

A. F. Zuur<sup>1</sup>, R. J. Fryer<sup>1</sup>, I.T. Jolliffe<sup>2</sup>, R. Dekker<sup>3</sup>, J. J. Beukema<sup>3</sup>

<sup>1</sup> FRS Marine Laboratory, PO Box 101, Victoria Road, Aberdeen AB11 9DB, UK.

<sup>2</sup> Department of Mathematical Science, University of Aberdeen, Aberdeen, UK.

<sup>3</sup> Netherlands Institute of Sea Research, Texel, The Netherlands.

Correspondence to A.F. Zuur: tel: +44 1224 295575; fax: +44 1224 295511; email: highstat@highstat.com

### *Abstract*

This paper discusses dynamic factor analysis, a technique for estimating common trends in multivariate time series. Unlike more common time series techniques such as spectral analysis and ARIMA models, dynamic factor analysis can analyse short, non-stationary time series containing missing values. Typically, the parameters in dynamic factor analysis are estimated by direct optimisation, which means that only small data sets can be analysed if computing time is not to become prohibitively long and the chances of obtaining sub-optimal estimates are to be avoided. This paper shows how the parameters of dynamic factor analysis can be estimated using the EM algorithm, allowing larger data sets to be analysed. The technique is illustrated on a marine environmental data set.

Keywords: Dynamic factor analysis, EM algorithm, multivariate time series analysis, common trends

## **1 Introduction**

Common questions in biological and environmental time series studies are (i) what are the general patterns over time in the measured variables, (ii) are there any interactions between the measured variables, and (iii) are the measured variables related to any explanatory variables. Unfortunately, most commonly used time series techniques, such as spectral analysis (Priestley, 1981), wavelet analysis (Shumway and Stoffer, 2000), ARIMA and Box-Jenkins models (Ljung, 1987), require stationary time series without missing values and are not particularly suitable for answering

such questions. Indeed, spectral analysis and wavelet analysis concentrate on cyclic patterns, and ARIMA and Box-Jenkins models are designed merely for prediction.

With modern computing power, more computing-intensive time series techniques can now be applied to short non-stationary multiple time series. One of these techniques is dynamic factor analysis, the subject of this paper. It is a dimension reduction technique that aims to model  $N$  observed time series in terms of  $M$  common trends. The aim is to choose  $M$  as small as possible, without losing too much information. The principle is the same as in other dimension reduction techniques, such as principal component analysis and factor analysis. Indeed, dynamic factor analysis can be thought of as a factor analysis in which the axes are restricted to be latent smoothing functions over time.

Dynamic factor analysis is not a new technique. It has been used in econometric and psychological related fields since the mid eighties, see for example Molenaar (1985, 1989, 1993), Molenaar and de Gooijer (1988), Molenaar *et al.* (1988, 1992, 1999), Harvey (1989), Lütkepohl (1991). However, most of the examples in these papers contain only a few time series. This is because the model parameters were estimated by direct numerical optimisation of a maximum likelihood criterion. Clearly, as the number of time series becomes large, so does the number of parameters to be estimated, and direct optimisation becomes harder and more time consuming. More recently, Aguilar *et al.* (1998), and West and Hall (1997), among others, have used Markov Chain Monte Carlo methods for parameter estimation. However, these methods require long time series, and the authors used examples with hundreds of observations in time.

This paper shows how the EM algorithm can be used to apply dynamic factor analysis to a larger number of time series. To do this, we write the dynamic factor model as a special form of a state-space model, and adapt the theory of Shumway and Stoffer (1982), and Wu *et al.* (1996), who discuss the EM algorithm for various types of state-space models. We also show how to deal with missing values, using techniques described in Digilakis *et al.* (1993) and Shumway and Stoffer (1982), and to incorporate explanatory variables, again following Wu *et al.* (1996). These result in a complete, unified description of dynamic factor analysis within the EM framework.

Section 2 describes the dynamic factor model. Section 3 develops the EM algorithm for the dynamic factor model and Section 4 discusses model identification. In Section 5, explanatory variables are added to the dynamic factor model, and missing values and other topics are discussed in Section 6. In Section 7, dynamic factor analysis is applied to a set of zoobenthic species measured at the Balgzand in the Netherlands (Beukema, 1998).

The mathematical algorithms were implemented in a user friendly software package and this is available from <http://www.brodgar.com>.

## 2 Dynamic Factor Analysis

Dynamic factor analysis is based on so-called structural time series models (Harvey, 1989). These model observations in terms of a trend, seasonal effects, a cycle, explanatory variables and noise, all of which are allowed to be stochastic. This means that one can have a seasonal component that changes slightly from year to year, a cyclic component that is not a cosine function, a trend that is not a straight line or a polynomial, or explanatory variables that have a significant influence in a certain period of the time series. In this paper, we only consider structural time series models of the form:

$$\text{data} = \text{trends} + \text{explanatory variables} + \text{noise} \quad (1)$$

The mathematical formulation of this model is discussed next.

First suppose that we have a univariate response variable  $y_t$  measured in year  $t$ , where  $t = 1, \dots, T$ . The most simple univariate structural time series model has no explanatory variables, and is given by:

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t \\ \alpha_t &= \alpha_{t-1} + \eta_t \end{aligned}$$

This model is called a random walk trend plus noise model. The term  $\alpha_t$  represents the unknown trend at time  $t$ . The components  $\epsilon_t$  and  $\eta_t$  are error components. It is assumed that  $\epsilon_t \sim N(0, h)$ ,  $\eta_t \sim N(0, q)$ , and  $\alpha_0 \sim N(a_0, v_0)$ . It is also generally assumed that  $\epsilon_t$ ,  $\eta_t$  and  $\alpha_0$  are independent of each other, but this is not strictly necessary. The variance  $q$  determines the smoothness of the trend component, with the trend becoming smoother as  $q$  decreases towards zero.

