# A state space imputation approach for missing data in non-stationary multivariate time series

Xiaoxuan Cai

joint work with Xinru Wang, Linda Valeri

Department of Biostatistics, Mailman School of Public Health
Columbia University

August 12, 2021
Joint Statistical Meetings 2021

# 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

The integration and translation of these cutting-edge technologies into rigorously evaluated health research have lagged behind.

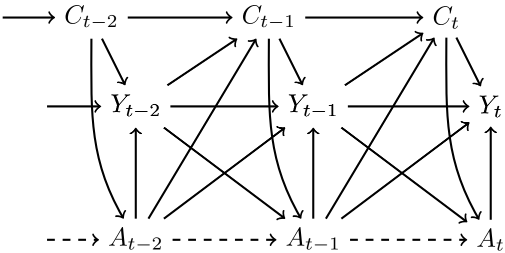
- Effect evaluation for dynamic exposure or intervention, mediation analysis, personalized treatment optimization, prediction, adaptive trail designs, ...

### Our research of interest

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

# Causal structure

- Outcome ($Y_t$): self-reported negative mood of the patient
- Exposure ($A_t$): social support (e.g., degree of calls and texts)
- Confounders ($C_t$): physical activity, medication, temperature, ...

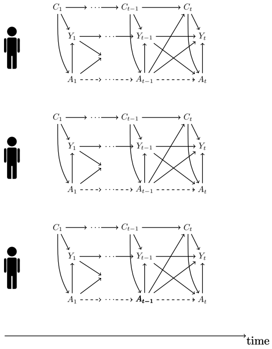


Causal effect: $\mathbb{E}[Y_t(A_t = 1)] - \mathbb{E}[Y_t(A_t = 0)]$

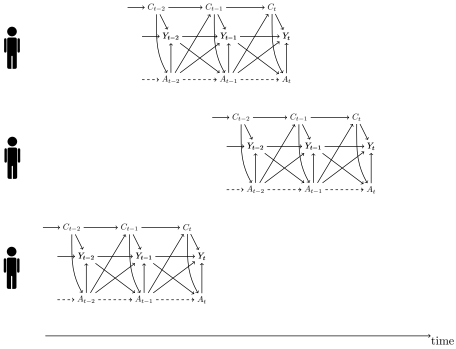
How to handle missing data for non-stationary time series?

# Observational N-of-1 study data

Mental health research studies heterogeneous patients, which follows up with a particular patient throughout the entire observation as in a N-of-1 study.



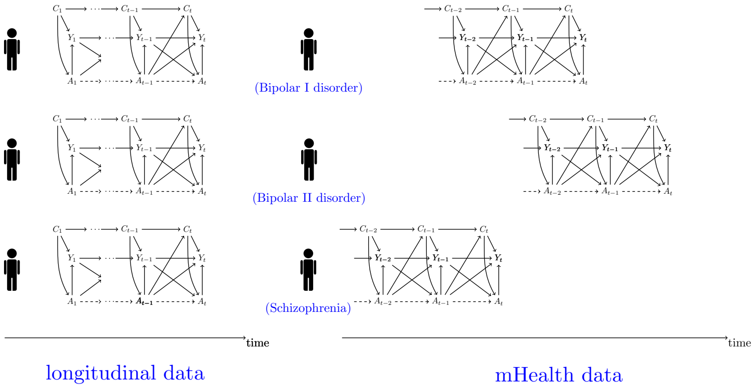
longitudinal data



mHealth data

# Observational N-of-1 study data

Mental health research studies heterogeneous patients, which follows up with a particular patient throughout the entire observation as in a N-of-1 study.



# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.

Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
. . .

Univariate time series

simple moving average
exponential weighted moving average
ARIMA model
. . .

Multivariate time series

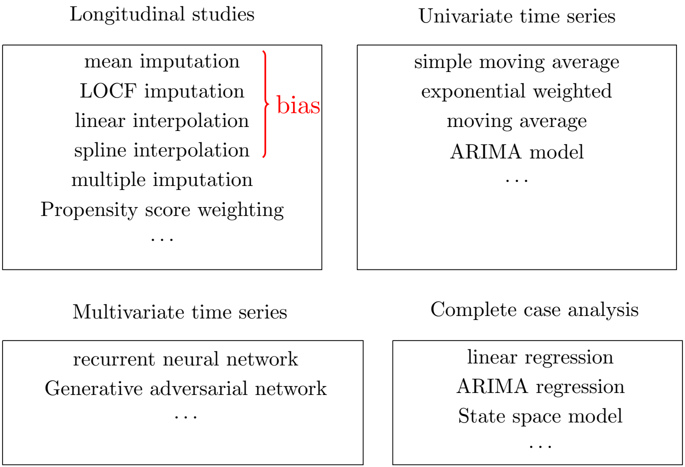
recurrent neural network
Generative adversarial network
. . .

Complete case analysis

linear regression
ARIMA regression
State space model
. . .

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.



Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
...

bias

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
...

Multivariate time series

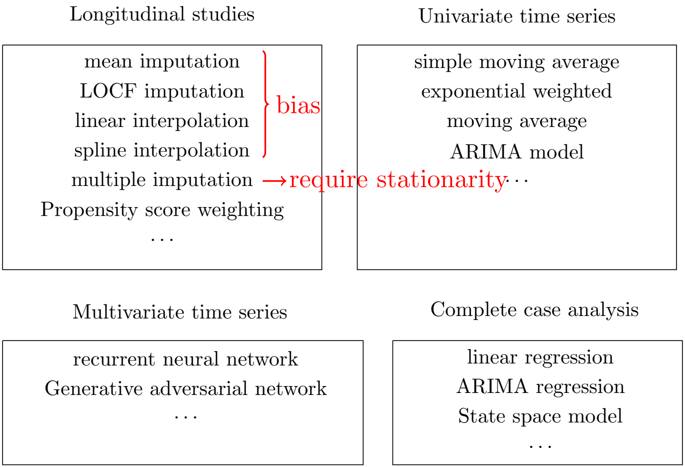
recurrent neural network
Generative adversarial network
...

