COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Testing unit root non-stationarity in the presence of missing data in univariate time series of mobile health studies

Charlotte Fowler, Xiaoxuan Cai, Justin Baker, Jukka-Pekka Onnela, Linda Valeri

March 20th, 2023

Motivation

Mobile health studies employ smartphones or wearable devices to collect information and present a new setting for time series analysis methods

- Plausible have missing data from different mechanisms including MNAR

Stationarity is a required assumption for many methods in longitudinal data analysis

- When non-stationarity is ignored, results can be spurious including detecting falsely significant relationships

Conventional methods for stationarity testing do not allow for missing data

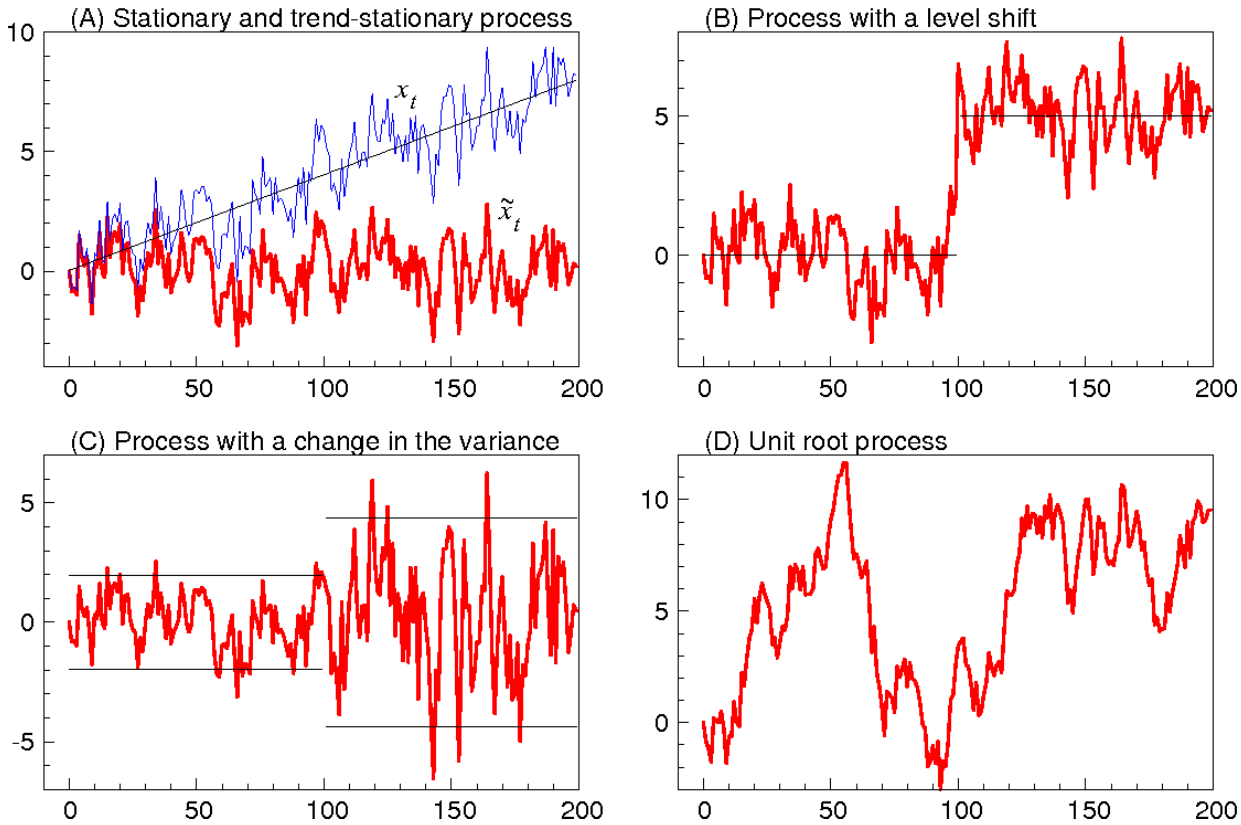
Missing Mechanisms

Missing Completely at Random (MCAR): probability of observation missing does not depend on any observed or unobserved variables

Missing at Random (MAR): probability of observation missing depends only on observed variables

Missing Not at Random (MNAR): probability of observation missing depends on unobserved variables

Non-Stationary Time Series



From Nielsen, H. B. (2007). Non-stationary time series and unit root testing. Engineering statistics hand book. Econometrics.

