Dynamic Factor Analysis for Panel Data: A Generalized Model*

Nikolaos Zirogiannis† and Yorghos Tripodis‡

Abstract

We develop a generalized dynamic factor model for panel data with the goal of estimating an unobserved performance index. While similar models have been developed in the literature of dynamic factor analysis, our contribution is three-fold. First, contrary to simple dynamic factor analysis where multiple attributes of the same cross sectional unit are measured at each time period, our model also accounts for multiple cross sectional units. It is therefore applicable to a panel data framework (i.e. multiple attributes for multiple cross sectional units observed over time). Second, our model estimates an unobserved index for every cross sectional unit for every time period, as opposed to previous work where a

*The authors would like to thank Alexander Danilenko for providing data from the International Benchmarking Network. We are grateful to John Buonaccorsi, Klaus Moeltner, Joe Moffitt and John Stranlund for their constructive feedback.Comments and suggestions from seminar participants at the University of Ottawa and the University of Massachusetts Amherst are greatly appreciated. This research was made possible with funding by the NIH grant AG13846. Any remaining errors are ours.

†School of Public and Environmental Affairs, Indiana University, nzirogia@indiana.edu.

‡Department of Biostatistics, Boston University, yorghos@bu.edu.

single unobserved index was estimated for all cross sectional units for every time period. Third, we address the complexity of the model by developing a novel iterative estimation process which we call the Two-Cycle Conditional Expectation-Maximization (2CCEM) algorithm. The 2CCEM algorithm is flexible enough to handle a variety of different datasets. The model is applied on a panel measuring attributes related to the operation of water and sanitation utilities. The goal is to estimate a dynamic benchmarking index that will capture the financial and operational performance of these utilities.

Keywords: Dynamic Factor Models, EM algorithm, Panel Data, State-Space models, Water utilities, IBNET

1 Introduction

Over the last several decades, technological developments in computer science have allowed the accumulation and storage of vast amounts of information. Many government agencies and research institutions around the world are continuously collecting data that are, more often than not, made publicly available. Examples include the Penn World Tables and the Open Data Services of the World Bank, which contain several time series variables for multiple countries. The emergence of this rich data environment creates the need for statistical methodologies that can summarize large databases into a few composite indicators which can be easily used and understood by policy makers and researchers alike.

Methods involving estimation of latent variables have been gaining increasing attention, with factor analysis being a prominent one. Until the late 1970s, the estimation of factor analytic models was limited to cross sectional datasets ignoring any dynamic analysis. Geweke (1977)

along with Sargent and Sims (1977) were the first to propose a new class of dynamic factor models (DFMs). Stock and Watson (1989) built on those contributions using maximum likelihood to estimate a dynamic factor model. The authors estimate unobserved coincident and leading economic indices for the US economy, where the estimation of the leading index is conducted conditional on the estimate of the unobserved coincident index. However, the model of Stock and Watson is limited by the fact that it cannot handle panel data, that is, multiple variables for multiple cross sectional units spanning several years. Forni et al. (2000) developed a dynamic factor model that could handle panel data. The authors used principal components to estimate one unobserved index for all cross sectional units for every time period in their dataset. The extension of factor analysis to a longitudinal setting greatly expanded the method's applicability. Apart from summarizing a large number of variables into a few coincident indicators, forecasts were also made possible. Bai (2003) contributes to this literature, by providing the inferential theory for DFMs of large dimensions. He discusses the convergence rates of factors and factor loadings and finds that stronger results are achieved when the errors of the idiosyncratic components are serially uncorrelated. Boivin and Ng (2006) suggest that when more data are used to extract factors and the idiosyncratic errors are correlated the forecasting power of the model can be reduced. In light of those findings, they question whether using a large set of variables increases the validity of the model. Doz et al. (2011) address the issue of the use of principle components in DFMs of large dimensions. They argue that, even though the principle components approach has been used extensively in the literature, maximum likelihood estimation can lead to greater efficiency gains, even when the DFM is misspecified.

Our work contributes to this literature by focusing on better exploitation of the panel nature of the data. We develop a novel iterative estimation process, which we call "Two-Cycle

Conditional Expectation-Maximization" (2CCEM) algorithm. Initially, the unobserved performance index is estimated (first cycle) and then the dynamic component of the index is incorporated into the estimation process (second cycle). The estimates of each cycle are updated with information from the estimates of the previous cycle until convergence is achieved. Our estimation strategy can account for multiple cross sectional units, making it flexible enough to be applicable to different types of datasets. Therefore, contrary to the model developed by Stock and Watson (1989), our model can be applied to a panel dataset. In addition, while Forni et al. (2000) estimate a single unobserved index common for all cross sectional units in their sample, we estimate one latent index for every cross sectional unit.

The paper is organized as follows. In section 2, we present the theoretical framework, and examine the various components of the model. We also discuss necessary conditions for identifiability. Section 3 presents the 2CCEM algorithm and illustrates the estimation process for each of the two cycles. In section 4, we apply our model to a longitudinal dataset of water and sanitation utilities from various developing countries. We discuss how we obtain initial values for the parameters and present estimation results. In the final section, we draw conclusions and discuss future extensions of our work.

2 A generalized dynamic factor model for panel data

The main contribution of our work lies in the development of the generalized dynamic factor model that accounts for correlations between cross sectional units and is applicable to a panel data setting. We begin by presenting the notation that will be used throughout the paper.

2.1 Notation

Denoting vectors with bold letters, we let $y_{ij,t}$ be the i^{th} indicator of the j^{th} cross sectional unit at time t with:

- $i = 1, ..., p$ denoting the number of observed variables (indicators) in the model;
- $j = 1, ..., m$ denoting the number of cross sectional units;
- $t = 1, ..., n$ denoting the time point of an observation;

To ease formulation of our model, we collect the observed data in vector form. Let:

- $\mathbf{Y}_{ij}$ be an $n \times 1$ vector with elements, $y_{ij,t}$, for i, j fixed and $t = 1, ..., n$;
- $\mathbf{Y}_t$ be a $mp \times 1$ vector with elements, $y_{ij,t}$, for t fixed with $i = 1, ..., p$ and $j = 1, ..., m$;
- $\mathbf{Y}$ be a $nmp \times 1$ vector of all p indicators for all m cross sectional units over all n years.

