Elizabeth A. Gibson
Environmental Health Sciences
Mailman School of Public Health
Columbia University

# Principal Component Pursuit for Pattern Identification in Environmental Health

*Joint Statistical Meetings*

August 4, 2020

# Why care about mixtures?

- We are exposed to hundreds (thousands?) of chemicals at any single time point
- Traditionally, epi studies have focused on single-chemical analyses
    - This does not represent reality
- The combination of exposures likely induces different responses

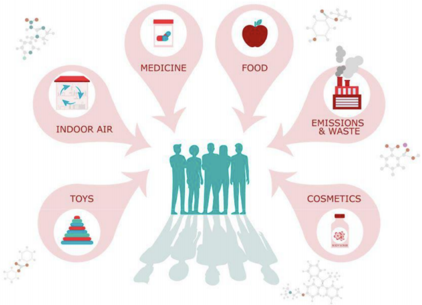
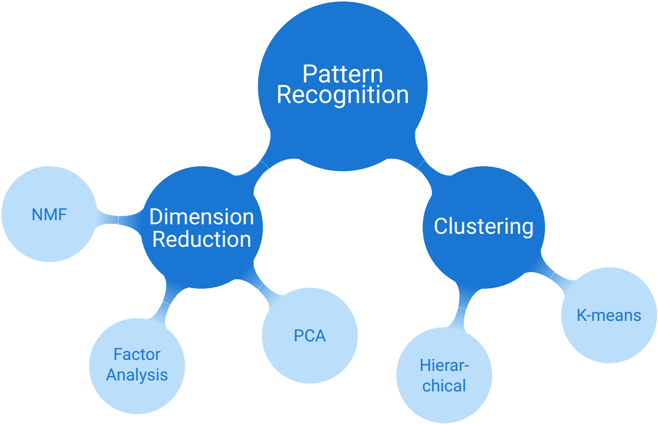



Image: ec.europa.eu via Yanelli Núñez

# Exposure pattern recognition

- Why should we care about identifying exposure patterns to chemicals in a population?
    - Sources
    - Behaviors
- If we link these patterns to (multiple) adverse health outcomes
    - Efficient regulations
    - Targeted interventions

# Some existing pattern recognition methods




*Not an exhaustive list of methods!!

# Problems with existing methods

- Choice of $k$ patterns/components/factors is subjective
- Local minima depend on initialization
- Outliers may affect solution
- Chemical concentrations may be <LOD

$\Longrightarrow$ Proposed solution: **Principal Component Pursuit**

# Principal Component Pursuit

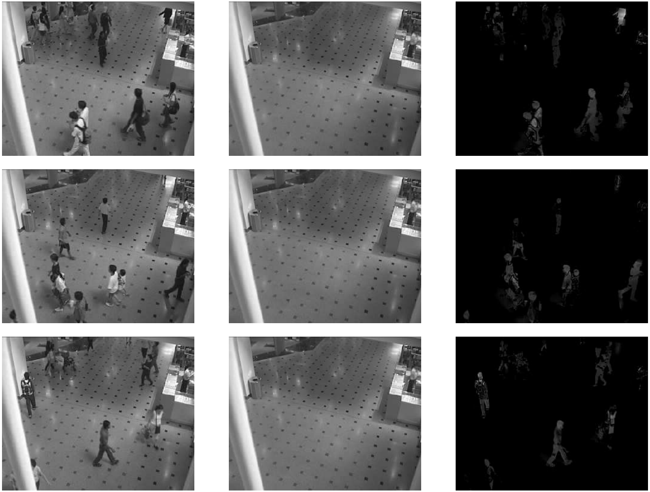
- Robust Principal Component Analysis (PCA)
- Unsupervised dimensionality reduction method adapted from computer vision
- Decomposes design matrix into low rank and sparse
  - Low rank matrix estimates consistent exposure patterns
  - Sparse matrix identifies unique events

$$\min_{L,S} \|L\|_\star + \lambda\|S\|_1 + \frac{\mu}{2}\|L + S - X\|_F^2$$

- Robust to noisy/corrupt data
- Global minimum

# PCP image example

Original | $\widehat{L_0}$ | $\widehat{S_0}$



## PCP extensions

- Non-negativity constraint on low rank matrix
- Novel penalties for values < LOD
  - Observed value < LOD & predicted value > LOD

$$\min_{L,S} \|L\|_\star + \lambda \|S\|_1 + \frac{\mu}{2} \|L + S - LOD\|_F^2 \tag{1}$$

  - Observed value < LOD & predicted value < 0

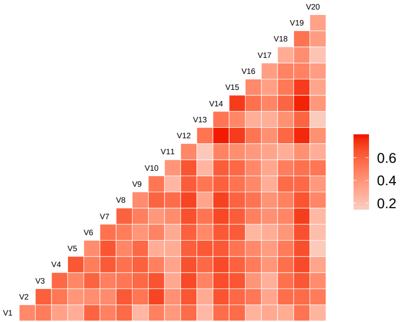
$$\min_{L,S} \|L\|_\star + \lambda \|S\|_1 + \frac{\mu}{2} \|L + S\|_F^2 \tag{2}$$