Unit Root Non-stationarity

Let $Y_0 = 0$, and for $t = 1, ..., T$

$$Y_t = \rho Y_{t-1} + \epsilon_t, \text{ where } \epsilon_t \text{ i.i.d. with } E(\epsilon_t) = 0$$

We say the time series Y_t has unit root (and thus is non-stationary) if $\rho = 1$.

ADF Test

The **augmented Dickey Fuller (ADF)** test is the conventional method to test $H_0 : \rho = 1$ versus $H_1 : \rho < 1$.

Letting $\hat{\rho}$ be the coefficient obtained from regressing Y_t by Y_{t-1}, we obtain test statistic

$$DF_{\hat{\rho}} = \frac{\hat{\rho} - 1}{SE(\hat{\rho})}$$

Under the null ($\rho = 1$) the test statistic follows the Dickey-Fuller distribution, no closed form

Current Methods for Missing data

Shin and Sarkar (1996, 1998) recommend last observation carry forward and complete case analysis to apply ADF test on ARMA(p, q) time series with observations MCAR.

Many methods for handling missing data:

- Complete case analysis (CC)
- Last observation carry forward (LOCF)
- Linear interpolation (IntL), Spline interpolation (IntS)
- Kalman smoothing imputation (K)
- Mean imputation (M)
- Multiple imputation with chained equations (MICE)

Maximum Likelihood Approach

1. Numerical Optimization

   - When data is MCAR or MAR, assuming $\epsilon_t \sim N(0, \sigma^2)$, the likelihood of the observed data is a function of the observed data and ρ, σ
   - Can use numerical optimization to solve for $\hat{\rho}$, $\hat{\sigma}$
   - Calculate conservative test statistic (MLEN): $DF_{\hat{\rho},c} = \frac{\hat{\rho}-1}{SE(\hat{\rho})}$
   - Calculate scaled test statistic (MLENS): $DF_{\hat{\rho},s} = \frac{\hat{\rho}-1}{SE(\hat{\rho})} \times \frac{\text{\# total time points}}{\text{\# observed time points}}$

Maximum Likelihood Approach

2. Expectation Maximization (MLEEM)

   - Assume $\epsilon_t \sim N(0, \sigma^2)$
   - Iteratively (1) impute expected values for missing y_t and (2) maximize likelihood w.r.t. $\hat{\rho}$, $\hat{\sigma}$ until convergence
   - Calculate test statistic from fully imputed time series
   - If data is assumed to be MNAR, can incorporate δ term to imputations within each iteration until convergence

State Space Model with Multiple Imputation

Assume time series is generated by lag order q

i.e. $Y_t = \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + ... + \rho_q Y_{t-q} + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma^2)$