If there are now  $N$  response variables, then these can still be analysed by univariate models by treating them as  $N$  separate time series. However, this results in  $N$  estimated trends that all have to be interpreted, and interactions between the response variables are ignored. The dynamic factor model aims to overcome these disadvantages by reducing the  $N$  univariate trends to  $M$  common trends, where  $1 \leq M < N$ . To illustrate, consider the dynamic factor model for two common trends:

$$\begin{aligned} \begin{bmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{bmatrix} &= \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \vdots & \vdots \\ \gamma_{N1} & \gamma_{N2} \end{bmatrix} \begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \end{bmatrix} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t \\ \begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \end{bmatrix} &= \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{bmatrix} + \begin{bmatrix} \eta_{1t} \\ \eta_{2t} \end{bmatrix} \end{aligned}$$

where  $y_{it}$  is the value of the  $i$ th response variable at time  $t$  ( $i = 1, \dots, N$  and  $t = 1, \dots, T$ ),  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)'$  is a level parameter and  $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{Nt})'$  is noise. The components  $\alpha_{1t}$  and  $\alpha_{2t}$  are the two common trends. Every response variable is modelled as the sum of (i) a linear combination of the two trends, (ii) a constant level parameter and (iii) a noise component. The multiplication factors  $\gamma_{i1}$  and  $\gamma_{i2}$  determine the linear combination for the  $i$ th variable and are called factor loadings.

A general formulation for the dynamic factor model with  $M$  common trends is given by:

$$\mathbf{y}_t = \mathbf{\Gamma}\boldsymbol{\alpha}_t + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t \quad (2)$$

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \quad (3)$$

The matrix  $\mathbf{\Gamma}$  is of dimension  $N \times M$  and contains the unknown factor loadings and  $\boldsymbol{\alpha}_t$  is a vector of dimension  $M$  containing the  $M$  common trends at time  $t$ . It is generally assumed that  $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{H})$ ,  $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q})$  and  $\boldsymbol{\alpha}_0 \sim N(\mathbf{a}_0, \mathbf{V}_0)$ , and that  $\boldsymbol{\epsilon}_t$ ,  $\boldsymbol{\eta}_t$  and  $\boldsymbol{\alpha}_0$  are independent of each other, but again the independence assumption is not strictly necessary. The unknown parameters in the model are the elements of  $\mathbf{\Gamma}$ ,  $\mathbf{H}$ ,  $\mathbf{Q}$ ,  $\boldsymbol{\mu}$ ,  $\mathbf{a}_0$  and  $\mathbf{V}_0$  and are called hyperparameters.

Conditional on the hyperparameters, the variance of  $\mathbf{y}_t$  is given by:

$$\text{var}(\mathbf{y}_t) = \mathbf{\Gamma}\text{var}(\boldsymbol{\alpha}_t)\mathbf{\Gamma}' + \mathbf{H}$$

In factor analysis, the variance of the observations has a similar form, so the model in equations (2) and (3) is called the dynamic factor model.

### 3 Dynamic Factor Analysis and the EM algorithm

The joint log likelihood function of the observations  $\mathbf{y}_1, \dots, \mathbf{y}_T$  and the trend components  $\boldsymbol{\alpha}_0, \dots, \boldsymbol{\alpha}_T$  is given by:

$$\begin{aligned} \log L(\mathbf{y}_1, \dots, \mathbf{y}_T, \boldsymbol{\alpha}_0, \dots, \boldsymbol{\alpha}_T) &= -\frac{1}{2} \log |\mathbf{V}_0| - \frac{1}{2} (\boldsymbol{\alpha}_0 - \mathbf{a}_0)' \mathbf{V}_0^{-1} (\boldsymbol{\alpha}_0 - \mathbf{a}_0) \\ &\quad - \frac{T}{2} \log |\mathbf{Q}| - \frac{1}{2} \sum_{t=1}^T (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1})' \mathbf{Q}^{-1} (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}) \\ &\quad - \frac{T}{2} \log |\mathbf{H}| + \text{constant} \\ &\quad - \frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{\Gamma}\boldsymbol{\alpha}_t - \boldsymbol{\mu})' \mathbf{H}^{-1} (\mathbf{y}_t - \mathbf{\Gamma}\boldsymbol{\alpha}_t - \boldsymbol{\mu}) \end{aligned}$$

This log likelihood function is also called the complete data likelihood. Because the trend components are unknown,  $\log L(\mathbf{y}_1, \dots, \mathbf{y}_T, \boldsymbol{\alpha}_0, \dots, \boldsymbol{\alpha}_T)$  can not be optimised directly. However, the EM algorithm provides a way to obtain maximum likelihood estimates of the hyperparameters based on the incomplete data  $\mathbf{y}_1, \dots, \mathbf{y}_T$ . This is done by successively maximising the conditional expectation of the complete data likelihood function. In the E-step, we calculate

$$\text{E} [ \log L(\mathbf{y}_1, \dots, \mathbf{y}_T, \boldsymbol{\alpha}_0, \dots, \boldsymbol{\alpha}_T) \mid \mathbf{y}_1, \dots, \mathbf{y}_T, \boldsymbol{\phi}^{j-1} ] \quad (4)$$

where  $\boldsymbol{\phi}$  contains all the hyperparameters estimated in the  $j-1^{\text{th}}$  iteration. In the M-step, (4) is maximised with respect to the hyperparameters. We now discuss the E-step and M-step in more detail.

### E-step

Shumway and Stoffer (2000) showed that the conditional expectation of the complete data likelihood function in (4) is equal to:

$$\begin{aligned}
& -\frac{1}{2} \log |\mathbf{V}_0| - \frac{1}{2} \text{tr} \{ \mathbf{V}_0^{-1} (\mathbf{V}_{0|T} + \boldsymbol{\alpha}_0 - \mathbf{a}_0) (\boldsymbol{\alpha}_0 - \mathbf{a}_0)' \} \\
& -\frac{1}{2} \log |\mathbf{Q}| - \frac{1}{2} \text{tr} \{ \mathbf{Q}^{-1} (\mathbf{C} - 2\mathbf{B} + \mathbf{A}) \} \\
& -\frac{1}{2} \log |\mathbf{H}| + \text{constant} \\
& -\text{tr} \{ \mathbf{H}^{-1} \frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_t - \boldsymbol{\mu})' (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_t - \boldsymbol{\mu}) + \boldsymbol{\Gamma} \mathbf{V}_{t|T} \boldsymbol{\Gamma}' \}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A} &= \sum_{t=2}^T \boldsymbol{\alpha}_{t-1|T} \boldsymbol{\alpha}'_{t-1|T} + \mathbf{V}_{t-1|T} \\
\mathbf{B} &= \sum_{t=2}^T \boldsymbol{\alpha}_{t|T} \boldsymbol{\alpha}'_{t-1|T} + \mathbf{V}_{t,t-1|T} \\
\mathbf{C} &= \sum_{t=2}^T \boldsymbol{\alpha}_{t|T} \boldsymbol{\alpha}'_{t|T} + \mathbf{V}_{t|T}
\end{aligned}$$

The terms  $\boldsymbol{\alpha}_{t|T}$  and  $\mathbf{V}_{t|T}$  represent the best linear estimator for  $\boldsymbol{\alpha}_t$ , using all observations, and the corresponding variance matrix respectively.  $\mathbf{V}_{t,t-1|T}$  is the covariance matrix of  $\boldsymbol{\alpha}_{t|T}$  and  $\boldsymbol{\alpha}_{t-1|T}$ . These are all obtained from the Kalman filter and smoother algorithm Shumway and Stoffer (2000) applied to the model in (2) and (3). Details of the algorithm for the dynamic factor model extended to include explanatory variables, are given in the Appendix.

