

Multidimensional time series model estimation in presence of partially missing observations

Giovanni Fonseca and Luca Grassetti

Abstract In multidimensional time series the presence of missing observations is very common. The model estimation problem in presence of missing data is usually solved by maximum likelihood approach or considering one of the many imputation methods introduced so far in the literature. In this paper we consider the case of multidimensional time series with a part of observations that are completely or partially missing. In order to recover as much as possible information we propose a model estimation procedure based on composite likelihood combined with a model based imputation method. The proposed method has been validated by simulation studies.

Key words: VAR model, missing data, pseudo-likelihood, model based imputation strategy

1 Introduction

The present paper aims at addressing the problem of model estimation in presence of partially missing observations in the multidimensional time series context. The presence of missing data in this framework is very common and, for this reason, this issue has been widely discussed in the existing literature (see, Little and Rubin, 2002). Nonetheless, most of the proposed methods are straightforward generalisations of the univariate techniques. In particular, the main literature about model estimation can be distinguished in “skipping” approaches and imputation based methods. The former are simply performed without considering the missing observations. There-

Giovanni Fonseca
Dipartimento di Scienze Statistiche - Università degli Studi di Udine, e-mail: fonseca@dss.uniud.it

Luca Grassetti
Dipartimento di Scienze Statistiche - Università degli Studi di Udine e-mail: grassetti@dss.uniud.it

fore, for example, in the state space model formulation, estimation is pursued skipping the updating step, corresponding to a missing value, in the Kalman filter recursion (see, for instance, Gómez, Maravall and Peña, 1999; Harvey, 1990; Harvey and Pierse, 1984; Proietti, 2008). On the other hand, imputation based procedures substitute the missing observations with suitable values and perform model estimation considering the “completed” dataset (see, Little and Rubin, 2002; Liu, Taylor and Belin, 2000). The generalisation of both approaches to the multidimensional case arises some problems. For instance, in the multiple time series context, it is common to find partially missing observations. In this case the “skipping” approaches consider the partially observed data as they were completely missing. In order to allow a partial filling in of the multidimensional observation, imputation methods based on regression estimation have been proposed in the multivariate independent sample case (see, Buck, 1960; Luceño, 1997).

In Fonseca and Grassetti (2010), we develop a pseudo-likelihood approach that allows to recover part of the information which is lost skipping the partially missing observations. Considering the pairwise likelihood (see Varin, 2008) we are able to include partially observed data as couples components in the composite likelihood computation. However, the conditional densities of data involving partially missing observations as conditional events still cannot be included in the likelihood. In the present work, we introduce a model-based imputation strategy in order to recover also this information.

A simulation study is developed in order to support the proposed approach. The obtained results show that the recovery of partially observed cases improves the multivariate time series models estimation.

The paper is organised as follows. In Section 2, we state the model specification and introduce the proposed estimation method. Section 3 presents a simulation study considering different cases of missingness patterns in the data. Finally, we report some conclusions and discuss the future developments.

2 Model estimation

For ease of exposition we introduce the estimation method considering the first order vector autoregressive VAR(1) model

$$y_t = Ay_{t-1} + u_t, \quad (1)$$

where $y_t = [y_{1,t}, \dots, y_{k,t}]^T$ is a $(k \times 1)$ vector of observations, A is a $(k \times k)$ fixed coefficients matrix, and $\{u_t\}_{t \in \mathbb{Z}}$ is a k -dimensional multivariate white noise sequence with variance-covariance matrix Σ . Under these assumptions, the estimation of multivariate autoregressive models can be faced by multivariate generalised least squares as shown in Lütkepohl (2005). Otherwise, as in the present work, one can assume that $u_t \sim N_k(0, \Sigma)$, for $t = 1, \dots, T$ and the maximum likelihood estimation procedure may be easily performed since the likelihood function assumes

a well known form which maximisation has been largely studied. The conditional likelihood of the sample $y_1, \dots, y_T$ is defined as