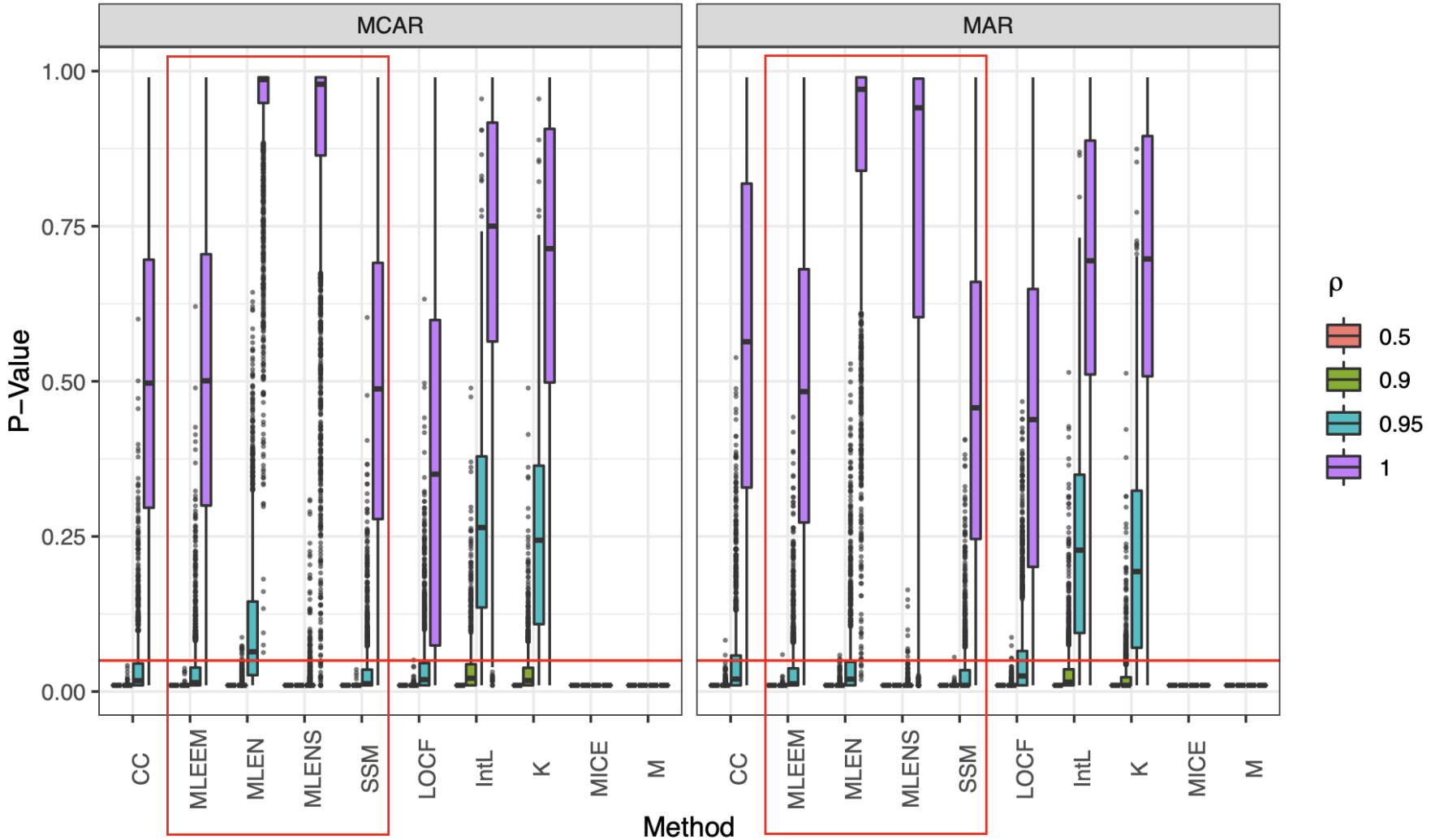
Inspired by algorithm from Cai et. al. (2022), uses an EM algorithm to iteratively impute lagged regressors and update posterior distribution.

Once convergence is reached, from the posterior obtain M draws of imputations.

