SISBayes - Padova

Multivariate Causal Effect: a Bayesian Regression Factor Model

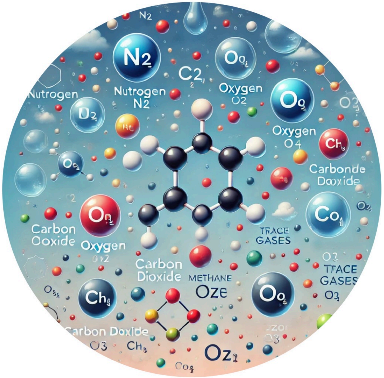
Dafne Zorzetto[1]

joint work with
Jenna Landy[2], Corwin Zigler[3], Giovanni Parmigiani[2,4], Roberta De Vito[1,3]

[1] Brown University - Data Science Institute,
[2] Harvard University - Department of Biostatistics,
[3] Brown University - Department of Biostatistics,
[4] Dana Farber Cancer Institute - Department of Data Science.

BROWN
Data Science Institute

- Epidemiological studies demonstrate that exposure to higher levels of $\mathrm{PM}_{2.5}$ corresponds to higher mortality/diseases risk.



- Epidemiological studies demonstrate that exposure to higher levels of $\mathrm{PM}_{2.5}$ corresponds to higher mortality/diseases risk.
- The fine particular matter $\mathrm{PM}_{2.5}$ is composed by different chemical components.

- Epidemiological studies demonstrate that exposure to higher levels of $\mathrm{PM}_{2.5}$ corresponds to higher mortality/diseases risk.
- The fine particular matter $\mathrm{PM}_{2.5}$ is composed by different chemical components.
- Some events, as wildfires, can change the chemical compositions.

Studying causal effects of wildfire smoke on $PM_{2.5}$ chemicals

Studying causal effects of wildfire smoke on $PM_{2.5}$ chemicals

Multivariate Outcome

Studying causal effects of wildfire smoke on $PM_{2.5}$ chemicals

Multivariate Outcome

Factor Models

Studying causal effects of wildfire smoke on $PM_{2.5}$ chemicals

Multivariate Outcome

Factor Models

Causality

Studying causal effects of wildfire smoke on $PM_{2.5}$ chemicals

Multivariate Outcome

Factor Models

Causality

Potential Outcome Framework

Outline

1. Potential Outcome Framework
2. Bayesian Causal Factor Model
3. Simulation Study
4. Environmental Application

Potential Outcome Framework

Rubin (1974)'s Potential Outcome Framework

$T_i \in \{0, 1\}$

T

Treatment

Rubin (1974)'s Potential Outcome Framework

$T_i \in \{0, 1\}$

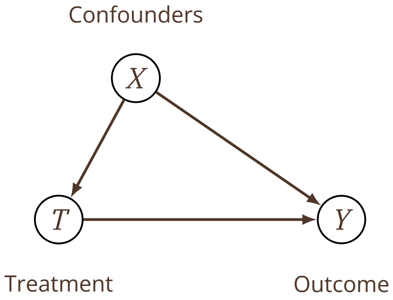
$$\begin{bmatrix} Y_{i1} \\ \dots \\ Y_{iq} \end{bmatrix} = \begin{cases} \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} & \textit{if } T_i = 0 \\ \begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} & \textit{if } T_i = 1 \end{cases}$$

