Infinite Hidden Markov Models for Multiple Multivariate Time Series with Missing Data

Lauren Hoskovec, Matthew D. Koslovsky, Kirsten Koehler, Nicholas Good, Jennifer L. Peel, John Volckens and Ander Wilson

ENAR 2023

Low-Cost, Real-Time Sensors

- Air pollution exposure causes a wide variety of negative health outcomes
- Most studies rely in area-level exposure such as from a centrally located monitor or modeled exposure surface
- Low-cost, real-time sensors offer the promise to measure air pollution at the individual level
- Measure exposure high temporal resolution
- Moves with individuals

[1] Figure source: https://finance.yahoo.com

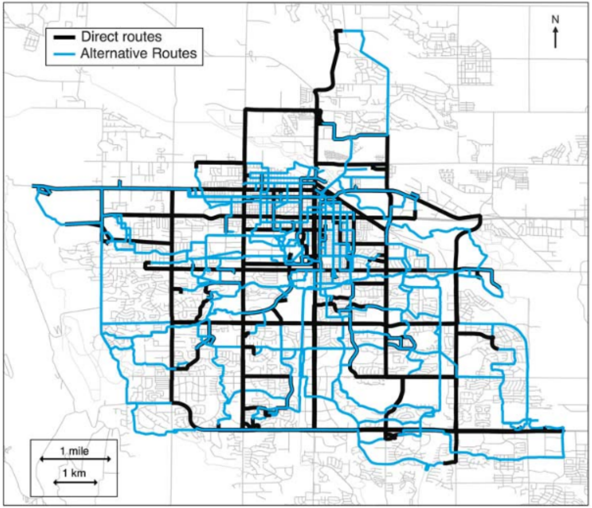
Low-Cost, Real-Time Sensors

- Lots of data
- Lots of promise
- Lots of challenges

[1]Figure source: https://finance.yahoo.com

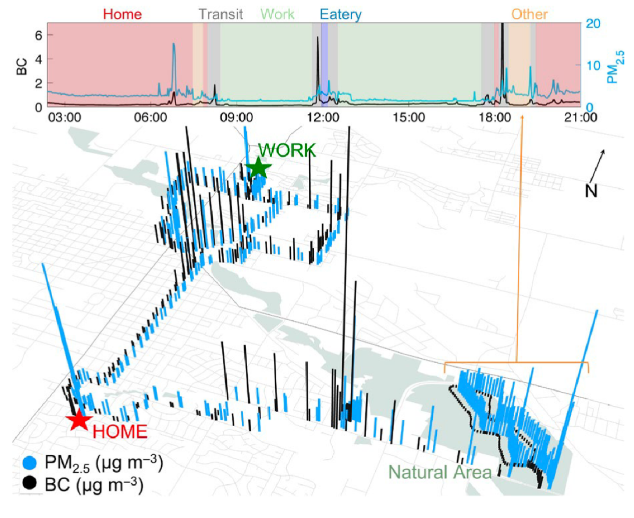
Fort Collins Commuter Study (FCCS)

- 45 individuals
- 1 to 13 non-consecutive days each
- Exposure measured for
  - black carbon (BC)
  - carbon monoxide (CO)
  - fine particulate matter ($PM_{2.5}$)
- Exposure at 10 second intervals



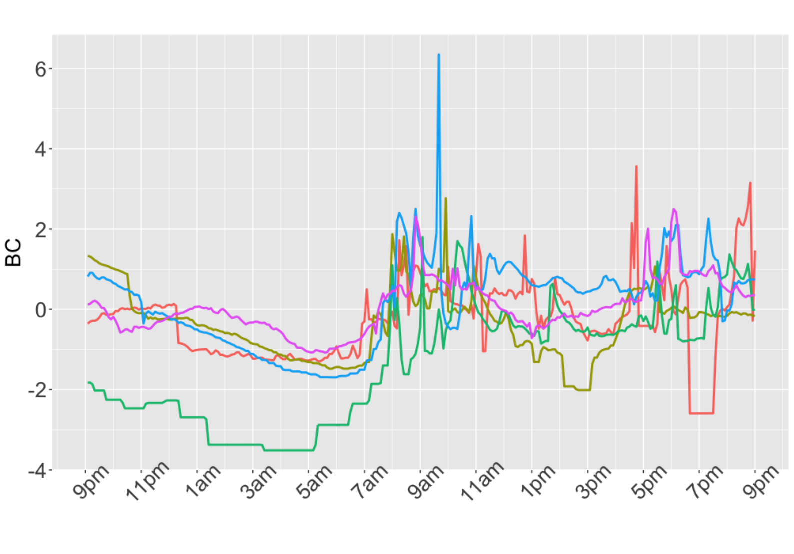
[1]Figure source: Good et al. (2016) *J. of Exposure Science & Environmental Epidemiology*.

