Missing data in non-stationary multivariate time series with application in digital psychiatry

Xiaoxuan Cai

joint work with Charlotte R. Fowler, Li Zeng, Habib Rahimi Eichi, Dost Ongur, Lisa Dixon, Justin Baker, Jukka-Pekka Onnela, Linda Valeri

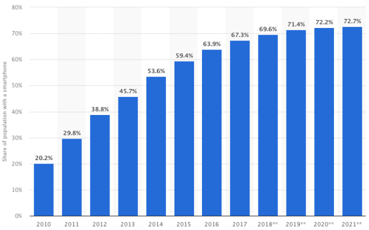
June 17, 2025

Department of Statistics, The Ohio State University
ICSA Applied Statistics Symposium 2025

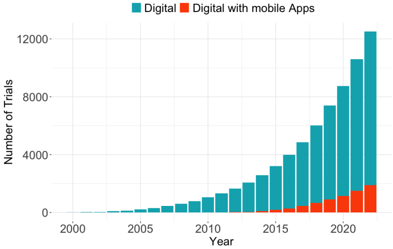
(Paper References: https://arxiv.org/abs/2506.14946.)

Mobile Health Research (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]



Clinical trials with digital devices[2]

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

[2] https://clinicaltrials.gov/. Accessed Oct 28, 2022. 2022 data may be incomplete due to delays in submitting registration.

Mobile Health Research for Bipolar Disease

Bipolar Longitudinal Study follows 73 patients with schizophrenia or bipolar illness over years, and explores how passive sensor data is linked to moods and cognitive status. (McLean Hospital)

- DSM-V diagnosis established once enrolled with monthly followup
-
-
-

Beiwe app

GEVEActiv watch

Samsung Galaxy Note 8

Mobile Health Research for Bipolar Disease

Bipolar Longitudinal Study follows 73 patients with schizophrenia or bipolar illness over years, and explores how passive sensor data is linked to moods and cognitive status. (McLean Hospital)

- DSM-V diagnosis established once enrolled with monthly followup
- User-reported survey data via the Beiwe app (mood, life-habits, ...)
-
-

Beiwe app

GEVEActiv watch

Samsung Galaxy Note 8

Mobile Health Research for Bipolar Disease

Bipolar Longitudinal Study follows 73 patients with schizophrenia or bipolar illness over years, and explores how passive sensor data is linked to moods and cognitive status. (McLean Hospital)

- DSM-V diagnosis established once enrolled with monthly followup
- User-reported survey data via the Beiwe app (mood, life-habits, ...)
- Passively collected telecommunication data (anonymized basic information of calls and texts), GPS data, and accelerometer data, using smartphones and fitness trackers
-

Beiwe app

GEVEActiv watch

Samsung Galaxy Note 8

Mobile Health Research for Bipolar Disease

Bipolar Longitudinal Study follows 73 patients with schizophrenia or bipolar illness over years, and explores how passive sensor data is linked to moods and cognitive status. (McLean Hospital)

- DSM-V diagnosis established once enrolled with monthly followup
- User-reported survey data via the Beiwe app (mood, life-habits, ...)
- Passively collected telecommunication data (anonymized basic information of calls and texts), GPS data, and accelerometer data, using smartphones and fitness trackers
- EHR data about medication use and psychotherapy

Beiwe app

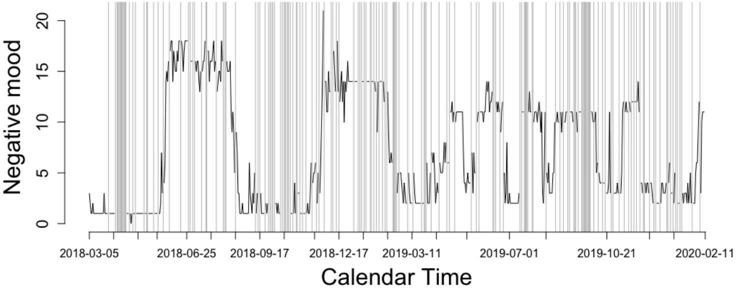
GEVEActiv watch

Samsung Galaxy Note 8

Daily self-evaluation of moods using Bewei survey

Focus on one female participant Bipolar II disorder, who has been followed up from 03/05/2018 to 02/10/2020 (708 days).

