Methods for causal inference under interference with applications to infectious disease and mobile health

Xiaoxuan Cai

Department of Biostatistics, Mailman School of Public Health
Columbia University

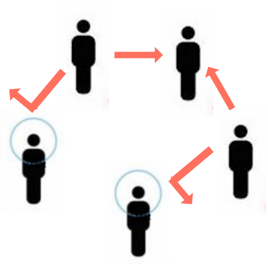
NYU Department of Population Health, Biostatistics Division

(References and slides are available on my personal website.)

Causal inference under interference

Classical causal inference assumes i.i.d realizations, which can be inappropriate for applications when dependence exist among observed data. This phenomenon is referred to as “interference”.

- Contagion of infectious outcomes
- Auto-correlation with past information in time series



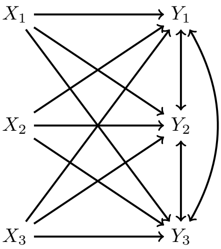
Infectious disease

Mobile health

Causal inference under interference

Classical causal inference assumes i.i.d realizations, which can be inappropriate for applications when dependence exist among observed data. This phenomenon is referred to as “interference”.

- Contagion of infectious outcomes
- Auto-correlation with past information in time series



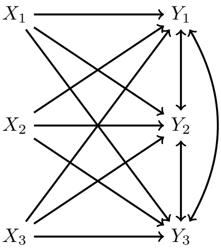
Interference across subjects

Mobile health

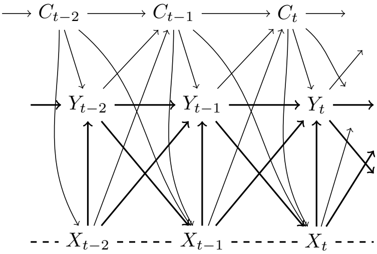
Causal inference under interference

Classical causal inference assumes i.i.d realizations, which can be inappropriate for applications when dependence exist among observed data. This phenomenon is referred to as “interference”.

- Contagion of infectious outcomes
- Auto-correlation with past information in time series



Interference across subjects



Interference across time points

Research Outline

Causal evaluation of infectious disease interventions

- Articulate causal structure of infectious outcomes in a stochastic and interactive transmission network, in which outcomes and treatments are all interdependent
- Propose novel causal estimands for individual-level direct and indirect vaccine effects under a general stochastic model, and provide non-parametric, semi-parametric, or parametric causal identification with adjustment for individual covariates.

Behavioral interventions for mental health using mobile devices

- Causal inference of time-varying exposures in short- and long-term in non-stationary multivariate time series of N-of-1 studies
- Missing data imputation for non-stationary multivariate time series
- Pathway decomposition and mediation analysis for non-stationary multivariate time series.

Causal inference for infectious disease interventions in a inter-connected population

joint work with Forrest W. Crawford, Wen Wei Loh, Eben Kenah

References:

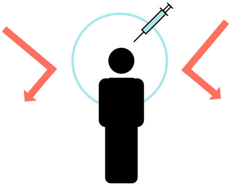
Xiaoxuan Cai, Wen Wei Loh, Forrest W. Crawford. (2021) Identification of Causal intervention effects under contagion. Journal of Causal Inference, 9, 9-38. (Winner of best paper award, ASA Section on Statistics in Epidemiology)

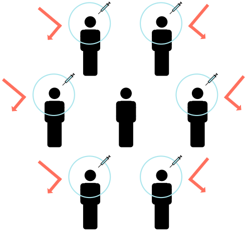
Xiaoxuan Cai, Eben Kenah, Forrest W. Crawford. (2020) Causal identification of infectious disease intervention effects in a clustered population. arXiv:2105.03493

Infectious disease and vaccination

- Direct protection for the treated individuals:
  - direct effect, vaccine efficacy, susceptibility effect, ...
- Indirect protection for the surrounding individuals:
  - indirect effect, herd immunity, contagion effect, infectiousness effect, ...

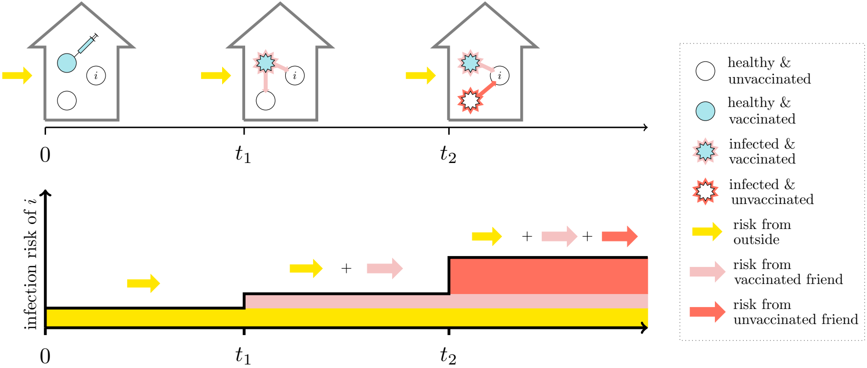
  (Examples: vaccines for Polio, Influenza, HIV/AIDS, Malaria, and etc).


Direct protection


Indirect protection

How epidemiologists understand disease transmission

For a focal individual *i*, the risk of infection increases as more neighbors become infectious and depend on neighbors' vaccination status.



One infection outcome depends on (i) its **own treatment**, (ii) **treatments of others**, and (iii) **infection times of others** (e.g. t_1, t_2).

Notation

- Subject i's treatment: X_i
- Others' treatments: $\mathbf{X}_{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$
- Others' infection history: $\boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)}) = \{T_j(\mathbf{x}_{(i)}) : j \neq i\}$

Potential outcome

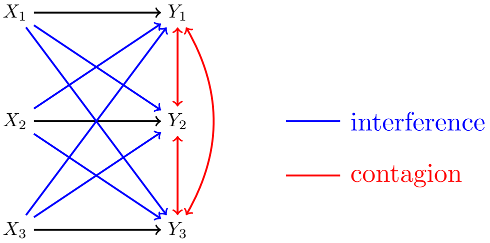
Identify $Y_i\big(t; x_i, \mathbf{x}_{(i)}, \boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)})\big)$ as the counterfactual infection outcome at time t under joint treatment $(x_i, \mathbf{x}_{(i)})$ and infection times of others $\boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)})$ under treatments $\mathbf{x}_{(i)}$,

- (i) own treatment: $X_i = x_i$
- (ii) others' treatments: $\mathbf{X}_{(i)} = \mathbf{x}_{(i)}$
- (iii) others' infection times: $\boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)})$

Why is infectious disease difficult to study?

1. Interdependence of outcomes and treatments across subjects
   - The infection outcome of one individual depends on others' treatments
   - The infection outcome of one individual depends on other's outcomes, since it is transmissible.

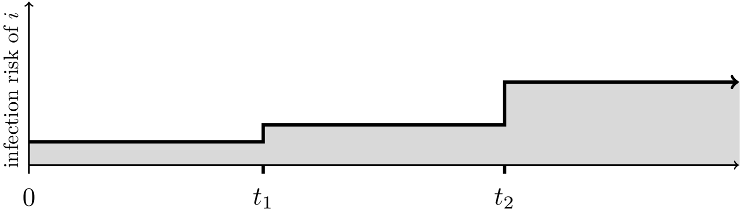
Consider a interconnected three individuals with treatment (X_1, X_2, X_3) and infection outcome (Y_1, Y_2, Y_3).