Complete case analysis

linear regression
ARIMA regression
State space model
...

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.



Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
} bias
multiple imputation →require stationarity
Propensity score weighting
. . .

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
· · ·

Multivariate time series

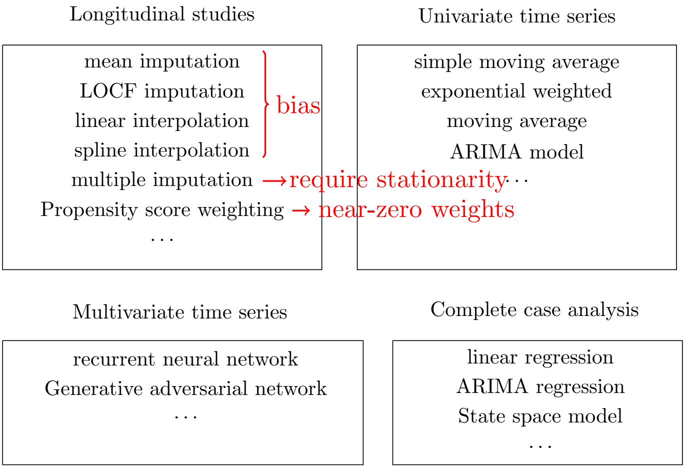
recurrent neural network
Generative adversarial network
. . .

Complete case analysis

linear regression
ARIMA regression
State space model
. . .

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.



Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
bias
multiple imputation → require stationarity
Propensity score weighting → near-zero weights
. . .

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
· · ·

Multivariate time series

recurrent neural network
Generative adversarial network
. . .

Complete case analysis

linear regression
ARIMA regression
State space model
. . .

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.

Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
. . .

Univariate time series

simple moving average
exponential weighted moving average
ARIMA model
. . .
(not for multivariate time series)

Multivariate time series

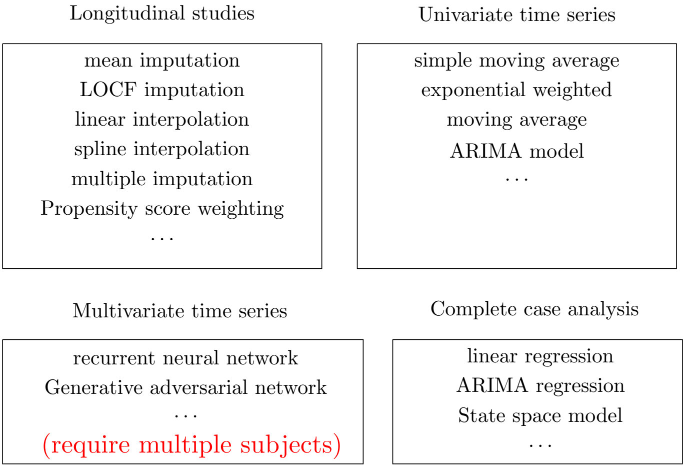
recurrent neural network
Generative adversarial network
. . .

Complete case analysis

linear regression
ARIMA regression
State space model
. . .

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.



Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
...

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
...

Multivariate time series

recurrent neural network
Generative adversarial network
...
(require multiple subjects)

Complete case analysis

linear regression
ARIMA regression
State space model
...

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.

Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
. . .

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
. . .

Multivariate time series

recurrent neural network
Generative adversarial network
. . .

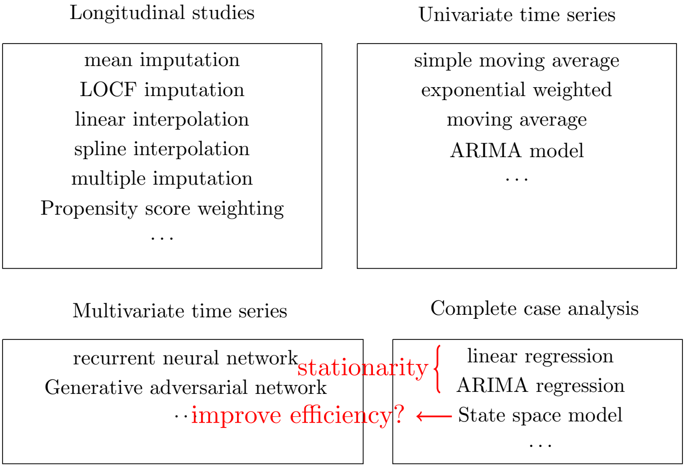
Complete case analysis

stationarity {
linear regression
ARIMA regression

State space model
. . .

# Existing methods in handling missing data

A variety of statistical methods have been developed to tackle missing data problem in various settings.



Longitudinal studies

mean imputation
LOCF imputation
linear interpolation
spline interpolation
multiple imputation
Propensity score weighting
. . .

Univariate time series

simple moving average
exponential weighted
moving average
ARIMA model
. . .

Multivariate time series

recurrent neural network
Generative adversarial network
· ·

Complete case analysis

linear regression
ARIMA regression
State space model
. . .

stationarity

improve efficiency? ⟵

# 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.

## 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 $Y_t$ is the observed outcome, $(A_t, A_{t-1}, C_t)$ are observed exposure and covariates, and $v_t \sim N(0, V_t)$.

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

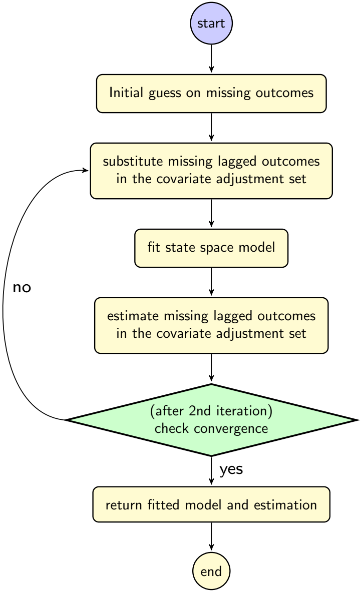
where $\beta_t$ denotes unknown coefficients to be estimated, $G_t$ is the transition matrix for how $\beta_t$ evolves over time, and $w_t \sim N_p(0, W_t)$.