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

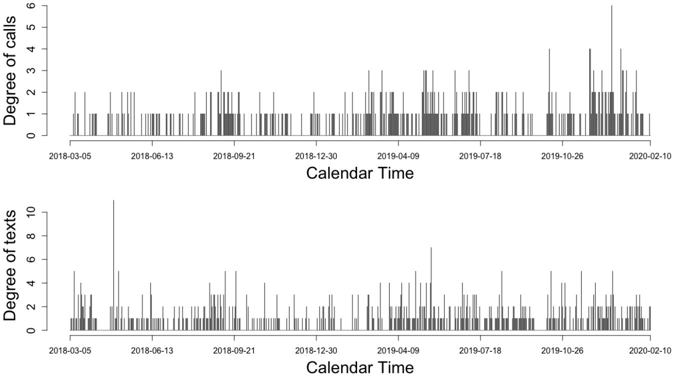


Missing rate is 23.31%.

Daily degrees of calls and texts collected by smartphone

Focus on one female participant Bipolar II disorder, who has been followed up from 03/05/2018 to 02/10/2020 (708 days).

- Exposures A_t: passively collected outgoing degrees of calls and texts on smartphone

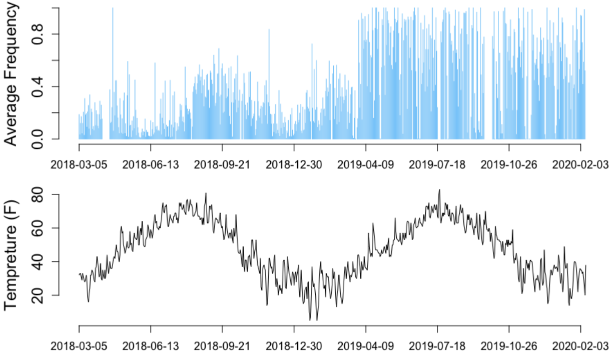


Missing rate is 0.00%.

Physical activity data and Weather temperature

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

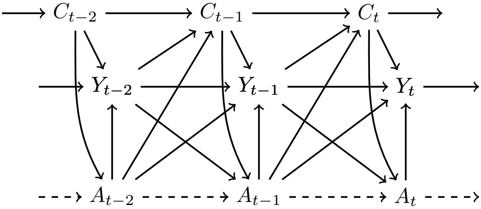
- Confounders C_t: passively collected accelerometer data and weather temperature



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

Causal Structure for the Bipolar Longitudinal Study

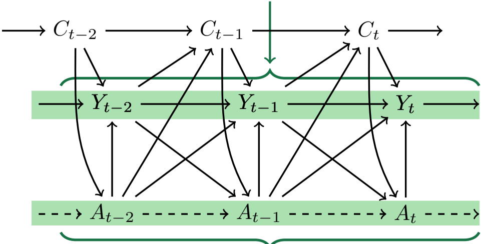
- Outcome (Y_t): self-reported negative mood of the patient
- Exposure (A_t): degree of calls and texts
- Confounders (C_t): physical activity, ...



Causal Structure for the Bipolar Longitudinal Study

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

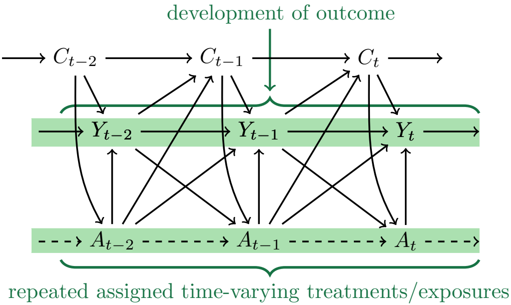
development of outcome



repeated assigned time-varying treatments/exposures

Causal Structure for the Bipolar Longitudinal Study

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



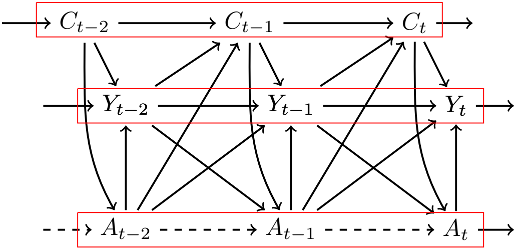
Our research of interest

How can we appropriately handle missing data to enable valid causal inference on the effect of **social support** on **negative mood** in non-stationary multivariate time series?