$$L(A, \Sigma) = \prod_{t=2}^T p_k(y_t | y_{t-1}; A, \Sigma)$$

where $p_k(\cdot | y_{t-1}; A, \Sigma)$ is the density function of a k -dimensional multinormal random variable with mean Ay_{t-1} and variance matrix Σ . The estimation of multivariate time series models in presence of missing data can be faced ignoring all time observations presenting at least one missing value (as, for example, in Proietti, 2008, and the citations therein). As just introduced, in the case of partially missing observations this approach causes the waste of available informations. In order to recover some of this information we define a surrogate of $p_k(\cdot)$ builded as the product of the bivariate margins of each pair of components of the vector, that is

$$p_k^*(y_t | y_{t-1}; A, \Sigma) = \prod_{i=1}^{k-1} \prod_{j=i+1}^k p_2(y_{t,i}, y_{t,j} | y_{t-1}; A, \Sigma)$$

with

$$p_2(y_{t,i}, y_{t,j} | y_{t-1}; A, \Sigma) = \frac{1}{2\pi} \frac{1}{\det(\Sigma^{ij})^{1/2}} \times \exp \left[-\frac{1}{2} \left(y_t^{ij} - A^{ij} y_{t-1} \right)^T (\Sigma^{ij})^{-1} \left(y_t^{ij} - A^{ij} y_{t-1} \right) \right].$$

where $y_t^{ij} = [y_{t,i}, y_{t,j}]^T$, Σ^{ij} is the 2-dimensional submatrix defined selecting the i th and j th rows and columns of the variance/covariance matrix Σ and A^{ij} is the matrix defined selecting the i th and j th rows of A . Finally, we define the pairwise likelihood

$$L_{pw}(A, \Sigma) = \prod_{t=2}^T p_k^*(y_t | y_{t-1}; A, \Sigma)$$

If a subset $y_{t,m}$ of y_t is missing, we can just exclude from the product the bivariate densities where elements of $y_{t,m}$ are involved. In this way the information contained in the non-missing components of y_t can be recovered. Results in Fonseca and Grassetti (2010) show that this method improves on the efficiency of the estimates $\hat{A}$ and $\hat{\Sigma}$, especially when the missingness patterns of the k univariate series are considerably differentiated. However, as in the other missing data treatment strategies, we are not able to recover the factor of the likelihood including y_t as conditioning event.

In order to solve this issue, we propose to use the model estimated by the pairwise likelihood function to fill the missing components in the vectors of partial observations. For example, the missing element $y_{t,i}$ of y_t is filled in with $\hat{y}_{t,i}$, that is the i th row of $\hat{A}y_{t-1}$ (pairwise likelihood imputation method). A residual bias may be due to the fact that in the model estimation are possibly considered vectors formed by both observed $y_{t,o}$ and imputed $\hat{y}_{t,m}$ components (subscripts o and m identify respectively

the observed and the missing components of the multivariate time series). In order to reduce this error we propose an imputation strategy based on the expected values of $y_{t,m}$ conditional to $y_{t,o}$. The multivariate conditional distribution of Y_t given $Y_{t-1} = y_{t-1}$ can be written as

$$Y_t = \begin{bmatrix} Y_{t,m} \\ Y_{t,o} \end{bmatrix} | Y_{t-1} = y_{t-1} \sim N \left(\begin{bmatrix} A_m y_{t-1} \\ A_o y_{t-1} \end{bmatrix}, \begin{bmatrix} \Sigma_{mm} & \Sigma_{mo} \\ \Sigma_{om} & \Sigma_{oo} \end{bmatrix} \right) \quad (2)$$

The estimated value of $y_{t,m}$ conditional to $y_{t,o}$ and y_{t-1} (pairwise likelihood adjusted imputation method) can finally be defined as

$$\hat{y}_{t,m} = E[Y_{t,m} | y_{t-1}, y_{t,o}; \hat{A}, \hat{\Sigma}] = \hat{A}_{t,m} y_{t-1} + \hat{\Sigma}_{t,mo} \hat{\Sigma}_{t,oo}^{-1} (y_{t,o} - \hat{A}_{t,o} y_{t-1}).$$

The estimated values are function of the past information and of the difference between observed and estimated values ($y_{t,o}$ and $\hat{A}y_{t-1}$ respectively). Matrices $\hat{\Sigma}_{t,mo}$ and $\hat{\Sigma}_{t,oo}$ are blocks of the estimated variance covariance matrix and are time dependent given the different missingness patterns at each time t . The completed vector is considered in the conditional mean computation.

