



SISBayes - Padova

Multivariate Causal Effect: a Bayesian Regression Factor Model

Dafne Zorzetto¹

joint work with

Jenna Landy², Corwin Zigler³, 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,

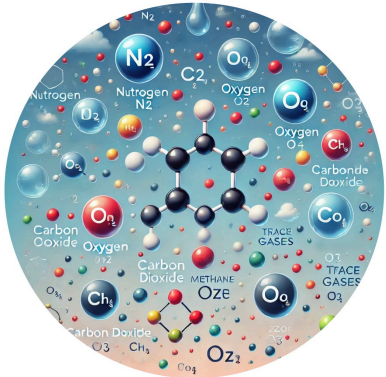
⁴ Dana Farber Cancer Institute - Department of Data Science.



BROWN
Data Science Institute



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





- Epidemiological studies demonstrate that exposure to higher levels of $PM_{2.5}$ corresponds to higher mortality/diseases risk.
- The fine particular matter $PM_{2.5}$ 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

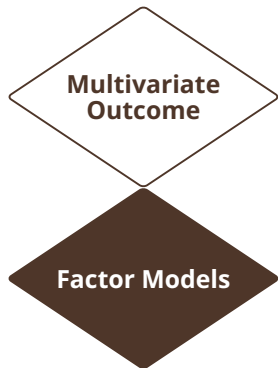


Studying causal effects of wildfire smoke on PM_{2.5} chemicals

**Multivariate
Outcome**



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



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

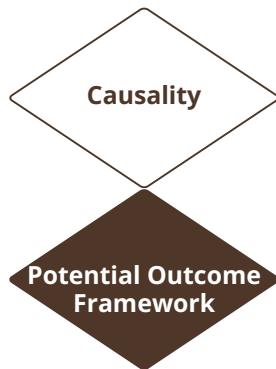
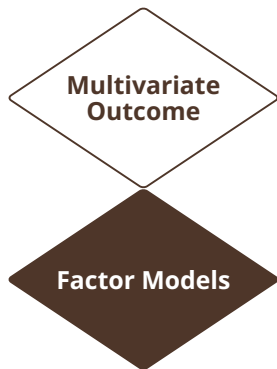
**Multivariate
Outcome**

Factor Models



Causality

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



Outline

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

Potential Outcome Framework

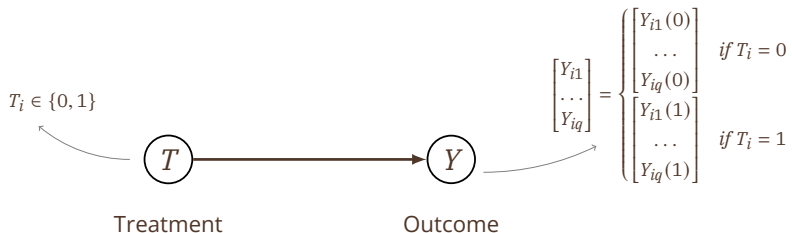
Rubin (1974)'s Potential Outcome Framework

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

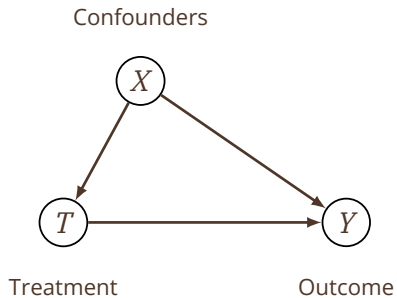


Treatment

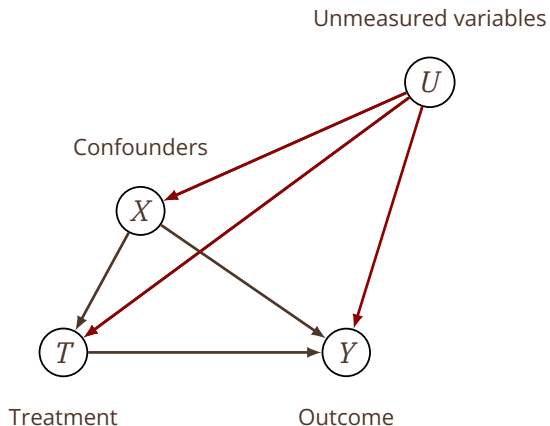
Rubin (1974)'s Potential Outcome Framework