How to deal with missing data for non-stationary multi-variate time series?

Problem due to missing data in mHealth

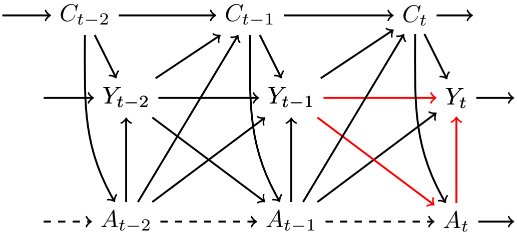
Denote outcome as Y_t, exposure as A_t, and other confounders as C_t. Assume true data generation process as



- High-autocorrelation with lagged values of the variables

Problem due to missing data in mHealth

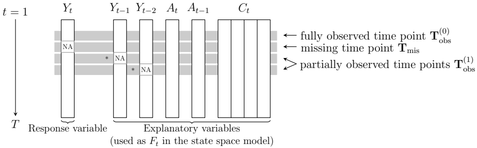
Denote outcome as Y_t, exposure as A_t, and other confounders as C_t. Assume true data generation process as



- 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

Problem due to missing data in mHealth

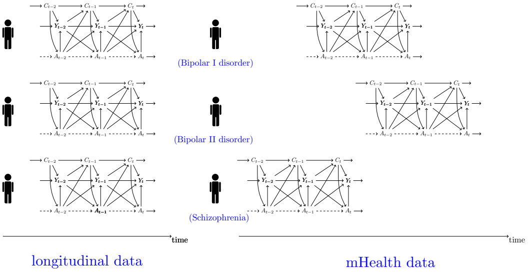
Denote outcome as Y_t, exposure as A_t, and other confounders as C_t. Assume true data generation process as


$t=1$
T
Y_t
Y_{t-1}
Y_{t-2}
A_t
A_{t-1}
C_t
NA
*
NA
*
NA
← fully observed time point $\mathbf{T}_{\text{obs}}^{(0)}$
← missing time point $\mathbf{T}_{\text{mis}}$
partially observed time points $\mathbf{T}_{\text{obs}}^{(1)}$
Response variable
Explanatory variables
(used as F_t in the state space model)

- 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

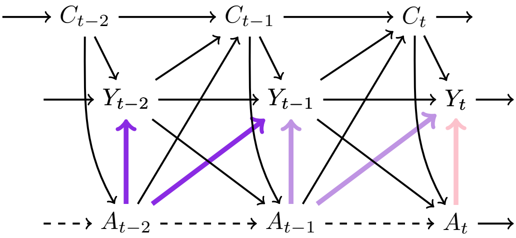
Denote outcome as Y_t, exposure as A_t, and other confounders as C_t. Assume true data generation process as



- 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 $Y_t \to Y_{t-1}$ is included
- Personalized monitoring of a single individual

Problem due to missing data in mHealth

Denote outcome as Y_t, exposure as A_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, ... → 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, ...
- 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 method for missing data in non-stationary multi-variate time series is needed.

State space model and Kalman Filter

State space model is widely used for navigation, location tracking, voice recognition, automotive control system, and parameter estimation for time series analysis in biostatistics, econometric, and many other areas.

Linear state space model

$$Y_t = (1, Y_{t-1}, A_t, A_{t-1}, C_t)^T \beta_t + v_t \quad \text{(Observational equation)}$$

where Y_t and Y_{t-1} are current and past outcomes, (A_t, A_{t-1}) are current and past exposures, C_t is current covariate, and $v_t \sim N(0, V_t)$.

$$\beta_t = G_t \beta_{t-1} + w_t \qquad \text{(State equation)}$$

where β_t denotes unknown coefficients to be estimated, G_t is the transition matrix for how β_t evolves over time, and $w_t \sim N_p(0, W_t)$.