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

# State space model imputation (SSMimpute)

Formula: $Y_t \sim A_t + A_{t-1} + C_t + Y_{t-1}$

| | | | | |
|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | NA |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | NA |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ |



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes in the covariate adjustment set

(after 2nd iteration) check convergence

no

yes

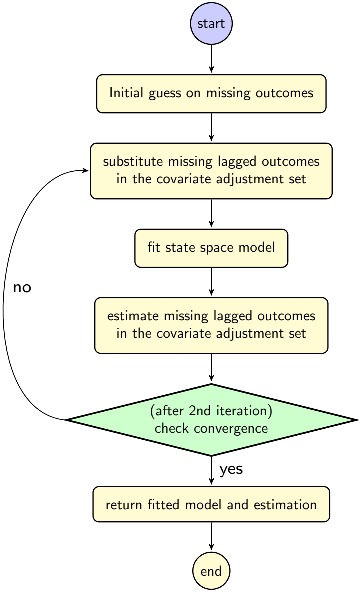
return fitted model and estimation

end

# State space model imputation (SSMimpute)

Formula: $Y_t \sim A_t + A_{t-1} + C_t + Y_{t-1}$

| | | | | |
|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | NA |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | NA |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ |



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes in the covariate adjustment set

(after 2nd iteration) check convergence

no

yes

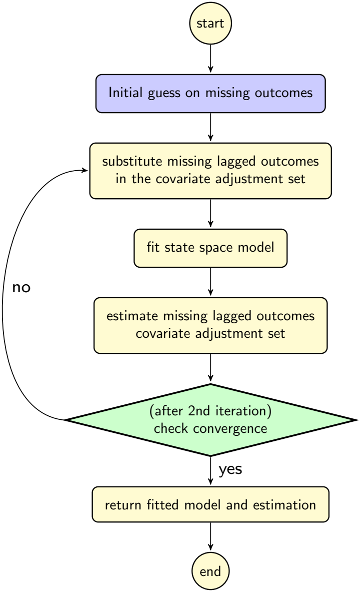
return fitted model and estimation

end

# State space model imputation (SSMimpute)

Formula: $Y_t \sim \quad A_t + A_{t-1} + C_t + Y_{t-1}$ initial guess of $Y_t$

$$\begin{array}{ccccc|c}
y_1 & a_1 & a_0 & c_0 & y_0 & y_1 \\
y_2 & a_2 & a_1 & c_1 & y_1 & y_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathrm{NA} & a_{t-1} & a_{t-2} & c_{t-2} & y_{t-2} & \tilde{y}_{t-1}^{(0)} \\
y_t & a_t & a_{t-1} & c_{t-1} & \mathrm{NA} & y_t \\
\mathrm{NA} & a_{t+1} & a_t & c_t & y_t & \tilde{y}_{t+1}^{(0)} \\
y_{t+2} & a_{t+2} & a_{t+1} & c_{t+1} & \mathrm{NA} & y_{t+2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_T & a_T & a_{T-1} & c_{T-1} & y_{T-1} & y_T
\end{array}$$



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes covariate adjustment set

(after 2nd iteration) check convergence

no

yes

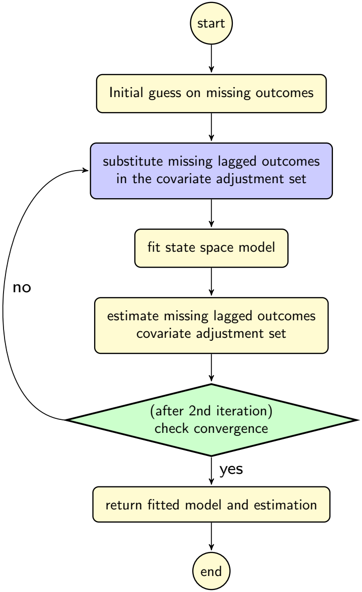
return fitted model and estimation

end

# State space model imputation (SSMimpute)

Formula:

| $Y_t$ $\sim$ | $A_t +$ | $A_{t-1}+$ | $C_t$ | $+Y_{t-1}$ | initial guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}_{t-1}^{(0)}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}_{t-1}^{(0)}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}_{t+1}^{(0)}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}_{t+1}^{(0)}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |


start
Initial guess on missing outcomes
substitute missing lagged outcomes in the covariate adjustment set
fit state space model
estimate missing lagged outcomes covariate adjustment set
(after 2nd iteration) check convergence
no
yes
return fitted model and estimation
end

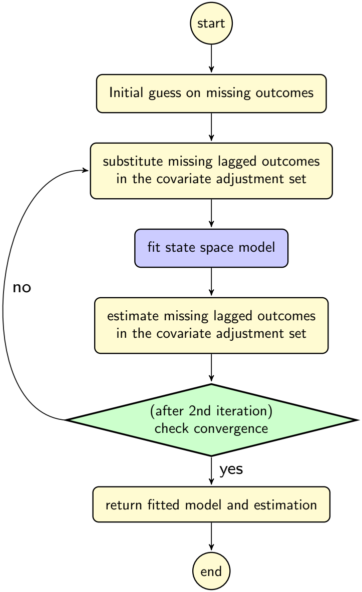
# State space model imputation (SSMimpute)