### M-step

Updating equations for  $\mathbf{Q}$ ,  $\mathbf{H}$ ,  $\boldsymbol{\Gamma}$ ,  $\boldsymbol{\mu}$ ,  $\boldsymbol{\alpha}_0$  and  $\mathbf{V}_0$  are obtained by maximising the conditional expectation of the complete data likelihood function in (4) with respect to these hyperparameters. Using basic calculus it can be shown that the following choices for  $\mathbf{H}$ ,  $\mathbf{Q}$ ,  $\boldsymbol{\mu}$ ,  $\mathbf{V}_0$  and  $\mathbf{a}_0$  maximise (4).

$$\mathbf{Q} = T^{-1} (\mathbf{C} - 2\mathbf{B} + \mathbf{A}) \quad (5)$$

$$\mathbf{H} = T^{-1} \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|T} - \boldsymbol{\mu}) (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|T} - \boldsymbol{\mu})' + \boldsymbol{\Gamma} \mathbf{V}_{t|T} \boldsymbol{\Gamma}' \quad (6)$$

$$\mathbf{a}_0 = \boldsymbol{\alpha}_{0|T}$$

$$\mathbf{V}_0 = \mathbf{V}_{0|T}$$

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|T}) \quad (7)$$

$$\boldsymbol{\Gamma} = \mathbf{E}_2 \mathbf{E}_1^{-1} \quad (8)$$

where

$$\mathbf{E}_1 = \sum_{t=1}^T (\boldsymbol{\alpha}_{t|T} \boldsymbol{\alpha}'_{t|T} + \mathbf{V}_{t|T}) \text{ and } \mathbf{E}_2 = \sum_{t=1}^T (\mathbf{y}_t \boldsymbol{\alpha}'_{t|T} - \boldsymbol{\mu} \boldsymbol{\alpha}'_{t|T})$$

## 4 Identification

The dynamic factor model is not identifiable (Harvey, 1989) since, if  $\mathbf{F}$  is a non-singular matrix, the (rotated) factor loadings  $\mathbf{\Gamma}\mathbf{F}^{-1}$  and common trends  $\mathbf{F}\boldsymbol{\alpha}_t$  will give the same model fit. This problem also exists in factor analysis, where it is resolved by imposing restrictions on the parameters. Harvey (1989) suggested using similar restrictions in dynamic factor analysis, namely that (i) the covariance matrix of the common factors is the identity matrix ( $\mathbf{Q} = \mathbf{I}$ ), (ii) the  $ij$ th element of  $\mathbf{\Gamma}$  is zero for  $j > i$ , where  $i = 1, \dots, M - 1$ , and (iii) the first  $M$  elements of the level parameter are equal to 0. Using these restrictions, a unique solution for the factor loadings exists. The drawback is that the first common trend is determined by the first response variable, the second common trend by the first two response variables, etc. However, once the parameters have been estimated, a factor rotation can be applied to the estimated factor loadings and common trends. For example, a varimax rotation (Basilevsky, 1994) will attempt to relate each time series to just one of the common trends. Implementing these restrictions within the EM algorithm is not trivial and is discussed below.

### 1. Restrictions on $\mathbf{\Gamma}$

Wu *et al.* (1996) extended the EM methodology to allow for restrictions on the hyperparameters. They used a model similar to (2)-(3) but with a known matrix  $\mathbf{\Gamma}$  and with restrictions placed on the covariance matrices. Their approach can be extended to allow for restrictions on  $\mathbf{\Gamma}$ . The derivation is based on complex algebra and we only present the final updating equation. Let  $\text{vec } \mathbf{\Gamma}$  be the  $NM \times 1$  vector containing all stacked columns of  $\mathbf{\Gamma}$ . The restrictions on  $\mathbf{\Gamma}$  can be written as:

$$\mathbf{G}\text{vec } \mathbf{\Gamma} = \mathbf{0}$$

where  $\mathbf{G}$  is a known matrix containing zeros and ones and  $\mathbf{0}$  is a vector containing zeros. The number of rows of  $\mathbf{G}$  (and  $\mathbf{0}$ ) is equal to the number of restricted elements in  $\mathbf{\Gamma}$ . Using this notation, the updating equation for the restricted matrix,  $\mathbf{\Gamma}_{restr}$  becomes:

$$\mathbf{\Gamma}_{restr} = \mathbf{\Gamma}_{unrestr} + (\mathbf{E}_1^{-1} \otimes \mathbf{H})\mathbf{G}'(\mathbf{G}(\mathbf{E}_1 \otimes \mathbf{H})\mathbf{G}')^{-1}(\mathbf{0} - \mathbf{G}\text{vec } \mathbf{\Gamma}_{unrestr})$$

where  $\mathbf{\Gamma}_{unrestr}$  is obtained from equation (8) and  $\otimes$  is the tensor product.

### 2. Restrictions on $\mathbf{Q}$

No updating equations for  $\mathbf{Q}$  are now needed, since we have assumed that  $\mathbf{Q} = \mathbf{I}$ . This means that we can omit equation (5).

### 3. Restrictions on the level parameters

Setting the first  $M$  level parameters to zero, as mentioned above, means that the first  $M$  elements of  $\boldsymbol{\mu}$  are set to zero. Another option is to set all elements of  $\mathbf{a}_0$  to zero and use no restrictions on  $\boldsymbol{\mu}$ . We followed another approach. After each Kalman smoothing iteration, we calculated the average of each common trend. Denote these by  $\bar{\boldsymbol{\alpha}}$ . This average is subtracted from  $\boldsymbol{\alpha}_{s|T}$  for all  $s$  (in each iteration, before the M-step). As a result, the common trends are centered around zero. This means that the estimated level parameter represents  $\mathbf{\Gamma}\bar{\boldsymbol{\alpha}} + \boldsymbol{\mu}$ . We found that this approach resulted in a much more stable algorithm.