Kalman Filter provides the optimal estimate for the latest (possibly time-varying) unknown parameters, given observations by time t, $\hat{\beta}_t | y_{1:t}$.

Notation and Likelihood

Likelihood-based method with EM algorithm have been extensively applied to handle missing data. Given complete data and omitting constant terms, the log-likelihood for the linear state space model can be written as:

$$
\begin{aligned}
&lnL(\Theta|\mathrm{Y}, \mathrm{A}_{1:T}, \mathrm{C}_{1:T}, \boldsymbol{\theta}_{1:T}) \\
&\quad = \frac{1}{2} ln(\det(\Sigma_0^{-1})) - \frac{1}{2}(\theta_0 - \mu_0)'\Sigma_0^{-1}(\theta_0 - \mu_0) \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T} ln(\det(Q_t^{-1})) - \frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})'Q_t^{-1}(\theta_t - G_t\theta_{t-1}) \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T} ln(\det(R_t^{-1})) - \frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)'R_t^{-1}(y_t - F_t\theta_t)
\end{aligned}
$$

where $(\theta_0, \ldots, \theta_t)$ are hidden states, F_t represents the design matrix, and G_t represents the state transition matrix. $(\mathrm{Y}_{1:T}, \mathrm{A}_{1:T}, \mathrm{C}_{1:T})$ refer to outcomes, exposures, and coavariates used in the F_t , respectively.

MCEM-SSM algorithm: the expectation (E) step

Since expectations in the E step are not directly computable, they are approximated using Monte Carlo simulation via MCMC as derived below:

$$
\begin{aligned}
&\tilde{\mathcal{Q}}\left(\Theta \mid \Theta^{(j-1)}\right) = \frac{1}{2}\ln(\det(\Sigma_0^{-1})) + \frac{T}{2}\ln(\det(Q^{-1})) + \frac{|\mathrm{T}_{obs}|}{2}\ln(\det(R^{-1})) - \frac{1}{2}\operatorname{tr}\left\{\Sigma_0^{-1}\left(\tilde{\mathrm{P}}_0^T + (\tilde{\theta}_0^T - \mu_0)(\tilde{\theta}_0^T - \mu_0)'\right)\right\} \\
&- \frac{1}{2}\operatorname{tr}\Big\{Q^{-1}\Big[\sum_{t=1}^{T}(\tilde{\mathrm{P}}_t^T + \tilde{\theta}_t^T\tilde{\theta}_t^{\prime T}) - \sum_{t=1}^{T}(\tilde{\mathrm{P}}_{t,t-1}^T + \tilde{\theta}_t^T\tilde{\theta}_{t-1}^{\prime T})\Phi' - \sum_{t=1}^{T}\Phi(\tilde{\mathrm{P}}_{t-1,t}^T + \tilde{\theta}_{t-1}^T\tilde{\theta}_t^{\prime T}) + \sum_{t=1}^{T}\Phi(\tilde{\mathrm{P}}_{t-1}^T + \tilde{\theta}_{t-1}^T\tilde{\theta}_{t-1}^{\prime T})\Phi'\Big] \\
&- \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(0)}} R^{-1}(y_t - F_t\tilde{\theta}_t^T)(y_t - F_t\tilde{\theta}_t^T)' + \sum_{t\in T_{obs}^{(0)}} R^{-1}F_t\tilde{\mathrm{P}}_t^T F_t'\Big\} \\
&- \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(1)}} R^{-1}(y_t - F_t^{(0)}\tilde{\theta}_t^{T(0)} - \mathbb{E}[\tilde{F}_t^{(1)}\tilde{\theta}_t^{(1)} \mid \mathcal{O}_{obs}])(y_t - F_t^{(0)}\tilde{\theta}_t^{T(0)} - \mathbb{E}[\tilde{F}_t^{(1)}\tilde{\theta}_t^{(1)} \mid \mathcal{O}_{obs}])'\Big\} \\
&- \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(1)}} R^{-1}F_t^{(0)}\tilde{\mathrm{P}}_t^{T(0)}F_t^{\prime(0)}\Big\} - \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(1)}} R^{-1}\tilde{\mathbb{E}}\left[F_t^{(0)}(\theta_t^{(0)} - \theta_t^{T(0)})(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])' \mid \mathcal{O}_{obs}\right]\Big\} \\
&- \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(1)}} R^{-1}\tilde{\mathbb{E}}\left[(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])(\theta_t^{(0)} - \theta_t^{T(0)})'F_t^{\prime(0)} \mid \mathcal{O}_{obs}\right]\Big\} \\
&- \frac{1}{2}\operatorname{tr}\Big\{\sum_{t\in T_{obs}^{(1)}} R^{-1}\tilde{\mathbb{E}}\left[(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])' \mid \mathcal{O}_{obs}\right]\Big\}.
\end{aligned}
$$