Fort Collins Commuter Study (FCCS)



[1]Figure source: Koehler et al. (2019) *Indoor Air*.

Fort Collins Commuter Study (FCCS)



Statistical Challenges of Low-Cost, Real-Time Sensors

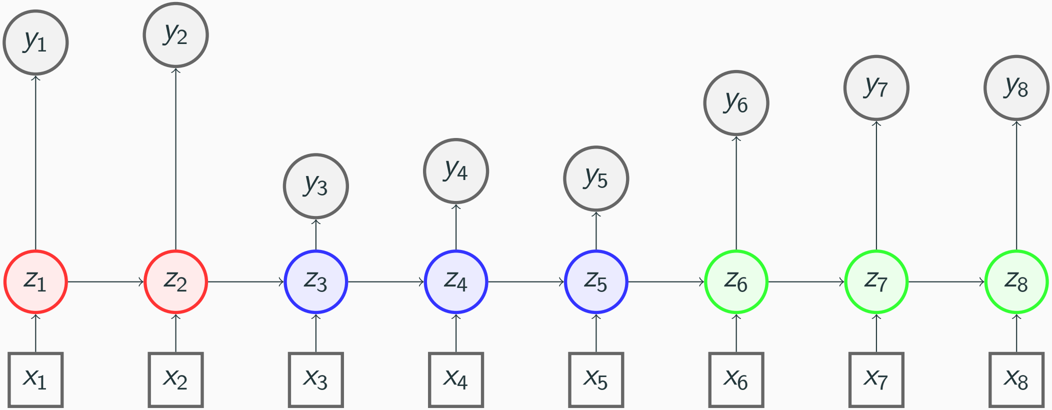
- Missing data due to
    - user non-compliance
    - device failure
    - levels below limit of detection
- How to classify exposure?
- How to relate to health outcomes?

Our Objectives

- Goals:
    - Find shared patterns in exposures
    - Impute missing data
- Want to do this in a way that allows for rapid changing of microenvironment and shared trends / common spaces
- Previous imputation methods either ignore temporal ordering of data or treat it as smoothly varying (Arroyo et al., 2018; Molitor et al., 2006; Krall et al., 2015; Houseman et al., 2017)

Hidden Markov Model

x = covariates, z = hidden states, y = exposure data



Infinite Hidden State Model: Notation

- Observed data
  - $\mathbf{y}_{ist}$ is the vector of exposures for individual i on day s at time t
  - $\mathbf{Y}_{is,1:T_{is}}$ is the full multivariate time series for individual i on sampling day s
  - $\mathbf{x}_{ist}$ is a set of covariates
    - time of day
    - individual characteristics
    - user reported activity or microenvironment (e.g. home, work, transit, etc.)
- Latent structure
  - z_{ist} is a categorical factor representing the latent state assignment
  - $z_{ist} = k$ if individual i is in latent state k at at time t on day s
  - iHMM allows for unknown number of hidden states

Infinite Hidden State Model: Key Assumptions

- Conditional independence of observed data conditional on the hidden states

$$f(\mathbf{y}_{it}|\mathbf{y}_{i,1:t-1}, \mathbf{z}_{i,1:t}) = f(\mathbf{y}_{it}|z_{it})$$