Formula: $Y_t \sim A_t + A_{t-1} + C_t + Y_{t-1}$ initial guess of $Y_t$

$$\begin{array}{|cccccc|c}
\hline
y_1 & a_1 & a_0 & c_0 & y_0 & & y_1 \\
y_2 & a_2 & a_1 & c_1 & y_1 & & y_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
\mathrm{NA} & a_{t-1} & a_{t-2} & c_{t-2} & y_{t-2} & & \tilde{y}_{t-1}^{(0)} \\
y_t & a_t & a_{t-1} & c_{t-1} & \tilde{y}_{t-1}^{(0)} & & y_t \\
\mathrm{NA} & a_{t+1} & a_t & c_t & y_t & & \tilde{y}_{t+1}^{(0)} \\
y_{t+2} & a_{t+2} & a_{t+1} & c_{t+1} & \tilde{y}_{t+1}^{(0)} & & y_{t+2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
y_T & a_T & a_{T-1} & c_{T-1} & y_{T-1} & & y_T \\
\hline
\end{array}$$

$$\Downarrow$$

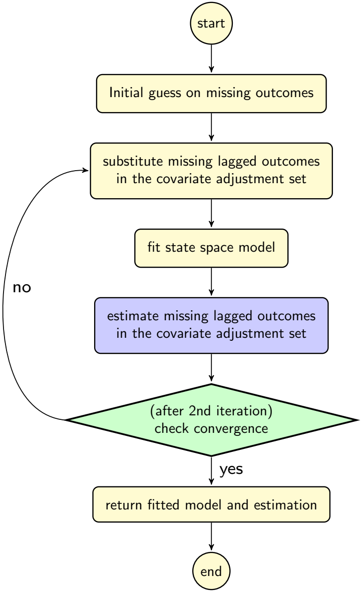
$$\boxed{\hat{\beta}_{0,t}^{(1)} \quad \hat{\beta}_{1,t}^{(1)} \quad \hat{\beta}_{2,t}^{(1)} \quad \hat{\beta}_{c,t}^{(1)} \quad \hat{\rho}_t^{(1)}}$$


start
Initial guess on missing outcomes
substitute missing lagged outcomes in the covariate adjustment set
fit state space model
estimate missing lagged outcomes in the covariate adjustment set
(after 2nd iteration) check convergence
no
yes
return fitted model and estimation
end

# State space model imputation (SSMimpute)

Formula: $Y_t \sim A_t + A_{t-1} + C_t + Y_{t-1}$ $1^{st}$ guess of $Y_t$

$$\left.\begin{array}{c}\boxed{\begin{array}{ccccc} y_1 & a_1 & a_0 & c_0 & y_0\\ y_2 & a_2 & a_1 & c_1 & y_1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathrm{NA} & a_{t-1} & a_{t-2} & c_{t-2} & y_{t-2}\\ y_t & a_t & a_{t-1} & c_{t-1} & \tilde{y}^{(0)}_{t-1}\\ \mathrm{NA} & a_{t+1} & a_t & c_t & y_t\\ y_{t+2} & a_{t+2} & a_{t+1} & c_{t+1} & \tilde{y}^{(0)}_{t+1}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ y_T & a_T & a_{T-1} & c_{T-1} & y_{T-1}\end{array}}\\ +\\ \boxed{\begin{array}{ccccc}\hat{\beta}^{(1)}_{0,t} & \hat{\beta}^{(1)}_{1,t} & \hat{\beta}^{(1)}_{2,t} & \hat{\beta}^{(1)}_{c,t} & \hat{\rho}^{(1)}_t\end{array}}\end{array}\right\} \Rightarrow \begin{array}{c} y_1\\ y_2\\ \vdots\\ \tilde{y}^{(1)}_{t-1}\\ y_t\\ \tilde{y}^{(1)}_{t+1}\\ y_{t+2}\\ \vdots\\ y_T\end{array}$$



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes in the covariate adjustment set

(after 2nd iteration) check convergence

no

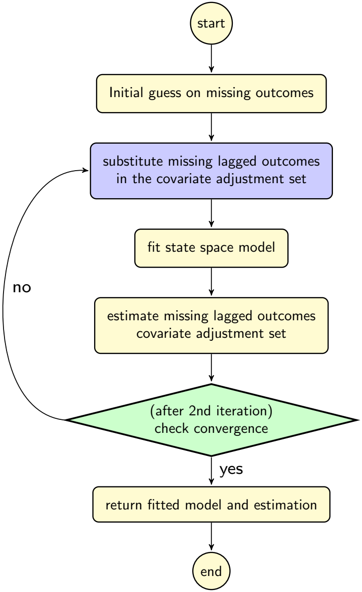
yes

return fitted model and estimation

end

# State space model imputation (SSMimpute)

| Formula: $Y_t$ ~ | $A_t$ + | $A_{t-1}$ + | $C_t$ | $+Y_{t-1}$ | $1^{st}$ guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}_{t-1}^{(1)}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}_{t-1}^{(1)}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}_{t+1}^{(1)}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}_{t+1}^{(1)}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes covariate adjustment set

(after 2nd iteration) check convergence

no

yes

return fitted model and estimation

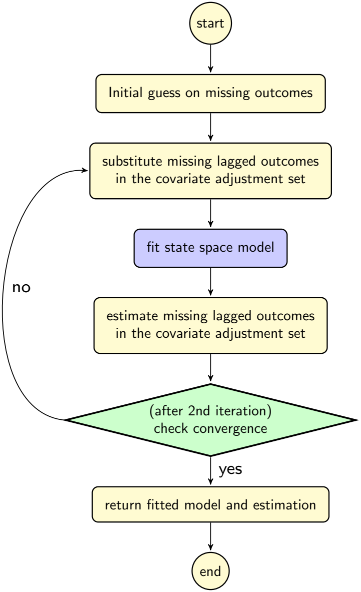
end

# State space model imputation (SSMimpute)

| Formula: $Y_t \sim$ | $A_t +$ | $A_{t-1} +$ | $C_t$ | $+Y_{t-1}$ | $1^{st}$ guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}_{t-1}^{(1)}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}_{t-1}^{(1)}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}_{t+1}^{(1)}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}_{t+1}^{(1)}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |

⬇

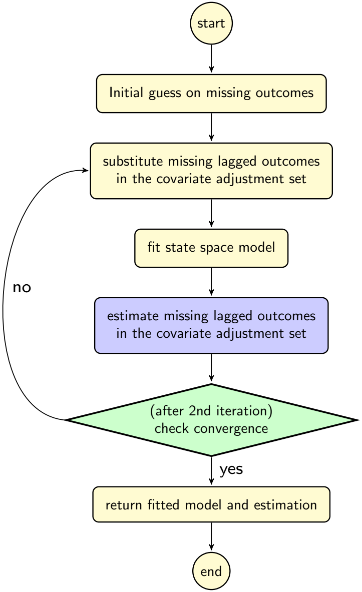
$\hat{\beta}_{0,t}^{(2)}$ $\hat{\beta}_{1,t}^{(2)}$ $\hat{\beta}_{2,t}^{(2)}$ $\hat{\beta}_{c,t}^{(2)}$ $\hat{\rho}_t^{(2)}$


start
Initial guess on missing outcomes
substitute missing lagged outcomes in the covariate adjustment set
fit state space model
estimate missing lagged outcomes in the covariate adjustment set
(after 2nd iteration) check convergence
no
yes
return fitted model and estimation
end

# State space model imputation (SSMimpute)

Formula: $Y_t \sim \quad A_t + A_{t-1} + \; C_t \; + Y_{t-1}$ $\quad 2^{nd}$ guess of $Y_t$

$$\left.\begin{array}{c}\boxed{\begin{array}{ccccc} y_1 & a_1 & a_0 & c_0 & y_0 \\ y_2 & a_2 & a_1 & c_1 & y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{NA} & a_{t-1} & a_{t-2} & c_{t-2} & y_{t-2} \\ y_t & a_t & a_{t-1} & c_{t-1} & \tilde{y}^{(1)}_{t-1} \\ \text{NA} & a_{t+1} & a_t & c_t & y_t \\ y_{t+2} & a_{t+2} & a_{t+1} & c_{t+1} & \tilde{y}^{(1)}_{t+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_T & a_T & a_{T-1} & c_{T-1} & y_{T-1} \end{array}} \\ + \\ \boxed{\begin{array}{ccccc} \hat{\beta}^{(2)}_{0,t} & \hat{\beta}^{(2)}_{1,t} & \hat{\beta}^{(2)}_{2,t} & \hat{\beta}^{(2)}_{c,t} & \hat{\rho}^{(2)}_t \end{array}} \end{array}\right\} \Rightarrow \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ \tilde{y}^{(2)}_{t-1} \\ y_t \\ \tilde{y}^{(2)}_{t+1} \\ y_{t+2} \\ \vdots \\ y_T \end{array}$$

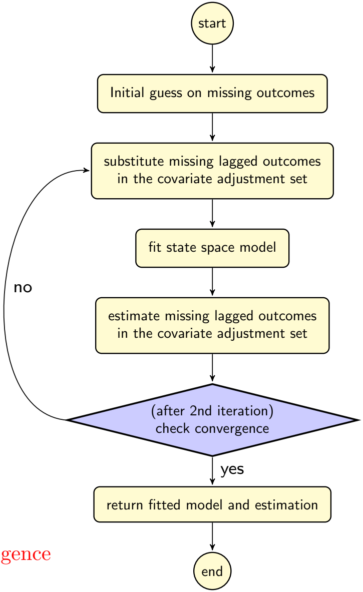


start
Initial guess on missing outcomes
substitute missing lagged outcomes in the covariate adjustment set
fit state space model
estimate missing lagged outcomes in the covariate adjustment set
(after 2nd iteration) check convergence
no
yes
return fitted model and estimation
end

# State space model imputation (SSMimpute)

| Formula: $Y_t \sim$ | $A_t +$ | $A_{t-1}+$ | $C_t$ | $+Y_{t-1}$ | $2^{nd}$ guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}^{(2)}_{t-1}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}^{(2)}_{t-1}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}^{(1)}_{t+1}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}^{(2)}_{t+1}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |

$$\left.\begin{array}{c}\boxed{\hat{\beta}^{(1)}_{0,t} \quad \hat{\beta}^{(2)}_{1,t} \quad \hat{\beta}^{(2)}_{2,t} \quad \hat{\beta}^{(2)}_{c,t} \quad \hat{\rho}^{(2)}_{t}} \\ + \\ \text{likelihood and other unknown parameters}\end{array}\right\} \text{check convergence}$$



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes in the covariate adjustment set

(after 2nd iteration) check convergence

no

yes

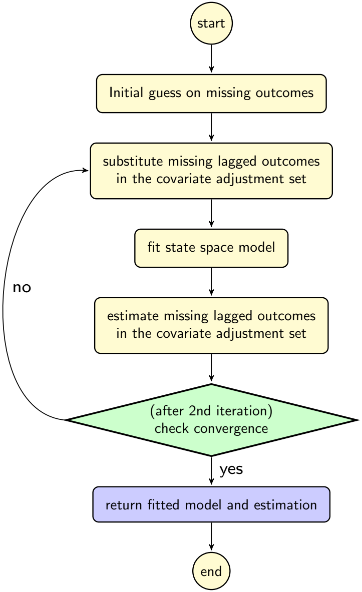
return fitted model and estimation

end

# State space model imputation (SSMimpute)

| Formula: $Y_t \sim$ | $A_t +$ | $A_{t-1} +$ | $C_t$ | $+Y_{t-1}$ | $3^{rd}$ guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}_{t-1}^{(3)}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}_{t-1}^{(2)}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}_{t+1}^{(3)}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}_{t+1}^{(2)}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |

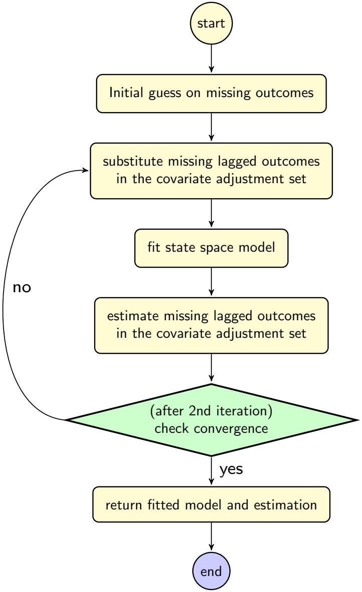
$$\boxed{\hat{\beta}_{0,t}^{(3)} \quad \hat{\beta}_{1,t}^{(3)} \quad \hat{\beta}_{2,t}^{(3)} \quad \hat{\beta}_{c,t}^{(3)} \quad \hat{\rho}_t^{(3)}}$$


start
Initial guess on missing outcomes
substitute missing lagged outcomes in the covariate adjustment set
fit state space model
estimate missing lagged outcomes in the covariate adjustment set
(after 2nd iteration) check convergence
no
yes
return fitted model and estimation
end

# State space model imputation (SSMimpute)

| Formula: $Y_t$ ~ | $A_t$ + | $A_{t-1}$ + | $C_t$ + | $Y_{t-1}$ | $3^{rd}$ guess of $Y_t$ |
|---|---|---|---|---|---|
| $y_1$ | $a_1$ | $a_0$ | $c_0$ | $y_0$ | $y_1$ |
| $y_2$ | $a_2$ | $a_1$ | $c_1$ | $y_1$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| NA | $a_{t-1}$ | $a_{t-2}$ | $c_{t-2}$ | $y_{t-2}$ | $\tilde{y}^{(3)}_{t-1}$ |
| $y_t$ | $a_t$ | $a_{t-1}$ | $c_{t-1}$ | $\tilde{y}^{(2)}_{t-1}$ | $y_t$ |
| NA | $a_{t+1}$ | $a_t$ | $c_t$ | $y_t$ | $\tilde{y}^{(3)}_{t+1}$ |
| $y_{t+2}$ | $a_{t+2}$ | $a_{t+1}$ | $c_{t+1}$ | $\tilde{y}^{(2)}_{t+1}$ | $y_{t+2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_T$ | $a_T$ | $a_{T-1}$ | $c_{T-1}$ | $y_{T-1}$ | $y_T$ |

$\hat{\beta}^{(3)}_{0,t}$ $\hat{\beta}^{(3)}_{1,t}$ $\hat{\beta}^{(3)}_{2,t}$ $\hat{\beta}^{(3)}_{c,t}$ $\hat{\rho}^{(3)}_{t}$



start

Initial guess on missing outcomes

substitute missing lagged outcomes in the covariate adjustment set

fit state space model

estimate missing lagged outcomes in the covariate adjustment set

(after 2nd iteration) check convergence

no

yes

return fitted model and estimation

end

# Simulation

- Stationary time series
- Non-stationary time series with varying variance
- Non-stationary time series with random-walk coefficients
- Non-stationary time series with coefficient having change points
- Non-stationary time series with multiple sources of non-stationarity

Methods to be compared:

- complete case analysis under both linear regression and state space model
- mean imputations, LOCF imputations, linear imputations, spline interpolation, multiple imputation under both linear regression and state space model
- Proposed state space model imputation (SSMimpute)

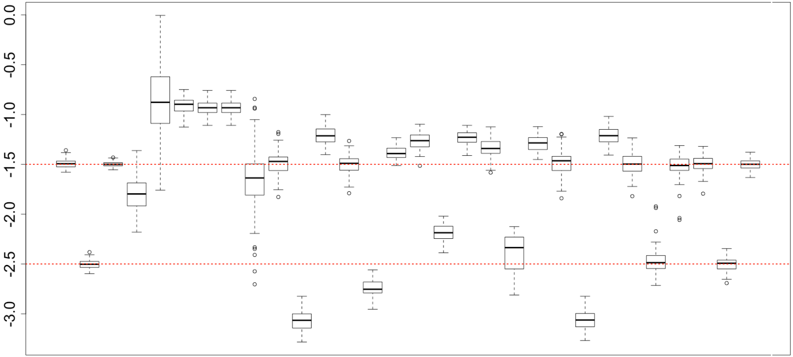
Missing mechanism:

- MCAR, MAR, MNAR

# Simulation: non-stationary with sudden change points

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

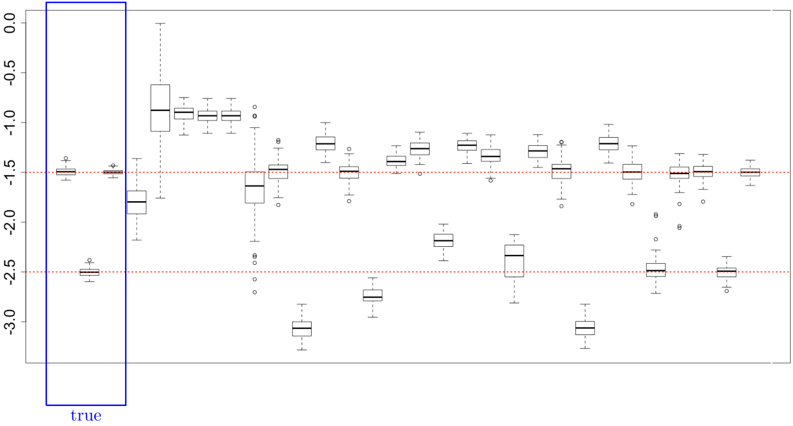
where $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$, and $t = 1, \ldots, 1000$. Missing rate is 50% under MCAR.