Bidirectional arrows causes problems in causal identification

Why is infectious disease difficult to study?

1. Interdependence of outcomes and treatments across subjects
2. Stochastic processes of "exposure to infection" $\boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)})$
   - "Exposure to infection" is determined by stochastic infection outcomes of others, whose distribution depends on their treatments.



Exposure to infection

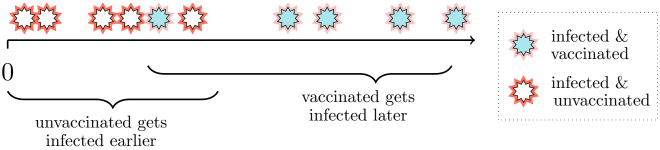
For example, earlier infection exposure (smaller t_1 and t_2) or a higher fraction of unvaccinated infectious members (red arrows) increases "exposure to infection" and consequently infection risk.

Why is infectious disease difficult to study?

1. Interdependence of outcomes and treatments across subjects
2. Stochastic processes of "exposure to infection" $\boldsymbol{\mathcal{H}}_{(i)}(\mathbf{x}_{(i)})$
3. Bias due to differential "exposure to infection"

Can we directly compare treated and untreated individuals using randomization?

$$E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$$


0
unvaccinated gets infected earlier
vaccinated gets infected later
infected & vaccinated
infected & unvaccinated

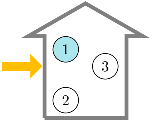
For example, vaccinated individuals endure higher exposure to infection, which is not a fair comparison. → Effect is under-estimated!

NO! It may be biased due to differential "exposure to infection". [1,2,3].

So how to solve the causal identification problem for infectious disease outcomes?

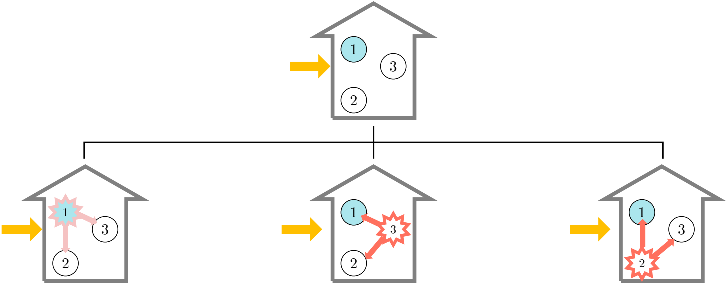
Decompose infection process regarding different orders

○ healthy & unvaccinated ● healthy & vaccinated ✷ infected & vaccinated ✷ infected & unvaccinated ➡ risk from outside ➡ risk from vaccinated ➡ risk from unvaccinated



Decompose infection process regarding different orders

○ healthy & unvaccinated ● healthy & vaccinated ✸ infected & vaccinated ✸ infected & unvaccinated ➡ risk from outside ➡ risk from vaccinated ➡ risk from unvaccinated



Decompose infection process regarding different orders

○ healthy & unvaccinated

● healthy & vaccinated

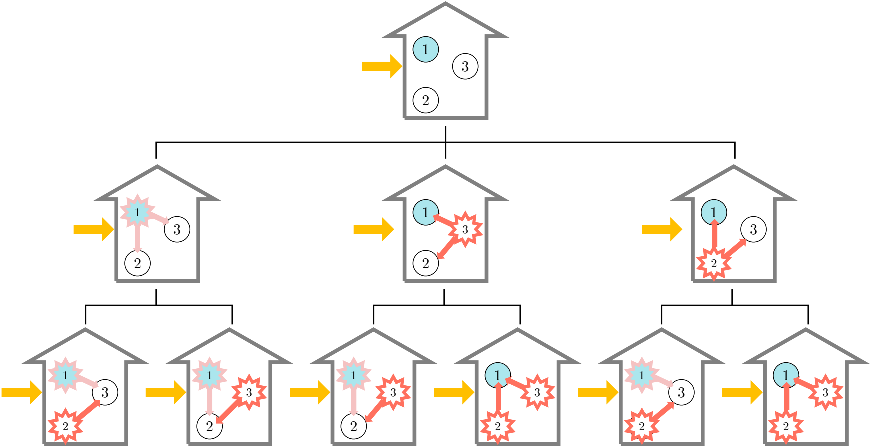
✸ infected & vaccinated

✸ infected & unvaccinated

➡ risk from outside

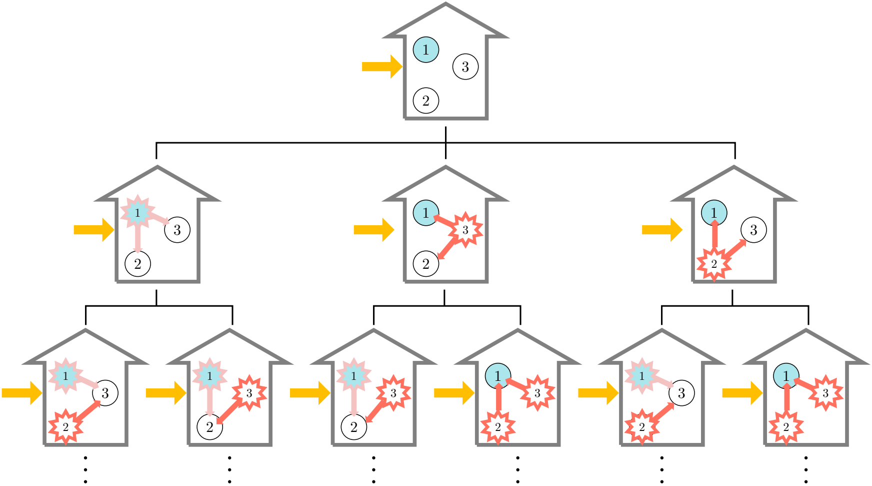
➡ risk from vaccinated

➡ risk from unvaccinated



Decompose infection process regarding different orders

○ healthy & unvaccinated ● healthy & vaccinated ✷ infected & vaccinated ✷ infected & unvaccinated ➡ risk from outside ➡ risk from vaccinated ➡ risk from unvaccinated



Decompose infection process regarding different orders

○ healthy & unvaccinated

● healthy & vaccinated

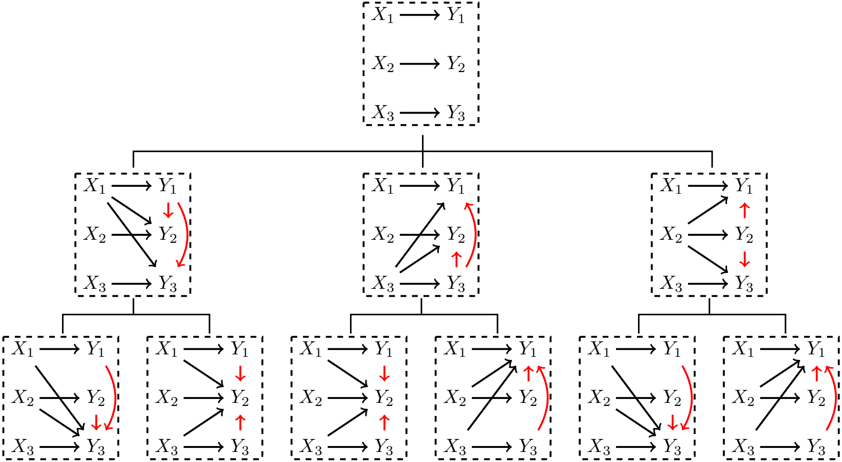
✸ infected & vaccinated

✸ infected & unvaccinated

➡ risk from outside

➡ risk from vaccinated

➡ risk from unvaccinated



Identification of exposure-marginalized potential outcomes

Causal identification

Under conventional assumptions in causal inference, the potential outcome $\mathbb{E}\big[Y_i\big(t; x_i, \mathbf{x}_{(i)}, \boldsymbol{\mathcal{H}}^*_{(i)}(\mathbf{x}'_{(i)})\big)\big|\mathbf{L} = \mathbf{l}\big]$ can be identified as

