models EH data

Non-convex objective function:

$$nc\sqrt{PCP} := \min_{L,S} \mathbf{1}_{rank(L)\leq r} + \lambda||S||_1 + \mu||L+S-X||_F + \mathbf{1}_{L\geq 0}$$

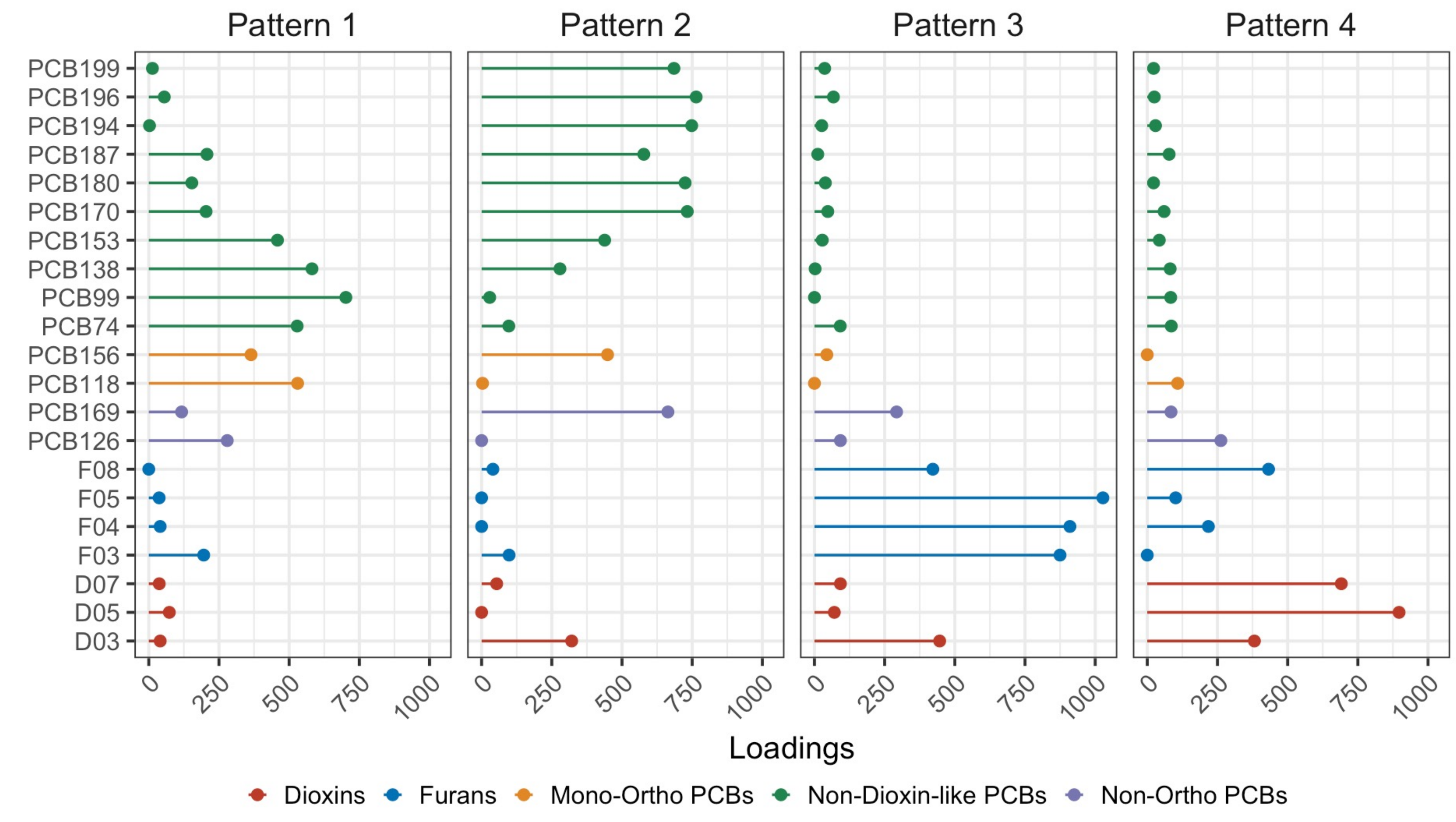
Applied $nc\sqrt{PCP}$ to NHANES:
- 2001 – 2002
- 1,000 U.S. adults
- Mixture of 21 persistent organic pollutants (POPs)
- % < LOD varies by POP