# Simulation: non-stationary with sudden change points

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

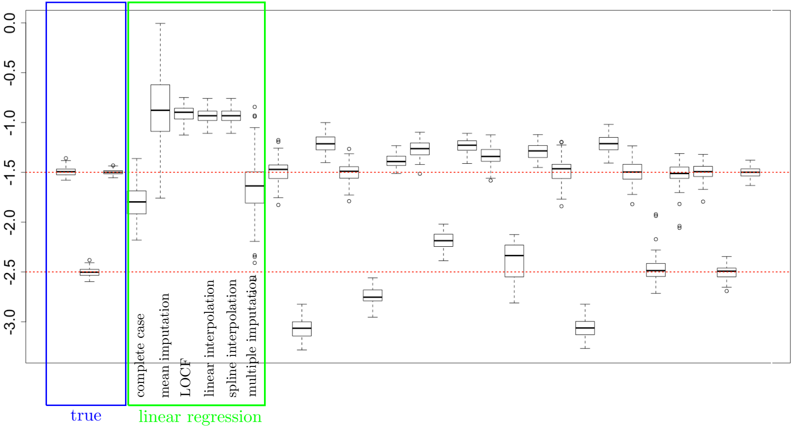
where $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$, and $t = 1, \ldots, 1000$. Missing rate is 50% under MCAR.



# Simulation: non-stationary with sudden change points

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

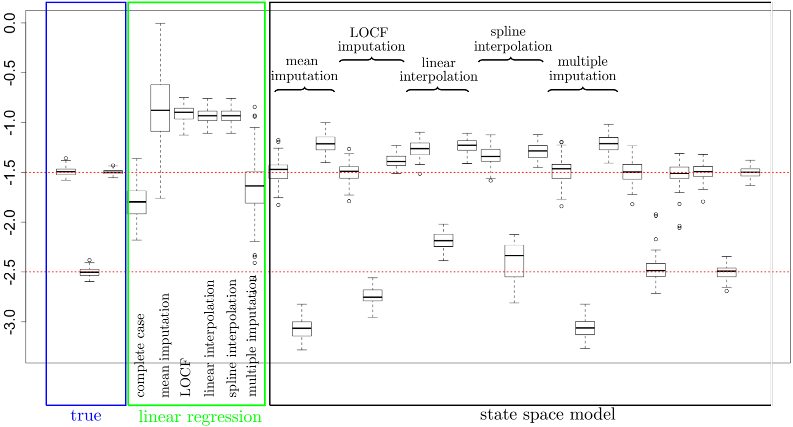
where $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$, and $t = 1, \ldots, 1000$. Missing rate is 50% under MCAR.



# Simulation: non-stationary with sudden change points

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

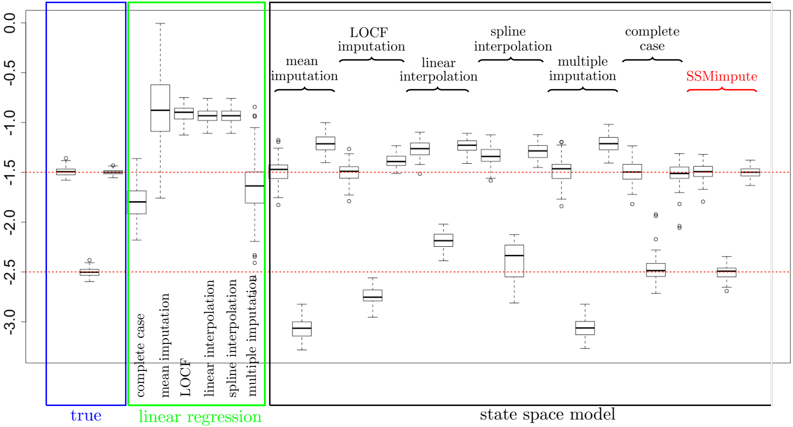
where $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$, and $t = 1, \ldots, 1000$. Missing rate is 50% under MCAR.



# Simulation: non-stationary with sudden change points

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

where $\beta_{1,t} = -1.5$ for $t = 1, \ldots, 400, 701, \ldots, 1000$ and $\beta_{1,t} = -2.5$ for $t = 401, \ldots, 700$, and $t = 1, \ldots, 1000$. Missing rate is 50% under MCAR.



# Conclusions: (see more results in the paper)

For stationary time series,

- Complete case using linear regression and state space model are equivalent in theory and unbiased.
- Multiple imputation and SSMimpute are unbiased and more efficient than complete case analysis.
- Mean imputation, LOCF, linear and spline imputations are all biased.

For non-stationary time series,

- Linear model is unable to handle time-varying coefficients, and thus all imputation methods based on linear model are biased.
- State space model handles time-varying coefficients and variance, and provides unbiased estimation under complete case analysis and SSMimpute. SSMimpute is more efficient than complete case analysis.
- Mean imputation, LOCF, linear and spline interpolation, and multiple imputation are all biased even using state space model.

# Bipolar Longitudinal Study (McLean Hospital)

The study explores how passive sensor data linked to moods and cognitive states in patients with Serious Mental Illness (SMI).
Following data is collected for 10 (out of 54) participants for up to 4 years.

- Self-reported survey data (moods, activities, life habits, in-person interactions, etc.)
- Passively collected telecommunication data (calls and texts)
- Passively collected accelerometer data
- ...