MCEM-SSM algorithm: the maximization (M) step

The M step maximizes the MCMC derived expected log-likelihood $\tilde{\mathcal{Q}}\left(\Theta \mid \Theta^{(j-1)}\right)$ at each iteration j.

$$\mu_0^{(j)} = \tilde{\theta}_0^T$$

$$\Sigma_0^{(j)} = \mathrm{P}_0^T + \left(\tilde{\theta}_0^T - \mu_0\right)\left(\tilde{\theta}_0^T - \mu_0\right)'$$

$$Q^{(j)} = \frac{1}{T}\left[\sum_{t=1}^{T}(\mathrm{P}_t^T + \tilde{\theta}_t^T\tilde{\theta}_t'^T) - \sum_{t=1}^{T}(\mathrm{P}_{t,t-1}^T + \tilde{\theta}_t^T\tilde{\theta}_{t-1}'^T)\Phi' - \sum_{t=1}^{T}\Phi(\mathrm{P}_{t-1,t}^T + \tilde{\theta}_{t-1}^T\tilde{\theta}_t'^T) + \sum_{t=1}^{T}\Phi(\mathrm{P}_{t-1}^T + \tilde{\theta}_{t-1}^T\tilde{\theta}_{t-1}'^T)\Phi'\right]$$

$$\begin{aligned}
R^{(j)} = \frac{1}{T_{obs}}\Bigg[& \sum_{t\in T_{obs}^{(0)}}(y_t - F_t\tilde{\theta}_t^T)(y_t - F_t\tilde{\theta}_t^T)' + \sum_{t\in T_{obs}^{(0)}}F_t\tilde{\mathrm{P}}_t^T F_t' \\
& + \sum_{t\in T_{obs}^{(1)}}(y_t - F_t^{(0)}\tilde{\theta}_t^{T(0)} - \mathbb{E}[\tilde{F}_t^{(1)}\tilde{\theta}_t^{(1)} \mid \mathcal{O}_{obs}])(y_t - F_t^{(0)}\tilde{\theta}_t^{T(0)} - \mathbb{E}[\tilde{F}_t^{(1)}\tilde{\theta}_t^{(1)} \mid \mathcal{O}_{obs}])' \\
& + \sum_{t\in T_{obs}^{(1)}}F_t^{(0)}\tilde{\mathrm{P}}_t^{T(0)}F_t'^{(0)} + \sum_{t\in T_{obs}^{(1)}}\tilde{\mathbb{E}}\left[F_t^{(0)}(\theta_t^{(0)} - \theta_t^{T(0)})(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])' \mid \mathcal{O}_{obs}\right] \\
& + \sum_{t\in T_{obs}^{(1)}}\tilde{\mathbb{E}}\left[(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])(\theta_t^{(0)} - \theta_t^{T(0)})'F_t'^{(0)} \mid \mathcal{O}_{obs}\right] \\
& + \sum_{t\in T_{obs}^{(1)}}\tilde{\mathbb{E}}\left[(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])(F_t^{(1)}\theta_t^{(1)} - \mathbb{E}[F_t^{(1)}\theta_t^{(1)} \mid \mathcal{O}_{obs}])' \mid \mathcal{O}_{obs}\right]\Bigg]
\end{aligned}$$