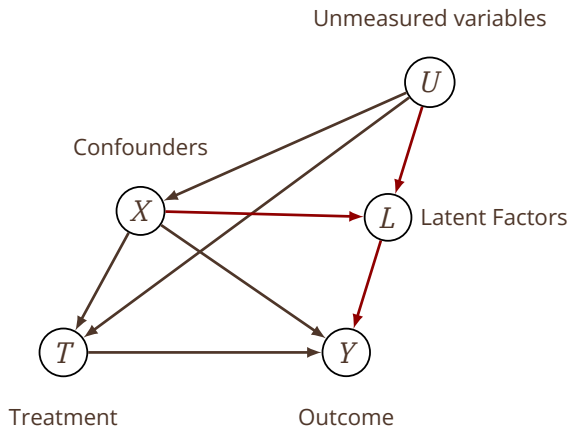
Rubin (1974)'s Potential Outcome Framework



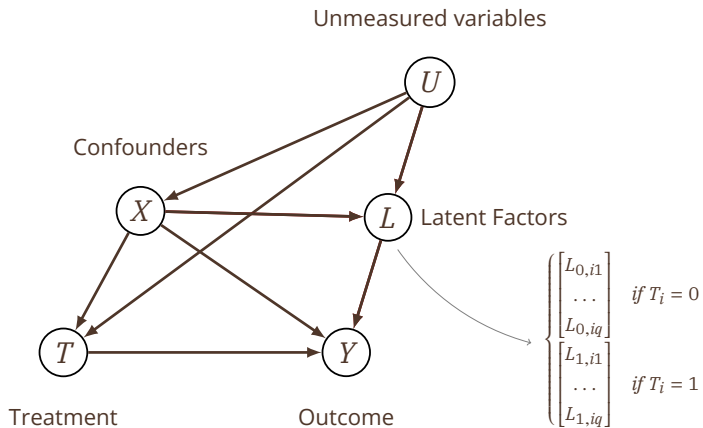
Rubin (1974)'s Potential Outcome Framework



Rubin (1974)'s Potential Outcome Framework



Rubin (1974)'s Potential Outcome Framework



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

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

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

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

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:

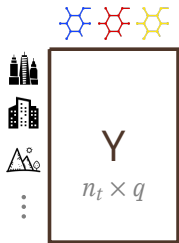
$$\{\mathbf{Y}_i(1), \mathbf{Y}_i(0)\} \perp T_i \mid \{\mathbf{X}_i = \mathbf{x}_i, \mathbf{L}_{i0} = \mathbf{l}_{i0}, \mathbf{L}_{i1} = \mathbf{l}_{i1}\}$$

4. Indirectness of unmeasured confounding effects:

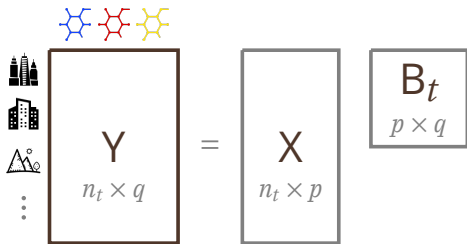
$$\mathbb{P}(\mathbf{Y}(1), \mathbf{Y}(0) \mid \mathbf{X}, \mathbf{L}_0, \mathbf{L}_1, U) = \mathbb{P}(\mathbf{Y}(1), \mathbf{Y}(0) \mid \mathbf{X}, \mathbf{L}_0, \mathbf{L}_1)$$