$$T \longrightarrow Y$$

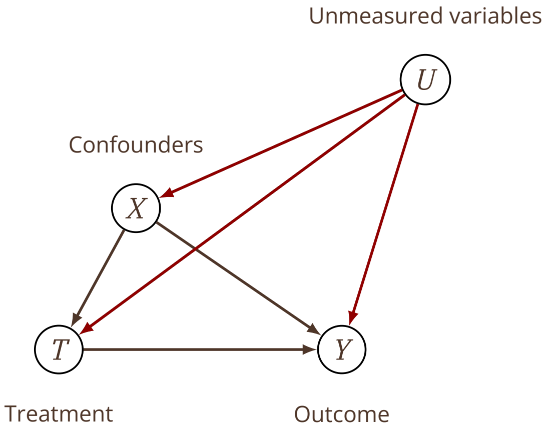
Treatment

Outcome

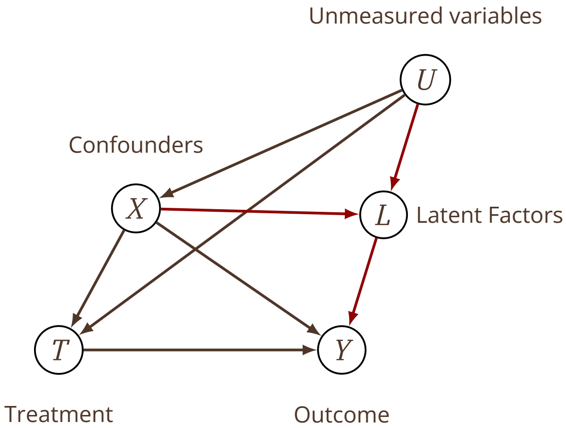
Rubin (1974)'s Potential Outcome Framework



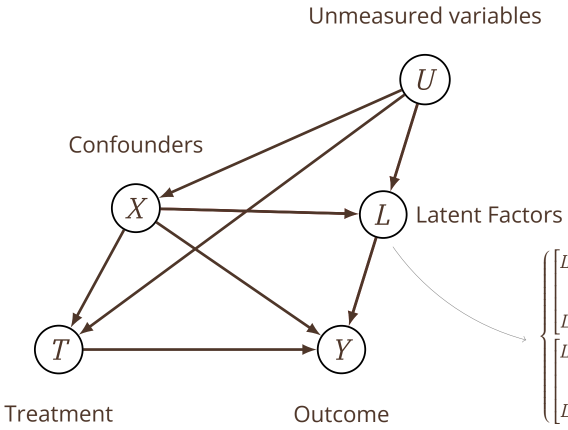
Rubin (1974)'s Potential Outcome Framework



Rubin (1974)'s Potential Outcome Framework



Rubin (1974)'s Potential Outcome Framework



$$\begin{cases} \begin{bmatrix} L_{0,i1} \\ \dots \\ L_{0,iq} \end{bmatrix} & \textit{if } T_i = 0 \\ \begin{bmatrix} L_{1,i1} \\ \dots \\ L_{1,iq} \end{bmatrix} & \textit{if } T_i = 1 \end{cases}$$

Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

Causal Assumptions

Causal Assumptions

Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

1. Stable Unit Treatment Value Assumption (SUTVA):

$$\mathbf{Y}_i = (1 - T_i)\,\mathbf{Y}_i(0) + T_i\,\mathbf{Y}_i(1)$$

Causal Assumptions

Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

1. Stable Unit Treatment Value Assumption (SUTVA):

$$\mathbf{Y}_i = (1 - T_i)\, \mathbf{Y}_i(0) + T_i\, \mathbf{Y}_i(1)$$

2. Positivity:

$$0 < \Pr(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) < 1$$

Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

Causal Assumptions

1. Stable Unit Treatment Value Assumption (SUTVA):

$$\mathbf{Y}_i = (1 - T_i)\, \mathbf{Y}_i(0) + T_i\, \mathbf{Y}_i(1)$$

2. Positivity:

$$0 < \Pr(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) < 1$$

3. Conditional ignorability:

$$\{\mathbf{Y}_i(1), \mathbf{Y}_i(0)\} \perp T_i \mid \{\mathbf{X}_i = \mathbf{x}_i, \mathbf{L}_{i0} = \mathbf{l}_{i0}, \mathbf{L}_{i1} = \mathbf{l}_{i1}\}$$

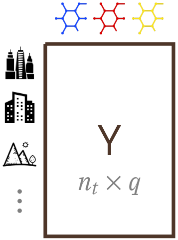
Average treatment effects: $\begin{bmatrix} \tau_1^P \\ \dots \\ \tau_q^P \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} Y_{i1}(1) \\ \dots \\ Y_{iq}(1) \end{bmatrix} - \begin{bmatrix} Y_{i1}(0) \\ \dots \\ Y_{iq}(0) \end{bmatrix} \right)$