MCEM-SSM algorithm summary

For each iteration j, repeat the following E step,

1. **Initialize:** Set $m = 0$ and initialize missing lagged outcome $\mathrm{Y}_{mis}^{[m,j]}$ (used in the design matrix $F_t^{[m,j]}$) and increase m by 1.
2. **Sample hidden states:** For each m, sample hidden states $\theta_t^{[m,j]}$ from the conditional distribution $f(\theta_t|\mathcal{O}_{obs}, F_t^{(1),[m-1,j]})$ for $t = 0, 1, \ldots, T$, using forward filtering backward sampling (BFBS).
3. **Sample missing outcome:** Sample missing lagged outcomes $\mathrm{Y}_{mis}^{[m,j]}$ (used in $F_t^{(1),[m,j]}$) from the conditional distribution $f(\mathrm{Y}_{mis}|\mathcal{O}_{obs}, \boldsymbol{\theta}_{0:T}^{[m,j]})$ and increase m by 1.
4. **Iterate:** Repeat steps 2 and 3 until $m = M$ to obtain MCMC dependent samples of missing outcomes and hidden states, $\{\mathrm{Y}_{mis}, \boldsymbol{\theta}_{1:T}|\mathcal{O}_{obs}\}$, to approximate expectations in $\mathcal{Q}\left(\Theta \mid \Theta^{(j-1)}\right)$.

and M step as,

1. **Maximize:** Maximizes the MCMC derived $\tilde{\mathcal{Q}}\left(\Theta \mid \Theta^{(j-1)}\right)$ at iteration j.

until convergence.

Simulation

Simulation Scenario

- Stationary time series with time-invariate coefficients
- Non-stationary time series with multiple sources of non-stationarity
  - time-varying baseline as a random walk
  - time-varying treatment effect as a periodic-stable process

Methods to be compared:

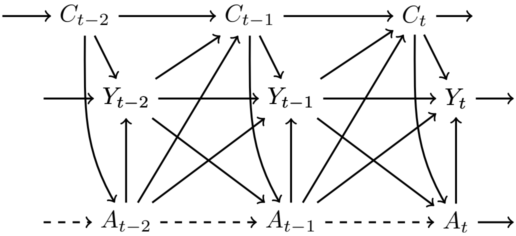
- complete case analysis, mean imputations, LOCF imputations, linear imputations, spline interpolation, multiple imputation
- Proposed MCEM-SSM

Missing mechanism:

- MCAR, MAR, MNAR

See additional more complex scenarios with details in papers.

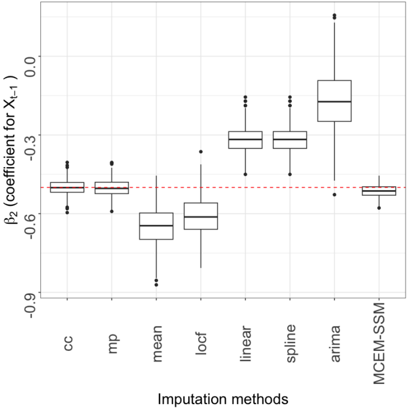
Simulation: stationary time series

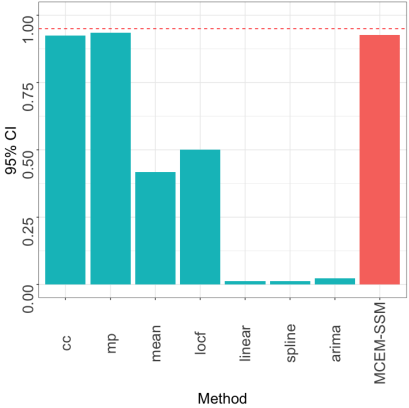


$$Y_t = \beta_0 + \rho Y_{t-1} + \beta_1 A_t + \beta_2 A_{t-1} + \beta_c C_t + v_t, \quad v_t \sim N(0, V)$$

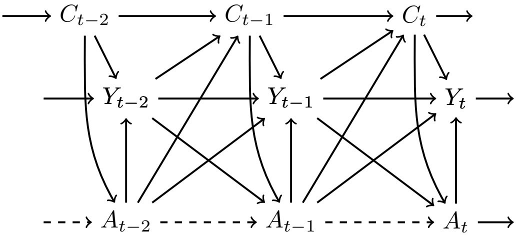
where $\beta_0 = 40$, $\rho = 0.5$, $\beta_1 = -1$, $\beta_2 = -0.5$, and $\beta_c = -1$.