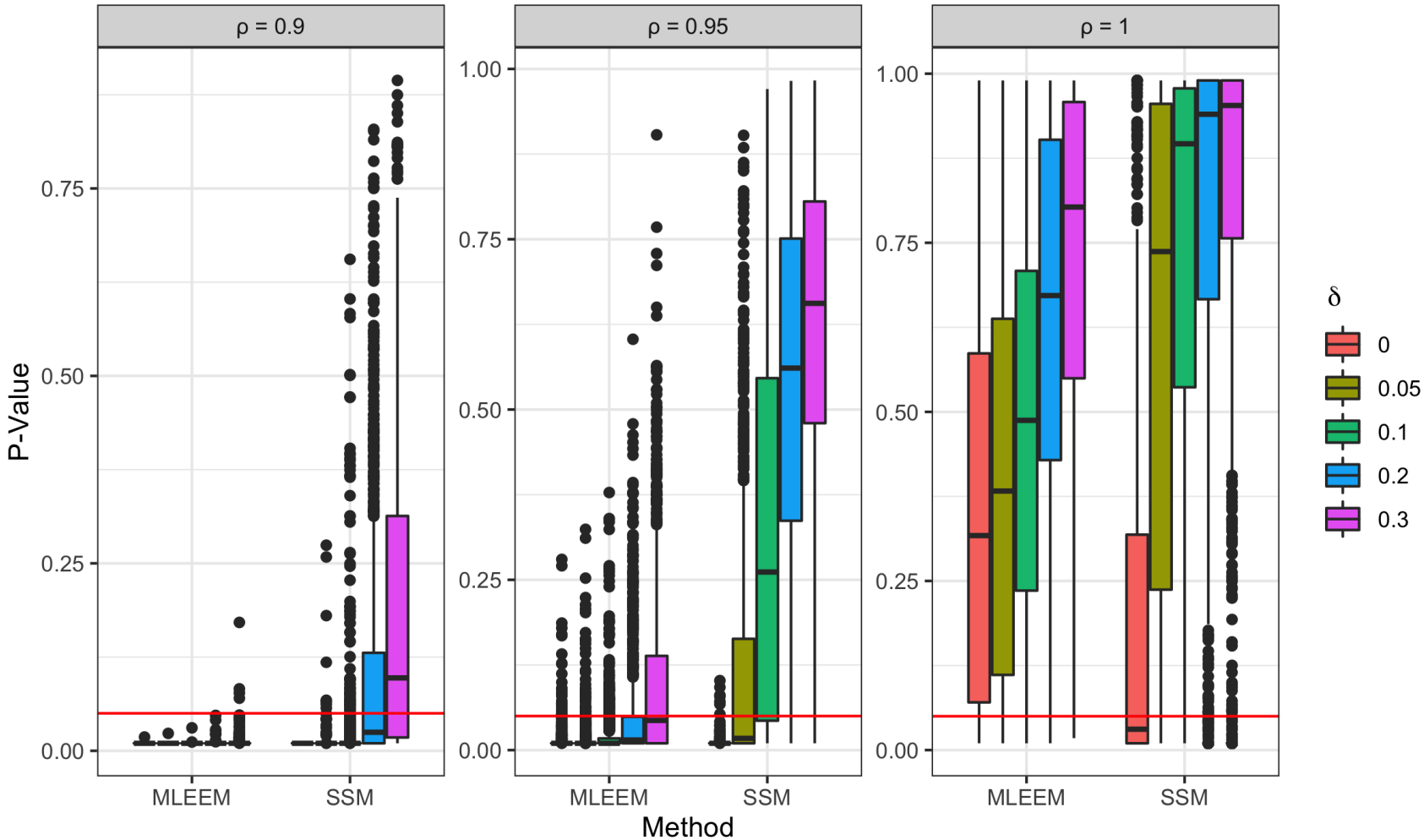
Apply ADF test to each imputation for $m = 1, ..., M$. Pool results by calculating the median test statistic and its p-value.

If data is assumed to be MNAR, can incorporate add δ term to imputations within each iteration until convergence