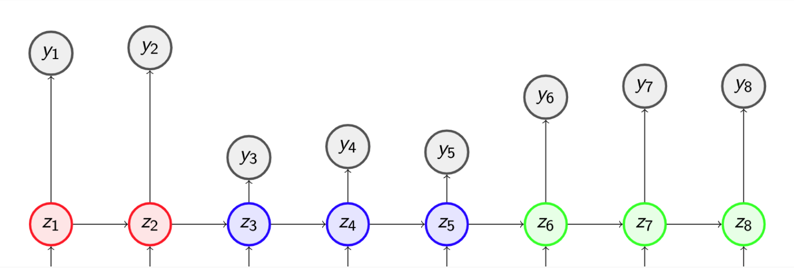
- Latent states follow the first-order Markov property

$$p(z_{it}|z_{i,1:t-1}) = p(z_{it}|z_{i,t-1})$$

Multivariate Normal Emission Distribution

Exposure data for individual i, sampling day s, and time point t is modeled

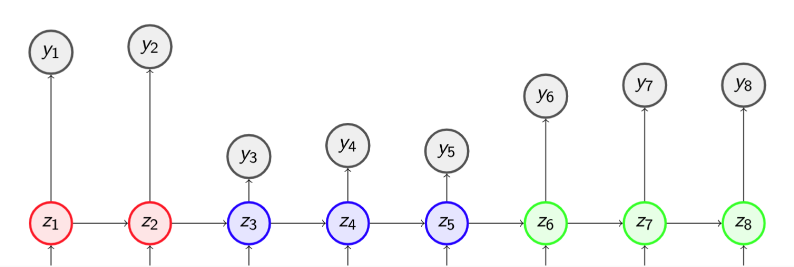
$$\begin{aligned} \mathbf{y}_{ist} | z_{ist} = k &\sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k &\sim \mathrm{N}\left(\mathbf{0}, \frac{1}{\lambda}\boldsymbol{\Sigma}_k\right) \\ \boldsymbol{\Sigma}_k &\sim \text{Inverse Wishart}\left(\nu, \mathbf{I}_p\right) \end{aligned}$$



Hidden State Model

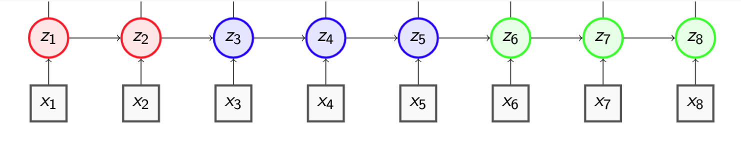
Model hidden states with hidden Markov model

$$z_{it} | z_{i,t-1} \sim \text{Categorical}\left(\pi_{z_{i,t-1}}\right)$$



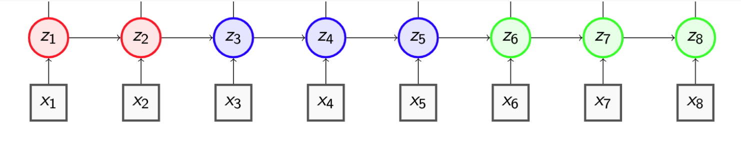
z = hidden states, y = exposure data

Probit Stick-Breaking Process on the Transition Distribution



- Want the state assignments and/or transitions to be covariate dependent
  - diurnal patterns
  - shared states among repeated sampling days for an individual
  - auxiliary information like time diary data
  - patterns in transitions

Probit Stick-Breaking Process on the Transition Distribution



- The probability of individual i on sampling day s transitioning from state j to state k at time t is

$$\begin{aligned}\pi_{jk}(\mathbf{x}_{ist}) &\equiv P(z_{ist} = k | z_{is,t-1} = j, \mathbf{x}_{ist}) \\ &= \Phi(\alpha_{jk} + \mathbf{x}'_{ist}\beta_k + \mathbf{x}'_{ist}\gamma_{ik}) \prod_{l<k} \{1 - \Phi(\alpha_{jl} + \mathbf{x}'_{ist}\beta_l + \mathbf{x}'_{ist}\gamma_{il})\}\end{aligned}$$

- α_{jk} controls state transitions at consecutive time points
- β_k and γ_{ik} control covariate-dependent and subject-specific trends

Prior Distributions on Transition Parameters

Transitions among states

$$\begin{aligned} \alpha_{jk} | \sigma_\alpha^2 &\sim \mathrm{N}(0, \sigma_\alpha^2) \text{ for } j \neq k \\ \sigma_\alpha^{-2} &\sim \mathrm{Gamma}(1, 1) \end{aligned}$$

