State space model multiple imputation for missing data in non-stationary multivariate time series from digital devices

Xiaoxuan Cai, Xinru Wang, Justin Baker, Jukka-Pekka Onnela, Linda Valeri

Department of Biostatistics, Mailman School of Public Health, Columbia University

3rd Annual Health Data Science Symposium at Harvard, Boston

Objectives

We propose innovative imputation method to deal with missing data in non-stationary multivariate time series from digital devices in N-of-1 studies.

- Missingness in both response and explanatory variables
- Auto-correlation with past values of variables
- Non-stationarity in multi-variate time series

Introduction

Missing data is an ubiquitous problem in almost all fields that collect data. Data imputation is commonly recommended to improve estimation efficiency for quantities of interest.

Mobile technology Mobile phones and wearable devices allows real-time monitoring of individuals' behavior, social interactions, symptoms, and other health conditions. resulting in the emergence of a new type of data – entangled multivariate time series of outcome, exposure, and covariates.

Existing imputation methods are either designed for longitudinal data with limited follow-up times

- Last-observation-carried-forward
- Linear/Spline interpolation
- Multiple Imputation
- Weighted estimation equations

or for stationary time series,

- Moving average techniques
- ARIMA regression model

or for multivariate time series of i.i.d samples

- Recurrent neural networks
- Generative adversarial networks

No available approaches address the issue of missing data in both the response variable and regressors of lagged values of the outcome for non-stationary multivariate time series in N-of-1 studies.

Methods – SSMimpute

SSMimpute We combine multiple imputation with state space model to iteratively impute missing data in potentially non-stationary multivariate time series

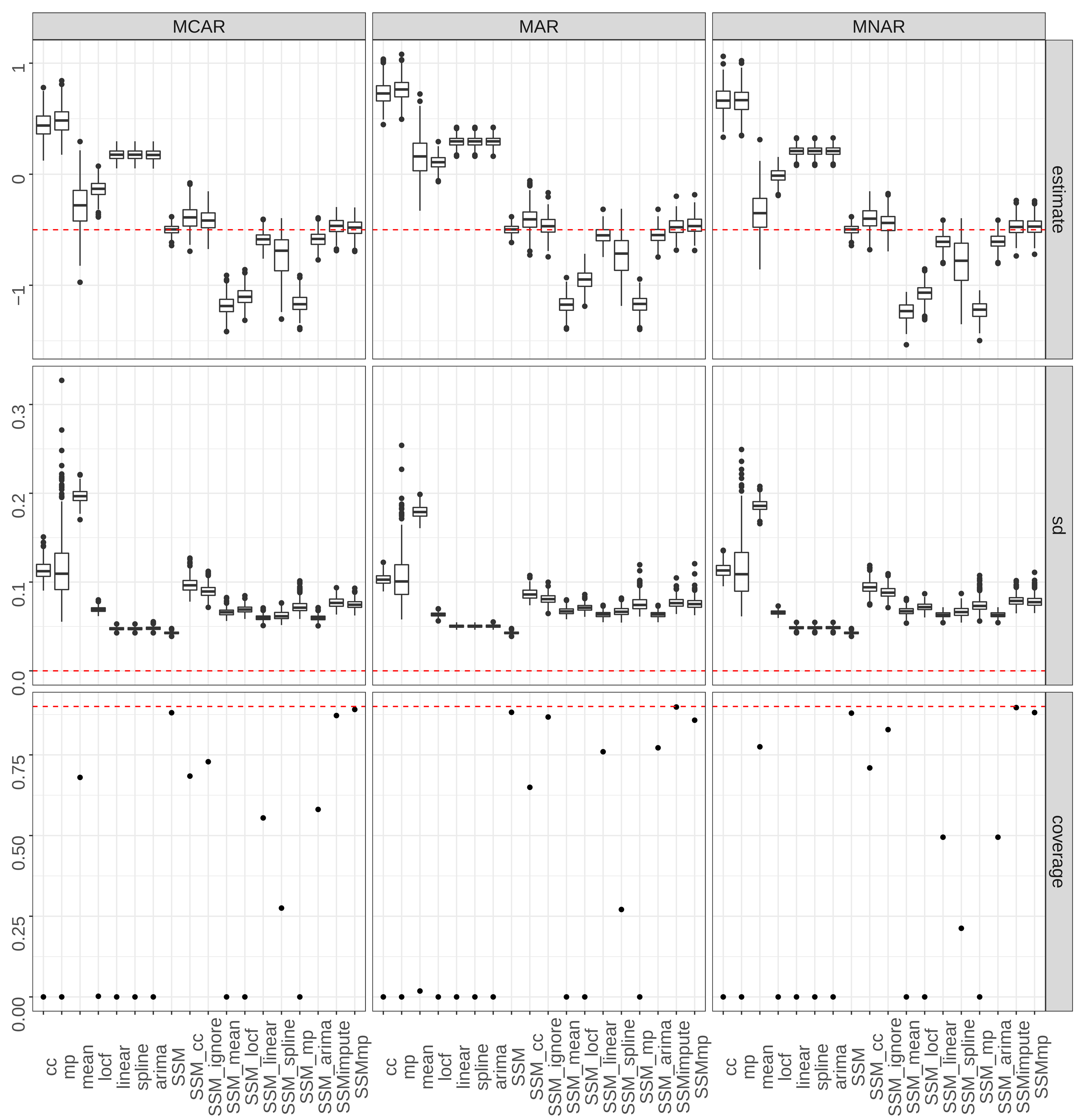
Theoretical properties of "SSMimpute" and its performance in extensive simulations of both stationary and non-stationary time series are evaluated under MCAR, MAR, and MNAR.