Bayesian Causal Factor Model

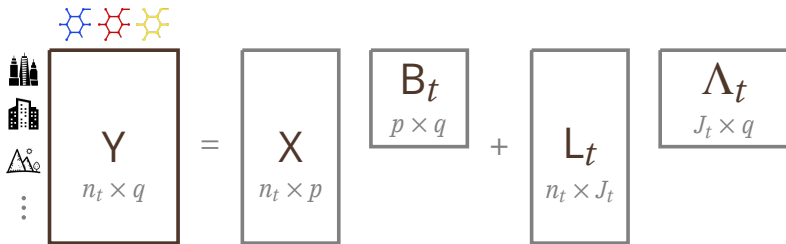
Bayesian Causal Regression Factor Model



Bayesian Causal Regression Factor Model



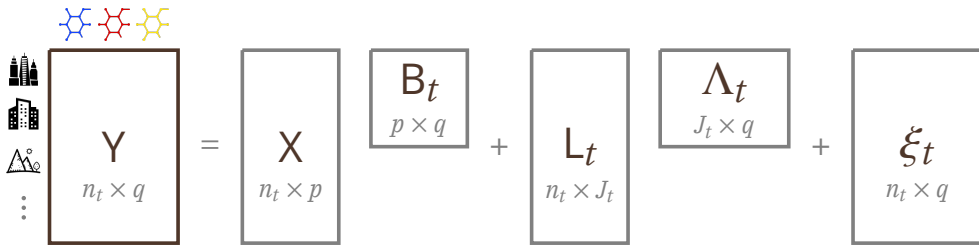
Bayesian Causal Regression Factor Model



The diagram illustrates the Bayesian Causal Regression Factor Model equation. On the left, a vertical stack of icons (city skyline, buildings, a graph, and an ellipsis) is positioned to the left of a large box labeled Y with dimensions $n_t \times q$. Above this box are three molecular structures colored blue, red, and yellow. An equals sign follows, then a box labeled X with dimensions $n_t \times p$. This is followed by a box labeled B_t with dimensions $p \times q$, a plus sign, a box labeled L_t with dimensions $n_t \times J_t$, and finally a box labeled Λ_t with dimensions $J_t \times q$.

$$\begin{matrix} \text{City skyline} \\ \text{Buildings} \\ \text{Graph} \\ \vdots \end{matrix} \begin{matrix} \text{Blue molecule} \\ \text{Red molecule} \\ \text{Yellow molecule} \end{matrix} \boxed{Y}_{n_t \times q} = \boxed{X}_{n_t \times p} \boxed{B_t}_{p \times q} + \boxed{L_t}_{n_t \times J_t} \boxed{\Lambda_t}_{J_t \times q}$$

Bayesian Causal Regression Factor Model



The diagram illustrates the Bayesian Causal Regression Factor Model equation. On the left, a vertical stack of icons (city skyline, building, mountain, and vertical ellipsis) is positioned to the left of a box labeled Y with dimensions $n_t \times q$. Above this box are three molecular structures colored blue, red, and yellow. The equation is represented as $Y = X B_t + L_t \Lambda_t + \xi_t$. The matrix X has dimensions $n_t \times p$. The matrix B_t has dimensions $p \times q$. The matrix L_t has dimensions $n_t \times J_t$. The matrix Λ_t has dimensions $J_t \times q$. The matrix ξ_t has dimensions $n_t \times q$.

$$\begin{matrix} \text{City skyline} \\ \text{Building} \\ \text{Mountain} \\ \vdots \end{matrix} \begin{matrix} \text{Blue molecule} & \text{Red molecule} & \text{Yellow molecule} \\ \boxed{Y} \\ n_t \times q \end{matrix} = \boxed{X} \begin{matrix} \boxed{B_t} \\ p \times q \end{matrix} + \boxed{L_t} \begin{matrix} \boxed{\Lambda_t} \\ J_t \times q \end{matrix} + \boxed{\xi_t} \begin{matrix} n_t \times q \end{matrix}$$