Self-transitions

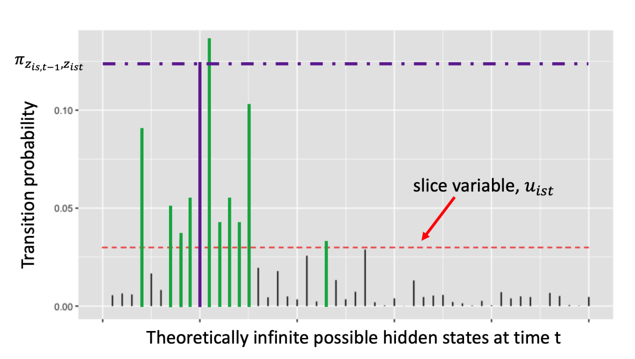
$$\begin{aligned} \alpha_{jj} | m_\alpha, v_\alpha &\sim \mathrm{N}(m_\alpha, v_\alpha) \\ m_\alpha &\sim \mathrm{N}(0, 1) \\ v_\alpha^{-1} &\sim \mathrm{Gamma}(1, 1) \end{aligned}$$

Covariate effects

$$\begin{aligned} \boldsymbol{\beta}_k &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\ \boldsymbol{\gamma}_{ik} | \kappa^2 &\sim \mathrm{N}(\mathbf{0}, \kappa^2 \mathbf{I}) \\ \kappa^{-2} &\sim \mathrm{Gamma}(1, 1) \end{aligned}$$

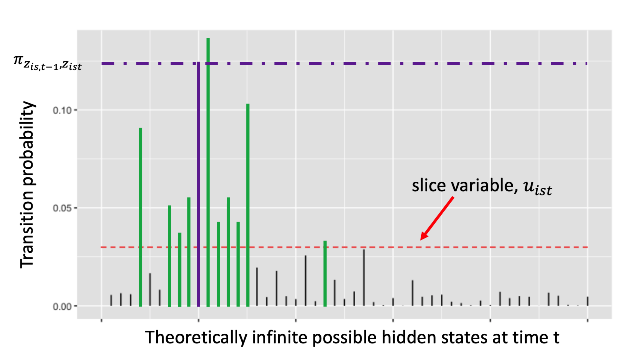
Posterior Sampling

- "Beam" sampling = slice sampling + dynamic programming (Van Gael et al., 2008)
- Sample entire hidden state trajectories at once
- Better mixing and faster convergence than Gibbs sampling



Posterior Sampling

- Slice sampling reduces problem to finite number of paths
- Forward pass calculates probabilities of each path
- Backwards step samples latent sequence



Imputation Model

- Impute missing data conditional on the state assignment using multivariate normal conditionals
- MAR for all components

$$\mathbf{y}_{it,\text{MAR}} | z_{it} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \text{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

- MAR for only some components use conditional distributions of

$$\begin{bmatrix} \mathbf{y}_{it,\text{obs}} \\ \mathbf{y}_{it,\text{MAR}} \end{bmatrix} \Bigg| z_{it} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \text{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{(k,\text{obs})} \\ \boldsymbol{\mu}_{(k,\text{MAR})} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{(k,\text{obs},\text{obs})} & \boldsymbol{\Sigma}_{(k,\text{obs},\text{MAR})} \\ \boldsymbol{\Sigma}_{(k,\text{MAR},\text{obs})} & \boldsymbol{\Sigma}_{(k,\text{MAR},\text{MAR})} \end{bmatrix} \right)$$

- For data missing below LOD similar but from a truncated multivariate distribution
  - We know the LOD and know if data is below LOD or MAR

Simulation Study

Purpose

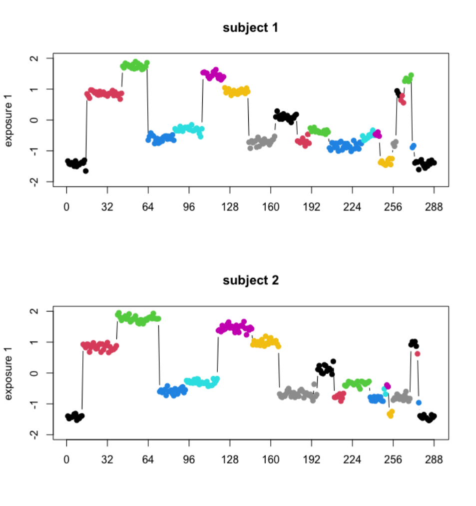
- Evaluate parameter estimation
- Evaluate imputations
- Compare with competing methods