Simulation Results 50% Missing



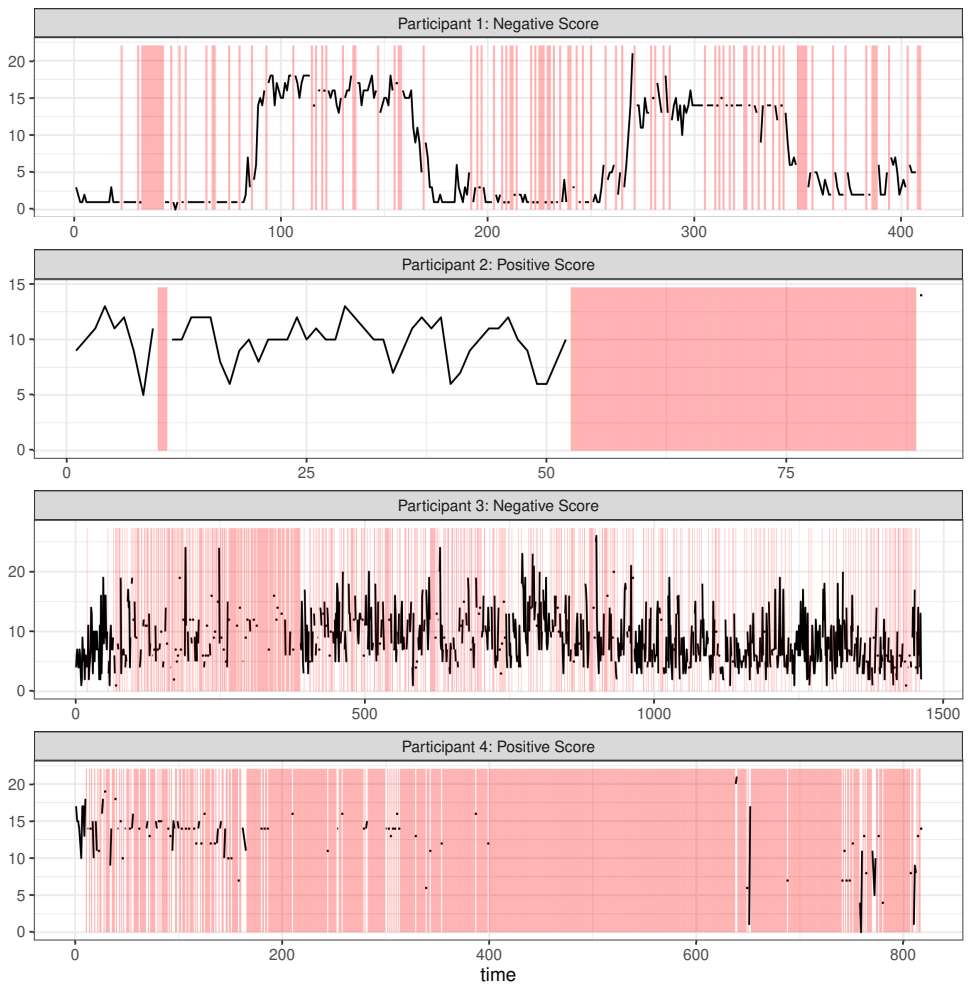
Sensitivity Simulation 50% Missing



Bipolar Longitudinal Study

- Ongoing study from McLean Hospital
- Participants with bipolar or schizophrenia followed for up to four years
- Outcome of interest: positive and negative mood scores
  - Aggregate summary score for daily survey responses

Bipolar Longitudinal Study



1. Participant 1
   - Appears non-stationary
   - MICE and M reject null
2. Participant 2
   - Appears stationary
   - IntS and K fail to reject
3. Participant 3
   - Appears stationary
   - All reject null
4. Participant 4
   - Too little information
   - Conflicting results

Conclusion

SSM, MLEEM, and MLENS achieve high power and acceptable type I error when data is MCAR or MAR

If data is MNAR, proposed sensitivity analysis can offer range of plausible results given δ values

Focus only on unit root non-stationarity, more research needed explore other types of non-stationarity

References

Shin, D. W., & Sarkar, S. (1996). Testing for a unit root in an AR (1) time series using irregularly observed data. Journal of Time Series Analysis, 17(3), 309-321.

Shin, D. W., & Sarkar, S. (1994). Unit root tests for ARIMA (0, 1, q) models with irregularly observed samples. Statistics & Probability Letters, 19(3), 189-194.

Cai, X., Wang, X., Eichi, H. R., Ongur, D., Dixon, L., Baker, J. T., ... & Valeri, L. (2022). State space model multiple imputation for missing data in non-stationary multivariate time series with application in digital Psychiatry. arXiv preprint arXiv:2206.14343.