Biometrika (2009), **96**, 2, *pp.* 494–496
© 2009 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asp009
Advance Access publication 28 March 2009

Dimension reduction in time series and the dynamic factor model

By DANIEL PEÑA

Departamento de Estadística, Universidad Carlos III de Madrid, c/Madrid 126, 28903 Getafe, Spain

daniel.pena@uc3m.es

Summary

This note shows that the dimension reduction method proposed by Li & Shedden (2002) is equivalent to the dynamic factor model introduced by Peña & Box (1987).

Some key words: Common factor; Covariance matrix; Dynamic principal component; Reduced rank.

1. Introduction

Dimension reduction is important in modelling vector time series because the number of parameters in the model grows very fast with the dimension m of the vector of time series. Linear models usually have a number of parameters which grow with m^2; for instance, a vector autoregressive moving average model of orders p and q contains $m^2(p+q)$ parameters. For such models, dimension reduction was analyzed by Box & Tiao (1977), Tiao & Tsay (1989), Velu et al. (1986), Ahn & Reinsel (1990), Ahn (1997) and Reinsel & Velu (1998). A second approach to dimension reduction is through using dynamic factor models; see Peña & Box (1987), Stock & Watson (1988), Forni et al. (2000), Hu & Chou (2004) and Peña & Poncela (2006b), among others. The relationship between these methods has been studied by Peña & Poncela (2006a).

A different approach for dimension reduction in time series was proposed by Li & Shedden (2002), whose proposal seems to work well in a large dataset. In this note we show that this procedure is equivalent to the model proposed by Peña & Box (1987) for stationary time series. This relationship increases our understanding of alternative representations of time series models.

2. The dynamic factor model and the Li–Shedden model

Let y_t be an m-dimensional vector of observed stationary time series, assumed to have mean zero in order to simplify the presentation. The dynamic factor model is defined by two equations. The first equation explains the relationship between the data and the factors, $y_t = Pf_t + e_t$, where f_t is the r-dimensional vector of common factors, P is an $m \times r$ factor-loading matrix and e_t is normally distributed with zero-mean and full-rank diagonal covariance matrix Σ_e. The second equation gives the model for the vector of common factors, $\Phi(B)f_t = \Theta(B)a_t$, where $\Phi(B)=I - \Phi_1 B - \cdots - \Phi_p B^p$ and $\Theta(B) = I - \Theta_1 B - \cdots - \Theta_q B^q$ are $r \times r$ polynomial matrices, B is the backshift operator satisfying $By_t = y_{t-1}$, the roots of the determinantal equations $|\Phi(B)| = 0$ and $|\Theta(B)| = 0$ are outside the unit circle, and $a_t \sim N_r(0, \Sigma_a)$ with $\text{rank}(\Sigma_a) \geqslant 0$ and $E(a_t a_{t-h}^{\mathrm{T}}) = 0$, for $h \neq 0$. The factors have mean zero, an identity covariance matrix and follow a stationary vector autoregressive moving average, VARMA(p, q), model. We assume that $E(a_t e_{t-h}^{\mathrm{T}}) = 0$, for all $h = 0, \pm 1, \pm 2, \ldots$. For identification we will use $P^{\mathrm{T}}P = I$. See Peña & Box (1987) for further details.

Let $\Gamma_y(k) = E(y_{t-k} y_t^{\mathrm{T}})$ denote the $m \times m$ lagged k covariance matrix of the observed variables and $\Gamma_f(k) = E(f_{t-k} f_t^{\mathrm{T}})$ denote the $r \times r$ lagged k covariance matrix of the common stationary factors. Then an important property of the model is that

$$\Gamma_y(k) = P\Gamma_f(k)P^{\mathrm{T}}, \quad k \neq 0, \tag{1}$$

and the observed covariance matrices $\Gamma_y(k)$ have rank r for $k < 0$. As the factors can be assumed to be uncorrelated, the matrices $\Gamma_f(k)$ can be assumed to be diagonal, and all the data covariance matrices have as eigenvectors the columns of the loading matrix. If we first standardize the series so that each component has mean zero and unit standard deviation, then (1) applies to the relationship between the autocorrelation matrices of the data and the factors.

In the model proposed by Li & Shedden (2002) each component of the vector of time series, y_{it} $(i = 1, \ldots, m)$ is generated by

$$y_{it} = \sum_{j=1}^{r} \lambda_{ij} Z_{ji}(t), \tag{2}$$

where the λ_{ij} are unknown scalar values and the components, $Z_{ji}(t)$ $(j = 1, \ldots, r)$, are independent. Also, for each source component the values $Z_{ji}(t)$ $(i = 1, \ldots, m)$ are assumed to follow a common stationary distribution. Assuming that the series are standardized, writing (2) as $y_{it} = \lambda_i^{\mathrm{T}} Z_i(t)$, where $\lambda_i^{\mathrm{T}} = (\lambda_{i1}, \ldots, \lambda_{ir})$ and $Z_i(t) = \{Z_{1i}(t), \ldots, Z_{ri}(t)\}$, and setting $\rho_i(k) = E(y_{it}y_{it+k})/E(y_{it}^2)$, we have that $\rho_i(k) = \lambda_i^{\mathrm{T}} C(k)\lambda_i/(\lambda_i^{\mathrm{T}} C(0)\lambda_i)$, where $C(k)$ is the autocorrelation matrix of lag k for the unobserved components $Z(t)$. As these components are standardized and uncorrelated, the matrix $C(0)$ is the identity matrix and the matrices $C(k)$ for $k > 0$ are diagonal. Letting $r_j(k) = E\{Z_{ij}(t)Z_{ij}(t+k)\}$ denote the diagonal terms of these matrices, which represent the autocorrelations of the process $Z(t)$, we have

$$\rho_i(k) = \frac{\sum_{j=1}^{r} \lambda_{ij}^2 r_j(k)}{\sum_{j=1}^{r} \lambda_{ij}^2}.$$