Simulation Scenarios

- Shared cyclical trends
- Distinct cyclical trends

Missing Data Levels

- 0%, 5%, 10%, 20%
- MAR and below LOD

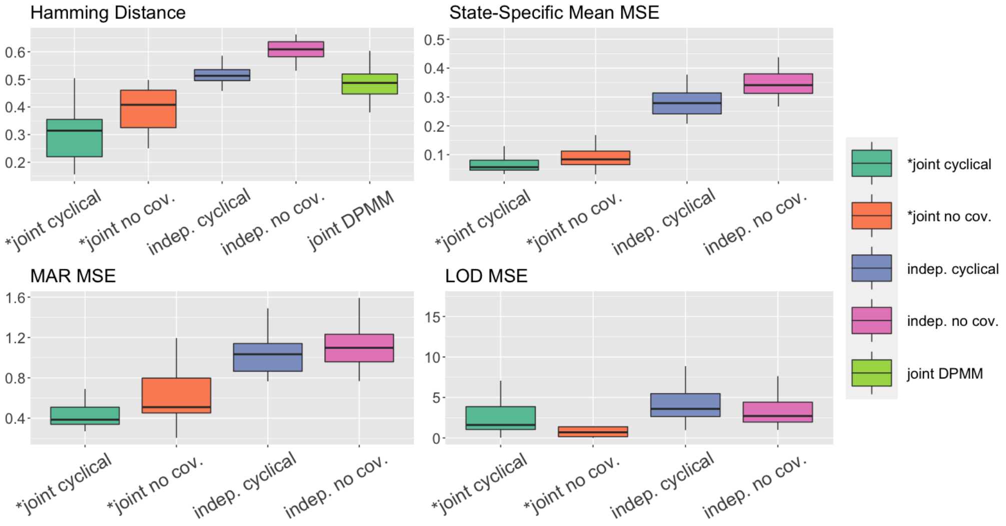


Models for Simulation Study

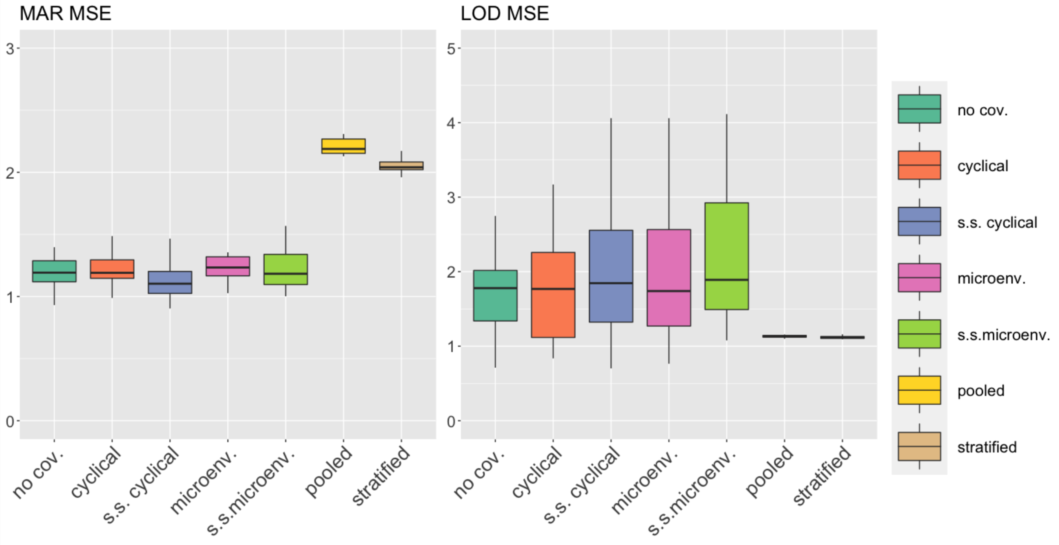
- ***Joint no covariates**: shared states, no covariates
- ***Joint cyclical**: shared states and shared cyclical trends
- **Independent no covariates**: individual states, no covariates
- **Independent cyclical**: individual states and individual cyclical trends
- **Joint DPMM**: Dirichlet process mixture model, shared states, no temporal dependency