2.2 The theoretical framework of the model

State space models have been used extensively, particularly in the early literature of DFMs, since they allow the study of unobserved factors over time through the use of the observed data (Stock and Watson 2010). We formulate our model using a state space approach, letting $\mathbf{U}_t$ denote the vector of m unobserved factors at time t. We assume that the dynamic properties of $\mathbf{U}_t$ can be captured by a Markov process. Thus, we form the following linear Gaussian state space model:

$$\mathbf{Y}_t = \mathbf{B}\mathbf{U}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{D}), \tag{2.1}$$

$$\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}), \tag{2.2}$$

where $\mathbf{B}$ is the matrix of factor loadings with dimensions $mp \times m$, $\mathbf{U}_t$ is the $m \times 1$ unobserved state vector at time t, $\mathbf{Y}_t$ is a $mp \times 1$ vector of observed variables at time t, $\mathbf{T}$ is a $m \times m$ transition matrix that describes the Markovian nature of the unobserved state vector, and $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ are error terms (Koopman 1993). Equation (2.1) is known as the observation equation (or measurement equation) and equation (2.2) is called the state equation (or transition equation) and represents the first order autoregressive nature of the model. The state space formulation described in (2.1) and (2.2) models the behavior of the unobserved state vector $\mathbf{U}_t$ over time using the observed values $\mathbf{Y}_1, ..., \mathbf{Y}_n$. The state vector $\mathbf{U}_t$ is assumed to be independent of the error terms $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ for all $t = 1, ..., n$. In addition, the error terms $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ are assumed to be independent, identically distributed (i.i.d.) and mutually uncorrelated (deJong 1991; Kohn and Ansley 1989).

We will describe the structure of each matrix, $\mathbf{B}$, $\mathbf{D}$, $\mathbf{T}$, $\mathbf{Q}$, separately, and its implications for model interpretation. The general form of the matrix of factor loadings $\mathbf{B}$ is:

$$\underset{mp \times m}{\mathbf{B}} = \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} & ... & \mathbf{b}_{1m} \\ \mathbf{b}_{21} & \mathbf{b}_{22} & ... & \mathbf{b}_{2m} \\ ... & ... & ... & ... \\ \mathbf{b}_{m1} & \mathbf{b}_{m2} & ... & \mathbf{b}_{mm} \end{bmatrix},$$

where $\mathbf{b}_{jj}$ $(j = 1, ..., m)$ is a $p \times 1$ vector of the factor loadings for the j^{th} cross sectional unit and $\mathbf{b}_{jj^*}$ $(j,\ j^* = 1, ..., m$ and $j \neq j^*)$ is also a $p \times 1$ vector representing the loadings of the indicators of cross sectional unit j to the factor of cross sectional unit j^*. For example, $\mathbf{b}_{11}$ contains the factor loadings of the first cross sectional unit, while $\mathbf{b}_{12}$ loads the indicators of the first cross sectional unit to the factor of the second cross sectional unit. There are four alternative formulations of $\mathbf{B}$ that we consider, which are illustrated in Table 1.

Table 1: Possible formulations of $\mathbf{B}$.

		Off-diagonal elements			
		$\mathbf{b}_{jj*} = \mathbf{0}$		$\mathbf{b}_{jj*} \neq \mathbf{0}$	
		Notation for $\mathbf{B}$	Parameters	Notation for $\mathbf{B}$	Parameters
Diagonal	$\mathbf{b}_{jj} = \mathbf{b}$	$\mathbf{B}_1$	m	$\mathbf{B}_2$	$mp \times [m-1]+1$
elements	$\mathbf{b}_{jj} \neq \mathbf{b}$	$\mathbf{B}_3$	mp	$\mathbf{B}_4$	$mp \times m$

Formulations $\mathbf{B}_1$ and $\mathbf{B}_2$ represent the cases where factor loadings are equal for all cross sectional units, while formulations $\mathbf{B}_3$ and $\mathbf{B}_4$ represent cases of unequal factor loadings across units. For example,

$$\underset{(m\times p)\times m}{\mathbf{B}_1} = \begin{bmatrix} \mathbf{b} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{b} & ... & \mathbf{0} \\ ... & ... & ... & ... \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{b} \end{bmatrix}. \tag{2.3}$$

The difference between $\mathbf{B}_1$ (or $\mathbf{B}_3$) and $\mathbf{B}_2$ (or $\mathbf{B}_4$) lies in the way cross sectional units interact with each other. In formulations $\mathbf{B}_1$ and $\mathbf{B}_3$ the indicators of each cross sectional unit do not load on the factors of other cross sectional units, since $\mathbf{b}_{jj*} = \mathbf{0}$. On the other hand, in formulations $\mathbf{B}_2$ and $\mathbf{B}_4$ $\mathbf{b}_{jj^*}$ is unconstrained, so that indicators of each cross sectional unit are allowed to load on factors other than their own.

Next, we consider the variance of the idiosyncratic errors in $\mathbf{D}$. The general form of $\mathbf{D}$ is $\underset{mp\times mp}{\mathbf{D}} = \text{diag}(\mathbf{d}_j)$, where $\mathbf{d}_j$ $(j = 1, ..., m)$ is a $p \times p$ diagonal matrix representing the variance of the error term for every cross sectional unit. Diagonality of $\mathbf{D}$ is required due to the factor analytic nature of (2.1). The matrix form of each $\mathbf{d}_j$ is $\mathbf{d}_j = \text{diag}(\sigma_{ij}^2)$, where σ_{ij}^2 is the variance of the error term of a specific cross sectional unit. We disitnguish between two alternative formulations of $\mathbf{D}$, namely $\mathbf{D}_1$ whereby each $\mathbf{d}_j$ is identical for every cross

sectional unit, and $\mathbf{D}_2$ where σ_{ij}^2 varies. Consequently, $\mathbf{D}_1$ and $\mathbf{D}_2$ have p and mp unknown parameters respectively.