Let $\rho_i = \{\rho_i(1), \ldots, \rho_i(k)\}^{\mathrm{T}}$ be the vector of autocorrelations for the ith observed series and let $R_j = \{r_j(1), \ldots, r_j(k)\}^{\mathrm{T}}$ be the same vector for the jth unobserved component. We have that $\rho_i = \sum_{j=1}^{r} w_{ij} R_j$, where $w_{ij} = \lambda_{ij}^2 / \sum_{j=1}^{r} \lambda_{ij}^2$ are weights and satisfy $\sum_{i=1}^{r} w_{ij} = 1$.

We now show that both models imply the same relationship for the observed autocorrelation matrices and therefore are equivalent. Assuming that the series are standardized, the factor model equation (1) implies that

$$\rho_i(k) = \sum_{j=1}^{r} p_{ij}^2 r_j(k),$$

where $r_j(k)$ are now the autocorrelations of the factors, and for the identification condition $P^{\mathrm{T}}P = I$ we have $\sum_{j=1}^{r} p_{ij}^2 = 1$. Thus, both models have the same empirical implications and are equivalent. In their procedure, Li & Shedden (2002) did not use the fact that their formulation implies that, letting $\rho_{ih}(k) = E(y_{it}y_{ht+k})/\{E(y_{it}^2)E(y_{ht}^2)\}^{1/2}$, we have

$$\rho_{ih}(k) = \frac{\lambda_i^{\mathrm{T}} C(k)\lambda_j}{\left(\lambda_i^{\mathrm{T}}\lambda_i\right)^{\frac{1}{2}} \left(\lambda_j^{\mathrm{T}}\lambda_j\right)^{\frac{1}{2}}},$$

which can also be obtained from (2).

3. Discussion

The equivalence of both formulations has several important implications for multiple time series analysis. First, in some applications of multivariate time series, as in signal processing problems, the time series are truly generated as linear combinations of r common components, which are generated by some stationary common distribution. In these models the components are not usually considered of interest and estimated because of their random nature. Understanding the equivalence shown in this note allows us to obtain factors which can be interpreted as average components, leading to a better understanding of their structure. Second, the comparison of both models has shown some properties of the observed autocorrelations which were not evident from either approach. Third, there are other multiple time series models that suppose that the series have been generated exactly by some components, such as independent component analysis (Hyvärinen et al., 2001). The equivalence shown in this note suggests that there is a close relationship between these models and non-Gaussian dynamic factor models. Fourth, the stationary dynamic factor model has been generalized to the nonstationary case by Peña & Poncela (2006b), thus enlarging the field of application of our results.

ACKNOWLEDGEMENT

This project was supported by a grant from the Ministerio de Educación y Ciencia, Spain.

REFERENCES

AHN, S. K. (1997). Inference of vector autoregressive models with cointegration and scalar components. *J. Am. Statist. Assoc.* **92**, 350–56.

AHN, S. K. & REINSEL, G. C. (1990). Estimation for partially nonstationary multivariate autoregressive models. *J. Am. Statist. Assoc.* **85**, 813–23.

BOX, G. & TIAO, G. (1977). A canonical analysis of multiple time series. *Biometrika* **64**, 355–65.

FORNI, M., HALLIN, M., LIPPI, M. & REICHLEIN, L. (2000). The generalized dynamic factor model: identification and estimation. *Rev. Econ. Statist.* **82**, 540–54.

HU, Y. & CHOU, R. (2004). On the Peña–Box model. *J. Time Ser. Anal.* **25**, 811–30.

HYVÄRINEN, A., KARHUNEN, J. & OJA, E. M. (2001). *Independent Component Analysis*. New York: John Wiley.

LI, K. C. & SHEDDEN, K. (2002). Identification of shared components in large ensembles of time series using dimension reduction. *J. Am. Statist. Assoc.* **97**, 759–65.

PEÑA, D. & BOX, G. (1987). Identifying a simplifying structure in time series. *J. Am. Statist. Assoc.* **82**, 836–43.

PEÑA, D. & PONCELA, P. (2006a). Dimension reduction in multivariate time series. In *Advances on Distribution Theory, Order Statistics and Inference, in Honor of B. C. Arnold*, Ed. N. Balakrishnan, E. Castillo and J. M. Sarabia. Boston: Birkhauser.

PEÑA, D. & PONCELA, P. (2006b). Nonstationary dynamic factor analysis. *J. Statist. Plan. Infer.* **136**, 1237–57.

REINSEL, G. C. & VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*. New York: Springer.

STOCK, J. H. & WATSON, M. W. (1988). Testing for common trends. *J. Am. Statist. Assoc.* **83**, 1097–107.

TIAO, G. C. & TSAY, R. S. (1989). Model specification in multivariate time series. *J. R. Statist. Soc.* B **51**, 157–213.

VELU, R. P., REINSEL, G. C. & WICHERN, D. W. (1986). Reduced rank models for multiple time series. *Biometrika* **73**, 105–18.

[*Received September* 2008. *Revised October* 2008]