$n_t \times p$

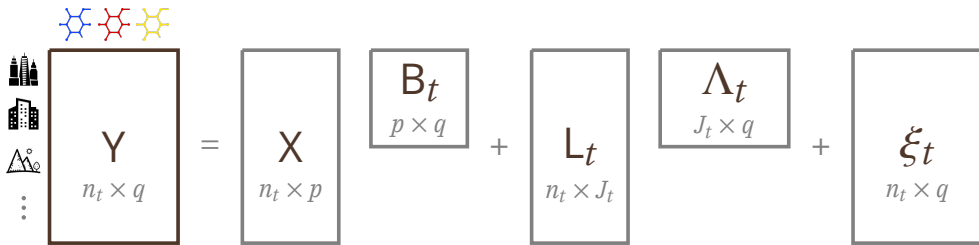
$p \times q$

$n_t \times J_t$

$J_t \times q$

$n_t \times q$

Bayesian Causal Regression Factor Model



The diagram illustrates the Bayesian Causal Regression Factor Model equation. On the left, a vertical rectangle labeled Y with dimensions $n_t \times q$ is preceded by three icons: a city skyline, a building, and a mountain with a peak, followed by a vertical ellipsis. Above the Y rectangle are three molecular structures in blue, red, and yellow. This is followed by an equals sign. To the right of the equals sign is a vertical rectangle labeled X with dimensions $n_t \times p$. This is followed by a plus sign and a vertical rectangle labeled B_t with dimensions $p \times q$. This is followed by another plus sign and a vertical rectangle labeled L_t with dimensions $n_t \times J_t$. This is followed by a third plus sign and a vertical rectangle labeled Λ_t with dimensions $J_t \times q$. Finally, there is a plus sign and a vertical rectangle labeled ξ_t with dimensions $n_t \times q$.

$$Y_{n_t \times q} = X_{n_t \times p} B_t_{p \times q} + L_t_{n_t \times J_t} \Lambda_t_{J_t \times q} + \xi_t_{n_t \times q}$$

$$\{Y_i \mid \mathbf{x}_i, t, \mathbf{l}_{it}\} = \mu_t + \mathbf{B}_t \mathbf{x}_i + \Lambda_t \mathbf{l}_{it} + \xi_{it}$$

$$\begin{array}{c} \mathbf{Y} \\ n_t \times q \end{array} = \begin{array}{c} \mathbf{X} \\ n_t \times p \end{array} \begin{array}{c} \mathbf{B}_t \\ p \times q \end{array} + \begin{array}{c} \mathbf{L}_t \\ n_t \times J_t \end{array} \begin{array}{c} \mathbf{\Lambda}_t \\ J_t \times q \end{array} + \begin{array}{c} \boldsymbol{\xi}_t \\ n_t \times q \end{array}$$

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n_t \times J_t \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \boldsymbol{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Observed variables

- multivariate outcomes: Y
- covariates: X

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n_t \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \boldsymbol{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Observed variables

- multivariate outcomes: Y
- covariates: X

Regression coefficients

$$\beta_{jt} \sim \mathcal{N}_q(\mu_\beta, \Sigma_\beta)$$

for $i \in \{1, \dots, n\}$ $t \in \{0, 1\}$

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \boldsymbol{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Treatment-specific factor scores

Dependent Dirichlet Process:

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

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Treatment-specific factor scores

Dependent Dirichlet Process:

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

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

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \boldsymbol{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Treatment-specific factor scores

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

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

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

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

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n_t \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \boldsymbol{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Treatment-specific factor loadings

Multiplicative gamma shrinkage prior (Bhattacharya and Dunson, 2011)

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

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

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

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

$$\begin{array}{|c|} \hline \mathbf{Y} \\ \hline n_t \times q \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{X} \\ \hline n_t \times p \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{B}_t \\ \hline p \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{L}_t \\ \hline n_t \times J_t \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{\Lambda}_t \\ \hline J_t \times q \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{\xi}_t \\ \hline n_t \times q \\ \hline \end{array}$$

Error matrix

$$\begin{aligned}
 \xi_{it} &\sim \mathcal{N}_p(0, \text{diag}(\psi_{i1t}, \dots, \psi_{ipt})) \\
 &\text{for } i \in \{1, \dots, n\} \ t \in \{0, 1\}
 \end{aligned}$$

Simulation Study

Table 1

Parameters and distributions for the simulation scenarios.