$$\mathbb{E}\big[Y_i\big(t; x_i, \mathbf{x}_{(i)}, \boldsymbol{\mathcal{H}}^*_{(i)}(\mathbf{x}'_{(i)})\big)\big|\mathbf{L} = \mathbf{l}\big] = \int \mathbb{E}[Y_i(t; x_i, \mathbf{x}_{(i)}, \mathbf{h}_{(i)})|\mathbf{L} = \mathbf{l}]\, \mathrm{d}G^*_{(i)}(\mathbf{h}_{(i)}|\mathbf{x}'_{(i)}, \mathbf{l}_{(i)})$$

where

$$\mathbb{E}\big[Y_i(t; \mathbf{h}_{(i)}, \mathbf{x})\,\big|\,\mathbf{L} = \mathbf{l}\big] = \sum_{j=0}^{n-1} \Big[F_{I_i^j}(\min\{t, t_{(i)}^{j+1}\} - t_{(i)}^j \,|\, \mathbf{x}, \mathbf{h}_{(i)}, \mathbf{l}) \prod_{k=0}^{j-1} \big(1 - F_{I_i^k}(t_{(i)}^{k+1} - t_{(i)}^k \,|\, \mathbf{x}, \mathbf{h}_{(i)}, \mathbf{l})\big)\Big]$$

$$\mathrm{d}\,G^*_{(i)}(\mathbf{h}_{(i)} \,|\, \mathbf{x}_{(i)},) = \prod_{j=1}^{n-1} \Big[f_{I^{j-1}_{\varphi_i^j}}\big(t_{(i)}^j - t_{(i)}^{j-1} \,\big|\, \mathbf{x}, \mathbf{h}^i_{(\varphi_i^j)}, \mathbf{l}\big) \prod_{k=j+1}^{n-1} S_{I^{j-1}_{\varphi_i^k}}\big(t_{(i)}^j - t_{(i)}^{j-1} \,\big|\, \mathbf{x}, \mathbf{h}^j_{(\varphi_i^k)}, \mathbf{l}\big)\Big]$$

$$F_{I_i^k}(s|\mathbf{x}, \mathbf{h}_{(i)}, \mathbf{l}) = 1 - \exp\left[-\int_{t_{(i)}^k}^{t_{(i)}^k + s} \frac{f_i^k(u|\mathbf{x}, \mathbf{h}_{(i)}, \mathbf{l})}{S_i^k(u|\mathbf{x}, \mathbf{h}_{(i)}, \mathbf{l})} du\right] \text{ for } k = 0, \ldots, n-1$$

Exposure-marginalized (natural) causal estimands

Exposure-marginalized (natural) causal estimands

- Susceptibility effect

$$SE_i(t, \mathbf{x}_{(i)}) = \mathbb{E}\big[Y_i\big(t; 1, \mathbf{x}_{(i)}, \mathcal{H}^*_{(i)}(\mathbf{x}_{(i)})\big) - Y_i\big(t; 0, \mathbf{x}_{(i)}, \mathcal{H}^*_{(i)}(\mathbf{x}_{(i)})\big)\big]$$

- Infectiousness effect

$$IE_i(t, x_i, \mathbf{x}_{(i)}) = \mathbb{E}\big[Y_i\big(t; x_i, \mathbf{1}, \mathcal{H}_{(i)}(\mathbf{x}_{(i)})\big) - Y_i\big(t; x_i, \mathbf{0}, \mathcal{H}_{(i)}(\mathbf{x}_{(i)})\big)\big]$$

- Contagion effect

$$CE_i(t, x_i, \mathbf{x}_{(i)}, \mathbf{x}'_{(i)}) = \mathbb{E}\big[Y_i\big(t; x_i, \mathbf{x}_{(i)}, \mathcal{H}^*_{(i)}(\mathbf{x}_{(i)})\big) - Y_i\big(t; x_i, \mathbf{x}_{(i)}, \mathcal{H}^*_{(i)}(\mathbf{x}'_{(i)})\big)\big]$$

- Susceptibility effect → shows if the vaccine protects treated individual
- Infectiousness effect → shows if the vaccine decreases transmission ability
- Contagion effect → shows if the disease is contagious

Traditional causal estimands in cluster studies

- Direct effect:

$$DE(t) = \mathbb{E}[Y_i(t)|X_i = 1] - \mathbb{E}[Y_i(t)|X_i = 0]$$

- Indirect effect:

$$\begin{aligned} IDE(t) = &\sum_{|\mathbf{x}_{(i)}|=\frac{n}{2}} \mathbb{E}[Y_i(t)|X_i = 0, \mathbf{X}_{(i)} = \mathbf{x}_{(i)}]p(\mathbf{x}_{(i)}) \\ &- \sum_{|\mathbf{x}_{(i)}|=0} \mathbb{E}[Y_i(t)|X_i = 0, \mathbf{X}_{(i)} = \mathbf{x}_{(i)}]p(\mathbf{x}_{(i)}) \end{aligned}$$

[1] Longini et al. Statistical inference for infectious diseases: risk-specific household and community transmission parameters. American Journal of Epidemiology, 128(4):845–859, 1988.
[2] Halloran et al. Direct and indirect effects in vaccine efficacy and effectiveness. American Journal of Epidemiology, 133(4):323–331, 1991.
[3] Halloran et al. Exposure efficacy and change in contact rates in evaluating prophylactic HIV vaccines in the field. Statistics in Medicine, 13(4):357–377, 1994.

Simulation: Estimations of causal estimands

Cluster	Treatment	Probability estimands				
		$\hat{CE}$	$\hat{SE}$	$\hat{IE}$	$DE(t)$	$IDE(t)$
2	Obs.	0.005	-0.015	-0.036	-0.013	-0.036
	Bernoulli	0.004	-0.015	-0.036	-0.014	-0.038
	Block	0.004	-0.013	-0.036	0.025	-
	Cluster	0.004	-0.013	-0.035	-0.048	-
4	Obs.	0.026	-0.014	-0.084	-0.012	-0.073
	Bernoulli	0.025	-0.013	-0.082	-0.012	-0.063
	Block	0.026	-0.015	-0.082	0.016	-
	Cluster.	0.025	-0.014	-0.083	-0.099	-
8	Obs.	0.068	-0.013	-0.131	-0.010	-0.088
	Bernoulli	0.069	-0.014	-0.133	-0.010	-0.096
	Block	0.069	-0.014	-0.132	0.010	-
	Cluster	0.070	-0.016	-0.132	-0.154	-

Simulation under $e^{\beta_1} = 0.9$, $e^{\beta_2} = 0.1$, $\alpha(t) = 0.3$, $\gamma(t) = 3$ and $e^{\theta_1} = e^{\theta_2} = 0.9$. Clusters of 2, 4, and 8 are observed at 0.4, 0.3 and 0.2 year.