The general form of $\mathbf{T}$ is illustrated as follows:

$$\underset{m\times m}{\mathbf{T}} = \begin{bmatrix} \phi_{11} & \phi_{12} & ... & \phi_{1m} \\ \phi_{21} & \phi_{22} & ... & \phi_{2m} \\ ... & ... & ... & .. \\ \phi_{m1} & \phi_{m2} & ... & \phi_{mm} \end{bmatrix},$$

where ϕ_{jj} is the autoregressive parameter that determines the effect through time of a cross sectional unit's own latent variable. The off-diagonal elements ϕ_{jj^*} (where $j,\ j^* = 1, ..., m$ and $j \neq j^*$), capture the correlation between latent variables of different cross sectional units. $\mathbf{T}$ is not symmetric and as a result we may have $\phi_{jj^*} \neq \phi_{j^*j}$. We distinguish between four cases for $\mathbf{T}$, illustrated in Table 2:

Table 2: Possible formulations of $\mathbf{T}$.

		Off-diagonal elements			
		$\phi_{jj^*} = 0$		$\phi_{jj^*} \neq 0$	
		Notation for $\mathbf{T}$	Parameters	Notation for $\mathbf{T}$	Parameters
Diagonal elements	$\mathbf{b}_{jj} = \mathbf{b}$	$\mathbf{T}_1$	1	$\mathbf{T}_2$	$m\,[m-1]+1$
	$\mathbf{b}_{jj} \neq \mathbf{b}$	$\mathbf{T}_3$	m	$\mathbf{T}_4$	m^2

In formulations $\mathbf{T}_1$ and $\mathbf{T}_2$, all cross sectional units share the same autoregressive parameter, while in $\mathbf{T}_3$ and $\mathbf{T}_4$, these parameters are allowed to vary by cross sectional unit. Formulations $\mathbf{T}_1$ and $\mathbf{T}_3$ imply that there is no correlation between the values of the state variable of different cross sectional units across time. In contrast, formulations $\mathbf{T}_2$ and $\mathbf{T}_4$ have unconstrained ϕ_{jj^*}, hence accounting for cross-temporal correlations between the state variables of different cross sectional units.

Finally, we focus on $\mathbf{Q}$, the covariance matrix of the error term in the state equation. The general form of the matrix is the following:

$$\underset{m\times m}{\mathbf{Q}} = \begin{bmatrix} \sigma_1^2 & \mathrm{E}\left(\eta_1\eta_2\right) & ... & \mathrm{E}\left(\eta_1\eta_m\right) \\ \mathrm{E}\left(\eta_2\eta_1\right) & \sigma_2^2 & ... & \mathrm{E}\left(\eta_2\eta_m\right) \\ ... & ... & ... & .. \\ \mathrm{E}\left(\eta_m\eta_1\right) & \mathrm{E}\left(\eta_m\eta_2\right) & ... & \sigma_m^2 \end{bmatrix},$$

where the diagonal elements σ_j^2 are the variances of the error term of the state equation. The off-diagonal elements $\mathrm{E}(\eta_j\eta_{j^*})$ (where j, $j^* = 1, ..., m$ and $j \neq j^*$) represent covariances, with $\mathrm{E}(\eta_j\eta_{j^*}) = \mathrm{E}(\eta_{j^*}\eta_j)$ by symmetry of $\mathbf{Q}$. In the following section, we impose certain restrictions on $\mathbf{Q}$ to ensure identifiability of the model.

2.3 Identifiability

A central issue in the literature of unobserved component models is identifiability. We explore identifiability directly using the order condition. The latter suggests that the number of parameters in an equation must be at least as great as the number of explanatory variables (Hamilton 1994, p.244). Hotta (1989) provides the order conditions for identifiability of a structural time series model. We follow a similar approach to derive the conditions for theoretical identifiability in the model specified in equations (2.1) and (2.2). In this section, we show the correlation structure of $\mathbf{Y}$ and derive the autocovariance equation of our model. Since the state vector $\mathbf{U}_t$ is unobserved, all the information in our model is contained in $\mathbf{Y}$.

The covariance matrix of $\mathbf{Y}$, denoted by $\mathbf{\Omega}$, has the following structure:

$$\mathrm{Var}(\mathbf{Y}) = \underset{nmp\times nmp}{\mathbf{\Omega}} = \begin{bmatrix} \mathrm{Var}(\mathbf{Y}_1) & \mathrm{Cov}(\mathbf{Y}_1\mathbf{Y}_2) & ... & \mathrm{Cov}(\mathbf{Y}_1\mathbf{Y}_n) \\ \mathrm{Cov}(\mathbf{Y}_2\mathbf{Y}_1) & \mathrm{Var}(\mathbf{Y}_2) & ... & ... \\ ... & ... & ... & ... \\ \mathrm{Cov}(\mathbf{Y}_n\mathbf{Y}_1) & ... & ... & \mathrm{Var}(\mathbf{Y}_n) \end{bmatrix}, \tag{2.4}$$

where $\mathrm{Cov}(\mathbf{Y}_t, \mathbf{Y}_t^*)$ is a $mp \times mp$ matrix, for $t, t^* = 1, ..., n$ and $t \neq t^*$. The off-diagonal elements of $\mathbf{\Omega}$ capture the covariance of $\mathbf{Y}_t$ across time. For ease of presentation, and without loss of generality, we assume that $\mathrm{E}(\mathbf{Y}_t) = \mathrm{E}(\mathbf{U}_t) = \mathbf{0}$. The unconditional covariance matrix of $\mathbf{Y}_t$, that is, the covariance matrix of all indicators for all cross sectional units at a given time period t, is denoted by $\mathbf{\Sigma}$. It follows from (2.1) and (2.2) that:

$$\mathbf{\Sigma} = \mathrm{Var}(\mathbf{Y}_t) = \mathbf{B}\mathrm{Var}(\mathbf{U}_t)\mathbf{B}' + \mathbf{D}, \tag{2.5}$$

and

$$\mathrm{E}(\mathbf{Y}_{t+1}\mathbf{Y}_t') = \mathbf{B}\mathbf{T}\mathrm{Var}(\mathbf{U}_t)\mathbf{B}'. \tag{2.6}$$

