A cyclostationary process with missing observations and a multivariate t-distribution error process

10th Conference on Cyclostationary Processes
Grodek, Poland, 2017

Chris Drake\textsuperscript{1}
Jacek Leskow\textsuperscript{2}
Aldo Garay\textsuperscript{3}

\textsuperscript{1}Department of Statistics, University of California, Davis, USA
\textsuperscript{2}Institute of Mathematics, Cracow University of Technology, Poland
\textsuperscript{3}Department of Statistics, University of Campinas, Brazil
Outline

(1) Missing data mechanisms
(2) Inference with missing data
(3) Missigness in cyclostationary processes
(4) The multivariate t-distribution
(5) Parameter estimation
(6) A simulation study
Missing Data Mechanisms

Consider a stochastic system that generates data $Y_i$. The goal is to model the data generating process $P_\theta$ and estimate the parameters $\theta$ that describe the process.

A second process $M_i$ of random variables exists with $M_i = 1$ if $Y_i$ is not observed and probability model $P_\psi$. If there are no missing data then $P_\psi(M_i = 1) = 0$ for all $i$ and the process is constant and can be ignored.

If there are missing data, then the inference needs to account for the missingness process except in special cases.
Complete model for the whole process

Joint distribution of \((Y, M)\) is given by

\[
P(y, m \mid \theta, \psi) = P(y \mid m, \tilde{\theta}) \times P(m \mid \psi)
\]

pattern-mixture model (PMM)

\[
= P(y \mid \theta) \times P(m \mid y, \tilde{\psi})
\]

selection model (SeM)
Missing Data Mechanisms

In general the parameters in the two models for the missingness patterns are not the same.

It is typically assumed that the parameters of the outcome model and the missingness distribution are distinct.

Missingness mechanisms are divided primarily into three different types, depending on the conditional distribution of the missingness given observed and missing data. These distinctions have implications for the analysis of the observed data and the inference that can be drawn.
Data is Missing Completely At Random (MCAR) if the probability of the observed missingness pattern does not depend on the observed or unobserved data. In terms of the missingness distribution

\[ P(m \mid y_{obs}, y_{mis}, \theta, \psi) = P(m \mid \psi) \]

Here the missingness distribution is a constant and selection and pattern-mixture models are identical.

Complete case analysis will lead to unbiased estimates since the complete cases are a random subsample of all cases.

Frequentist inference is possible.
Data is Missing At Random (MAR), if the missingness distribution depends only on observed data.

The missingness distribution in this case is given by

\[ P(m \mid y_{obs}, y_{mis}, \psi) = P(m \mid y_{obs}, \psi) \]

so that the selection model simplifies to

\[ P(y, m \mid \theta, \psi) = P(y \mid \theta) \times P(m \mid y_{obs}, \tilde{\psi}) \]

When data is missing at random, unbiased estimation based on the likelihood or Bayesian inference is possible. Frequentist inference is not possible.
In the context of a process \( \{y_t; \ t \in Z\} \), we have a missingness process \( \{m_t; \ t \in Z\} \)

**Missing completely at random** implies

\[
P(m_t = 1 \mid y_t, y_{t-1}, y_{t-2}, \ldots y_{t-h}, \ldots; \psi) = p
\]

In this model, missing completely at random means that observations are missing in a manner that does not depend on any observed and unobserved values.

An observation is as likely to be missing anywhere in the cycle and the cycles should not exhibit any pattern regarding missingness.
**Missing at random** implies

\[ P(m_t = 1 \mid y_{t-1}, y_{t-2}, \ldots, y_{t-h}; \psi) = p(y_{t-1}, \ldots, y_{t-h}; \psi) \]

If we assume an \( AR(1) \) Gaussian process with \( \phi > 0 \) and if we further assume

\[ P(m_t = 1 \mid y_{t-1}) = \frac{\exp(\alpha_0 + \alpha_1 y_{t-1})}{1 + \exp(\alpha_0 + \alpha_1 y_{t-1})} \]

with \( \alpha_1 > 0 \) the missing values are more likely to occur at large values of \( y_t \). Note, that \( y_{t-1} \) has to be observed for the process to be \( MAR \)
Missing Not at Random in the AR(1) process implies

\[ P(m_t = 1 \mid y_t, y_{t-1}; \psi) = p(y_t, y_{t-1}; \psi) \]

Note, it is sufficient for the missingness to depend on either an unobserved previous observation or the value that would have occurred for missingness to be not at random.
A Cyclostationary Model

We consider an observed process \( \{y_t, t \in \mathbb{Z}\} \) where

\[
y_t = c_t \times x_t
\]

1. \( x_t \) is a zero-mean \( K \)-dependent stationary sequence with a bounded and continuous spectral density.
2. \( c_t \) is a deterministic process with period \( H \), \( c_t > 0 \ \forall t \in \{1, \ldots, T\} \); and \( K << H \).
3. We assume furthermore that the model is identifiable by imposing a suitable set of conditions.