Other relevant work and future direction

- We further apply a generalized Cox-type transmission hazard model to facilitate the inference of causal estimands parametrically or semi-parametrically.
- We promote hazard ratio as alternative causal estimands for the susceptibility and infectiousness effect, and compared them to existing estimands for vaccine efficacy.
- Extend current research on causal identification for contagious outcomes to more realistic scenarios, for example, relaxing requirement on accurate infection times, accommodating incomplete knowledge of transmission network, allowing recovering and re-infection of outcomes.

Causal inference and missing data imputation for non-stationary time series data in mobile health

joint work with Xinru Wang, Dost Ongur, Lisa Dixon, Justin T. Baker, Jukka-Pekka Onnela, Linda Valeri

References:

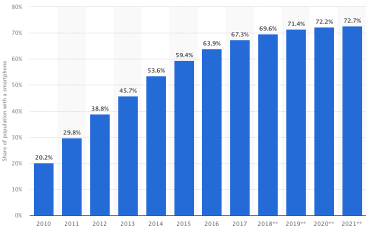
Xiaoxuan Cai, Xinru Wang, Lisa Dixon, Justin T. Baker, Jukka-Pekka Onnela, Linda Valeri (2021) State space model multiple imputation for missing data in non-stationary multivariate time series. (Manuscript accepted by NeurIPS 2021 Workshop on Causal Inference Challenges in Sequential Decision Making: Bridging Theory and Practice)

Xiaoxuan Cai, Jukka-Pekka Onnela, Justin T. baker, Linda Valeri (2021) Causal inference for non-stationary multivariate time series data from mobile devices in N-of-1 studies.

Linda Valeri, Xiaoxuan Cai, Aijin Wang, Zixu Wang, Habiballah Rahimi Eichi, Einat Liebenthal, Scott Rauch, Dost Ongur, Russell Schutt, Lisa Dixon, Justin Baker, Jukka-Pekka Onnela (2021). Smartphone-based markers of social networks in schizophrenia and bipolar disorder.

Causal inference in mHealth

> “mHealth is the use of mobile and wireless devices (cell phones, tablets, etc.) to improve health outcomes, health care services, and health research.” – NIH



Smartphone penetration[1]

Mobile health[2]

[1] https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/
[2] https://www.shutterstock.com/g/maschatace

Bipolar Longitudinal Study (McLean Hospital)

The study follows 73 patients with severe mental illness (SMI), and explores how passive sensor data is linked to moods and cognitive status.

- DSM-V diagnosis established once enrolled
- Monthly assessment of clinical symptoms (e.g., PANSS, MADRS, ...)
- User-reported survey data via the Beiwe app (mood, life-habits, in-person interactions, ...)
- Passively collected telecommunication data (call and text logs), GPS data, and accelerometer data using smartphones and fitness trackers
- EHR data about medication use and psychotherapy

Beiwe app

GEVEActiv watch

Samsung Galaxy Note 8

Bipolar Longitudinal Study (McLean Hospital)

The study follows 73 patients with severe mental illness (SMI), and explores how passive sensor data is linked to moods and cognitive status.

- DSM-V diagnosis established once enrolled
- Monthly assessment of clinical symptoms (e.g., PANSS, MADRS, ...)
- User-reported survey data via the Beiwe app (mood, life-habits, in-person interactions, ...)
- Passively collected telecommunication data (call and text logs), GPS data, and accelerometer data using smartphones and fitness trackers
- EHR data about medication use and psychotherapy

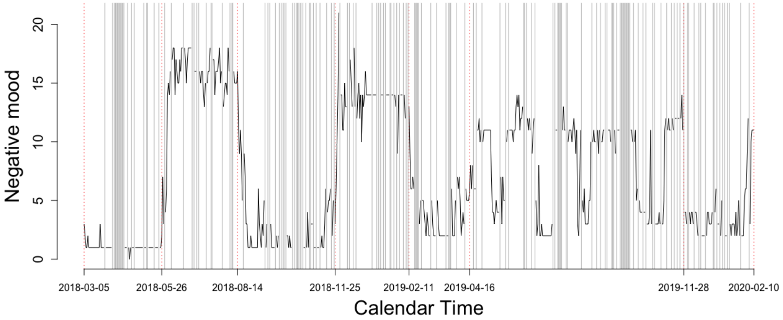
Our research of interest

Evaluate the causal effect of **social support** on **mood improvement** in patients with serious mental illness in an <u>observational N-of-1 study</u>.

Bipolar Longitudinal Study (McLean Hospital)

Focus on one female participant Bipolar I disorder, who has been followed up from 03/05/2018 to 2020/20/10 (708 days).

- Outcome (Y_t): a self-reported composite index for negative moods, including being afraid, anxious, ashamed, hostile, stressed, upset, irritable and lonely.

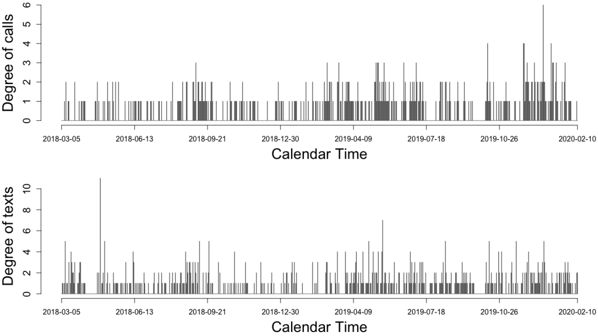


Missing rates: 23.31%

Bipolar Longitudinal Study (McLean Hospital)

Focus on one female participant Bipolar I disorder, who has been followed up from 03/05/2018 to 2020/20/10 (708 days).

- Outcome (Y_t): a self-reported composite index for negative moods, including being afraid, anxious, ashamed, hostile, stressed, upset, irritable and lonely.
- Exposures (X_t): passively collected degree of calls and texts

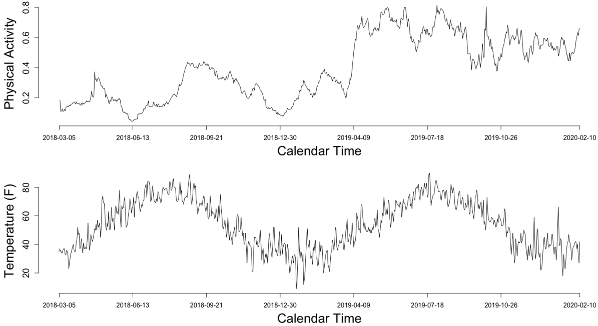


Missing rates: 0%

Bipolar Longitudinal Study (McLean Hospital)

Focus on one female participant Bipolar I disorder, who has been followed up from 03/05/2018 to 2020/20/10 (708 days).

