ESTIMATING AUTOCORRELATIONS AND CROSS-CORRELATIONS WHERE MISSING VALUES ARE PRESENT: A SAS MACRO

Kenneth M. Portier, University of Florida
Pan-Yu Lai, University of Florida

INTRODUCTION.

In most statistical data analysis methods, missing values are either thrown away or replaced by estimated values. Both of these methods are not adequate in dealing with time series data since the latter may influence the estimation of parameters and the former may affect the stationarity of the series.

All SAS® software related to time series data such as the ARIMA, AUTOBOX, SPECTRA and STATESPACE procedures in the SAS/ETS™ system at the beginning of each series until a nonmissing value is encountered, then use only data from this point until the next missing data occurs. In this way, only part of nonmissing data can be utilized when missing values are embedded in the series. A program which allows estimation of some basic parameters of time series analysis, such as autocorrelations and cross-correlations, utilizing all nonmissing values in the series is presented here.

Estimation of autocorrelations of a discrete stationary stochastic process having missing values and the properties of these estimators have been discussed in Dunsmuir and Robinson (1981a, 1981b) and Marshall (1980). Similarly, properties of spectral density estimation can be found in Bloomfield (1970), Dunsmuir and Robinson (1981b), Jones (1962), and Porzen (1963). An application of these findings to Yule-Walker estimators assuming an autoregressive model is also given in Marshall (1980).

Although the estimate of cross-correlations of discrete stationary stochastic processes when missing values are present in at least one of the series has not been discussed in the literature, intuitively one could use an approach similar to that given for univariate series. An outline for estimating the cross-correlations is presented here along with a SAS macro for estimating autocorrelations and cross-correlations.

ESTIMATION.

Autocorrelations - Univariate case.

Let \( \{X_t, t = 1, 2, ..., n\} \) be the original series which may contain some missing values. Define a series \( \{A_t, t = 1, 2, ..., n\} \) by

\[
A_t = \begin{cases} 
1 & \text{if } X_t \text{ is observed} \\
0 & \text{otherwise.} 
\end{cases}
\]

Using this indicator series, the amplitude modulated sequence \( U_t \), which was first introduced by Parzen (1963), is then defined as

\[
U_t = A_t X_t, \quad t = 1, 2, ..., n.
\]

Define the cross products sum for lag \( k \) for the indicator series and the amplitude modulated sequence as follows:

Let \( C_a(t) = n^{-1} \sum_{t=1}^{n-k} A_t A_{t+k} \quad 0 \leq k < n \)

\[
C_a(t) = n^{-1} \sum_{t=1}^{n-k} U_t U_{t+k} \quad 0 \leq k < n.
\]

Then, the sample autocovariances are estimated by

\[
\hat{C}_X(t) = C_a(t)/C_a(0), \quad \text{when } C_a(t) \neq 0 \text{ and } EX_n = 0,
\]

and the sample autocorrelations are estimated by

\[
\hat{P}_X(t) = \hat{C}_X(t)/\hat{C}_X(0).
\]

When \( EX_n \neq 0 \), then \( EX_n \) is estimated by \( \hat{U}_X = n^{-1} \sum_{t=1}^{n} U_t A_t \). In this case the original \( X_t \) is replaced by \( X'_t = X_t - \hat{U}_X \) and \( U'_t = A_t X'_t \) in the equations for \( \hat{C}_X(t) \) and \( \hat{P}_X(t) \). Let the serial autocovariances \( EX_n \) be denoted by \( \gamma_X(k) \) and \( C_X(t) = \gamma_X(t)/\gamma_X(0) \). Dunsmuir and Robinson (1981a) prove that if \( u = \lim_{n \to \infty} n^{-1} A(t) \) a.s. and \( \gamma_X(t) = \gamma_X(t)/\gamma_X(0) \) for all \( k \) a.s., plus certain stationary conditions on the \( E(t) \) term, then \( C_X(t) \) is asymptotically normal with asy. var. \( \gamma_X(t) \) for each \( k \) for which \( V(t) \neq 0 \).

If asymptotic stationarity to the fourth order for the \( A(n) \) can be assumed, then a central limit theorem for \( \sqrt{n}(C_X(t) - C_X(0)) \) is also given by Dunsmuir and Robinson. From this, the central limit theorem can be deduced for \( \sqrt{n}(C_X(t) - \gamma_X(t)) \) or equivalently \( \sqrt{n}(\hat{C}_X(t) - \gamma_X(t)) \) in particular, if the \( X(n) \) are i.i.d with finite fourth moment, then the \( \hat{P}_X(t) \) are asymptotically independent normal with asy. var. \( (\gamma_X(t))^{-1} \) which can be estimated by \( (nC_a(0))^{-1} \). One problem with using the \( C_a(t) \) as estimates of covariances is that the sequence \( \{C_a(k) : |k| < n\} \) is not necessarily (for a given finite sample size) nonnegative definite. An alternative approach which does not fully use all the nonmissing information but does guarantee a nonnegative definite sequence was given by Burg (1975). This method is not included in our program.