## 5 Explanatory variables

To include explanatory variables in the dynamic factor model, equations (2)-(3) can be extended to:

$$\mathbf{y}_t = \mathbf{\Gamma}\boldsymbol{\alpha}_t + \mathbf{D}\mathbf{x}_t + \boldsymbol{\epsilon}_t \quad (9)$$

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \quad (10)$$

where  $\mathbf{D}$  is a  $N \times K$  matrix containing unknown regression parameters and  $\mathbf{x}_t$  is a  $K \times 1$  vector containing the values of the  $K$  explanatory variables at time  $t$ . To simplify notation, we let the first explanatory variable equal 1 for all  $t$ , so the first column of  $\mathbf{D}$  represents the level parameter  $\boldsymbol{\mu}$ . This modification means that  $\boldsymbol{\mu}$  must be replaced by  $\mathbf{D}\mathbf{x}_t$  in the expression for the conditional expectation of the complete data likelihood function, and in the updating equations for  $\mathbf{H}$  and  $\mathbf{\Gamma}$ . Wu *et al.* (1996) derived an updating equation for  $\mathbf{D}$ :

$$\mathbf{D} = \sum_{t=1}^T (\mathbf{y}_t \mathbf{x}_t' - \mathbf{\Gamma} \boldsymbol{\alpha}_{t|T} \mathbf{x}_t') \left( \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \quad (11)$$

It is also possible to put restrictions on particular elements of  $\mathbf{D}$ . This might be necessary if it is believed that a certain explanatory variable only affects a subgroup of response variables. Mathematical details can be found in Wu *et al.* (1996).

To assess whether the response variables are related to the explanatory variables, confidence intervals for  $\mathbf{D}$  need to be calculated. Unfortunately, the standard EM algorithm does not provide the means to do this. Meng and Rubin (1991) developed SEM, an extension of the EM method to estimate confidence intervals of hyperparameters. Alternatively, bootstrapping methods can be used (e.g. Stoffer and Wall (1991)). Our approach, slightly easier and less computationally demanding, was to rerun the EM algorithm once more after convergence, but with the state-space model rewritten so that the new state-vector  $\boldsymbol{\alpha}_t^*$  contains the common trends and the stacked columns of  $\mathbf{D}$ . The variance matrix of  $\boldsymbol{\alpha}_t^*$  can then be used to calculate confidence intervals for the elements of  $\mathbf{D}$ . This approach works as follows.

Apply the EM algorithm as described in the appendix. Define a new vector  $\boldsymbol{\alpha}_t^*$  of dimension  $M + NK$ , where the first  $M$  elements of  $\boldsymbol{\alpha}_t^*$  contain the estimated values of  $\boldsymbol{\alpha}_{t|T}$  and the remaining elements the stacked columns of the estimated matrix  $\mathbf{D}$ . Define new matrices  $\mathbf{\Gamma}_t^*$  of dimension  $N \times (M + NK)$ , where the first  $M$  columns contain the estimated values of  $\mathbf{\Gamma}$  and the remaining columns contain  $K$  diagonal matrices of dimension  $N \times N$ . The first diagonal matrix contains the element  $x_{1t}$  (each element on the diagonal is the same), the last diagonal matrix the element  $x_{Kt}$ . By doing this, we have rewritten the dynamic factor model in (9)-(10) as :

$$\begin{aligned} \mathbf{y}_t &= \mathbf{\Gamma}_t^* \boldsymbol{\alpha}_t^* + \boldsymbol{\epsilon}_t \\ \boldsymbol{\alpha}_t^* &= \boldsymbol{\alpha}_{t-1}^* + \boldsymbol{\eta}_t^* \end{aligned}$$

where  $\boldsymbol{\eta}_t^* \sim N(\mathbf{0}, \mathbf{Q}^*)$  and  $\mathbf{Q}^*$  is a diagonal matrix. Set the first  $M$  elements of  $\mathbf{Q}^*$  equal to 1 and the remaining elements to 0. Using all estimated hyperparameters, run the Kalman smoothing algorithm. Starting values for  $\boldsymbol{\alpha}_0^*$  are obtained from the estimated values of  $\boldsymbol{\alpha}_0$  and the stacked columns of the estimated values of  $\mathbf{D}$ . The estimated covariance matrix of  $\boldsymbol{\alpha}_{t|T}^*$  can be used to determine confidence intervals for the estimated values of  $\mathbf{D}$ .

## 6 Miscellaneous

### Modifications for missing values

The algorithm for Kalman filtering and smoothing can easily be adapted for missing values. To account for missing values at time  $t$  say, a  $N_t \times N$  design matrix  $\mathbf{W}_t$  is used, where  $N_t$  is the number of non-missing observations at time  $t$ . For example, if  $N = 5$  and at time  $t = 1$  the second and fourth observations are missing,  $\mathbf{W}_1$  is equal to:

$$\mathbf{W}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The matrix  $\mathbf{\Gamma}$  is then replaced by  $\mathbf{W}_t\mathbf{\Gamma}$  in the Kalman filter and smoother algorithm.

Updating the hyperparameters in the case of missing values is more complex. Let  $\mathbf{L}_t$  be the  $N \times N$  diagonal matrix with the  $i$ , $i$ th element equal to 0 if the  $i$ th element of  $\mathbf{y}_t$  is missing and 1 otherwise. The variance matrix  $\mathbf{H}$  is updated via:

$$\begin{aligned} \mathbf{H} &= T^{-1} \sum_{t=1}^T \mathbf{L}_t [(\mathbf{y}_t - \mathbf{\Gamma}\boldsymbol{\alpha}_{t|T} - \mathbf{D}\mathbf{x}_t)(\mathbf{y}_t - \mathbf{\Gamma}\boldsymbol{\alpha}_{t|T} - \mathbf{D}\mathbf{x}_t)' + \mathbf{\Gamma}\mathbf{V}_{t|T}\mathbf{\Gamma}'] \mathbf{L}_t \\ &\quad + (\mathbf{I} - \mathbf{L}_t)\mathbf{H}_{prev}(\mathbf{I} - \mathbf{L}_t) \end{aligned}$$

where  $\mathbf{H}_{prev}$  is the matrix  $\mathbf{H}$  obtained in the previous EM iteration and missing values in  $\mathbf{y}_t$  are replaced by zeros. This approach uses information from the previous EM-iteration for rows and columns that correspond to the missing values in  $\mathbf{y}_t$  (Shumway and Stoffer, 1982). For updating  $\mathbf{D}$  in the case of missing values, we followed Digalakis *et al.* (1993), and replaced the missing values in  $\mathbf{y}_t$  by the corresponding values in  $\mathbf{\Gamma}\boldsymbol{\alpha}_{t|T} + \mathbf{D}\mathbf{x}_t$ . We followed the same approach for updating  $\mathbf{\Gamma}$ .