  - Observed value < LOD & predicted value [0 – LOD]

$$\min_{L,S} \|L\|_\star + \lambda \|S\|_1 \tag{3}$$

# Simulations

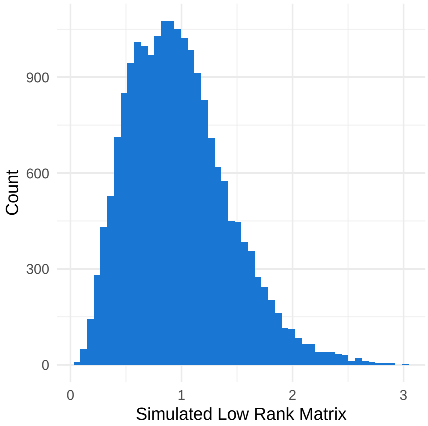
- Matrix size
  - 1,000 $\times$ 20
- Low rank structure
  - Uniform distributions
  - Matrix product
  - Rank: 4
- Added noise
  - Gaussian
  - 0.6 $\times$ low rank SD
- Values <LOD
  - 0%–90%



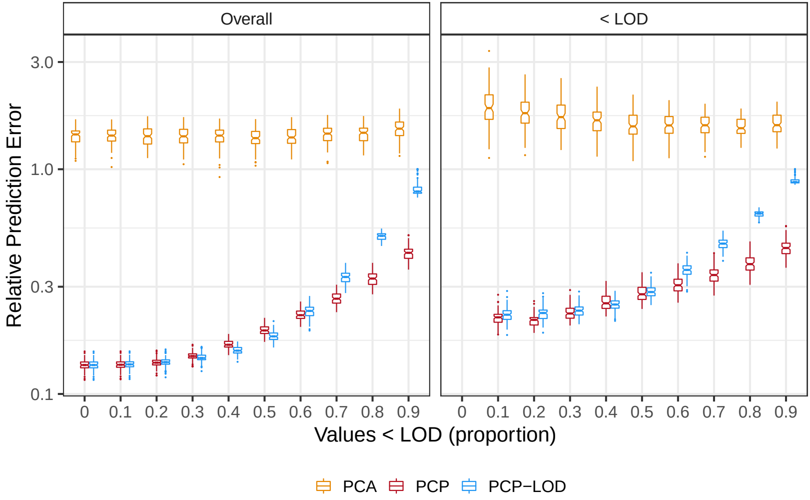

Simulated Correlation Matrix

# Simulations

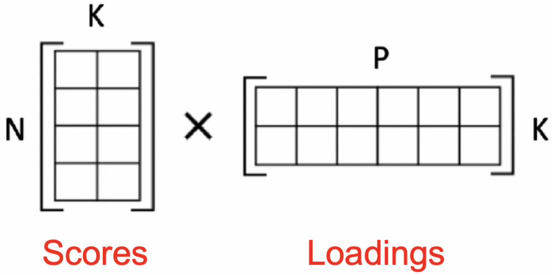
- Matrix size
  - $1{,}000 \times 20$
- Low rank structure
  - Uniform distributions
  - Matrix product
  - Rank: 4
- Added noise
  - Gaussian
  - $0.6 \times$ low rank SD
- Values <LOD
  - 0%–90%



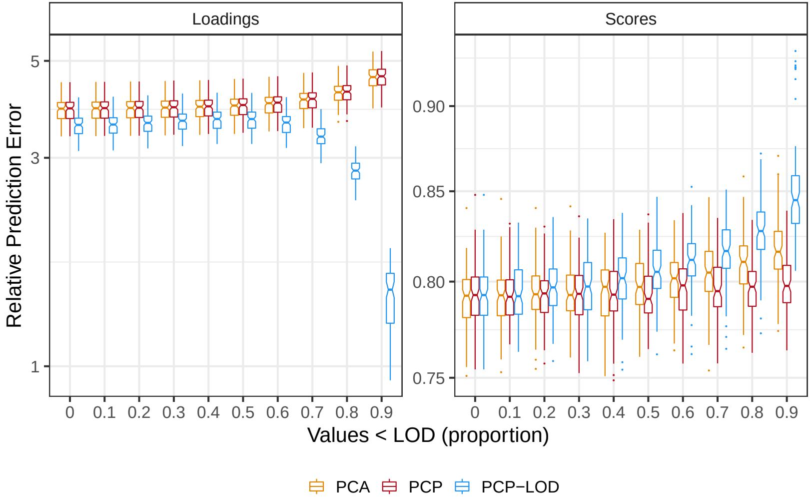
# Relative error overall & <LOD



# Relative error in loadings and scores



# Relative error in loadings and scores



## Results

- PCP-LOD outperforms PCA
- PCP-LOD outperforms PCP imputed with LOD/$\sqrt{2}$ under these conditions:
  - True underlying low-rank structure exists
  - Proportion < LOD is low

# Conclusion

Benefits of PCP:

- Researcher does not need to choose $k$
- Global minimum
- Improved predictive accuracy over PCA
- Information on extreme events not lost / does not influence patterns

Benefits of PCP-LOD:

- Do not need to impute values < LOD
- Outperforms PCP imputed with LOD/$\sqrt{2}$ when LOD is low

# Next steps

Immediate next steps:

- Add penalty for known values <LOD
- Determine optimal $\mu$ parameter value / range
- Apply to real environmental data

Where to take the method:

- What to do with S?
- Non-negative pattern identification in L
- User-friendly R package

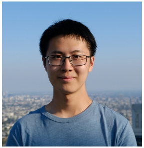
# Acknowledgements

## Columbia University PRIME Team:

Marianthi-Anna Kioumourtzoglou
Environmental Health

John Wright
Electrical Engineering

Jeff Goldsmith
Biostatistics



Jingkai Yan
Electrical Engineering

Robert Colgan
Computer Science

Lawrence Chillrud
Computer Science

makLab

Mike He
Maggie Li
Yanelli Núñez
Robbie Parks
Sebastian Rowland
Jenni Shearston
Rachel Tao

PhD Advisor:
Julie B. Herbstman

Supported by:
NIEHS F31 ES030263 &
PRIME R01 ES028805

e.a.gibson@columbia.edu