Assumption (2) is necessary so that the process is not degenerate at zero; the dimension of the parameter vector \( c_t \) is \( H \).
We assume that the process $x_t$ has a multivariate t-distribution with the following functional form for the density of $x_t, \ldots, x_{t+K}$

$$f(x_t, \ldots, x_{t+K}) = \frac{\Gamma((\nu + K)/2)}{(\pi \nu)^{(K/2)} \Gamma(\nu/2) | \Sigma |^{(K/2)}} 
\begin{bmatrix} 1 + \frac{1}{\nu} x^T \Sigma^{-1} x \\ \end{bmatrix}^{(-\nu+K)/2}$$

The variance of $y_t$ is given by

$$V(y_t) = c_t^2 V(x_t) = c_t^2 \frac{\nu}{\nu - 2}$$
Relevant facts about the multivariate t:

1. If \( x \) is \( K \)-variate t with \( df = \nu \) then \( x_q \) is \( q \)-variate t with \( df = \nu \); variance-covariance matrix partitions same as in multivariate normal.

2. Each \( x_t \) is univariate t with \( df = \nu \).

3. If \( x_q \) has dimension \( q \) and \( x_{K-q} \) has dimension \( K-q \) then \( x_q \mid x_{K-q} \) is conditionally also \( q \)-variate t with \( df = \nu \) if \( x_{q+1}, \ldots, x_K = \pm 1 \).

4. It can be shown where \( r_{jk}^* \) is the \((j,k)^{th}\) element of the inverse of the correlation matrix

\[
E[x_K \mid x_1, \ldots, x_{K-1}] = \frac{1}{r_{KK}^*} \sum_{j=0}^{K-1} r_{jK}^* x_j
\]
The correlation structure

We specify the correlation matrix for $x_t$ and we assume $x_t, x_{t+k}$ are correlated whenever $1 \leq k \leq K$.

The periodic behavior of the observed process $y_t$ is entirely determined by the $c_t$ and the correlation structure of $x_t$.

Note, that the correlation matrix of the $x_t$ process, which is unobserved, equals the correlation matrix of the $y_t$ process, which is observed. This correlation matrix has the following form:
\[ \Sigma \text{ is a } T \times T \text{ matrix, block-diagonal as follows:} \]

\[ \Sigma = \begin{pmatrix}
A & C & 0 & \ldots & \ldots & 0 \\
B & A & C & \ldots & \ldots & 0 \\
0 & B & A & C & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \ldots & \ldots & B & A & C \\
0 & \ldots & \ldots & 0 & B & A \\
\end{pmatrix} \]
The matrix $A$ is $K \times K$ 

$$A = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \ldots & \ldots & \rho_{K-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \ldots & \rho_{K-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \ldots & \rho_{K-3} \\ \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ \rho_{K-2} & \ldots & \ldots & \rho_1 & 1 & \rho_1 \\ \rho_{K-1} & \ldots & \ldots & \rho_2 & \rho_1 & 1 \end{pmatrix}$$
and $B$ is a lower triangular matrix with elements below the diagonal given by:

$$C = \begin{pmatrix}
\rho_1 \\
\rho_2 & \rho_1 \\
\vdots & \ddots \\
\rho_{K-1} & \cdots & \rho_2 & \rho_1 
\end{pmatrix}$$

The matrix $C$ is upper triangular and symmetric to $B$. 
Estimation of the model parameters

We will need to estimate:

1. $c_1, \ldots, c_H$
2. the parameters of the correlation matrix $\rho_1, \ldots, \rho_{K-1}$ for the $K$-dependent process $x_t$.
3. We assume the period $H$ is known; then $c_t = c_{t+H} = \ldots = c_{t+lH}$ for $t = 1, \ldots, H$ and $l = 1, \ldots, L - 1$. The parameters $c_1, \ldots, c_H$ determine the variance of the process $y_t$. Since we assume that $x_t$ is $K$-dependent then $x_t, x_u$ are independent if $|t - u| > K$ for each $t$ and $u$.
4. We also assume the period $H$ to be much larger than $K$. 
Based on the preceding assumptions:

1. The observed subsequence \( y_t, y_{t+H}, y_{t+2H}, \ldots, y_{t+(L-1)H} \) consists of independent and identically distributed random variables for \( t = 1, \ldots, H \).

2. We use these sequences to obtain \( \hat{c}_t \).

3. We calculate \( \tilde{x}_{h+jH} = \frac{y_{h+jH}}{\hat{c}_h} \) for \( h = 1, \ldots, H \).

4. We use \( \tilde{X} \) to estimate the sample correlation function as

\[
\hat{\rho}_\tau = \frac{1}{T} \sum_{t=1}^{T-\tau} (\tilde{x}_{t-\tau} - \bar{x})(\tilde{x}_t - \bar{x})
\]

5. Based on the asymptotic normality of \( \hat{\rho}_\tau(T) \) we will determine \( K \) (Shumway and Stoffer (2011)).
6 Once $K$ has been determined we will estimate the correlation matrix $R$ by methods of moments.