### Number of common trends

Just as in dimension reduction techniques such as principal component analysis and factor analysis, one has to choose the number of axes or common trends. The more common trends are used, the better the fit will be, but the more parameters that have to be estimated. The number of parameters in the dynamic factor model is given in Table 1. A convenient choice is to use 2 common trends since factor loadings can then be plotted versus each other. Obviously, it is desirable to have a more formal procedure for choosing  $M$ . We used Akaike's information criterion (AIC), defined as twice the difference between the log likelihood function (measure of fit) and the number of parameters (penalty). The AIC can be calculated for models containing any number of common trends, and the model with the smallest AIC is taken to be the 'best' model. Once the common trends have been estimated, various model validation tools are also available. For example, various types of residual plots can be considered, as in classical linear regression (Johnston, 1984).

### Starting values

Shumway and Stoffer (2001) showed that  $\mathbf{V}_0$  and  $\boldsymbol{\alpha}_0$  can not be estimated simultaneously within the EM algorithm. He suggested that  $\mathbf{V}_0$  should be kept fixed

and  $\alpha_0$  estimated with the EM algorithm. We followed this suggestion and used a diagonal matrix for  $\mathbf{V}_0$  with all diagonal elements set to a fixed value  $v$ . There are no general guidelines for choosing  $v$ . We noticed that if  $v$  is chosen large relative to  $\alpha_0$ , but not too large, the estimation process resulted in similar results for different choices of  $v$ . For standardised response variables (see below), we used  $v = 5$ . Other values (between 1 and 15) were tested but negligible differences were detected in the final model estimates.

### Covariance matrix $\mathbf{H}$

Various choices are possible for the covariance matrix  $\mathbf{H}$  of the error term  $\epsilon_t$ . For example Liang and Zeger (1986) discuss different covariance structures in a generalised estimation equation context, all of which can be used in dynamic factor analysis. The simplest approach is to use a diagonal matrix. Wu *et al.* (1996) showed that the diagonal elements of  $\mathbf{H}$  can then be updated with equation (6), with the off-diagonal elements being set to 0. We have also used a general symmetric, non-diagonal covariance matrix, where off-diagonal elements represent information in the response variables which cannot be explained by the common trends or explanatory variables.

### Standardising

It is often convenient to standardise each time series by subtracting its mean and dividing by its standard deviation. This can be written as:

$$\mathbf{y}_t^* = \Sigma^{-1}(\mathbf{y}_t - \bar{\mathbf{y}})$$

where  $\mathbf{y}_t^*$  is the transformed series,  $\bar{\mathbf{y}}$  is a  $N \times 1$  vector containing the mean values of the  $N$  time series and  $\Sigma$  is a  $N \times N$  diagonal matrix containing standard deviations. As a result, the first equation of the dynamic factor model for standardised data can be rewritten as:

$$\mathbf{y}_t^* = \Gamma^* \alpha_t + \mathbf{D} \mathbf{x}_t^* + \epsilon_t^*$$

where  $\Gamma^* = \Sigma^{-1} \Gamma$ ,  $\mathbf{D} \mathbf{x}_t^* = \Sigma^{-1} \mathbf{D} \mathbf{x}_t - \Sigma^{-1} \bar{\mathbf{y}}$  and  $\epsilon_t^* = \Sigma^{-1} \epsilon_t$ . This shows that the only effect of standardising the time series is a matrix multiplication of the factor loadings, regression coefficients and noise term. However, if the time series are on different scales it is an advantage to standardise them and interpret the factor loadings  $\Sigma^{-1} \Gamma$  instead of  $\Gamma$ . If this is not done,  $\Gamma$  might only reflect differences in the scale of the series, rather than interactions between series.

## 7 Case study: biomass of macro zoobenthic species in the Balgzand

The Balgzand is a 50 km<sup>2</sup> tidal flat area in the western part of the Wadden Sea (Figure 1). In this area, fifteen stations were selected and have been sampled at least annually since 1970. At each station, numbers and biomass of macro zoobenthic species have been counted and measured. Details of the sampling are given in Beukema (1988).

To illustrate dynamic factor analysis, we use the biomass (summed over the 15 stations) of 12 important species sampled in March of each year. The species are

given in Table 2. Because various species had biomass series which fluctuated considerably over years, a square root transformation was applied. Water temperature has been measured at a nearby pier and winter averages were used as an explanatory variable. The species and temperature time series were standardised. Underlying questions in this study are whether there are any common patterns in the biomass time series, and what are the relationships between biomass time series and water temperature. The standardised species and water temperature time series are presented in Figure 2.

## Results

The following two dynamic factor analysis models were applied.

Model I: data = common trends + noise

Model II: data = common trends + explanatory variable + noise

Model I is given by equations (2)-(3) and model II by equations (9)-(10). In all models, a symmetric, non-diagonal matrix  $\mathbf{H}$  was used for the noise component. The AIC, log likelihood function and number of parameters using  $M$  common trends for models I and II are given in Table 3. Models with  $N$  univariate trends were used as well. The AIC values indicated that the model containing 3 common trends and temperature as an explanatory variable is the best model. Differences between AIC values using 2 and 3 common trends in model II were small. However, a model validation on both models, which included residual plotting, model fits, etc., indicated that the model containing 3 common trends was slightly better.

The estimated parameters and  $t$ -values for the explanatory variable temperature, using 3 common trends in model II, are given in Table 4. Results indicate that temperature had a significant influence on the species *N. hombergii*, *L. conchilega*, *C. edule*, *P. mucosa* and in a lesser extend *H. ulvae*, *M. balthica* and *M. edulis*. Values of water temperature were well above or below average in 1979, 1988-1990 and 1996 (Figure 2). The values for *N. hombergii*, *L. conchilega* were below or above average in the same periods as temperature, explaining why temperature was important for these species.

The estimated three common trends are presented in Figure 3 and the factor loadings in Figure 4. Factor loadings with values smaller than 0.1 were not displayed. To simplify interpretation of the results, coefficients of temperature with  $t$ -values larger than 3 (in absolute sense) and the largest factor loadings for each axis are given again in Table 5. Presenting the results in this way shows clearly which common trends is related to which species and vice versa. For example, *H. filiformis* and *H. ulvae* were both mainly determined by the second trend and *C. edule* and *L. conchilega* were only related to temperature.