In addition, the variance of the state variable $\mathbf{U}_t$ is given by:

$$\mathrm{E}(\mathbf{U}_t\mathbf{U}_t') = \mathbf{T}\mathrm{Var}(\mathbf{U}_{t-1})\mathbf{T}' + \mathbf{Q}, \tag{2.7}$$

while $\mathrm{E}(\mathbf{Y}_t\mathbf{U}_t)$ is:

$$\mathrm{E}(\mathbf{Y}_t\mathbf{U}_t') = \mathrm{E}\left[(\mathbf{B}\mathbf{U}_t + \mathbf{e}_t)\mathbf{U}_t'\right] = \mathbf{B}\mathrm{Var}(\mathbf{U}_t). \tag{2.8}$$

Therefore, the joint multivariate normal vector $(\mathbf{Y}_t^T, \mathbf{U}_t^T)^T$ has zero mean and a covariance matrix that can be calculated recursively, using equations (2.5)-(2.8). In order to obtain the

necessary conditions for indentifiability, we first derive the autocovariance function of $\mathbf{Y}_t$ in the following lemma.

Lemma 2.1. *The autocovariance function of* $\mathbf{Y}_t$ *is:*

$$\begin{aligned} \text{vec}[\Gamma_{\mathbf{Y}}(0)] &= \mathbf{B} \otimes \mathbf{B}\{[\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1}\text{vec}(\mathbf{Q})\} + \text{vec}(\mathbf{D}) \quad (2.9) \\ \text{vec}[\Gamma_{\mathbf{Y}}(1)] &= \mathbf{B} \otimes (\mathbf{BT})\{[\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1}\text{vec}(\mathbf{Q})\} \quad (2.10) \\ \text{vec}[\Gamma_{\mathbf{Y}}(h)] &= \mathbf{B} \otimes (\mathbf{BT})\{[\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1}, \text{ for } h > 1 \quad (2.11) \end{aligned}$$

Proof. The proof is provided in the Appendix. □

Theorem 2.2 provides the necessary conditions for the model to be identifiable.

Theorem 2.2. *The necessary conditions for the model in* (2.1) *and* (2.2) *to be identifiable are:*

1. $$\Gamma_{\mathbf{U}}(0) = \mathbf{C}, \quad (2.12)$$

   where $\mathbf{C}$ *is a known symmetric positive definite matrix, and*

2. $$m > \frac{1}{3p - 2 - \frac{2}{p}} \quad (2.13)$$

Proof. The proof is provided in the Appendix. □

The choice of $\mathbf{C}$ is arbitrary as long as the conditions for a symmetric positive definite matrix are satisfied.

Remark 2.1. For $\mathbf{C} = \mathbf{I}$ we obtain the dynamic version of the factor analytic model of McLachlan and Peel (2000, p.243). It follows from the proof of Theorem 2.2 that, when $\mathbf{C} = \mathbf{I}$, the necessary conditions for identifiability imply that $\mathbf{Q} = \mathbf{I} - \mathbf{T}\mathbf{T}'$.

3 The 2CCEM algorithm

Another contribution of our work is the development of the 2CCEM algorithm, which is a novel approach to the estimation of dynamic factor models. The high dimensionality of the data vector $\mathbf{Y}_t$ makes estimation of our model rather problematic. Usual Newton-type gradient methods do not work in this situation creating the need for a novel estimation approach. The likelihood function of the model described in (2.1) and (2.2) is:

$$L(\mathbf{B},\, \mathbf{D},\, \mathbf{T},\, \mathbf{Q}; \mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{t=2}^{n} f(\mathbf{Y}_1) f_{\mathbf{Y}}(\mathbf{Y}_t; [\mathbf{B},\, \mathbf{D},\, \mathbf{T},\, \mathbf{Q}]\, |\mathbf{Y}_{t-1}), \tag{3.1}$$

where $\mathbf{Y}_{t-1}$ represents the set of past observations $\mathbf{Y}_1, ..., \mathbf{Y}_{t-1}$ and the model parameters to be estimated are $\mathbf{B}$, $\mathbf{D}$, $\mathbf{T}$ and $\mathbf{Q}$. We showed in Theorem 2.2 that the parameterization of $\mathbf{Q}$ depends on $\mathbf{T}$ for identifiability of the model. Therefore, although $\mathbf{Q}$ is not estimated by the model, for ease of presentation we continue to include it in the parameter space.

We introduce the 2CCEM algorithm that makes estimation of the model specified in (2.1) and (2.2) feasible through an iterative two-cycle process. The 2CCEM algorithm is an extension of the EM algorithm developed by Dempster et al. (1977). The EM algorithm has been widely used in cases where maximization of the likelihood function cannot occur because of missing or unobserved data. Shumway and Stoffer (1982) were the first to use the EM algorithm to estimate state space models, similar to the one specified in (2.1) and (2.2). The algorithm is comprised of an Expectation and a Maximization step, referred to as E-step and

M-step respectively. The former replaces the unobserved quantities with their expected values while the latter maximizes the likelihood conditional on those expectations (McLachlan and Krishnan 1996, p.13).

We let the complete-data log likelihood function of $\mathbf{\Psi}$, if $\mathbf{Y}_t$ and $\mathbf{U}_t$ were fully observable, be:

$$\log L_c(\mathbf{\Psi}) = \log f_c(\mathbf{Y}_t,\, \mathbf{U}_t; \mathbf{\Psi}), \tag{3.2}$$

where the subscript c denotes the complete-data likelihood.