"SSMimpute" to existing methods in non-stationary time series

We illustrate one scenario of non-stationary multivariate time series with periodic treatment effect, random-walk baseline intercept, and time-variant effect of other variable.

$$Y_t = \beta_{0,t} + \rho Y_{t-1} + \beta_{1,t} A_{1,t} + \beta_{2,t} A_{1,t-1} + \beta_{c,t} C_t + v_t, \quad v_t \sim N(0, V)$$

where $\beta_{0,t}$ follows a random walk, and $\beta_{1,t}$ is periodic-stable with change points at $t = 400$ and $t = 700$, with missing rate of 50%.



Acknowledgement

This work was supported by grant K01MH118477. We thank Crystal Blankenbaker, Habib Rahimi, Aijin Wang, Zixu Wang, and Li Zeng for their great help and suggestions.

Bipolar Longitudinal Smartphone Study

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 (Temp$_t$), past physical activity (PA$_t$)

$$Y_t = \beta_{0,t} + \rho_t Y_{t-1} + \beta_{11,t} A_{\text{calls},t} + \beta_{12,t} A_{\text{calls},t-1} + \beta_{21,t} A_{\text{texts},t} + \beta_{22,t} A_{\text{texts},t-1} + \beta_{temp,t}\text{Temp}_t + \beta_{PA,t}\text{PA}_t + v_t$$

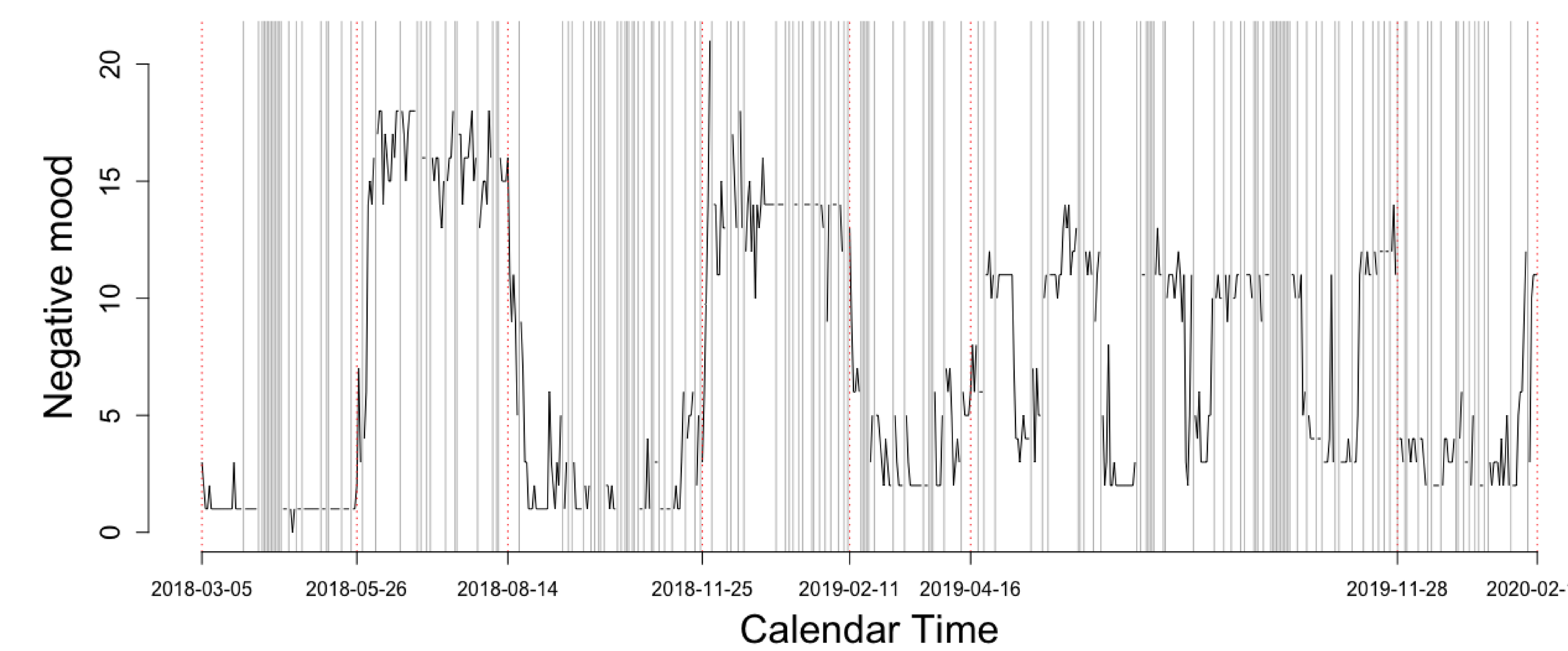


Figure: Negative mood (Y_t) of one bipolar patients, followed from from 03/05/2018 to 2020/20/10 (708 days)

	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)

- "degree of outgoing calls" is significantly associated with decrease in negative mood.
- Identified a changepoint for the effect of "degree of outgoing texts" on negative mood around 06/14/19: no effect in period I and negative effect in stage II.
- Identified two changepoints for the effect of "physical activity" on negative mood around 08/15/18 and 11/27/18: no effect in period I and III and negative effect in period II.

Conclusions

We proposed a novel state-space model based multiple imputation method for non-stationary multivariate time series, which is able to identify changepoints in the effect of covariates over time. The proposed imputation method provides unbiased and more efficient estimation for non-stationary time series with missing outcomes.