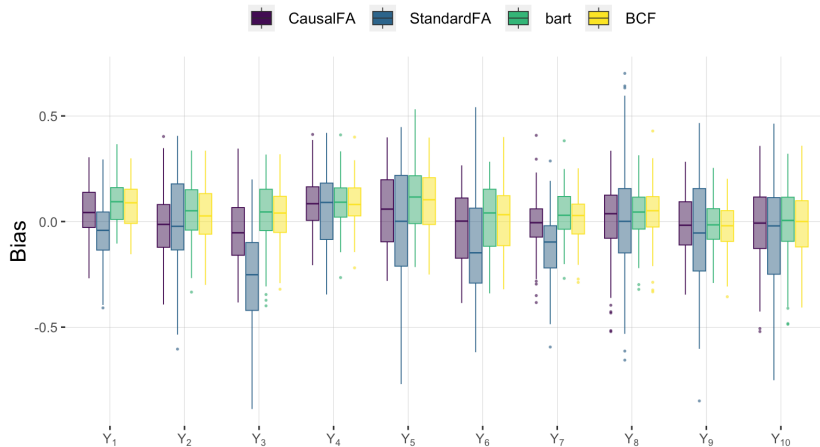
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Dimensions:				
n	500	500	500	3426
J_t	(3, 3)	(3, 3)	(3, 3)	(3, 3)
q	10	10	10	27
p	4	4	4	28
Variables:				
U	$U \sim N(0, 2)$	$U \sim N(f_U(X_{1:4}), 0.5)$	$U \sim N(0, 2)$	-
$\mathbf{X}$	$X_{1:2} \sim N(0, 1)$	$X_{1:2} \sim N(0, 1)$	$X_k \sim N(f_k(U), 1), k = 1, 2$	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 by $\lambda_{th} \sim \pi Uni f[-1, -0.8] + (1 - \pi)Uni f[0.8, 1]$, with $\pi \sim Be(0.5)$			estimated Λ_t in observed data
Outcome mode:				
B_t	$\beta_{ht} \sim Uni f[-3 + t, 2 + t], t = 0, 1, h = 1, \dots, p$			
ϵ_t	$\xi_{it} \sim Uni f_p[0, 1], t = 0, 1$			
$\mathbf{Y}$	$Y_i(t) = B_t X + \Lambda_t l_{it} + \xi_{it}, t = 0, 1 \ i = 1, \dots, n$			

Simulation study

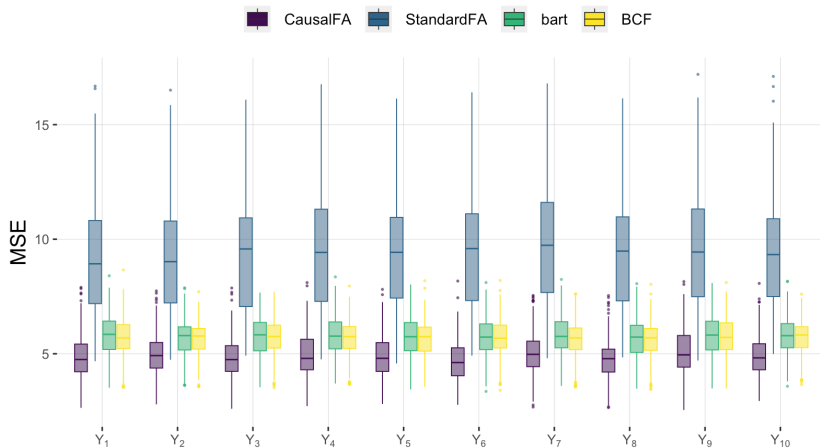
Models comparison:

- **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

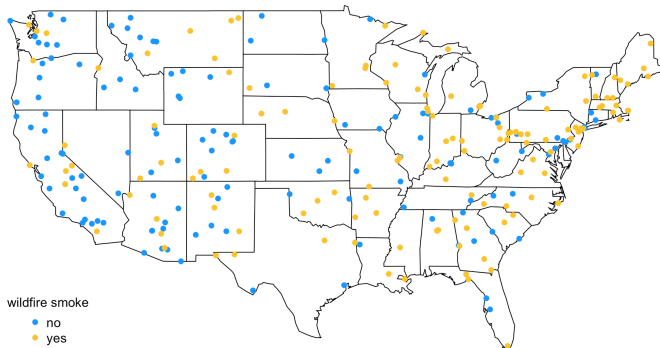


Comparison results: MSE



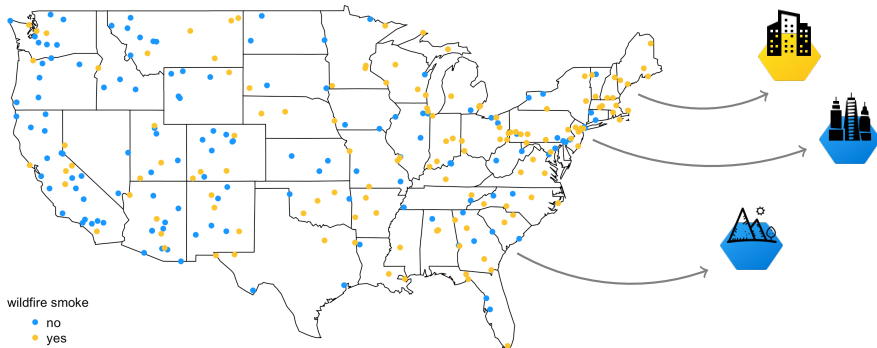
Environmental Application

Causal effect of wildfire smoke on ambient chemical



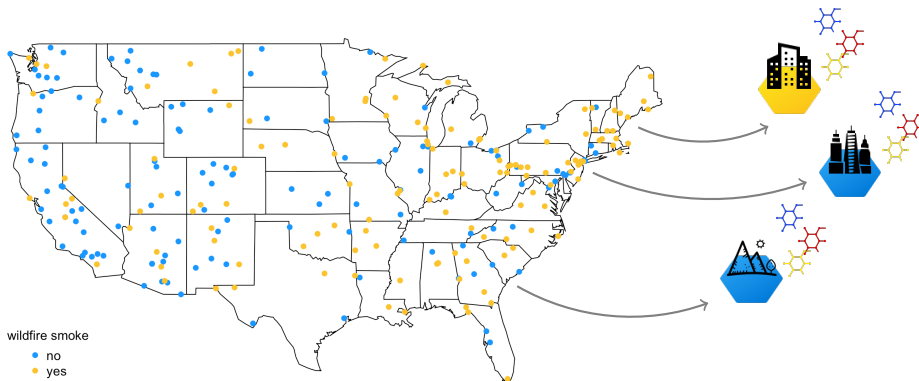
Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository echolab-stanford/daily-10km-smokePM (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidycensus R package (Walker et al., 2021).

Causal effect of wildfire smoke on ambient chemical



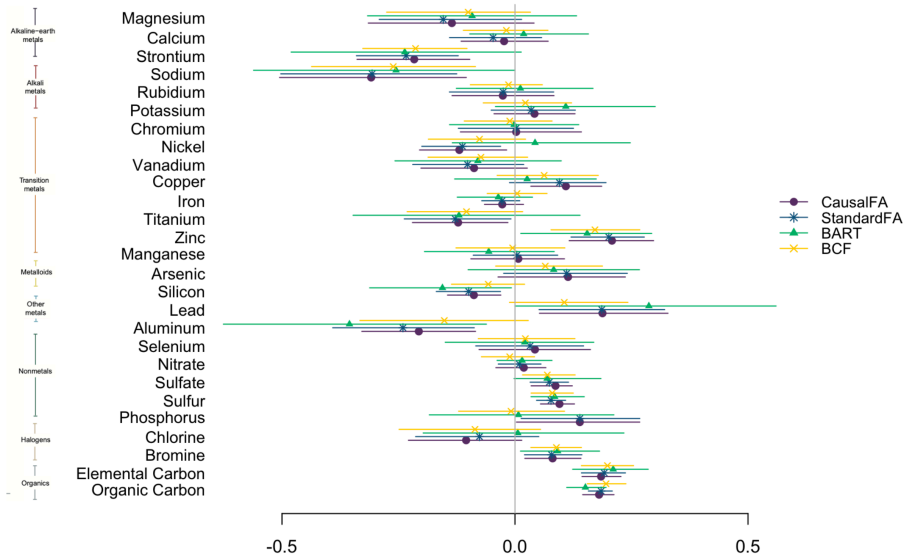
Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository [echolab-stanford/daily-10km-smokePM](#) (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidyrcensus R package (Walker et al., 2021).