The first common trend is mainly related to the species *A. marina*, *M. balthica*, *Nereis spp.* and *P. mucosa*. This trend is characterised by relatively constant values between 1970-1977, a rapid decrease (except for 1984) until 1989, followed by a rapid increase. Because all the large factor loadings for this axis are negative, the effect of this trend is actually the opposite. The second common trend is related to the species *H. filiformis*, *H. ulvae*, *S. armiger*, *M. arenaria* and *Nereis spp.* (all have positive loadings), and *A. marina* and *N. hombergii*. The last two species have a

negative sign, which means that their pattern is opposite to the second common trend. This trend is approximately constant between 1970 and 1976, followed by an increase up to 1981, no great changes between 1981-1989, a small dip in 1990, and a rapid increase from there onwards. The third common trend was related to the species *A. marina* (negative loading) and *M. edulis* and it shows an increase from 1970 until 1981, and a decrease thereafter.

A few species were determined by two common trends, namely *A. marina* and *M. arenaria*. Comparing the fitted values with the common trend and factor loadings, one can clearly see how the fitted curves are composed and why. The same holds if a species is determined by temperature and a common trend. For example, *N. hombergii* was determined by both temperature and the (opposite) second common trend. *L. conchilega* was determined by temperature only. Comparing these two fits, one can recognise the decline in *N. hombergii* which is caught by the second common trend. Similar conclusions can be made for other species.

Because the trends are estimated simultaneously, it cannot be said which common trend is more important. However, a feel for this can be obtained by first fitting a model with one common trend and then a model with two common trends, and comparing the trends. For example, the common trend estimated by model II with  $M = 1$  looks very similar to the second common trend in Figure 3, indicating that the second common trend in model II with  $M = 3$ , is the most important one. Model II with  $M=2$  resulted in 2 common trends which were very similar to the first two common trends in Figure 3. Hence, the third common trend in this figure is the least important one. The fitted curves are given in Figure 5. Most species are fitted reasonably well.

## 8 Discussion and Conclusions

In this paper, we have shown how the EM method can be used to estimate the parameters and common trends of the dynamic factor model. As a result, relative large numbers of short, non-stationary time series with missing values can be analysed. Parameter estimation time with the EM method for the Balgzand data took only a few minutes on a 400 MhZ notebook. With direct optimisation of the likelihood, parameter estimation for data sets of this size would take several hours. Our approach makes it possible for applied scientists to use dynamic factor analysis as a standard time series analysis tool.

Dynamic factor analysis was applied to 12 zoobenthic species monitored in the Balgzand. Results indicated that temperature was strongly related to the species *N. hombergii*, *L. conchilega*, *C. edule* and *P. mucosa*. The relationship between temperature and *N. hombergii* and *L. conchilega* was reported in Beukema (1979). Other species known to be influenced by temperature are the epi-benthic species *H. ulvae* and *M. edulis* (Blegvad, 1929), and we also found a weak relationship with temperature for these species. The second common trend was the most important trend. This trend was related to the species *H. filiformis*, *H. ulvae*, *M. arenaria* and *S. armiger* (all have positive loadings) and *A. marina* and *N. hombergii* (negative loadings). The fit of the first four species was very similar. The species *S. armiger*

and *H. filiformis* are prey for *N. hombergii*, which might explain why *N. hombergii* has a negative loading for this trend. There is no clear biological reason why *A. marina* has a negative loading for this trend as well. However, the fitted curve for this species seem to be determined mainly by the first common trend.

The Balgzand data set has been the subject of a large number of publications, e.g. Beukema (1974, 1979, 1984, 1992), Beukema *et al.* (2000) among others. However, these papers have typically presented results for pre-selected subsets of species. Dynamic factor analysis on the other hand, does not require any major pre-selection of species, and this provides an objective assessment of trends in the series and relationships with explanatory variables throughout the entire data set.

There were also various aspects obtained by dynamic factor analysis which have not been reported in the literature so far and are food for thought, for example, the joint signal of temperature and the second common trend in the *N. hombergii* time series, among others

All models discussed in this paper were based on normality assumptions. Fahrmeir and Tutz (1994) and Durbin and Koopman (2000) used state-space models in which the errors terms were Poisson and Binomial distributed. These extensions can also be used in dynamic factor analysis to analyse count data or presence/absence data.

Another interesting extension is the use of time lags. Molenaar (1985) and Molenaar *et al.* (1992) used dynamic factor models of the form:

$$\mathbf{y}_t = \mathbf{\Gamma}_1 \boldsymbol{\alpha}_t + \mathbf{\Gamma}_2 \boldsymbol{\alpha}_{t-1} + \boldsymbol{\epsilon}_t$$

The matrix  $\mathbf{\Gamma}_1$  refers to factor loadings corresponding to  $\boldsymbol{\alpha}_t$  and the elements of  $\mathbf{\Gamma}_2$  indicate which of the response variables are related to  $\boldsymbol{\alpha}_{t-1}$ . Further time lags can be modelled by including terms of the form  $\mathbf{\Gamma}_j \boldsymbol{\alpha}_{t-j}$  where  $j$  represents a time lag of  $j$  time units. The interpretation of such a model is that the effects of the common trends can occur with a time delay. The parameters were estimated by direct optimisation of a likelihood function. However, both papers show how these dynamic factor models can be written as state-space models. This means that the EM algorithm can also be used for parameter estimation and that much larger data sets can therefore be analysed. Molenaar *et al.* (1992) also use a trend formulation which allows for common seasonal components as well. We are currently working on these topics.

#### ACKNOWLEDGEMENT

The first author would like to thank Han Lindeboom for providing the initial motivation for this research.