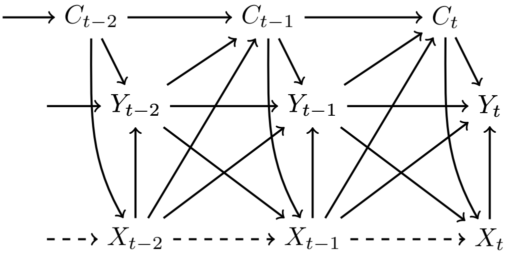
- Outcome (Y_t): a self-reported composite index for negative moods, including being afraid, anxious, ashamed, hostile, stressed, upset, irritable and lonely.
- Exposures (X_t): passively collected degree of calls and texts
- Confounders (C_t): passively collected accelerometer data and temperature



Temperature is obtained from National Centers for Environmental Information (NOAA) Database. Physical activity is processed following Bai (2013,2014). Missing rates: 0%

Causal structure for the Bipolar Longitudinal Study

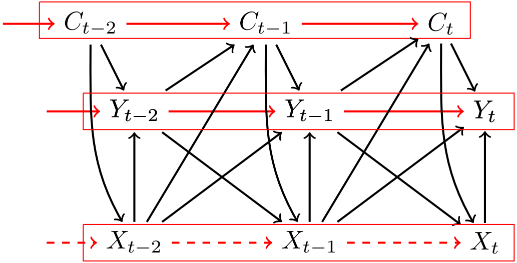
- Outcome (Y_t): self-reported negative mood of the patient
- Exposure (X_t): degree of calls and texts
- Confounders (C_t): physical activity, temperature, ...



How to deal with missing data for non-stationary multi-variate time series in N-of-1 studies?

Problem due to missing data in mHealth

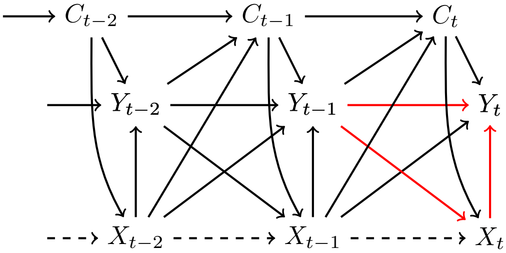
Denote outcome as Y_t, exposure as X_t, and other confounders as C_t. Assume true data generation process as



- High-autocorrelation with lagged values of the variables

Problem due to missing data in mHealth

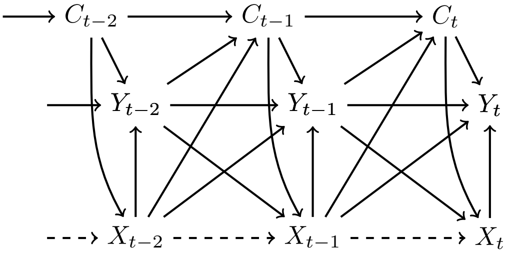
Denote outcome as Y_t, exposure as X_t, and other confounders as C_t. Assume true data generation process as



- High-autocorrelation with lagged values of the variables
- Elevated missing rate due to including previous values of variables study the effect of X_t on Y_t → Y_{t-1} is included
  → increase missing rate 50.1% → 74.1%

Problem due to missing data in mHealth

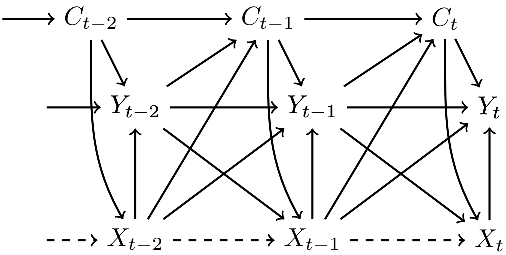
Denote outcome as Y_t, exposure as X_t, and other confounders as C_t. Assume true data generation process as



- High-autocorrelation with lagged values of the variables
- Elevated missing rate due to including previous values of variables study the effect of X_t on Y_t → Y_{t-1} is included
- Personalized monitoring of a single individual

Problem due to missing data in mHealth

Denote outcome as Y_t, exposure as X_t, and other confounders as C_t. Assume true data generation process as



- High-autocorrelation with lagged values of the variables
- Elevated missing rate due to including previous values of variables study the effect of X_t on $Y_t \rightarrow Y_{t-1}$ is included
- Personalized monitoring of a single individual
- Non-stationary multi-variate time series

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ...
  - Multiple imputation
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
- Complete case analysis

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... → Biased
  - Multiple imputation
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
- Complete case analysis

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... → Biased
  - Multiple imputation → Based on static models and stationarity
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
- Complete case analysis

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... $\rightarrow$ Biased
  - Multiple imputation $\rightarrow$ Based on static models and stationarity
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
  $\rightarrow$ not appropriate for multivariate time series
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
- Complete case analysis

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... → Biased
  - Multiple imputation → Based on static models and stationarity
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
  → not appropriate for multivariate time series
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
  → require multiple subjects, not appropriate for N-of-1 studies
- Complete case analysis

Existing methods for missing data

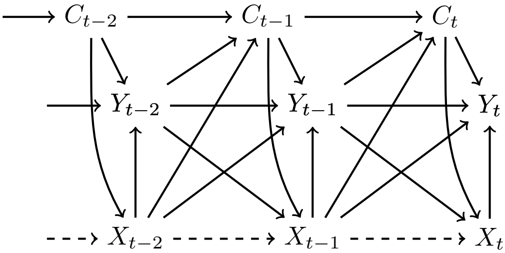
- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... → Biased
  - Multiple imputation → Based on static models and stationarity
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
  → not appropriate for multivariate time series
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
  → require multiple subjects, not appropriate for N-of-1 studies
- Complete case analysis → break temporal structure, to be evaluated

Existing methods for missing data

- Longitudinal studies:
  - Mean imputation, Last-observation-carried-forward (LOCF) imputation, Linear or spline imputation, ... → Biased
  - Multiple imputation → Based on static models and stationarity
- Univariate time series:
  Simple moving average, Exponential weighted moving average, ARIMA, ...
  → not appropriate for multivariate time series
- Multivariate time series:
  Recurrent neural network, Generative adversarial network, ...
  → require multiple subjects, not appropriate for N-of-1 studies
- Complete case analysis → break temporal structure, to be evaluated

New imputation method for non-stationary multi-variate time series is needed.

Simulation: non-stationary with change points and random walk

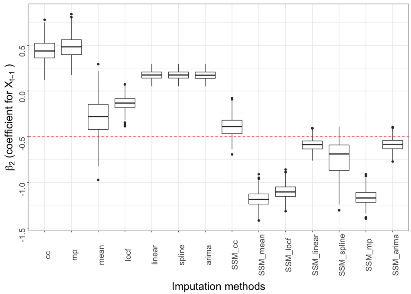


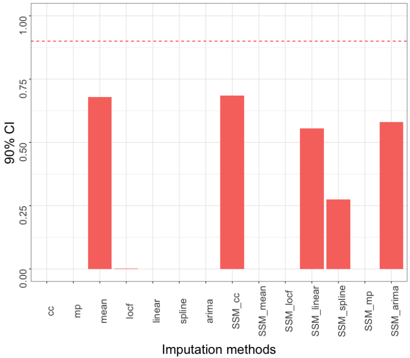
$$Y_t = \color{red}{\beta_{0,t}} + \rho Y_{t-1} + \color{red}{\beta_{1,t}} X_t + \beta_2 X_{t-1} + \beta_c C_t + v_t, \quad v_t \sim N(0, V)$$