Causal Assumptions

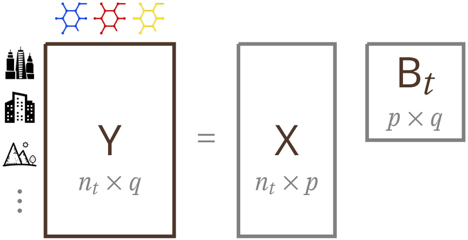
1. Stable Unit Treatment Value Assumption (SUTVA):
$$\mathbf{Y}_i = (1 - T_i)\,\mathbf{Y}_i(0) + T_i\,\mathbf{Y}_i(1)$$
2. Positivity:
$$0 < \Pr(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) < 1$$
3. Conditional ignorability:
$$\{\mathbf{Y}_i(1), \mathbf{Y}_i(0)\} \perp T_i \mid \{\mathbf{X}_i = \mathbf{x}_i, \mathbf{L}_{i0} = \mathbf{l}_{i0}, \mathbf{L}_{i1} = \mathbf{l}_{i1}\}$$
4. Indirectness of unmeasured confounding effects:
$$\mathbb{P}(\mathbf{Y}(1), \mathbf{Y}(0) \mid \mathbf{X}, \mathbf{L}_0, \mathbf{L}_1, U) = \mathbb{P}(\mathbf{Y}(1), \mathbf{Y}(0) \mid \mathbf{X}, \mathbf{L}_0, \mathbf{L}_1)$$

Bayesian Causal Factor Model

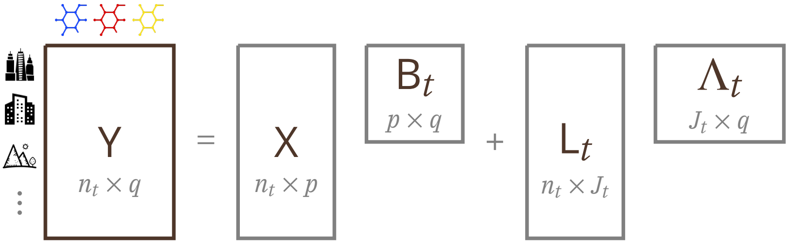
Bayesian Causal Regression Factor Model



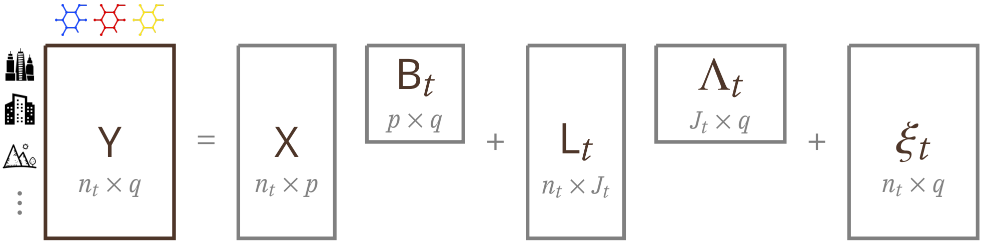
Bayesian Causal Regression Factor Model



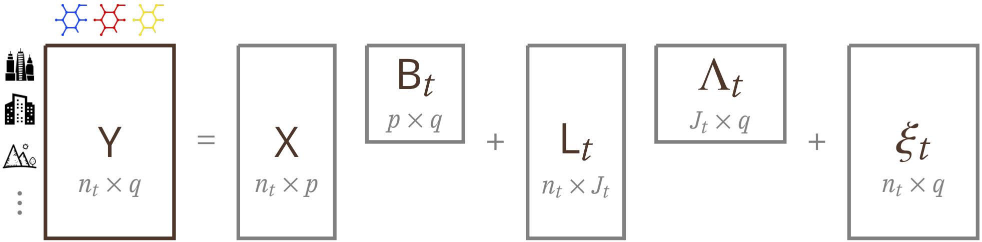
Bayesian Causal Regression Factor Model



Bayesian Causal Regression Factor Model



Bayesian Causal Regression Factor Model



$$\{\mathbf{Y}_i \mid \mathbf{x}_i, t, \mathbf{l}_{it}\} = \boldsymbol{\mu}_t + \mathbf{B}_t\mathbf{x}_i + \Lambda_t\mathbf{l}_{it} + \boldsymbol{\xi}_{it}$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}}\,\underset{p \times q}{\mathrm{B}_t} + \underset{n_t \times J_t}{\mathrm{L}_t}\,\underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

$$\underset{n_t \times q}{Y} = \underset{n_t \times p}{X}\,\underset{p \times q}{B_t} + \underset{n_t \times J_t}{L_t}\,\underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Observed variables