## APPENDIX

### Algorithm for Kalman filter and smoother

The Kalman filter and smoother algorithm for the model in equations (9)-(10) is given by:

$$\begin{aligned}
 & \text{Initialise} && \boldsymbol{\alpha}_{0|0} \text{ and } \mathbf{V}_{0|0} \\
 & && \text{Repeat the prediction and correction step for } t = 1, \dots, T \\
 & \text{Prediction step:} && \boldsymbol{\alpha}_{t|t-1} = \boldsymbol{\alpha}_{t-1|t-1} \\
 & && \mathbf{V}_{t|t-1} = \mathbf{V}_{t-1|t-1} + \mathbf{Q} \\
 & \text{Correction step:} && \mathbf{K}_t = \mathbf{V}_{t|t-1} \boldsymbol{\Gamma}' (\boldsymbol{\Gamma} \mathbf{V}_{t|t-1} \boldsymbol{\Gamma}' + \mathbf{H})^{-1} \\
 & && \boldsymbol{\alpha}_{t|t} = \boldsymbol{\alpha}_{t|t-1} + \mathbf{K}_t (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|t-1} - \mathbf{D} \mathbf{x}_t) \\
 & && \mathbf{V}_{t|t} = \mathbf{V}_{t|t-1} - \mathbf{K}_t \boldsymbol{\Gamma} \mathbf{V}_{t|t-1} \\
 & \text{Initialise} && \mathbf{V}_{T, T-1|T} = (\mathbf{I} - \mathbf{K}_T \boldsymbol{\Gamma}) \mathbf{V}_{T-1|T-1} \\
 & && \text{Repeat the smoothing step for } t = T, \dots, 2 \\
 & \text{Smoothing step:} && \boldsymbol{\alpha}_{t-1|T} = \boldsymbol{\alpha}_{t-1|t-1} + \mathbf{B}_{t-1} (\boldsymbol{\alpha}_{t|T} - \boldsymbol{\alpha}_{t|t-1}) \\
 & && \mathbf{V}_{t-1|T} = \mathbf{V}_{t-1|t-1} + \mathbf{B}_{t-1} (\mathbf{V}_{t|T} - \mathbf{V}_{t|t-1}) \mathbf{B}'_{t-1} \\
 & && \mathbf{V}_{t, t-1|T} = \mathbf{V}_{t-1|t-1} \mathbf{B}'_{t-2} + \mathbf{B}_{t-1} (\mathbf{V}_{t, t-1|T} - \mathbf{V}_{t-1|t-1}) \mathbf{B}'_{t-2} \\
 & \text{where} && \mathbf{B}_{t-1} = \mathbf{V}_{t-1|t-1} \mathbf{V}_{t|t-1}^{-1}
 \end{aligned}$$

The value of the log likelihood function of the incomplete data  $\mathbf{y}_1, \dots, \mathbf{y}_T$  is given by:

$$\begin{aligned}
 \log G &= \text{constant} - \frac{1}{2} \sum_{t=1}^T \log |\boldsymbol{\Gamma} \mathbf{V}_{t|t-1} \boldsymbol{\Gamma}' + \mathbf{H}| \\
 &\quad - \frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|t-1} - \mathbf{D} \mathbf{x}_t)' \mathbf{F}_t^{-1} (\mathbf{y}_t - \boldsymbol{\Gamma} \boldsymbol{\alpha}_{t|t-1} - \mathbf{D} \mathbf{x}_t)
 \end{aligned}$$

where  $\mathbf{F}_t = \boldsymbol{\Gamma} \mathbf{V}_{t|t-1} \boldsymbol{\Gamma}' + \mathbf{H}$ . This function can easily be calculated within the Kalman filter algorithm.

### EM Algorithm

The EM algorithm for dynamic factor analyses with missing values and explanatory variables is given by:

1. Choose starting values for  $\mathbf{H}$ ,  $\boldsymbol{\Gamma}$ ,  $\mathbf{a}_0$ ,  $\mathbf{D}$  and set all diagonal elements of  $\mathbf{V}_0$  to 5. Denote the hyperparameters by  $\mathbf{H}^0$ ,  $\boldsymbol{\Gamma}^0$ ,  $\mathbf{a}_0^0$ ,  $\mathbf{D}^0$ .
2. For  $p = 0, 1, 2, \dots$ 
  - E-step: estimate  $\boldsymbol{\alpha}_{t|T}^{(p)}$  and  $\mathbf{V}_{t|T}^{(p)}$  by Kalman smoothing. Let  $\bar{\mathbf{a}}$  be the average of  $\boldsymbol{\alpha}_{t|T}^{(p)}$ . Set  $\boldsymbol{\alpha}_{t|T}^{(p)} := \boldsymbol{\alpha}_{t|T}^{(p)} - \bar{\mathbf{a}}$ .
  - M-step: update the parameters  $\mathbf{a}_0$ ,  $\boldsymbol{\Gamma}$ ,  $\mathbf{D}$  and  $\mathbf{H}$  as described in the text. Denote these by  $\mathbf{a}_0^{(p+1)}$ ,  $\boldsymbol{\Gamma}^{(p+1)}$ ,  $\mathbf{D}^{(p+1)}$  and  $\mathbf{H}^{(p+1)}$ .

- Stop on convergence (change in  $\log G < 0.000001$ ).
3. Apply a varimax rotation on factor loadings and apply an inverse factor rotation on the common trends and modify the corresponding covariance matrix.

## REFERENCES

- Aguilar O, Huerta G, Prado R, West M. 1998. Bayesian Inference on Latent Structure in Time Series. *Bayesian Statistics* **6**:1-16.
- Basilevsky A. 1994. *Statistical Factor Analysis and Related Methods. Theory and Applications*. John Wiley & Sons, Inc.
- Beukema JJ. 1974. Seasonal changes in the biomass of the macro-benthos of a tidal flat area in the Dutch Wadden Sea. *Neth. J. Sea Res.* **8**:94-107.
- Beukema JJ. 1979. Biomass and species richness of the macrobenthic animals living on a tidal flat area in the Dutch Wadden Sea; effects of a severe winter. *Neth. J. Sea Res.* **13**:203-223.
- Beukema JJ. 1984. Zoobenthos survival during severe winters on high and low tidal flats in the Dutch Wadden Sea. In *Marine biology of polar regions and effects of stress on marine organisms*, Gray JS, Christiansen ME (eds); 351-361.
- Beukema JJ. 1988. An evaluation of the ABC-method (abundance/biomass comparison) as applied to macrozoobenthic communities living on tidal flats in the Dutch Wadden Sea. *Mar. Biol.* **99**:425-433.
- Beukema JJ. 1992. Long-term and recent changes in the benthic macrofauna living on tidal flats in the western part of the Wadden Sea. *Neth. J. Sea Res.* **20**:135-141.
- Beukema JJ, Essink K, Dekker R. 2000. Long-term observations on the dynamics of three species of polychaetes living on tidal flats of the Wadden Sea: the role of weather and predator-prey interactions. *J. Anim. Ecol.* **69**:31-44.
- Blegvad H. 1929. Mortality among animals of the littoral region in ice winters. *Re. Dan. biol. Station* **35**:49-62.
- Digalakis V, Rohlicek JR, Ostendorf M. 1993. ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition. *IEEE Transactions on speech and audio processing* **1**:431-442.
- Durbin J, Koopman SJ. 2000. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc. B* **62**:3-56.
- Fahrmeir L, Tutz G. 1994. *Multivariate statistical modelling based on generalized linear models*. Springer-Verlag: New York.
- Harvey AC. 1989. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
- Johnston J. 1984. *Econometric methods*. McGraw-Hill, Inc.
- Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**:13-22.
- Ljung L. 1987. *System Identification: Theory for the User*. Prentice-Hall: New York.
- Lütkepohl H. 1991. *Introduction to multiple time series analysis*. Springer-Verlag: Berlin.
- Meng X, Rubin DB. 1991. Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm. *JASA* **86**:899-909.
- Molenaar PCM. 1985. A dynamic factor model for the analysis of multivariate time series. *Psychometrika* **50**:181-202.
- Molenaar PCM. 1989. Aspects of dynamic factor analysis. In *Analysis of statistical information*. pp. 183-199. Tokyo: The Institute of Statistical Mathematics.