*proposed methods

Shared Scenario Results 5% Missing Data



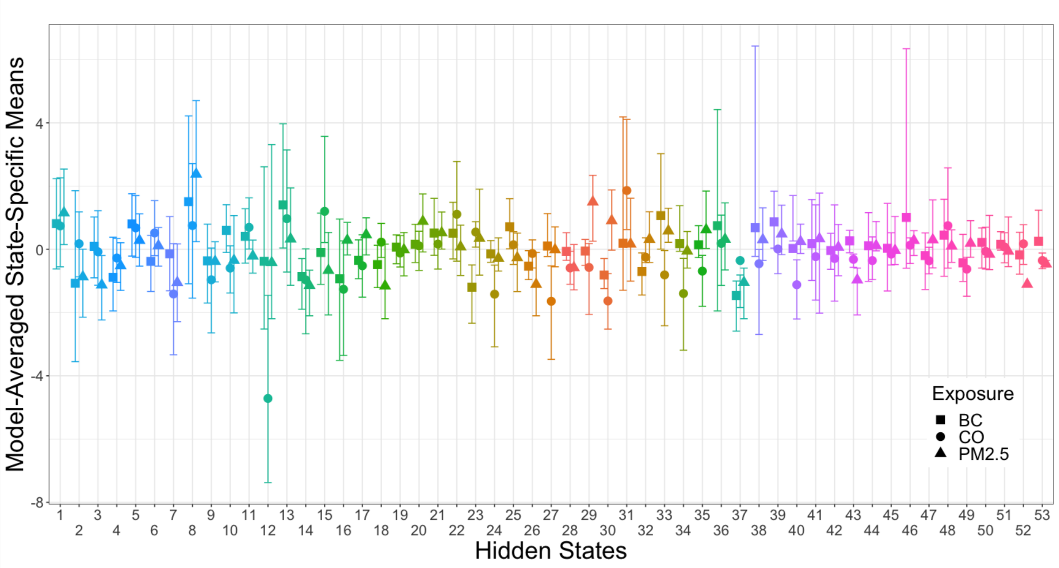
Validation Study Results



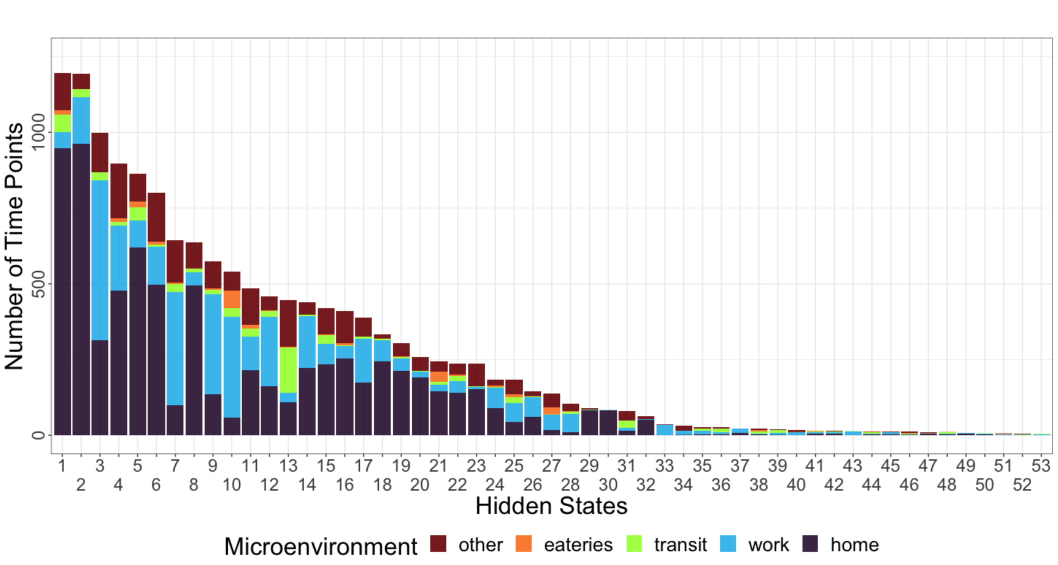
Fort Collins Commuter Study (FCCS)

- Fit joint model with cyclical trends to the data
- Considered average exposure over five minute intervals
- Trimmed data to be 24 hour segments