The 2CCEM algorithm starts by partitioning the vector of unknown parameters $\mathbf{\Psi}$ into $(\mathbf{\Psi}_1, \mathbf{\Psi}_2)$ where $\mathbf{\Psi}_1$ contains the elements of $\mathbf{B}$ and $\mathbf{D}$ that need to be estimated, while $\mathbf{\Psi}_2$ contains the relevant elements of $\mathbf{T}$ and $\mathbf{Q}$. Partitioning the parameter space is a common practice in the EM algorithm literature (Meng and Van Dyk 1997; McLachlan and Peel 2000, p.245) since it facilitates the maximization process. We let $\mathbf{\Psi}_1^{(k-1)}$ and $\mathbf{\Psi}_2^{(k-1)}$ denote the initial values of $\mathbf{\Psi}$ where k denotes the number of iterations in the estimation process with $k = 1, ..., m$. Following the terminology of Meng and Van Dyk (1997) we use the term "cycle" as an intermediary between a "step" and an "iteration". In the case of our 2CCEM algorithm, every iteration is comprised of two cycles. The first cycle includes three steps (one E-step and two M-steps) and estimates $\mathbf{\Psi}_1$, while the second cycle is composed of two steps (one E-step and one M-step) and estimates $\mathbf{\Psi}_2$.

3.1 First cycle of the 2CCEM

During the k^{th} iteration of the first cycle, the E-step of the 2CCEM algorithm requires the following calculation:

$$\mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) = \mathrm{E}_{\mathbf{\Psi}_1}\left\{\mathrm{logL}_c\left(\mathbf{\Psi}_1\right)|\mathbf{Y}, \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}\right\}. \tag{3.3}$$

The first M-step involves differentiating $\mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)})$ with respect to $\mathbf{\Psi}_1$ in order to obtain $\mathbf{\Psi}_1^{(k/2)}$:

$$\mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1^{(k/2)}; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}), \tag{3.4}$$

The second M-step maximizes $\mathbf{Z}_{\mathbf{\Psi}_1}$ with respect to $\mathbf{B}$ and $\mathbf{D}$ using $\mathbf{\Psi}_1^{(k/2)}$ as the initial value of the parameters. Our goal, in this step, is to obtain $\mathbf{\Psi}_1^{(k)}$ such that:

$$\mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1^{(k)}; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1^{(k/2)}; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) \tag{3.5}$$

3.1.1 Estimation in the first cycle

As mentioned in section 2.3 since the state variable is unobserved, all the information that is observed is contained in $\mathbf{Y}$. Following the notation presented in McLachlan and Peel (2000, p.242) the sample covariance matrix of $\mathbf{Y}$, $\mathbf{\Sigma}$, is denoted by $\mathbf{C}_{yy}$. The latter is the main building block in the E-step of the first cycle of the 2CCEM algorithm described in equation (3.3) and treats the unobserved state vector $\mathbf{U}_t$ as missing data while iteratively maximizing $\mathbf{Z}_{\mathbf{\Psi}_1}$ assuming that $\mathbf{U}_t$ is observed (Rubin and Thayer 1982). This first E-step of the 2CCEM

algorithm requires the calculation of the expected value of the sufficient statistics, namely:

$$\begin{aligned} \mathrm{E}(\mathbf{Y}\mathbf{Y}^T|\mathbf{Y}) &= \mathbf{C}_{yy}, \\ \mathrm{E}(\mathbf{Y}^T\mathbf{U}|\mathbf{Y}) &= \mathbf{C}_{yy}\boldsymbol{\gamma}, \\ \mathrm{E}(\mathbf{U}^T\mathbf{U}|\mathbf{Y}) &= \boldsymbol{\gamma}^T\mathbf{C}_{yy}\boldsymbol{\gamma} + n\boldsymbol{\omega}, \end{aligned} \tag{3.6}$$

where:

$$\boldsymbol{\gamma} = \left(\mathbf{B}\mathbf{B}^T + \mathbf{D}\right)^{-1}\mathbf{B} \text{ and } \boldsymbol{\omega} = \mathbf{I} - \boldsymbol{\gamma}^T\mathbf{B}. \tag{3.7}$$

The distribution of the unobserved state vector $\mathbf{U}_t$, conditional on $\mathbf{Y}_t$, is given by:

$$\mathbf{U}_t|\mathbf{Y}_t \sim N(\boldsymbol{\gamma}^T\mathbf{Y}_t,\, \mathbf{I} - \boldsymbol{\gamma}^T\mathbf{B}). \tag{3.8}$$

Equations (3.6) and (3.7) constitute the E-step of the first cycle of the 2CCEM algorithm illustrated in (3.3). The subsequent first M-step, illustrated in equation (3.4), is identical to the M-step of the traditional EM algorithm which involves replacing the sufficient statistics in (3.6) into $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ and differentiating with respect to $\boldsymbol{\Psi}_1$. The functional form of $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ is:

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}_1) &= \frac{n}{2}\log\{|\mathbf{D}^{-1}| + \log|\mathbf{Q}^{-1}|\} - \frac{1}{2}\sum_{t=1}^{n}\{(\mathbf{y}_t - \mathbf{B}\hat{\boldsymbol{u}}_t)^T\mathbf{D}^{-1}(\mathbf{y}_t - \mathbf{B}\hat{\boldsymbol{u}}_t) \\ &\quad -(\hat{\boldsymbol{u}}_{t+1} - \mathbf{T}\hat{\boldsymbol{u}}_t)^T\mathbf{Q}^{-1}(\hat{\boldsymbol{u}}_{t+1} - \mathbf{T}\hat{\boldsymbol{u}}_t)\}. \end{aligned} \tag{3.9}$$

Equation (3.9) is the complete data log likelihood; complete both in terms of data and parameters. Setting the first derivatives of $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ equal to zero yields the following first order

conditions:

$$\mathbf{B}^{(k/2)} = \mathbf{C}_{yy}\boldsymbol{\gamma}\left\{\boldsymbol{\gamma}^T\mathbf{C}_{yy}\boldsymbol{\gamma} + n\boldsymbol{\omega}\right\}^{-1}, \tag{3.10}$$

$$\mathbf{D}^{(k/2)} = n^{-1}\mathrm{diag}\left\{\mathbf{C}_{yy} - \mathbf{C}_{yy}\boldsymbol{\gamma}\mathbf{B}^T\right\}, \tag{3.11}$$

where $\mathbf{B}^{(k/2)}$ and $\mathbf{D}^{(k/2)}$ represent the updated values $\boldsymbol{\Psi}_1^{(k/2)}$. We introduce a second M-step, where (3.9) is maximized, through a Newton-Raphson algorithm, with respect to $\boldsymbol{\Psi}_1$, using (3.10) and (3.11) as initial values. Upon convergence of this maximization we obtain the final updated values for $\boldsymbol{\Psi}_1^{(k)}$.