Cross-correlations.

Cross-correlations are used to identify the form...
of the relationship (transfer function) between two time series. Although the methodology of estimating cross-correlations when at least one series contains missing values has not been discussed in the literature, an intuitive approach which is similar to the one used to estimate autocorrelations of univariate series leads to the following equations.

Let \( \{X_t\}, t = 1, 2, \ldots, n \) and \( \{Y_t\}, t = 1, 2, \ldots, n \) be the original series in which one or both may contain some missing values. Let

\[
A_t = \begin{cases} 1 & \text{if } X_t \text{ is observed} \\ 0 & \text{otherwise} \end{cases}
\]

\[
B_t = \begin{cases} 1 & \text{if } Y_t \text{ is observed} \\ 0 & \text{otherwise} \end{cases}
\]

be the indicator functions as defined for univariate series. Define

\[
U_t = A_t X_t, \quad V_t = B_t Y_t
\]

1 \leq t \leq n to be the amplitude modulated series for \( X_t \) and \( Y_t \), respectively. Then, using

\[
C_{ab}(t) = n^{-1} \sum_{t=1}^{n} A_t B_{t+\tau} 0 \leq \tau < n
\]

and

\[
C_{UV}(t) = n^{-1} \sum_{t=1}^{n} U_t V_{t+\tau} 0 \leq \tau < n
\]

the sample cross-correlations are estimated as

\[
\hat{\rho}_{XY}(\tau) = \frac{C_{UV}(\tau)}{\sqrt{C_{X}(0) \cdot C_{Y}(0)}}
\]

where \( C_{ab}(t) \neq 0 \), \( \bar{E}X = EY = 0 \) and \( \bar{C}_X(0), \bar{C}_Y(0) \) are defined as in the univariate series. If \( \bar{E}X \) \neq 0 (or \( \bar{E}Y \) \neq 0), then define \( \hat{\mu}_X = \frac{1}{n} \sum_{t=1}^{n} X_t, \hat{\mu}_Y = \frac{1}{n} \sum_{t=1}^{n} Y_t \), and the original \( X_t \) (\( Y_t \)) is replaced by \( \hat{\mu}_X X_t \) \( \hat{\mu}_Y Y_t \) respectively in the equations for \( C_{ab}(t) \) and \( C_{UV}(t) \).

Again, we like to remind the reader that to our knowledge, no work has been published on the properties of these estimators.

Yule-Walker estimates for the stationary autoregressive model. Let \( \{X_t\} \) are from a stationary autoregressive model of order \( K \), that is \( X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_K X_{t-K} + A_t \). Then, its autocorrelation function satisfies the difference equations

\[
\phi^\tau = \phi_1 \phi_{K-1} + \phi_2 \phi_{K-2} + \ldots + \phi_K \phi_{1-K}
\]

In matrix notation, define

\[
\phi = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_K \\ \phi_2 & \phi_3 & \ldots & \phi_K \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K-1} & \phi_{K-2} & \ldots & 1 \end{bmatrix}
\]

then \( \phi = \phi^\tau \) \( K \) \(*\) where the \( \phi \) are Yule-Walker estimates of autoregressive parameters. According to Dunsmuir and Robinson (1981a), one can simply replace autocorrelations by sample autocorrelations in \((*)\) and solve the equation to get the Yule-Walker estimates of autoregressive parameters.

MISSCORR MACRO.

The program included with this paper allows users to include any number of variables and obtain the autocorrelation estimates. The program calculates sample autocorrelations along with their sample variance for the i.i.d case for each variable and sample cross-correlations between the first variable and the other variables in the input list up to lag \( K \), where \( K \) is specified by the users. The sample autocorrelations are then used to solve the Yule-Walker equations to estimate the parameters of the autoregressive model of each variable. Detailed information of how to use this program is given in the comments in the program listing.

EXAMPLE RUN.

A plot of the average number of eggs laid by the mosquito Aedes aegypti in oviposition traps over four years in New Orleans, La., are given in Figure 1. Note that during the late fall and early spring the level of oviposition by Aedes Aegypti is too low to be measured by the sample traps. In attempting to describe this series, the zero counts are considered missing. The relationship between egg counts and three environmental parameters, average weekly minimum daily temperature, relative humidity and total weekly rainfall, is also of interest. A plot of minimum temperature is given in Figure 2.