Simulation: estimated $\hat{\beta}_{2,t}$ for stationary time series





Simulation: non-stationary with change points and random walk

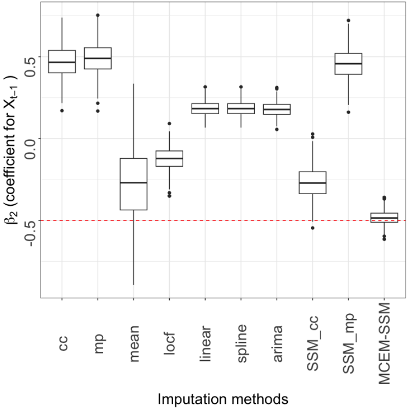


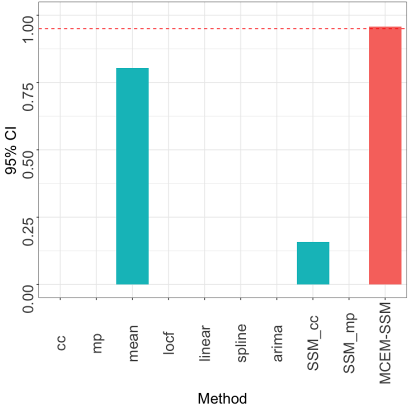
$$Y_t = \beta_{0,t} + \rho Y_{t-1} + \beta_{1,t} A_t + \beta_2 A_{t-1} + \beta_c C_t + v_t, \quad v_t \sim N(0, V)$$

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 $\hat{\beta}_{2,t}$ for non-stationary time series





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

For stationary time series,

- Mean imputation, LOCF, linear and spline imputations are significantly biased.
- Complete case analysis, multiple imputation, and MCEM-SSM achieve satisfactory results.

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 significantly biased.
- MCEM-SSM achieve satisfactory results in coefficient estimation.

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 ($A_{1,t}$) and texts ($A_{2,t}$)
- Covariates: temperature ($C_{\text{temp},t}$), physical mobility ($C_{\text{pm},t}$)

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

Missing rate before imputation → 40.4%
Missing rate after imputation → 23.3%

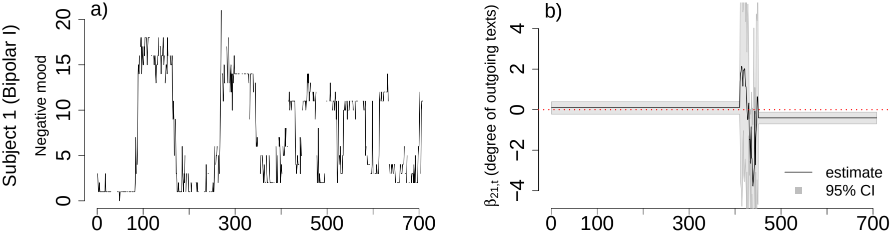
Estimation result

$$Y_t = \text{intercept}_t + \rho_t Y_{t-1} + \beta_{11,t} A_{\text{calls},t} + \beta_{12,t} A_{\text{calls},t-1} + \beta_{21,t} A_{\text{texts},t} + \beta_{22,t} A_{\text{texts},t-1} + \beta_{temp,t} C_{\text{temp},t} + \beta_{pm,t} C_{\text{pm},t} + v_t$$