- multivariate outcomes: Y
- covariates: X

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}} \; \underset{p \times q}{\mathrm{B}_t} + \underset{n_t \times J_t}{\mathrm{L}_t} \; \underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Observed variables

- multivariate outcomes: Y
- covariates: X

Regression coefficients

$$\beta_{jt} \sim \mathcal{N}_q(\mu_\beta, \Sigma_\beta)$$
$$\text{for } i \in \{1, \ldots, n\} \; t \in \{0, 1\}$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}}\ \underset{p \times q}{\mathrm{B}_t} + \underset{n \times J_t}{\mathrm{L}_t}\ \underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Treatment-specific factor scores

Dependent Dirichlet Process:

$$l_{itj} \mid \mathbf{x}_i, \mu, \sigma^2 \sim \sum_{r \geq 1} \pi_{tjr}(\mathbf{x}_i) \mathcal{N}(\mu_r, \sigma_r^2)$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}} \; \underset{p \times q}{\mathrm{B}_t} + \underset{n \times J_t}{\mathrm{L}_t} \; \underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Treatment-specific factor scores

Dependent Dirichlet Process:

$$l_{itj} \mid \mathbf{x}_i, \mu, \sigma^2 \sim \sum_{r \geq 1} \pi_{tjr}(\mathbf{x}_i) \mathcal{N}(\mu_r, \sigma_r^2)$$

$$\pi_{tjr}(\mathbf{x}_i) = V_{tjr}(\mathbf{x}_i) \prod_{g<r} \{1 - V_{tjg}(\mathbf{x}_i)\}$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}} \; \underset{p \times q}{\mathrm{B}_t} + \underset{n \times J_t}{\mathrm{L}_t} \; \underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Treatment-specific factor scores

Dependent Dirichlet Process: probit stick-breaking process (Rodriguez and Dunson, 2011)

$$l_{itj} \mid \mathbf{x}_i, \mu, \sigma^2 \sim \sum_{r \geq 1} \pi_{tjr}(\mathbf{x}_i) \mathcal{N}(\mu_r, \sigma_r^2)$$

$$\pi_{tjr}(\mathbf{x}_i) = V_{tjr}(\mathbf{x}_i) \prod_{g<r} \{1 - V_{tjg}(\mathbf{x}_i)\}$$

$$V_{tjr}(x_i) = \Phi(a_{tjr}(x_i)) \quad a_{tjr}(x_i) \sim \mathcal{N}(x_i^T \alpha_{tjr}, 1)$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}}\;\underset{p \times q}{\mathrm{B}_t} + \underset{n_t \times J_t}{\mathrm{L}_t}\;\underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Treatment-specific factor loadings

Multiplicative gamma shrinkage prior (Bhattacharya and Dunson, 2011)

$$\lambda_{tjh} \mid \kappa_{tjh}, \iota_{th} \sim N\left(0, \kappa_{tjh}^{-1} \iota_{th}^{-1}\right)$$

$$\forall j = \{1 \ldots J_t\} \quad t = \{0, 1\} \quad k = 1 \ldots \infty$$

$$\kappa_{tjh} \sim Ga(\nu_t/2, \nu_t/2)$$

$$\iota_h = \prod_{l=1}^{h} \delta_{tl} \quad \delta_{t1} \sim Ga\left(a_{t1}, 1\right) \quad \delta_{tl} \sim Ga\left(a_{t2}, 1\right) \quad l \geq 2$$

$$\underset{n_t \times q}{\mathrm{Y}} = \underset{n_t \times p}{\mathrm{X}}\,\underset{p \times q}{\mathrm{B}_t} + \underset{n_t \times J_t}{\mathrm{L}_t}\,\underset{J_t \times q}{\Lambda_t} + \underset{n_t \times q}{\xi_t}$$

Error matrix

$$\xi_{it} \sim \mathcal{N}_p(0, \mathrm{diag}(\psi_{i1t}, \ldots, \psi_{ipt}))$$
$$\text{for } i \in \{1, \ldots, n\} \; t \in \{0, 1\}$$

Simulation Study

Table 1

Parameters and distributions for the simulation scenarios.