7 Once $\hat{R}$ has been calculated we will perform imputation procedures to fill in missing values. We will use the linearity of the conditional mean. Note, that we will only need the $K - 1$ preceding and following values in the time series for a missing value.

We evaluated our procedure with a simulation study
Simulation Study

A Data generation

1 Simulate $T + 2$ different dependent processes of order $k = 2p + 1$ using moving averages of independent and identically distributed random variables from a student t distribution, denoted by $u_i$ for $i = 1, ..., T + 2$

2 Generate the process $x_t$ as follows:

$$x_t = \frac{1}{k} \sum_{i=t-p}^{t+p} u_i$$

where $u_i \sim t(0, 1; \nu)$

the $u_i$ independent and $t = 1, ..., T$. 
3 We generate a series of periodic functions $c_t$ with period $H >> K$ as follows:

$$c_h = \sin\left(\frac{2\pi h}{H}\right) + 1.1 \quad \text{for} \quad h = 1, \ldots, H$$

Note, adding the constant 1.1 ensures that $c_h > 0$

4 Generate the cyclostationary time series $y_t$ for $t = 1, \ldots, T$
B Estimation of $c_h$

1 We calculate

$$\hat{c}_h = S_h \sqrt{3} \left( \frac{\nu}{\nu - 2} \right)^{-\frac{1}{2}} \quad \text{for} \quad h = 1, \ldots, H$$

2 where

$$S_h^2 = \frac{1}{H - 1} \sum_{i=1}^{H} \left( y_{(i-1)H+h} - \bar{y}_h \right)^2$$

3

$$\bar{y}_h = \frac{1}{H} \sum_{h=1}^{H} y_{(i-1)H+h}$$
C Imputation of $y_t$

We will use the the complete data, which if data is missing completely at random (MCAR) constitute a random subsample of each sequence of iid variables to calculate $\hat{c}_h$. We use the marginal-conditional decomposition of the multivariate t distribution and the fact that the t-distribution is closed under addition to impute missing observations. We propose 4 different imputations.
1 **Backward** substitution

\[ \tilde{y}_{Fi} = E[y_i \mid y_{i+1}, \ldots, \min(y_{i+2(k-1)}, y_T)] \]

2 **Forward** substitution

\[ \tilde{y}_{Bi} = E[y_i \mid y_{i-1}, \ldots, \max(y_{i-2(k-1)}, y_1)] \]

3 **Average** substitution

\[ \tilde{y}_{Mi} = \frac{1}{2}(\tilde{y}_{Fi} + \tilde{y}_{Bi}) \]

4 **Generalized** substitution where \( \tilde{y}_{Gi} \) is obtained by sampling from a t-distribution

\[ f(y_i \mid y_{i-1}, \ldots, \max(y_{i-2(k-1)}, y_1), y_{i+1}, \ldots, \min(y_{i+2(k-1)}, y_T)) \]
D Evaluation of methods

1 We calculated the mean absolute error (MAE) defined as

\[ \text{MAE} = \frac{1}{I} \sum_{i=1}^{I} | y_i - \tilde{y}_i | \]

2 We also calculated the mean squared error (MSE) calculated as

\[ \text{MSE} = \frac{1}{I} \sum_{i=1}^{I} (y_i - \tilde{y}_i)^2 \]

3 We generated \( M = 100 \) different processes, each with length \( T = 15000 \), order of dependence \( K = 3 \) and period \( H = 15 \). We imputed \( \gamma = 1\%, 3\%, 5\% \) and \( 10\% \) missing values.
Plots 1 and 2 in the following pages are comparisons between the true and estimated $c_t$ for two random samples.

Plots 3 and 4 compare MAE and MSE as a function of percent imputed values.
<table>
<thead>
<tr>
<th>Impute</th>
<th>Backward</th>
<th>Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>1%</td>
<td>.42962</td>
<td>.43228</td>
</tr>
<tr>
<td>3%</td>
<td>.45266</td>
<td>.45154</td>
</tr>
<tr>
<td>5%</td>
<td>.47523</td>
<td>.47637</td>
</tr>
<tr>
<td>10%</td>
<td>.53847</td>
<td>.53928</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>Generalized</td>
</tr>
<tr>
<td>1%</td>
<td>.50903</td>
<td>.48155</td>
</tr>
<tr>
<td>3%</td>
<td>.53261</td>
<td>.53681</td>
</tr>
<tr>
<td>5%</td>
<td>.60392</td>
<td>.61190</td>
</tr>
<tr>
<td>10%</td>
<td>.82201</td>
<td>.81727</td>
</tr>
</tbody>
</table>
Real and Estimated Curves

Plot for randomly sampled curve from simulated curves generated
Real and Estimated Curves

Plot for randomly sampled curve from simulated curves generated.
Mean Squared Error (MSE)
Mean Absolute Error (MAE)