Summarising, we propose a three step estimation procedure.

- **Step 1.** Model parameters A and Σ are estimated using pairwise likelihood specification in order to recover the most possible information included in the partially missing observations.
- **Step 2.** Model-based imputation of the missing values is derived using the estimates $\hat{A}, \hat{\Sigma}$ obtained in Step 1.
- **Step 3.** The “completed” observations are then considered in the pairwise likelihood estimation procedure in order to recover some more informations about lagged variables.

Steps 2 and 3 of the procedure are implicitly iterated in the maximisation procedure until convergence is reached.

3 Simulation studies

Simulation studies are developed in order to show the improvements of the pairwise likelihood approach in comparison to a classical maximum likelihood estimation method. The following results are an integration of the simulation reported in Fonseca and Grassetti (2010). In particular, the original data generating process is adopted here with a slightly modification. The multivariate time series ($k = 3$) are, in fact, generated following the VAR(1) model with deterministic linear trend and non zero mean

$$y_t = A y_{t-1} + M X_t + u_t, \quad t = 2, \dots, T, \quad (3)$$

with $u_t \sim N(0, \Sigma_u)$. The initial points y_0 are generated from a multivariate standard gaussian distribution. Constant and time trend are both introduced by the term $M X_t$.

The fixed components coefficient (A and M) and the error correlation matrices (Σ) are as follows

$$A = \begin{bmatrix} 0.2 & -0.7 & 0.9 \\ 0.4 & 0.5 & -0.3 \\ -0.2 & 0.6 & 0.1 \end{bmatrix}, M = \begin{bmatrix} 2.5 & 0.10 \\ 1.5 & 0.07 \\ 0.5 & 0.05 \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} 1.0 & -0.8 & 0.4 \\ -0.8 & 1.0 & -0.6 \\ 0.4 & -0.6 & 1.0 \end{bmatrix}.$$

The simulation design is developed under four different sample sizes ($T = \{50, 100, 200, 500\}$) and the following missing proportion mix are considered: $\{0.05, 0.10, 0.25\}$. A more realistic case (where the proportions of missing data vary between time series) is, finally, considered. The present study considers a proportion of missing cases in which the whole vector is missing. This proportion is fixed at the 40% of the missing cases. When this proportion tends to 1 our proposal is completely equivalent to the benchmark likelihood methods.

3.1 Comparative results and estimation improvements

The first simulation study is performed in order to show how the proposed pseudo likelihood procedure behaves in comparison with likelihood estimation based on Kalman filtering skipping approach. For the sake of simplicity we analyse the general model behaviour by a subset of the estimated model parameters. In particular we selected the coefficients of the first vector component (y_1) model. The omitted informations are redundant.

The comparative simulation results are collected in Table 1. As one can easily notice from both the bias and root mean square error points of view the pairwise likelihood results overcome the likelihood ones. This effect is even more evident when the sample size increases.

The estimators behaviour is not particularly affected by the adjustments introduced in Section 2. The pairwise likelihood imputation method presents good efficiency and bias results for the fixed effects coefficients but it introduces some bias in the variance covariance matrix at least when the percentage of missing values is high. This issue is at least partially solved by considering the adjusted imputation. In comparison to pairwise likelihood, the adjusted imputation method presents a generalised minimum gain in estimators' efficiency at least when the proportion of missing is relevant.

The same simulation study has been replicated for larger sample size ($T = \{200, 500\}$). The simulation results confirm the model behaviour observed with $T = 100$ but give more stable evidence about the obtainable gains. These results have been omitted for space reason.