Our approach builds on the Expectation Conditional Maximization (ECM) algorithm introduced by Meng and Rubin (1993) which is itself an extension of the EM algorithm (Dempster et al. 1977). The ECM algorithm uses the same first M-step as we do, but in the second M-step maximizes the log likelihood with respect to one parameter, holding the value of the other parameter fixed to the estimate of the first M-step.

3.2 Second cycle of the 2CCEM

In the E-step of the second cycle we estimate $\boldsymbol{\Psi}_2^{(k)}$. We proceed by calculating:

$$\mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}) = \mathrm{E}_{\boldsymbol{\Psi}_2}\left\{\mathrm{logL}_c\left(\boldsymbol{\Psi}_2\right)|\mathbf{Y}, \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}\right\}. \tag{3.12}$$

The E-step involves forming the expected complete-data log likelihood by conditioning $\mathbf{Z}_{\boldsymbol{\Psi}_2}$ on the estimates $\boldsymbol{\Psi}_1^{(k)}$. The subsequent M-step involves differentiating $\mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)})$ with respect to $\boldsymbol{\Psi}_2$. We choose $\boldsymbol{\Psi}_2^{(k)}$ such that:

$$\mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2^{(k)}; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}). \tag{3.13}$$

Upon maximization of $\mathbf{Z}_{\mathbf{\Psi}_2}$, the estimate $\mathbf{\Psi}_2^{(k)}$ is used in the E-step of the first cycle. This iterative maximization process will continue until convergence of both likelihood functions $\mathbf{Z}_{\mathbf{\Psi}_1}$ and $\mathbf{Z}_{\mathbf{\Psi}_2}$ is achieved.

3.2.1 Estimation in the second cycle

The functional form of $\mathbf{Z}_{\mathbf{\Psi}_2}$ is:

$$\log \mathbf{L}_c(\mathbf{\Psi}_2) = n - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\left[\log|\mathbf{F}_t| + \mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t\right], \tag{3.14}$$

where $\mathbf{v}_t$ is the one step ahead forecast error and $\mathbf{F}_t$ is the variance of the one step ahead forecast error. Quantities, $\mathbf{v}_t$ and $\mathbf{F}_t$ can be estimated with the use of the Kalman filter, which is a set of recursions that allow the information we have about the system to be updated every time an additional observation $\mathbf{Y}_t$ is introduced into the model (Kalman 1960; Durbin and Koopman 2001, p.11). Let $\mathbf{Y}_{t-1}$ be the set of past observations $\mathbf{Y}_1, ..., \mathbf{Y}_{t-1}$ and assume that $\mathbf{U}_t|\mathbf{Y}_{t-1} \sim N(\hat{\mathbf{U}}_t, \mathbf{P}_t)$, where $\hat{\mathbf{U}}_t$ and $\mathbf{P}_t$ are to be determined. If we assume that $\hat{\mathbf{U}}_t$ and $\mathbf{P}_t$ are known, then our goal is to calculate $\hat{\mathbf{U}}_{t+1}$ and $\mathbf{P}_{t+1}$ when $\mathbf{Y}_t$ is introduced. Once $\mathbf{v}_t$ and $\mathbf{F}_t$ are calculated, (3.14) is maximized with respect to $\mathbf{\Psi}_2$, as illustrated in (3.13).

In contrast to the filtering process described above, smoothing considers both prior information as well as information after time period t. In other words, the smoothed estimate of $\mathbf{U}_t$ incorporates information from the entire sample, $\mathbf{Y}_1, ..., \mathbf{Y}_n$ (deJong 1989; Koopman 1993).

4 Application

We apply our model to a dataset of water and sanitation utilities (hence forth referred to as water utilities) in order to estimate a dynamic performance index. The data are obtained from the International Benchmarking Network (IBNET) of Water and Wastewater Utilities (IBNET 2005). IBNET was launched in 1996 with the goal of facilitating a standardized comparison amongst water utilities with respect to their financial and operational performance. For illustration purposes we apply our model to a random sample of eight IBNET utilities (one from Armenia, three from Moldova and four from Peru), with each utility measured over a period of ten years (i.e. 1998-2007).

A critical issue in constructing indices is the weighting scheme applied to the aggregated variables. Those weights are often determined based on expert knowledge, which makes the resulting index rather subjective. In the case of water utilities such a subjective index was created by the World Bank (van den Berg and Danilenko 2010). The authors estimate a static index called the "APGAR score" whose aim is to assess the health of a water utility based on a weighted sum of six indicators. The "APGAR score" considers six continuous indicators for its formulation, namely 1) water coverage (percentage of the urban population with access to drinking water), 2) sewerage coverage (population with access to sanitation services), 3) non revenue water (water provided to the network that is not being paid for), 4) affordability (money spent paying for water services), 5) collection period and 6) operating cost coverage. For ease of estimation, we transform the indicators so that they are positively correlated. We easily accomplish this by multiplying non-revenue water, affordability and collection period by -1, since those three indicators were negatively correlated to the performance index. Furthermore, to enable comparisons between the factor loadings, all indicators are standardized.

Our sample of eight water utilities from the IBNET database considers the same six indicators that van den Berg and Danilenko use in their APGAR score. Our goal is to estimate a performance index using the model in (2.1) and (2.2) whose contribution is twofold: 1) It is dynamic since performance in every time period is assessed using information from the entire sample; 2) We do not use subjective weighting schemes for the six components of the index. Instead the estimated factor loadings are used to rank the components of the index with regards to their importance.

The development of such a dynamic performance index serves several purposes. It can be used as a benchmarking tool for utility managers and policy makers since it succinctly communicates whether the utility has been performing well or not. Furthermore, it allows managers to compare their company's effectiveness vis a vis other water providers at the national, regional or even international level.