where

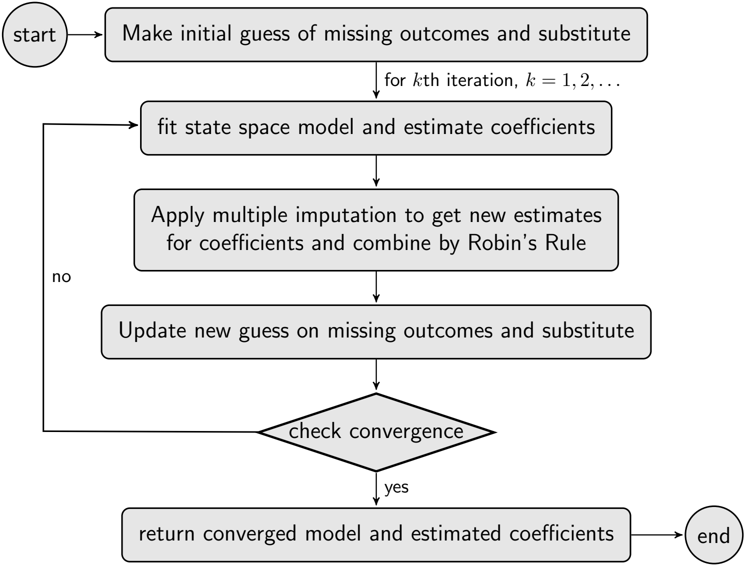
- Random walk intercept $\beta_{0,t} = 40 + \beta_{0,t-1} + w_t$, $w_t \sim N(0, 1)$.
- Periodic coefficient $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$

Simulation: estimated $\hat{\beta}_{2,t}$ under existing methods





State space model multiple imputation (SSMimpute)



start
Make initial guess of missing outcomes and substitute
for kth iteration, $k = 1, 2, \ldots$
fit state space model and estimate coefficients
Apply multiple imputation to get new estimates for coefficients and combine by Robin's Rule
Update new guess on missing outcomes and substitute
check convergence
no
yes
return converged model and estimated coefficients
end

State-space model imputation (SSMmp)

Remark1

The state space model reveals its structure as well as its unknown parameters along with iterations until convergence.

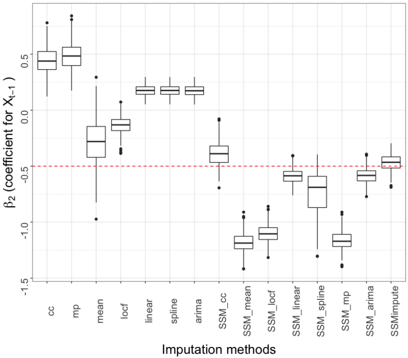
Remark2

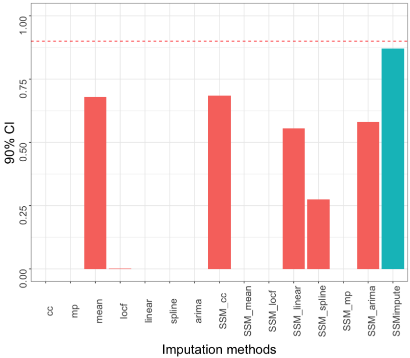
Missing values are only imputed for missing lagged outcomes in the confounder adjustment set, not for the missing outcome in the response variable.

Assumption

We require state space model to be correctly specified with no unmeasured confounders for unbiased estimation of the causal effect.

Simulation: estimated $\hat{\beta}_{2,t}$ under existing methods





Conclusions of simulations: (see more results in the paper)

For stationary time series,

- Mean imputation, LOCF, linear and spline imputations are biased.
- Complete case analysis, multiple imputation, and SSMimpute are unbiased. Multiple imputation and SSMimpute are more efficient than complete case analysis.

For non-stationary time series,

- Complete case analysis breaks temporal structure and induces bias in estimation.
- Mean imputation, LOCF, linear and spline interpolation, and multiple imputation are biased.
- SSMimpute provides unbiased estimation for time-varying coefficients, and is more efficient than complete case analysis.

Estimation for the Bipolar Longitudinal Study

We estimate the association between the degree of calls and texts and the negative mood, controlling for physical activity and temperature.

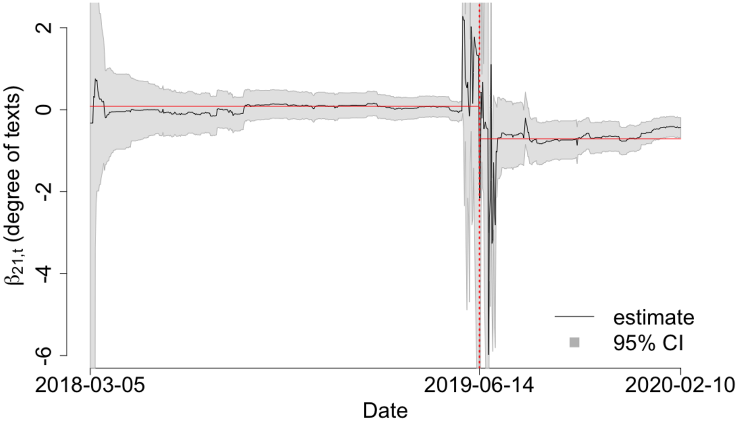
- Outcome: negative mood (Y_t)
- Exposures: degree of calls ($X_{1,t}$) and texts ($X_{2,t}$)
- Covariates: temperature (Temp_t), past physical activity (PA_t)

$$\begin{aligned} Y_t = {} & \text{intercept}_t + \rho_t Y_{t-1} + \beta_{11,t} X_{\text{calls},t} + \beta_{12,t} X_{\text{calls},t-1} \\ & + \beta_{21,t} X_{\text{texts},t} + \beta_{22,t} X_{\text{texts},t-1} + \beta_{temp,t} \text{Temp}_t + \beta_{PA,t} \text{PA}_t + v_t \end{aligned}$$

Missing rate before imputation → 40.4%

Missing rate after imputation → 23.3%

Estimation result: estimated coefficient for degree of outgoing texts



Estimation result: compared to multiple imputation

$$
\begin{aligned}
Y_t = \text{intercept}_t + \rho_t Y_{t-1} + \beta_{11,t} X_{\text{calls},t} + \beta_{12,t} X_{\text{calls},t-1} \\
+ \beta_{21,t} X_{\text{texts},t} + \beta_{22,t} X_{\text{texts},t-1} + \beta_{temp,t} \text{Temp}_t + \beta_{PA,t} \text{PA}_t + v_t
\end{aligned}
$$

	SSMimpute (n=542)		multiple imputation (n=542)	
	Estimate	90% CI	Estimate	90% CI
intercept_t	(random walk)		(random walk)	
ρ_t (for Y_{t-1})	0.64	(0.57,0.71)	0.11	(-0.14,0.36)
$\beta_{11,t}$	-0.14	(-0.27,0.00)	-0.11	(-0.23,0.01)
$\beta_{12,t}$	0.00	(-0.12,0.12)	-0.05	(-0.16,0.07)
$\beta_{21,t}$ (period 1)	-0.03	(-0.30,0.24)	-0.02	(-0.27,0.23)
$\beta_{21,t}$ (period 2)	-0.49	(-0.78,-0.21)	-0.38	(-0.65,-0.1)
$\beta_{22,t}$	-0.17	(-0.37,0.03)	-0.23	(-0.42,-0.05)
$\beta_{\text{PA},t}$ (period 1)	-5.87	(-16.73,5.00)	-3.94	(-18.65,10.76)
$\beta_{\text{PA},t}$ (period 2)	-12.19	(-21.27,-3.11)	-16.96	(-32.94,-0.98)
$\beta_{\text{PA},t}$ (period 3)	2.31	(-1.00,5.62)	1.64	(-3.97,7.25)
$\beta_{\text{temp},t}$	-0.01	(-0.03,0.01)	-0.01	(-0.03,0.01)