Bewei app

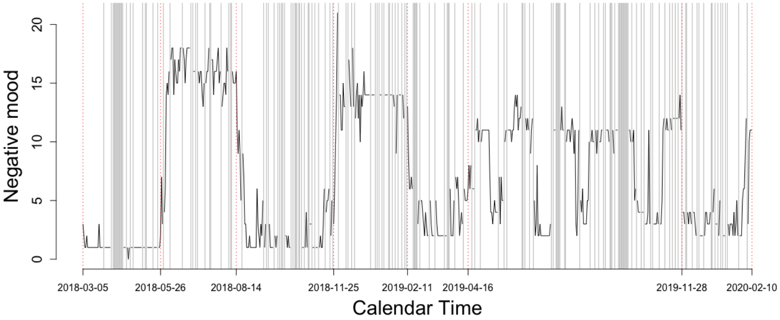
GEVEActiv watch

Samsung Galaxy Note 8

# Bipolar Longitudinal Study (Outcome $Y_t$)

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

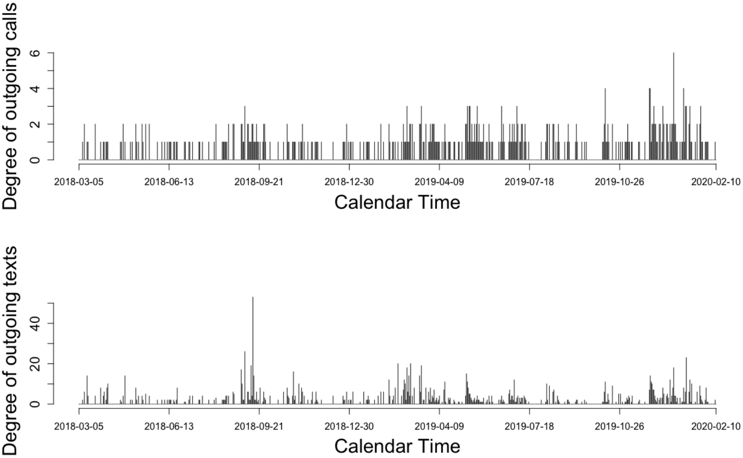
Outcome is a composite index for negative mood of being afraid, anxious, ashamed, hostile, stressed, upset, irritable and lonely.



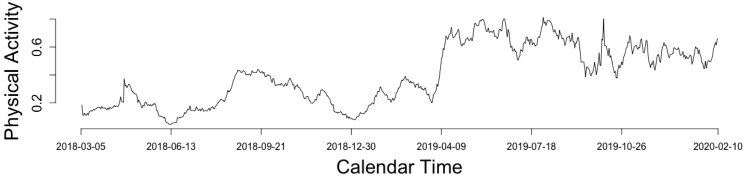
Longest consecutive missing days: 11

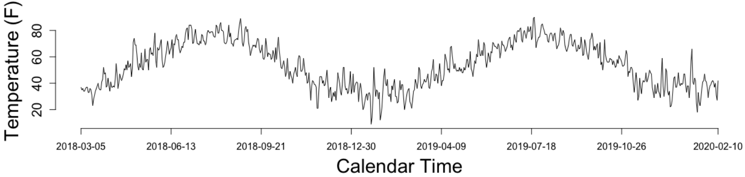
Missing rates: 23.31%

# Bipolar Longitudinal Study (Telecommunication $A_t$)



# Bipolar Longitudinal Study (Other covariates $C_t$)





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

# Estimation model

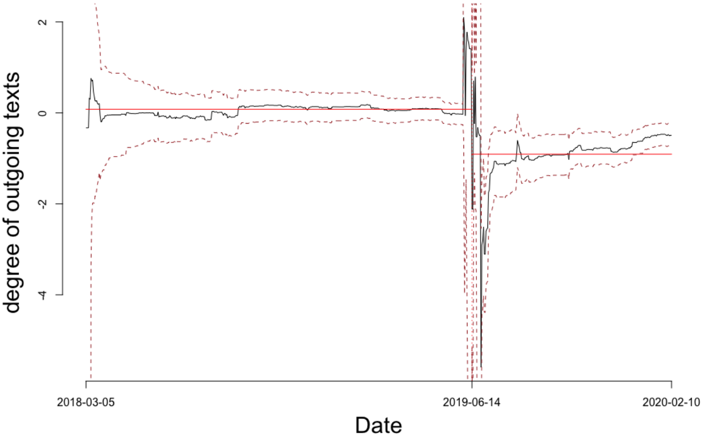
We estimate the association between the degree of outgoing calls and texts and the negative mood, controlling physical activity and temperature.

- Outcome: negative mood ($Y_t$)
- Exposures: degree of outgoing calls ($A_{1,t}$) and outgoing texts ($A_{2,t}$)
- Covariates: temperature ($\mathrm{Temp}_t$), past physical activity ($\mathrm{PA}_t$)

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

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

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

| | SSMimpute (n=542) | | multiple imputation (n=542) | |
|---|---|---|---|---|
| | Estimate | 90% CI | Estimate | 90% CI |
| $\text{intercept}_t$ | (random walk) | | (random walk) | |
| $\rho_t$ (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

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

| | SSMimpute (n=542) | | multiple imputation (n=542) | |
|---|---|---|---|---|
| | Estimate | 90% CI | Estimate | 90% CI |
| $\text{intercept}_t$ | (random walk) | | (random walk) | |
| $\rho_t$ (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

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

| | SSMimpute (n=542) | | multiple imputation (n=542) | |
|---|---|---|---|---|
| | Estimate | 90% CI | Estimate | 90% CI |
| $\text{intercept}_t$ | (random walk) | | (random walk) | |
| $\rho_t$ (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) |

# Conclusions

- Estimated coefficient of the intercept is non-stationary and modeled as a random walk (→ unknown systematic changes over time)
- "degree of outgoing calls" is significantly negatively associated with the negative mood.
- Estimated coefficient of "degree of outgoing texts" shows two periods: no significant association in stage I, and significant negative association in stage II.
- Estimated coefficient of "previous physical activity" shows three periods: no significant association in stage I and III, and significant negative association in stage II.

# Summary

- Existing imputation methods mostly assume the time series to be stationary.
- We proposed a EM 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 and more efficient estimation for non-stationary time series with missing outcomes.

Limitation and future work:

- Extend the SSMimpute method to missing values in exposures and covariates.
- Apply more flexible state space modeling than linear regression
- Evaluate lagged effect or long-term effect of exposure, combining g-methods for confounding adjustment
- The current model may suffer from unmeasured confounders.

# Acknowledgement

This work was supported by K01MH118477 grand (PI: Linda Valeri).

We thank Crystal Blankenbaker, Justin Baker, Zilan Chai, Charly Fowler, Jouni Helske, Melanie Mayer, Jukka-Pekka Onnela, 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!