4.1 Initial values

In section 2.2 we specified several scenarios with respect to parameter formulation. In this application we estimate a specification of the model that includes formulations $\mathbf{B}_1$ and $\mathbf{D}_1$. For parameter $\mathbf{T}$ we estimate specifications $\mathbf{T}_1$ and $\mathbf{T}_3$ as well as a scenario whereby all water utilities from the same country share the same autoregressive parameter. Each of the parameters and their initial values is discussed below.

The choice of $\mathbf{B}_1$ suggests that:

1. The factor loadings for every utility are identical. This is a plausible assumption, since we estimate an index that can be used as a benchmarking tool among utilities. Having a different set of factor loadings for each utility would not allow comparisons between

utilities.

2. The indicators of utility j do not load on the factors of utility j^*. This assumption is made to facilitate the interpretation of the factor loadings with regards to their effect on the performance index.

The initial value of $\mathbf{B}$ is denoted by $\mathbf{B}^0$ and has the matrix form specification of $\mathbf{B}_1$ illustrated in (2.3). The initial value of the block diagonal vector $\mathbf{b}$ is denoted by $\mathbf{b}^0$ where $\mathbf{b}^0 = \left(\frac{1}{p}\right)\mathbf{i}_p$. With regards to the covariance matrix of the idiosyncratic errors, we specify formulation $\mathbf{D}_1$ such that $\mathbf{D} = \text{diag}(\mathbf{d}_j)$ where each $\mathbf{d}_j$ matrix is diagonal and identical for all j utilities. This formulation suggests that the idiosyncratic errors of the indicators are the same for each utility. The initial value of $\mathbf{D}$ denoted by $\mathbf{D}^0$ is calculated as follows:

$$\mathbf{D}^0 = \text{diag}\left\{\mathbf{C}_{yy} - \left(\mathbf{B}^0 \times \left(\mathbf{B}^0\right)^T\right)\right\}. \tag{4.1}$$

Given the specification of $\mathbf{B}^0$ and $\mathbf{D}^0$, the first cycle of the 2CCEM algorithm outlined in (3.3)-(3.5) will yield ML estimates of $\mathbf{B}$ and $\mathbf{D}$. During the first iteration of the first cycle of the 2CCEM algorithm we set $\mathbf{T} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$. The ML estimates of $\mathbf{B}$ and $\mathbf{D}$ from the first cycle of the 2CCEM algorithm are used to obtain the initial value of $\mathbf{T}$ by running the following Vector Autoregression (VAR): $\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t$. In order to initialize the Kalman filter we need to make some assumption about the distribution of $\mathbf{U}_1$, the value of the state vector during the first period. deJong (1991) proposes the use of a diffuse prior density whereby $\mathbf{U}_1 \sim N(\breve{\mathbf{U}}_1, \mathbf{P}_1)$ with $\breve{\mathbf{U}}_1$ fixed at an arbitrary value and $\mathbf{P}_1 \to \infty$. We retain the assumption that $\mathbf{P}_1 \to \infty$ but substitute $\breve{\mathbf{U}}_1$ with the mean of $\mathbf{U}_1|\mathbf{Y}_1$ which, from (3.8), is equal to $\boldsymbol{\gamma}^T\mathbf{Y}_1$. Finally to ensure that the model is identifiable we make use of Remark 2.1 and set $\Gamma_{\mathbf{U}}(0) = \mathbf{I}$.

4.2 Results

Table 3 presents the likelihood at convergence of the three versions of the model, that is, specification $\mathbf{T}_1$, $\mathbf{T}_3$ and the specification where utilities of the same country share the same ϕ. In addition, we present the total number of estimated parameters in each specification as well as the resulting AIC.

Table 3: Results of alternative specifications

	Likelihood	Number of parameters	AIC
$\mathbf{T}_1$	-439.50	13	905.02
$\mathbf{T}_3$	-423.27	20	886.35
Same ϕ per country	-425.05	15	880.10

Based on the AIC the preferred model is the one where a different ϕ is estimated for every different country. The results of that specification are presented in Table 4.

Table 4: Factor loadings, idiosyncratic variance and AR(1) coefficient estimates

Indicators	**B**	**D**	Countries	**T**
Water Coverage	.504	.869	Armenia (utility 1)	.661
Sewerage Coverage	.183	.983	Moldova (utilities 2-4)	.821
Non Revenue Water	.359	.933	Peru (utilities 5-8)	.758
Affordability	.455	.893		
Collection period	.442	.899		
Operating Cost Coverage	.218	.975		

Our results indicate that water coverage, affordability and collection period are the three indicators that affect the performance index the most. Water coverage is ranked as the most important indicator, suggesting that providing water access to as many people as possible should be the primary focus of a water utility. The second most important priority should be keeping water provision affordable. Collection period ranks third, highlighting the importance of being able to promptly collect payments from customers. The fourth most important

indicator is non-revenue water. By minimizing leakages through the network as well as reducing the amount of water for which it is not getting any compensation a utility can help bolster its operational performance and increase the value of the smoothed index. Operating cost coverage ranks fifth. This result underlines the fact that due to the nature of the industry, public water utilities can be expected to operate at a loss. Sewerage coverage is the least important out of the six indicators suggesting that provision of sanitation services is not critical in judging a utility's performance.

Figure 1 illustrates the smoothed estimate of the performance index for each of the eight utilities in the sample. When referring to the smoothed index it is implied that the estimate includes information from the entire sample. For example, the performance of utility 1 in the year 2000 is assessed both with respect to how that utility did on that specific year, but also with respect to its performance before and after 2000.