Estimation result: compared to multiple imputation

$$
\begin{aligned}
Y_t = \text{intercept}_t + \rho_t Y_{t-1} + \beta_{11,t} X_{\text{calls},t} + \beta_{12,t} X_{\text{calls},t-1} \\
+ \beta_{21,t} X_{\text{texts},t} + \beta_{22,t} X_{\text{texts},t-1} + \beta_{temp,t} \text{Temp}_t + \beta_{PA,t} \text{PA}_t + v_t
\end{aligned}
$$

	SSMimpute (n=542)		multiple imputation (n=542)	
	Estimate	90% CI	Estimate	90% CI
intercept_t	(random walk)		(random walk)	
ρ_t (for Y_{t-1})	0.64	(0.57,0.71)	0.11	(-0.14,0.36)
$\beta_{11,t}$	-0.14	(-0.27,0.00)	-0.11	(-0.23,0.01)
$\beta_{12,t}$	0.00	(-0.12,0.12)	-0.05	(-0.16,0.07)
$\beta_{21,t}$ (period 1)	-0.03	(-0.30,0.24)	-0.02	(-0.27,0.23)
$\beta_{21,t}$ (period 2)	-0.49	(-0.78,-0.21)	-0.38	(-0.65,-0.1)
$\beta_{22,t}$	-0.17	(-0.37,0.03)	-0.23	(-0.42,-0.05)
$\beta_{\text{PA},t}$ (period 1)	-5.87	(-16.73,5.00)	-3.94	(-18.65,10.76)
$\beta_{\text{PA},t}$ (period 2)	-12.19	(-21.27,-3.11)	-16.96	(-32.94,-0.98)
$\beta_{\text{PA},t}$ (period 3)	2.31	(-1.00,5.62)	1.64	(-3.97,7.25)
$\beta_{\text{temp},t}$	-0.01	(-0.03,0.01)	-0.01	(-0.03,0.01)

Estimation result: compared to multiple imputation

$$
\begin{aligned}
Y_t = \text{intercept}_t + \rho_t Y_{t-1} + \beta_{11,t} X_{\text{calls},t} + \beta_{12,t} X_{\text{calls},t-1} \\
+ \beta_{21,t} X_{\text{texts},t} + \beta_{22,t} X_{\text{texts},t-1} + \beta_{temp,t}\text{Temp}_t + \beta_{PA,t}\text{PA}_t + v_t
\end{aligned}
$$

	SSMimpute (n=542)		multiple imputation (n=542)	
	Estimate	90% CI	Estimate	90% CI
intercept_t	(random walk)		(random walk)	
ρ_t (for Y_{t-1})	0.64	(0.57,0.71)	0.11	(-0.14,0.36)
$\beta_{11,t}$	-0.14	(-0.27,0.00)	-0.11	(-0.23,0.01)
$\beta_{12,t}$	0.00	(-0.12,0.12)	-0.05	(-0.16,0.07)
$\beta_{21,t}$ (period 1)	-0.03	(-0.30,0.24)	-0.02	(-0.27,0.23)
$\beta_{21,t}$ (period 2)	-0.49	(-0.78,-0.21)	-0.38	(-0.65,-0.1)
$\beta_{22,t}$	-0.17	(-0.37,0.03)	-0.23	(-0.42,-0.05)
$\beta_{\text{PA},t}$ (period 1)	-5.87	(-16.73,5.00)	-3.94	(-18.65,10.76)
$\beta_{\text{PA},t}$ (period 2)	-12.19	(-21.27,-3.11)	-16.96	(-32.94,-0.98)
$\beta_{\text{PA},t}$ (period 3)	2.31	(-1.00,5.62)	1.64	(-3.97,7.25)
$\beta_{\text{temp},t}$	-0.01	(-0.03,0.01)	-0.01	(-0.03,0.01)

Summary

- Existing imputation methods mostly assume the time series to be stationary.
- We proposed a multiple imputation algorithm based on state-space model, which applies to non-stationary multi-variate time series of a single subject.
- The proposed imputation method provides unbiased coefficient estimation for non-stationary time series with missing outcomes.

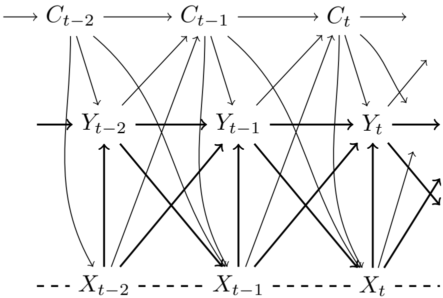
Limitation and future work:

- Extend the SSMimpute method to missing data in exposures and covariates.
- Apply more flexible state space modeling than linear regression
- The current model may suffer from unmeasured confounders.

How to quantify the causal effects of exposure time series on the outcome time series in short- and long-term? (Brief)

Causal structure of the Bipolar Longitudinal Study

- Outcome (Y_t): self-reported negative mood of the patient
- Exposure (X_t): degree of calls and texts
- Confounders (C_t): physical activity, temperature, ...



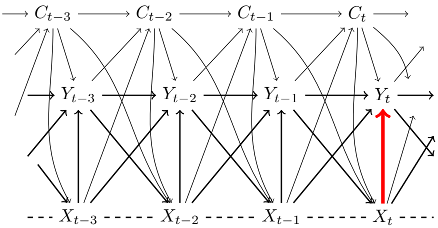
Our research of interest

Evaluate the causal effect of **social support** on **mood improvement** in patients with serious mental illness in an observational N-of-1 study.

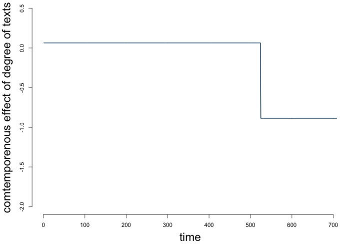
Causal estimand

Causal Quantities of interest: contemporaneous effect

$$\mathbb{E}[Y_t(x_t = 1)] - \mathbb{E}[Y_t(x_t = 0)]$$



Causal diagram

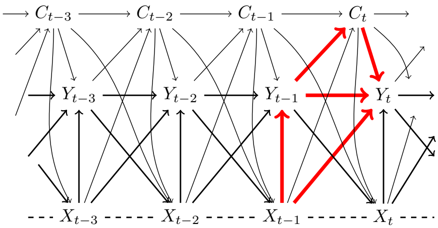


Estimated effect in BLS study

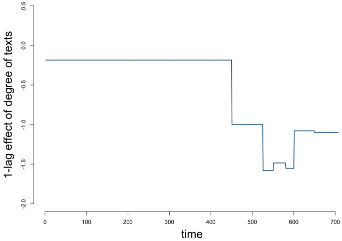
Causal estimand

Causal Quantities of interest: 1-lag total effect

$$\mathbb{E}[Y_t(x_{t-1} = 1, X_t)] - \mathbb{E}[Y_t(x_{t-1} = 0, X_t)]$$



Causal diagram

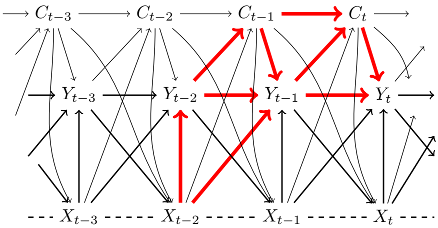


Estimated effect in BLS study

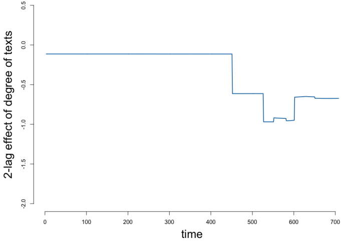
Causal estimand

Causal Quantities of interest: 2-lag total effect

$$\mathbb{E}[Y_t(x_{t-2}=1, X_{(t-1):t})] - \mathbb{E}[Y_t(x_{t-2}=0, X_{(t-1):t})]$$