Table 1 Results for simulation design considering A and Σ (correlated error terms with homogeneous missing proportions). The simulation is developed considering different sample sizes $T = \{50, 100, 200, 500\}$ and $k = 3$. The missingness patterns are defined by different proportions in the three time series $\{0.05, 0.10, 0.25\}$. A subset of coefficients bias and RMSE (in parentheses) are given. The simulation is based on 200 replications.

T=50		
Coefficients	Likelihood	PW Likelihood
γ_{11}	-0.0399 (0.0220)	-0.0279 (0.0161)
γ_{12}	0.0188 (0.0420)	0.0117 (0.0298)
γ_{13}	-0.0119 (0.0383)	-0.0173 (0.0263)
β_{11}	0.0934 (1.0420)	0.0825 (0.8442)
β_{12}	0.0027 (0.0047)	0.0033 (0.0037)
σ_{11}	-0.2858 (0.1586)	-0.2013 (0.0968)
σ_{22}	-0.2712 (0.1479)	-0.1985 (0.0949)
σ_{33}	-0.2573 (0.1428)	-0.2184 (0.1175)
σ_{12}	0.2277 (0.1128)	0.1664 (0.0724)
σ_{13}	-0.0999 (0.0591)	-0.0675 (0.0510)
σ_{23}	0.1555 (0.0781)	0.1149 (0.0599)
T=100		
Coefficients	Likelihood	PW Likelihood
γ_{11}	-0.0152 (0.0116)	-0.0130 (0.0088)
γ_{12}	-0.0023 (0.0249)	-0.0037 (0.0200)
γ_{13}	-0.0059 (0.0182)	-0.0081 (0.0130)
β_{11}	0.0622 (0.6889)	0.0623 (0.5964)
β_{12}	0.0031 (0.0026)	0.0034 (0.0021)
σ_{11}	-0.1532 (0.0688)	-0.1064 (0.0463)
σ_{22}	-0.1631 (0.0725)	-0.1130 (0.0467)
σ_{33}	-0.1477 (0.0684)	-0.1189 (0.0557)
σ_{12}	0.1308 (0.0537)	0.0904 (0.0372)
σ_{13}	-0.0502 (0.0306)	-0.0322 (0.0266)
σ_{23}	0.0932 (0.0440)	0.0640 (0.0350)
T=200		
Coefficients	Likelihood	PW Likelihood
γ_{11}	-0.0158 (0.0074)	-0.0135 (0.0058)
γ_{12}	-0.0128 (0.0162)	-0.0129 (0.0130)
γ_{13}	-0.0167 (0.0082)	-0.0142 (0.0066)
β_{11}	0.1101 (0.5328)	0.1013 (0.4272)
β_{12}	0.0065 (0.0016)	0.0059 (0.0013)
σ_{11}	-0.0694 (0.0338)	-0.0505 (0.0222)
σ_{22}	-0.0655 (0.0286)	-0.0481 (0.0185)
σ_{33}	-0.0756 (0.0372)	-0.0622 (0.0304)
σ_{12}	0.0544 (0.0247)	0.0398 (0.0161)
σ_{13}	-0.0198 (0.0163)	-0.0142 (0.0136)
σ_{23}	0.0403 (0.0208)	0.0294 (0.0158)
T=500		
Coefficients	Likelihood	PW Likelihood
γ_{11}	-0.0061 (0.0033)	-0.0046 (0.0027)
γ_{12}	-0.0074 (0.0075)	-0.0064 (0.0059)
γ_{13}	-0.0097 (0.0040)	-0.0073 (0.0033)
β_{11}	0.0577 (0.2561)	0.0469 (0.2092)
β_{12}	0.0034 (0.0008)	0.0027 (0.0006)
σ_{11}	-0.0248 (0.0108)	-0.0189 (0.0076)
σ_{22}	-0.0274 (0.0108)	-0.0198 (0.0079)
σ_{33}	-0.0287 (0.0111)	-0.0246 (0.0097)
σ_{12}	0.0219 (0.0081)	0.0164 (0.0059)
σ_{13}	-0.0117 (0.0049)	-0.0074 (0.0045)
σ_{23}	0.0166 (0.0072)	0.0125 (0.0061)