Causal effect of wildfire smoke on ambient chemical

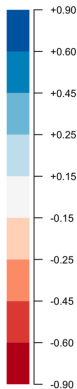
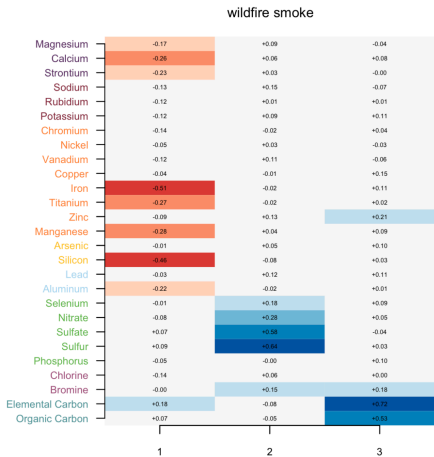
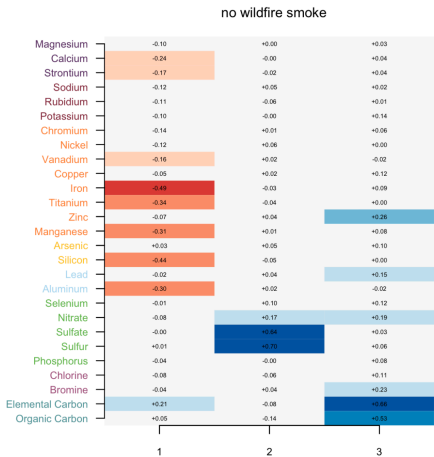


Dataset sources: Krasovich Southworth et al. (2025), Federal Land Manager Environmental Database (F.E.D. IMPROVE, 2024), GitHub repository [echolab-stanford/daily-10km-smokePM](#) (Child, et al 2022), Harvard Dataverse (Audirac, 2024), tidycensus R package (Walker et al., 2021).

Causal effects of wildfire smoke

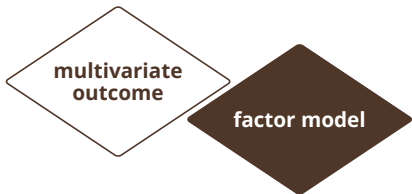


Treatment-specific factor loadings



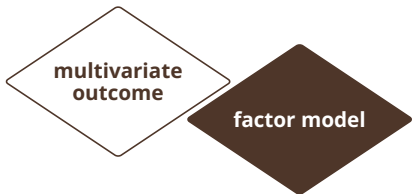
Wrapping up

Wrapping up



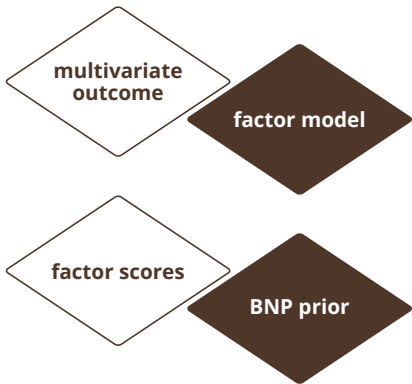
- We define a causal inference framework for multivariate potential outcomes.

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