Causal diagram

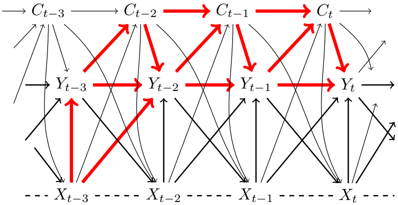


Estimated effect of in BLS study

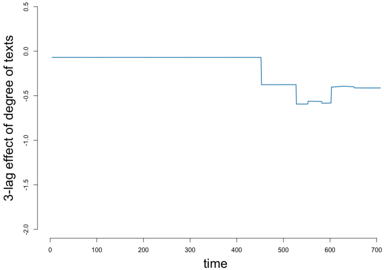
Causal estimand

Causal Quantities of interest: 3-lag total effect

$$\mathbb{E}[Y_t(x_{t-3}=1, X_{(t-2):t})] - \mathbb{E}[Y_t(x_{t-3}=0, X_{(t-2):t})]$$

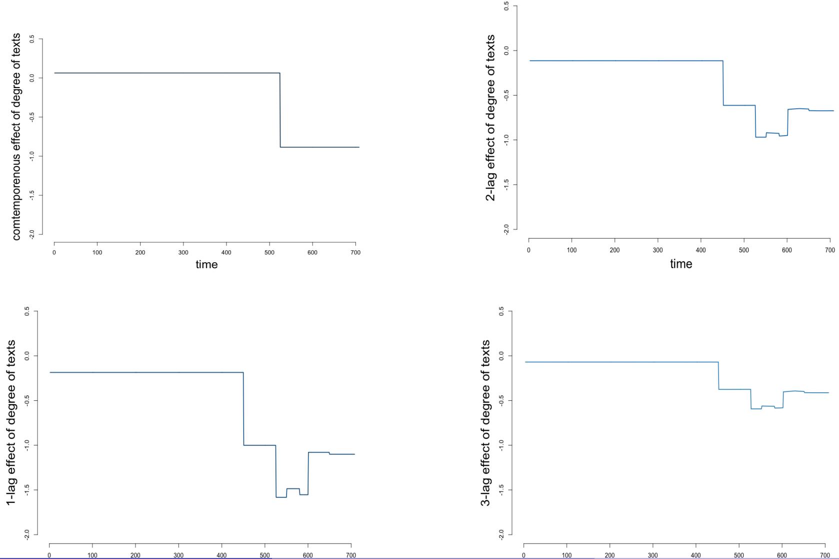


Causal diagram

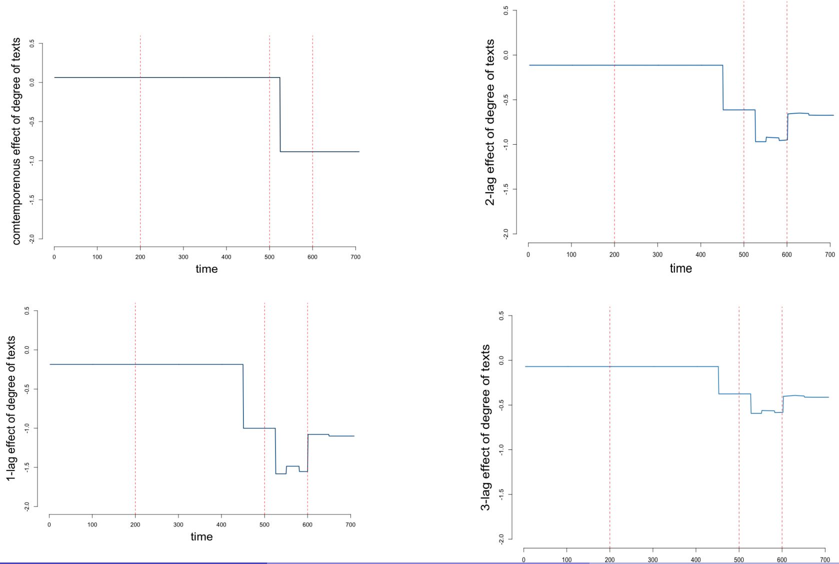


Estimated effect in BLS study

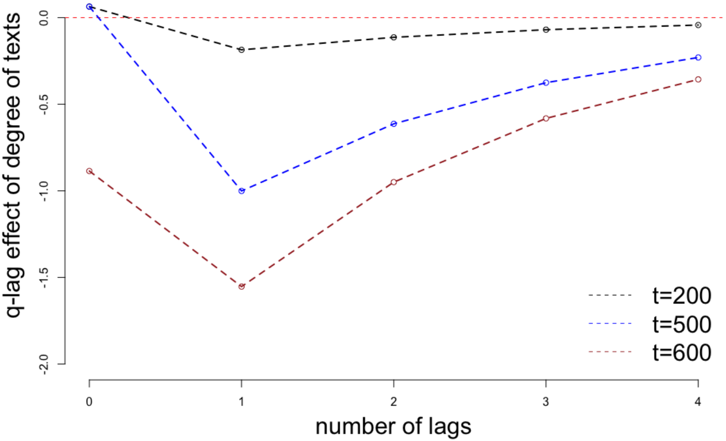
Estimated contemporaneous and 1-, 2- and 3-lag total effect of degree of texts in the BLS study



Estimated contemporaneous and 1-, 2- and 3-lag total effect of degree of texts in the BLS study



Estimated contemporaneous and 1-, 2- and 3-lag total effect of degree of texts in the BLS study



Summary

- We propose a collection of causal estimands for non-stationary multivariate time series in N-of-1 studies, summarizing how time-varying exposures affect outcomes in the short- and long- term
- We provide causal identification for dynamic exposure effects in the presence of feedback between exposures, outcomes, and covariates using g-formula with the state space model.
- We propose graphical tools for checking positivity assumption over different length of exposures, and design optimal treatment strategy under constrains from positivity assumption.

Future direction and limitations

Limitation and future work:

- Extend causal identification for continuous outcome to binary or ordinal outcome.
- Employ machine learning algorithms for more flexible model fitting.
- Apply mediation analysis to decompose long-term effects into different mechanism.
- The current model may suffer from unmeasured confounders.

Acknowledgement

Part 1 research was supported by grant NICHD/BD2K 1DP2HD09179 (PI: Forrest W. Crawford).
Part 2 research was supported by K01MH118477 grand (PI: Linda Valeri).

We thank Peter Aronow, Crystal Blankenbaker, Justin Baker, Zilan Chai, Charly Fowler, Jouni Helske, Zehang Li, Melanie Mayer, Olga Morozova, Jukka-Pekka Onnela, Virginia Pitzer, Habib Rahimi, Aijin Wang, Zixu Wang, Weijia Xiong, and Li Zeng for their great help and suggestions.

xc2577@cumc.columbia.edu
https://xiaoxuan-cai.github.io/

Thank you!