Using MISSCORR, autocorrelations \((\hat{\rho}_X(\tau))\), cross-correlations \((\hat{\rho}_{XY}(\tau))\), asymptotic variances \((nV(\tau)-1)\) and Yule-Walker parameter estimate for first and second order autoregressive models are computed. The data is stored in an SAS data set called EGGS.BASE. The analysis program is as follows:

```sas
*-------------------------------------DATA SAMPLE;
SET EGGS.BASE;
IF YEAR>1981;
IF AVGEGGS=0 THEN AVGEGGS=0.1;
KEEP MINTEMP RAIN RH AVGEGGS;
MACRO DATAIN SAMPLE %VARCARD
AVGEGGS MINTEMP RAIN RH;
OPTCARD
```

51
The authors wish to thank Dr. Dana A. Focks, Research Entomologist, Insects Affecting Man and Animals Research Laboratory, ARS, USDA for introducing us to this problem and supplying the example data.

References


SAS is the registered trademark and SAS/ETS the trademark of SAS Institute Inc., Cary, NC, USA.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>SAMPLE EFF DATA</th>
<th>AUTOCORRELATIONS OF THE SAMPLE EFF DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>MEAN</td>
<td>VAR</td>
</tr>
<tr>
<td>X1</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
<tr>
<td>X2</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
<tr>
<td>X3</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
<tr>
<td>X4</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
<tr>
<td>X5</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
<tr>
<td>X6</td>
<td>2.1416</td>
<td>1.824</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 1 (continued)</th>
<th>AMPLITUDE MODULATED OBSERVATIONS</th>
<th>TABLE 1</th>
<th>SAMPLE EFF DATA</th>
<th>AUTOCORRELATIONS OF THE SAMPLE EFF DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>MEAN</td>
<td>VAR</td>
<td>R</td>
<td>B</td>
</tr>
<tr>
<td>X1</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>X2</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>X3</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>X4</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>X5</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>X6</td>
<td>2.1416</td>
<td>1.824</td>
<td>0.98</td>
<td>0.99</td>
</tr>
</tbody>
</table>

The results of the analysis are given in Table 1.
PROGRAM LISTING

OPTIONS LS779;
MACRO ACRO calculates auto-correlations.
MACRO ISSCORR calculates cross-correlations and the coefficients.
THE USER IS REQUIRED TO SPECIFY A MACRO WHICH IDENTIFIES THE DATASET TO BE USED. IT'S FORMATTED AS FOLLOWS:

MACRO DATA MACRO
WHERE NNENN IS THE DATA SET TO BE USED.
DATA SET IS NNENN.

MACRO OPTIONS MACRO
WHERE OPTIONS REQUESTED BY THE USER.

EX: OPTIONS "LAGO" LAG1 LAG2 LAG3 LAG4 LAG5 LAG6 LAG7 LAG8 LAG9 LAG10 LAG11 LAG12 LAG13 LAG14 LAG15 LAG16 LAG17 LAG18 LAG19 LAG20 LAG21 LAG22 LAG23 LAG24 LAG25 LAG26 LAG27 LAG28 LAG29 LAG30"

EX: OPTIONS "LAGS"

MACRO VARSPEC MACRO
WHERE VARSPEC IDENTIFIES THE VARIABLES WHICH WILL BE USED IN THE MACRO ISSCORR.

EX: VARSPEC "X1 X2 X3 X4 X5 X6"

MACRO VARIABLES MACRO
WHERE VARIABLES IDENTIFIED FROm VARSPEC TO THE SPECIFIED DATA SET.

EX: VARIABLES "X1 X2 X3 X4 X5 X6"

MACRO OPTIONS MACRO
WHERE OPTIONS REQUESTED BY THE USER.

EX: OPTIONS "LAGO" LAG1 LAG2 LAG3 LAG4 LAG5 LAG6 LAG7 LAG8 LAG9 LAG10 LAG11 LAG12 LAG13 LAG14 LAG15 LAG16 LAG17 LAG18 LAG19 LAG20 LAG21 LAG22 LAG23 LAG24 LAG25 LAG26 LAG27 LAG28 LAG29 LAG30"

EX: OPTIONS "LAGS"

MACRO VARSPEC MACRO
WHERE VARSPEC IDENTIFIES THE VARIABLES WHICH WILL BE USED IN THE MACRO ISSCORR.

EX: VARSPEC "X1 X2 X3 X4 X5 X6"

MACRO VARIABLES MACRO
WHERE VARIABLES IDENTIFIED FROm VARSPEC TO THE SPECIFIED DATA SET.

EX: VARIABLES "X1 X2 X3 X4 X5 X6"