NHANES data: 1000 U.S. adults x 21 persistent organic pollutants



Results

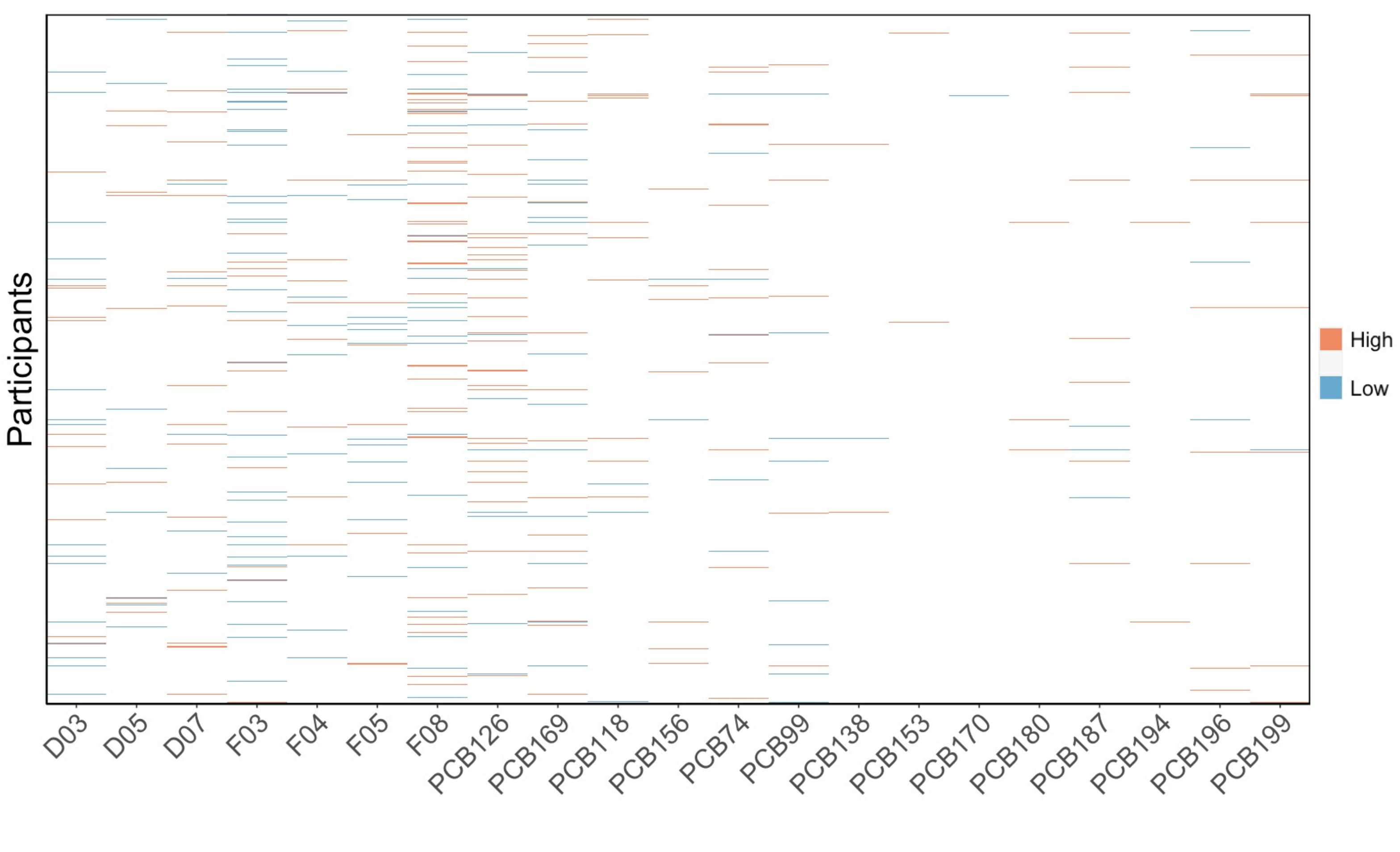
NHANES: patterns identified from PCP $\hat{L}$ matrix



PCP extracted 4 patterns from NHANES data, each with a distinct profile:
1. Lower weight PCBs
2. Higher weight PCBs
3. Furans
4. Dioxins

Conclusions

NHANES: unique exposure events in the $\hat{S}$ matrix



Summary:
- Identified exposure patterns not influenced by outliers
- Extreme events separated, not discarded
- Novel < LOD penalty
- Enhanced interpretability
- Can handle missingness

PCP results can be used in health models to identify those sources or behaviors that are harmful to human health.

References:
1. Zhang J, Yan J, Wright J. Square Root Principal Component Pursuit: Tuning-Free Noisy Robust Matrix Recovery. arXiv:2106.09211.

Acknowledgments:
NIEHS R01 ES028805, P30 ES009089, and F31 ES030263