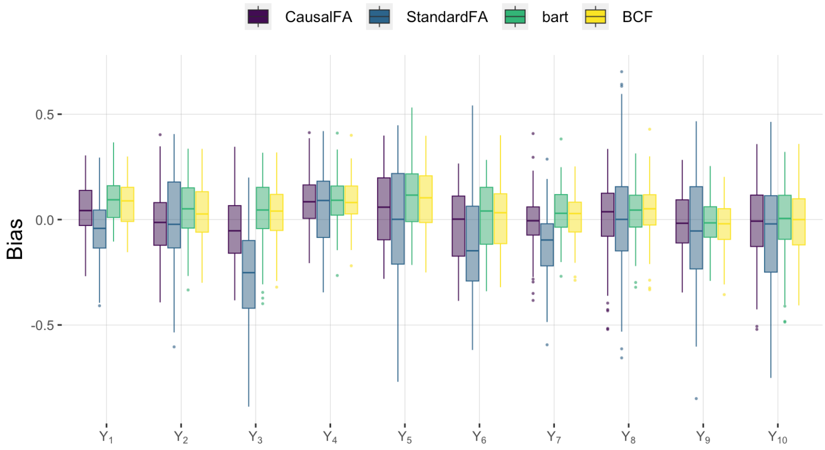
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Dimensions:				
n	500	500	500	3426
J_t	$(3,3)$	$(3,3)$	$(3,3)$	$(3,3)$
q	10	10	10	27
p	4	4	4	28
Variables:				
U	$U \sim N(0,2)$	$U \sim N(f_U(X_{1:4}), 0.5)$	$U \sim N(0,2)$	-
$\mathbf{X}$	$X_{1:2} \sim N(0,1)$	$X_{1:2} \sim N(0,1)$	$X_k \sim N(f_k(U), 1)$, $k=1,2$	observed $\mathbf{X}$
	$X_k \sim Be(\pi_k)$, $k=3,4$	$X_k \sim Be(\pi_k)$, $k=3,4$	$X_k \sim Be(\pi_k)$, $k=3,4$	
T	$T \sim Be(f_T(X_{1:4}, U))$	$T \sim Be(f_T(X_{1:4}, U))$	$T \sim Be(f_T(X_{1:4}, U))$	observed T
Factors:				
Clusters	3 or 2 cluster for each treatment-specific factors: $C_{ith} := f_C(\mathbf{X}_{1:4})$			$C_{ith} := f_C(\mathbf{X}_{1:27})$
l_{it}	$\{l_{ith} \mid C_{ith} = c\} \sim N(\mu_{tc} + \gamma_{tc}U, 1),\ h = \{1, \dots, p\}$			$\{l_{itj} \mid C_{ith} = c\} \sim N(\mu_{tc}, 1)$
Λ_t	matrices with 25% of zeros and nonzeros elements generated			estimated Λ_t
	by $\lambda_{th} \sim \pi Unif[-1, -0.8] + (1-\pi) Unif[0.8, 1]$, with $\pi \sim Be(0.5)$			in observed data
Outcome mode:				
B_t	$\beta_{ht} \sim Unif[-3+t, 2+t],\ t = 0,1,\ h = 1, \dots, p$			
ϵ_t	$\xi_{it} \sim Unif_p[0,1],\ t = 0,1$			
$\mathbf{Y}$	$Y_i(t) = B_t X + \Lambda_t l_{it} + \xi_{it},\ t = 0,1\ i = 1, \dots, n$			

Simulation study

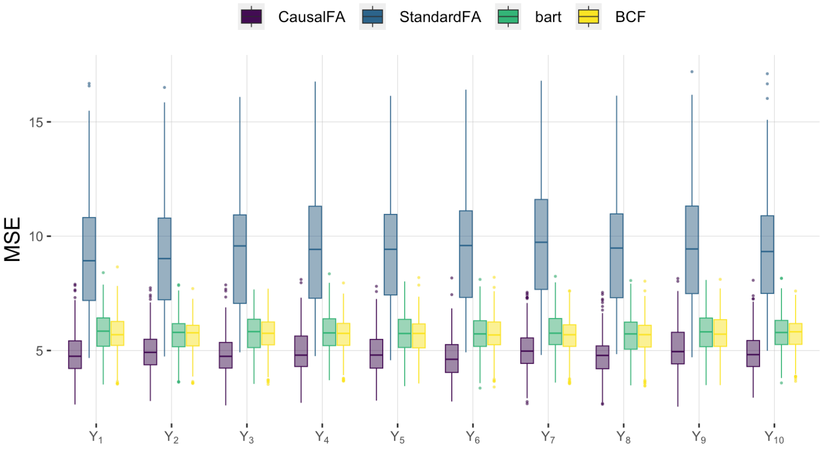
Models comparison:

- **CausalFA**: our regression factor model with mixture distribution for the factor score;
- **StandardFA**: regression factor model with standard prior: $l_t \sim \mathcal{N}(0, \mathbb{I}_{J_t})$;
- **BART** (Hill, 2011): estimated for each component of the multivariate outcome independent;
- **BCF** (Hahn, et al., 2020): estimated for each component of the multivariate outcome independent.