State-Specific Mean Estimation

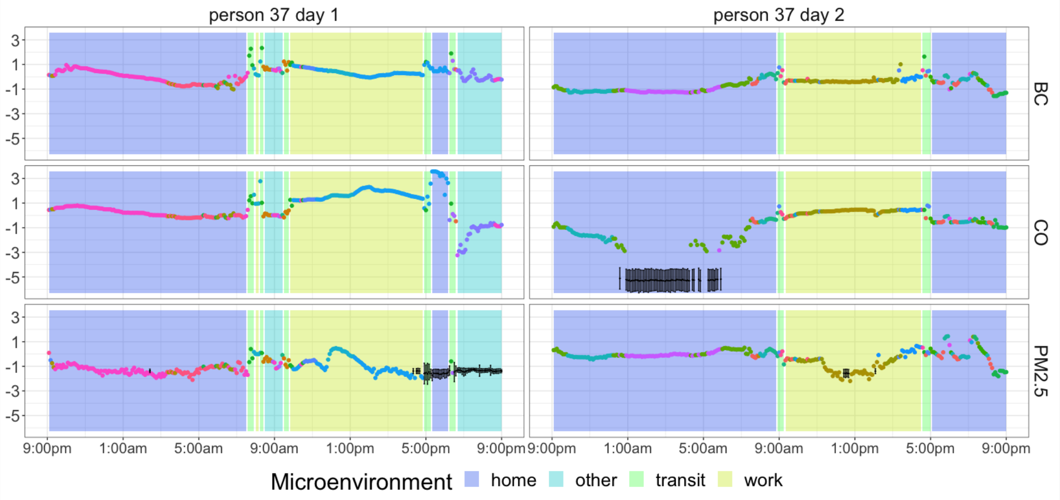


Hidden State Correspondence with Microenvironments



Fort Collins Commuter Study (FCCS)

Fort Collins Commuter Study (FCCS)



Summary

- Developed statistical method for analyzing multiple multivariate time series with missing data
- PSBP on transition distribution to estimate an unknown number of hidden states
- Multiple imputation for data that are MAR and below the LOD
- Demonstrated our method's estimation and imputation gains over competing approaches in simulation and validation studies
- Applied method to FCCS data to impute missing exposure data and identify time-activity patterns associated with exposures
- Many more challenges with low-cost, real-time sensor data

Thank You

anderwilson.github.io

ander.wilson@colostate.edu

@ander_wilson

Hoskovec, L., Koslovsky, M. D., Koehler, K., Good, N., Peel, J. L., Volckens, J., Wilson, A. (2022). Infinite hidden Markov models for multiple multivariate time series with missing data. *Biometrics*. arxiv.org/abs/2204.06610

NIEHS grant for FCCS: ES020017
NIEHS grant for statistical methods: ES028811

References

Arroyo, A, A Herrero, V Tricio, E Corchado, and M Wozniak (2018). "Neural models for imputation of missing ozone data in air-quality datasets". *Complexity* 2018.

Good, N, A Mölter, C Ackerson, A Bachand, T Carpenter, ML Clark, KM Fedak, A Kayne, K Koehler, B Moore, C L'Orange, C Quinn, V Ugave, AL Stuart, JL Peel, and J Volckens (2016). "The Fort Collins Commuter Study: Impact of route type and transport mode on personal exposure to multiple air pollutants". *u* 26.4.

Hoskovec, L, MD Koslovsky, K Koehler, N Good, JL Peel, J Volckens, and A Wilson (2022). "Infinite hidden Markov models for multiple multivariate time series with missing data". *Biometrics*.

Houseman, EA and MA Virji (2017). "A Bayesian approach for summarizing and modeling time-series exposure data with left censoring". *Annals of Work Exposures and Health* 61.7.

Koehler, K, N Good, A Wilson, A Mölter, BF Moore, T Carpenter, JL Peel, and J Volckens (2019). "The Fort Collins commuter study: Variability in personal exposure to air pollutants by microenvironment". *Indoor Air* 29.2.

Krall, JR, CH Simpson, and RD Peng (2015). "A model-based approach for imputing censored data in source apportionment studies". *Environmental and Ecological Statistics* 22.4.

Molitor, J, NT Molitor, M Jerrett, R McConnell, J Gauderman, K Berhane, and D Thomas (2006). "Bayesian modeling of air pollution health effects with missing exposure data". *American Journal of Epidemiology* 164.1.

Van Gael, J, Y Saatci, YW Teh, and Z Ghahramani (2008). "Beam sampling for the infinite hidden Markov model". *Proceedings of the 25th International Conference on Machine Learning*.