Estimate(SE)	Complete Case	Multiple Imputation	MCEM-SSM
$\beta_{0,t}$	(random walk)	(random walk)	(random walk)
ρ_t	0.71(0.03)**	0.84(0.05)**	0.42(0.04)**
$\beta_{11,t}$	-0.17(0.08)*	-0.13(0.07)†	-0.13(0.06)*
$\beta_{12,t}$	-0.08(0.07)	-0.08(0.07)	-0.02(0.07)
$\beta_{21,t}(1)$	0.04(0.14)	-0.05(0.15)	0.11(0.17)
$\beta_{21,t}(2)$	-0.41(0.17)*	-0.23(0.19)	-0.41(0.15)**
$\beta_{22,t}$	-0.20(0.11)†	-0.23(0.11)*	-0.19(0.09)**
$\beta_{pm,t}(1)$	-9.43(4.26)*	-4.62(3.27)	-11.11(3.81)**
$\beta_{pm,t}(2)$	1.13(1.6)	0.45(1.21)	2.71(1.97)
$\beta_{temp,t}$	0.00(0.01)	0.00(0.01)	-0.01(0.01)

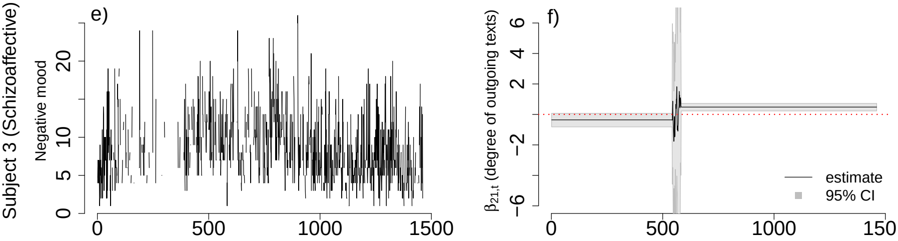
† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$

Estimation result: estimated coefficient for degree of outgoing texts



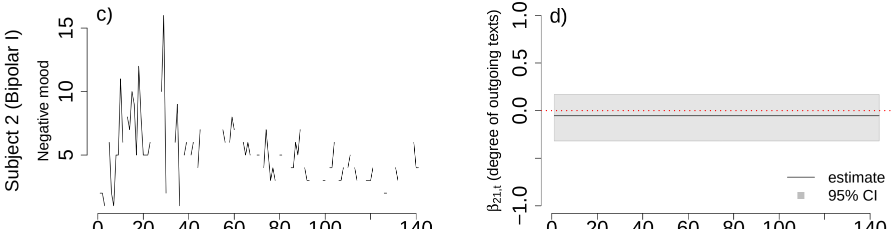
Self-reported negative moods (left) and estimated $\beta_{2,t}$ (right) in BLS.

Estimation result: estimated coefficient for degree of outgoing texts



Self-reported negative moods (left) and estimated $\beta_{2,t}$ (right) in BLS.

Estimation result: estimated coefficient for degree of outgoing texts



Self-reported negative moods (left) and estimated $\beta_{2,t}$ (right) in BLS.

Summary

- Existing imputation methods mostly assume stationarity and induce significant bias in coefficient estimation for non-stationary case.
- We proposed a Monte Carlo Expectation-Maximization algorithm of the state space model (MCEM-SSM) to effectively handle missing data in non-stationary entangled multivariate time series.
- The current model rely on correct model specification.

We have another work using multiple imputation to address this issue. Xiaoxuan Cai, et al. State space model multiple imputation for missing data in non-stationary multivariate time series. (https://arxiv.org/abs/2206.14343).

Limitation and ongoing work

- Explore more flexible models, such as nonlinear state-space frameworks, to account for potential model misspecification.
- Extend to Counts or Ordinary response variables.
- Extend to missingness in exposures and confounders as well.
- Incorporate explicit model of missingness mechanism to handle MNAR scenarios.
- Explore latent disease for conducting group inferences for participants with heterogeneous severe mental diseases.

Acknowledgement

Research was supported by K01MH118477 grant. R packages are under development for MCEM-SSM for MCAR/MAR.

We thank Justin Baker, Zilan Chai, Charly Fowler, Melanie Mayer, Jukka-Pekka Onnela, Habib Rahimi, Junzhe Shao, Aijin Wang, Zixu Wang, and Li Zeng for their great help and suggestions.

cai.1083@osu.edu
https://xiaoxuan-cai.github.io/

Thank you!