Comparison results: bias

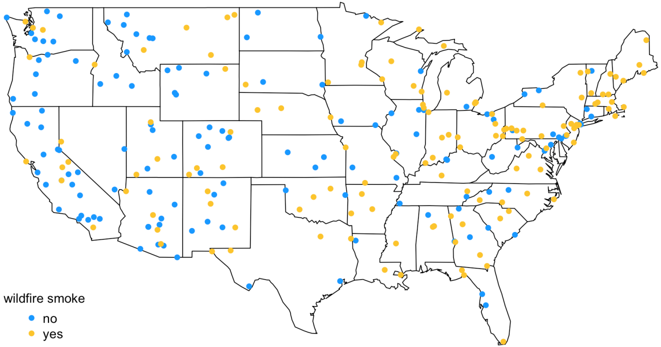


Comparison results: MSE



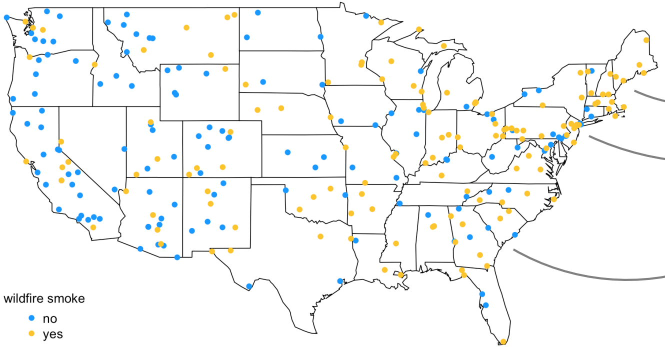
Environmental Application

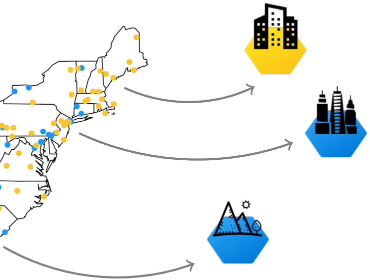
Causal effect of wildfire smoke on ambient chemical



Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository echolab-stanford/daily-10km-smokePM (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidycensus R package (Walker et al., 2021).

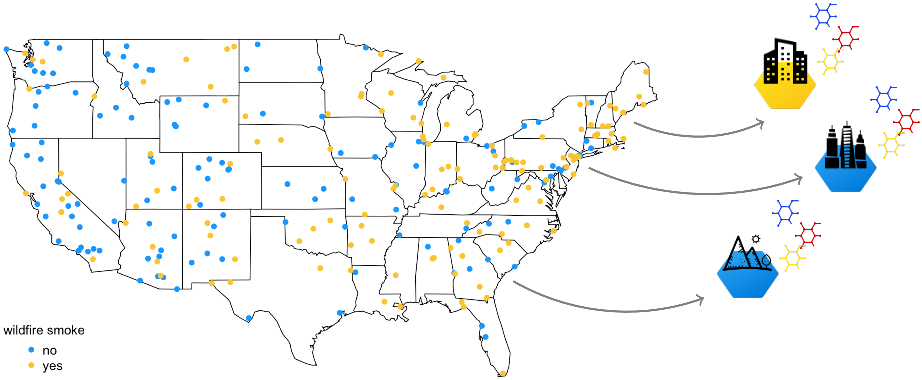
Causal effect of wildfire smoke on ambient chemical