- Molenaar PCM. 1993. Dynamic factor analysis of psychophysiological signals. In *Advances in Psychophysiology*. Jennings JR, Ackels P, Coles MGH (eds). Vol 5, pp. 229-302. Jessica Kingsley Publishers: London.
- Molenaar PCM, de Gooijer JG. 1988. On the identification of the latent covariance structure in dynamic nonstationary factor models. In *The many faces of multivariate analysis*. Jansen MGH, van Schuur WH (eds). Vol 1, pp. 196-209. Groningen: Society for multivariate analysis in the behavioral sciences.
- Molenaar PCM, de Gooijer JG, Schmitz B. 1992. Dynamic factor analysis of non-stationary multivariate time series. *Psychometrika* **57**:333-349.
- Molenaar PCM, Rovine MJ, Corneal SE. 1999. Dynamic factor analysis of emotional dispositions of adolescent stepsons towards their stepfathers. In *Growing up in times of social change*. Silbereisen RK (ed), pp. 287-318. De Gruyter: Berlin.
- Priestley MB. 1991. *Spectral Analysis and Time Series Analysis*. Academic Press: London.
- Shumway RH, Stoffer DS. 2000. *Time Series Analysis and Its Applications*. Springer-Verlag: New York.
- Shumway RH, Stoffer DS. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis* **3**:253-264.
- Stoffer DS, Wall KD. 1991. Bootstrapping State-Space Models: Gaussian Maximum Likelihood Estimation and the Kalman Filter. *JASA* **86**:1024-1033.
- West M, Harrison PJ. 1997. *Bayesian Forecasting and Dynamic Models, 2nd edition*. Springer-Verlag: New York.
- Wu L S-Y, Pai JS, Hosking JRM. 1996. An Algorithm for Estimating Parameters of State-Space Models. *Statistics and Probability Letters* **28**:99-106.

Table 1: Number of non-restricted parameters in the dynamic factor model.  $N$  is the number of response variables,  $M$  the number of common trends and  $K$  is the number of explanatory variables plus 1 (level parameters).

matrix	number of parameters
$\mathbf{\Gamma}$	$M(N - \frac{1}{2}(M - 1))$
$\mathbf{D}$	$NK$
$\mathbf{H}$ (diagonal)	$N$
$\mathbf{H}$ (symmetric, non-diagonal)	$N + N(N - 1)/2$

Table 2: Twelve macro zoobenthic species used in the analysis

<i>Arenicola marina</i>	<i>Heteromastus filiformis</i>
<i>Mya arenaria</i>	<i>Phyllodoce mucosa</i>
<i>Cerastoderma edule</i>	<i>Hydrobia ulvae</i>
<i>Mytilus edulis</i>	<i>Scoloplos armiger</i>
<i>Nephtys hombergii</i>	<i>Lanice conchilega</i>
<i>Nereis spp.</i>	<i>Macoma balthica</i>

Table 3: AIC, log likelihood and numbers of parameters for models I and II using  $M$  common trends applied to 12 zoobenthic species (using temperature as an explanatory variable in model II).  $N^*$  refers to the models with  $N$  univariate trends.

$M$	Model I			Model II		
	AIC	log likelihood	parameters	AIC	log likelihood	parameters
1	846.52	-321.26	102	802.66	-287.33	114
2	824.22	-299.11	113	773.67	-261.84	125
3	821.75	-287.87	123	770.25	-250.12	135
4	821.30	-278.65	132	775.07	-243.51	144
5	832.37	-276.18	140	781.66	-238.83	152
$N^*$	921.03	-358.52	102	885.00	-328.50	114

Table 4: Estimated parameters and  $t$ -values for temperature using 3 common trends.

species	estimated parameter	$t$ -value
<i>A. marina</i>	0.11	0.85
<i>C. edule</i>	0.50	3.29
<i>H. filiformis</i>	0.10	1.25
<i>H. ulvae</i>	0.23	2.25
<i>L. conchilega</i>	0.69	5.89
<i>M. balthica</i>	-0.24	-2.92
<i>M. arenaria</i>	-0.09	-0.75
<i>M. edulis</i>	0.38	2.38
<i>N. hombergii</i>	0.73	9.42
<i>Nereis spp.</i>	0.14	1.28
<i>P. mucosa</i>	0.47	3.36
<i>S. armiger</i>	0.00	-0.01

Table 5: Summary of results. Values in the temperature are the estimated parameters with  $t$ -values larger (in absolute sense) than 3. P and M refers to positive (P) and negative (M) factor loadings larger than 0.1 in absolute sense.

species	temperature	trend 1	trend 2	trend 3
<i>A. marina</i>		M	M	
<i>C. edule</i>	0.50			
<i>H. filiformis</i>			P	
<i>H. ulvae</i>			P	
<i>L. conchilega</i>	0.69			
<i>M. balthica</i>		M		
<i>M. arenaria</i>			P	M
<i>M. edulis</i>				P
<i>N. hombergii</i>	0.73		M	
<i>Nereis spp.</i>		M	P	
<i>P. mucosa</i>	0.47	M		
<i>S. armiger</i>			P	

## FIGURE LEGENDS

Figure 1. The Balgzand.

Figure 2. Standardised species and temperature time series.

Figure 3. Three common trends obtained by model II.

Figure 4. Factor loadings obtained by the dynamic factor model using three common trends and temperature as explanatory variable.

Figure 5. Fitted values obtained by a dynamic factor model containing three common trends and temperature as explanatory variable.