Table 2 Comparison of simulation results under a fixed sample size ($T = 100$). Results are obtained considering A and Σ (correlated error terms) with different missing proportion configurations $\{0.05; 0.10; 0.25\}$. Bias and RMSE (in parentheses) are reported for the coefficients estimators. The simulation is based on 1000 replications.

proportion= 0.05				
Coefficients	PW Likelihood	Imputation PW Likelihood	Adj. Imput. PW Likelihood	
γ_{11}	-0.0103 (0.0058)	-0.0005 (0.0057)	-0.0064 (0.0058)	
γ_{12}	-0.0031 (0.0123)	-0.0079 (0.0125)	-0.0050 (0.0130)	
γ_{13}	-0.0104 (0.0078)	-0.0010 (0.0077)	-0.0078 (0.0079)	
β_{11}	0.0533 (0.4011)	0.0373 (0.4124)	0.0501 (0.4272)	
β_{12}	0.0033 (0.0013)	0.0013 (0.0013)	0.0027 (0.0014)	
σ_{11}	-0.0541 (0.0258)	0.0334 (0.0278)	0.0120 (0.0259)	
σ_{22}	-0.0581 (0.0258)	-0.0245 (0.0226)	-0.0325 (0.0227)	
σ_{33}	-0.0604 (0.0271)	-0.0215 (0.0227)	-0.0296 (0.0224)	
σ_{12}	0.0432 (0.0201)	0.0067 (0.0184)	0.0137 (0.0183)	
σ_{13}	-0.0179 (0.0136)	-0.0311 (0.0147)	-0.0294 (0.0166)	
σ_{23}	0.0351 (0.0162)	0.0285 (0.0155)	0.0251 (0.0155)	
proportion= 0.10				
Coefficients	PW Likelihood	Imputation PW Likelihood	Adj. Imput. PW Likelihood	
γ_{11}	-0.0101 (0.0068)	-0.0036 (0.0063)	-0.0037 (0.0065)	
γ_{12}	-0.0002 (0.0145)	-0.0204 (0.0145)	-0.0027 (0.0142)	
γ_{13}	-0.0081 (0.0087)	-0.0005 (0.0090)	-0.0029 (0.0092)	
β_{11}	0.0443 (0.4449)	0.0802 (0.4663)	0.0333 (0.4670)	
β_{12}	0.0024 (0.0016)	0.0038 (0.0016)	0.0011 (0.0016)	
σ_{11}	-0.0710 (0.0315)	0.1198 (0.0519)	0.0705 (0.0389)	
σ_{22}	-0.0716 (0.0328)	-0.0059 (0.0257)	-0.0201 (0.0252)	
σ_{33}	-0.0729 (0.0310)	-0.0065 (0.0246)	-0.0125 (0.0245)	
σ_{12}	0.0553 (0.0253)	-0.0219 (0.0240)	-0.0045 (0.0219)	
σ_{13}	-0.0262 (0.0161)	-0.0392 (0.0192)	-0.0492 (0.0218)	
σ_{23}	0.0451 (0.0200)	0.0218 (0.0171)	0.0218 (0.0178)	
proportion= 0.25				
Coefficients	PW Likelihood	Imputation PW Likelihood	Adj. Imput. PW Likelihood	
γ_{11}	-0.0094 (0.0126)	0.0190 (0.0096)	-0.0128 (0.0116)	
γ_{12}	-0.0021 (0.0249)	-0.0392 (0.0220)	-0.0202 (0.0250)	
γ_{13}	-0.0158 (0.0180)	0.0237 (0.0149)	-0.0068 (0.0163)	
β_{11}	0.0573 (0.7581)	0.0443 (0.6680)	0.1194 (0.8438)	
β_{12}	0.0036 (0.0026)	0.0005 (0.0022)	0.0058 (0.0028)	
σ_{11}	-0.1529 (0.0827)	0.4054 (0.2499)	0.2659 (0.1400)	
σ_{22}	-0.1496 (0.0731)	0.0497 (0.0414)	0.0141 (0.0347)	
σ_{33}	-0.1286 (0.0655)	0.0590 (0.0435)	0.0294 (0.0378)	
σ_{12}	0.1128 (0.0554)	-0.1041 (0.0530)	-0.0562 (0.0388)	
σ_{13}	-0.0488 (0.0383)	-0.1144 (0.0469)	-0.1049 (0.0451)	
σ_{23}	0.0867 (0.0427)	0.0352 (0.0264)	0.0196 (0.0241)	
proportion= {0.05, 0.10, 0.25}				
Coefficients	PW Likelihood	Imputation PW Likelihood	Adj. Imput. PW Likelihood	
γ_{11}	-0.0092 (0.0077)	0.0043 (0.0064)	-0.0066 (0.0065)	
γ_{12}	-0.0008 (0.0160)	-0.0552 (0.0158)	-0.0028 (0.0141)	
γ_{13}	-0.0108 (0.0115)	0.0029 (0.0089)	-0.0032 (0.0094)	
β_{11}	0.0402 (0.4762)	0.1619 (0.4411)	0.0326 (0.4568)	
β_{12}	0.0029 (0.0017)	0.0082 (0.0015)	0.0018 (0.0015)	
σ_{11}	-0.0946 (0.0431)	0.1917 (0.0762)	0.0862 (0.0391)	
σ_{22}	-0.0947 (0.0461)	-0.0023 (0.0269)	-0.0335 (0.0253)	
σ_{33}	-0.1156 (0.0553)	-0.0151 (0.0303)	-0.0366 (0.0284)	
σ_{12}	0.0741 (0.0348)	-0.0525 (0.0285)	-0.0053 (0.0218)	
σ_{13}	-0.0331 (0.0260)	-0.0387 (0.0214)	-0.0438 (0.0219)	
σ_{23}	0.0614 (0.0317)	0.0392 (0.0199)	0.0285 (0.0178)	