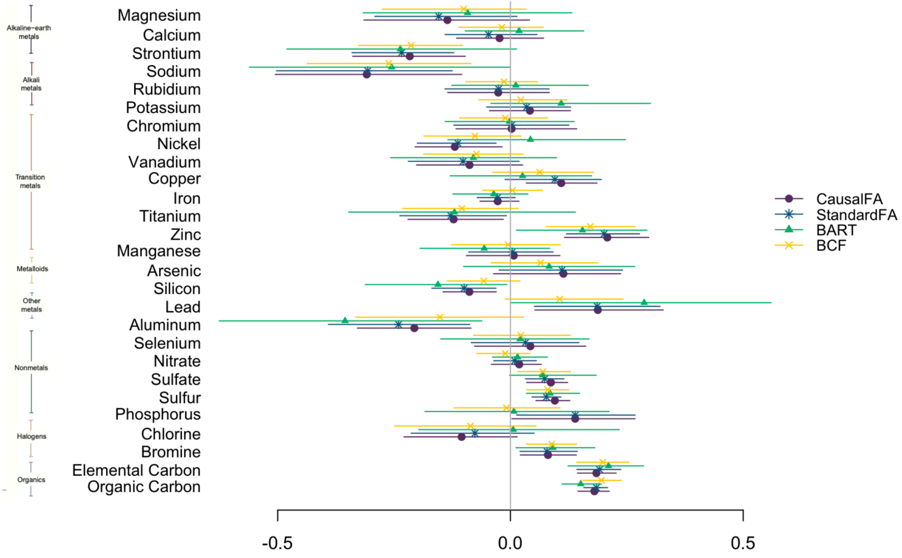
Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository echolab-stanford/daily-10km-smokePM (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidycensus R package (Walker et al., 2021).

Causal effect of wildfire smoke on ambient chemical

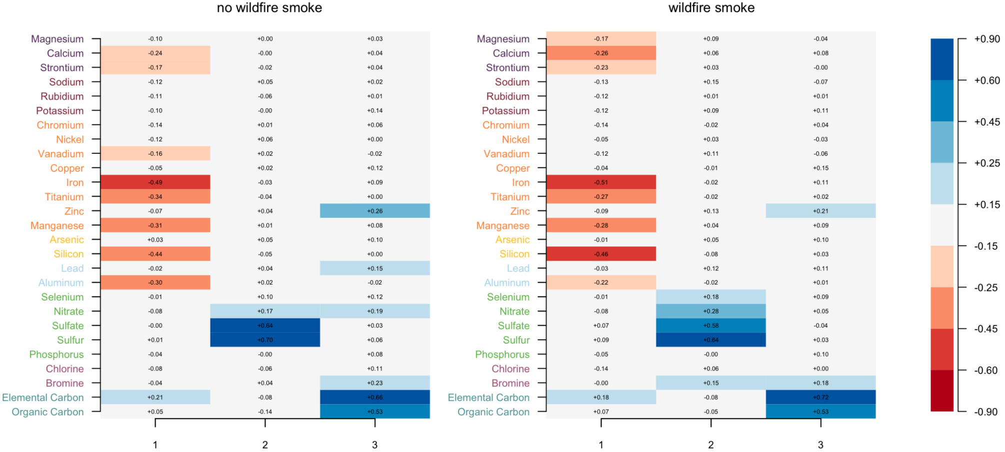


Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository echolab-stanford/daily-10km-smokePM (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidycensus R package (Walker et al., 2021).

Causal effects of wildfire smoke



Treatment-specific factor loadings



Wrapping up

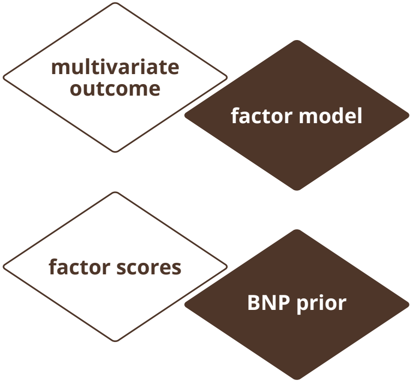
Wrapping up

- We define a causal inference framework for multivariate potential outcomes.

Wrapping up

- We define a causal inference framework for multivariate potential outcomes.
- We introduce a Bayesian regression factor model for causal inference framework.

Wrapping up



- We define a causal inference framework for multivariate potential outcomes.
- We introduce a Bayesian regression factor model for causal inference framework.
- We employs a dependent Dirichlet process as distribution for treatment-specific factor scores for overcome missing data problem.

Thanks for your attention

dafne_zorzetto@brown.edu