4 Conclusion

The present article extends the results of pairwise likelihood estimation in vector autoregression model introduced in Fonseca and Grassetti (2010). In order to recover some more information from the partially missing observations, we propose two possible imputation based approaches.

The estimators behaviour is evaluated by simulation studies. The obtained results enlighten the effect of imputation strategies on model estimation. The adjusted estimation of missing values obtains better results at least for the variance covariance parameters.

Next steps in the research regard the analytic analysis of efficiency gain, the study of asymptotic properties of the proposed method and a full comparison with existing imputation strategies.

References

- Buck, S. F.: A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer. *J Roy Stat Soc B Met*, 22 (2), 302-306 (1960)
- Fonseca, G. and Grassetti, L.: Pairwise likelihood for missing data treatment in VAR models. Working Paper N. 1, Dept. of Statistics (2010)
- Gómez, V., Maravall, A. and Peña, D.: Missing observations in ARIMA models: Skipping approach versus additive outlier approach. *J Econometrics*, 88 (2), 341–363 (1999)
- Harvey, A.C.: *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge (1990)
- Harvey, A.C. and Pierse, R.G.: Estimating missing observations in economic time series. *J Am Stat Assoc*, 79 (385), 125–131 (1984)
- Little, R.J.A. and Rubin, D.B.: *Statistical Analysis with Missing Data*, 2nd edition. New York; John Wiley and Sons (2002)
- Liu, M., Taylor, J.M.G. and Belin, T.R.: Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, 56, 1157–1163 (2000)
- Luceño, A.: Estimation of missing values in possibly partially nonstationary vector time series, *Biometrika*, 84 (2), 495–499 (1997)
- Lütkepohl H.: *New Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin (2005)
- Proietti, T.: Missing data in time series: A note on the equivalence of the dummy variable and the skipping approaches. *Stat Probabil Lett*, 78, 257–264 (2008)
- Varin, C.: On composite marginal likelihoods. *ASTA-Adv Stat Anal*, 92 (1), 1–28 (2008)