5 Conclusion

Our paper contributes to the literature of DFMs by introducing a generalized dynamic factor model for panel data. Traditionally, DFMs have considered multiple attributes over several time periods for a single cross sectional unit, firm or economy (Stock and Watson, 1989). Even when multiple cross sectional units are considered (Forni et al. 2000) only a single unobserved index, common for all cross sectional units, is estimated for every time period. We develop a model that estimates one index for every cross sectional unit in every time period. In addition, we introduce the 2CCEM algorithm which is a novel estimation process that can handle panels of large dimensions. Previous dynamic factor models have used similar estimation algorithms that relied on two separate cycles. In the first cycle of those models, the parameters are

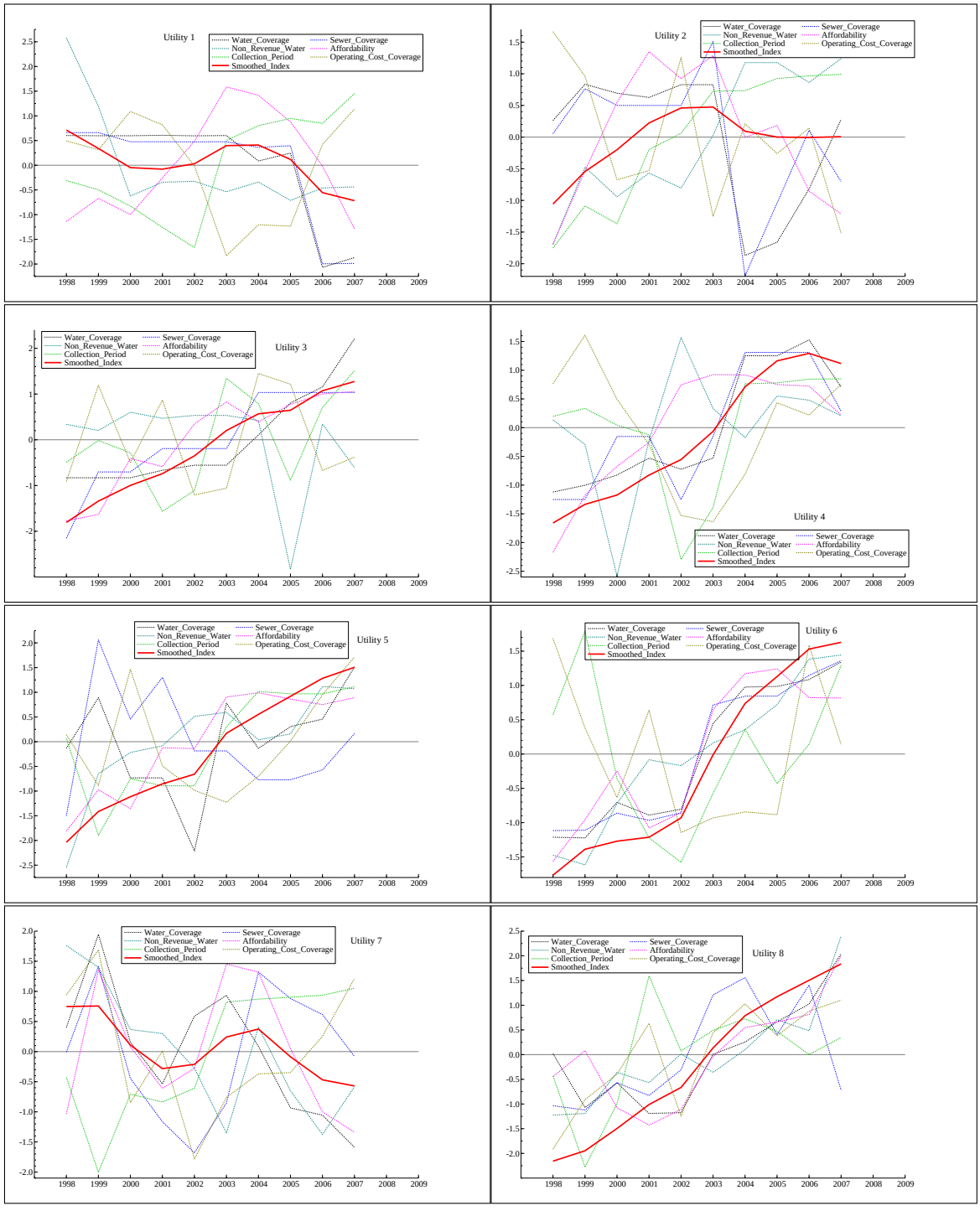

Figure 1: Standardized indicators and smoothed index for the 8 water utilities.

estimated using the EM algorithm. Then, conditional on those results, dynamic estimates of the parameters are obtained using the Kalman filter (Stock and Watson, 2010). However, those models achieve, at best, a conditional local maximum. The algorithm that we propose has the advantage of iteratively searching for an unconditional global maximum. Within every iteration each cycle is conditioned on the results of the previous cycle. Each iteration updates the estimated parameters until convergence is achieved. Therefore, the convergence point of previous estimation processes in the dynamic factor literature is, in principle, equivalent to the convergence point of only the first iteration of the 2CCEM algorithm.

In this paper, we apply the model on data from the IBNET database and estimate a performance index for water utilities. Future applications where our model could be applied include rankings of public institutions such as hospitals and universities (Grosskopf and Valdmanis 1987; Marginson 2007). In addition, our model can be used to estimate dynamic alternatives of existing static indices such as the Human Development Index (Sen and Anand 1994) or the Sustainability Index recently developed by FEEM (FEEM 2011).

Appendix

A Proof of Lemma 2.1

Assuming stationarity of the state variable we have:

$$\mathrm{Var}(\mathbf{U}_t) = \mathrm{Var}(\mathbf{U}_{t-1}) = \Gamma_{\mathbf{U}}(0), \tag{A.1}$$

Under assumption (A.1), we can rewrite (2.5), (2.6) and (2.7) and as follows:

$$\begin{aligned}
\Gamma_{\mathbf{Y}}(0) &= \mathbf{B}\Gamma_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{D}, && \text{(A.2)}\\
\Gamma_{\mathbf{Y}}(1) &= \mathbf{B}\mathbf{T}\Gamma_{\mathbf{U}}(0)\mathbf{B}'. && \text{(A.3)}\\
\Gamma_{\mathbf{U}}(0) &= \mathbf{T}\Gamma_{\mathbf{U}}(0)\mathbf{T}' + \mathbf{Q}, && \text{(A.4)}
\end{aligned}$$

A closed form solution for (A.4) can be obtained with the use of the vec operator as shown by Hamilton (1994, p.265):

$$\begin{aligned}
\text{vec}[\Gamma_{\mathbf{U}}(0)] &= \text{vec}[\mathbf{T}\Gamma_{\mathbf{U}}(0)\mathbf{T}' + \mathbf{Q}]\\
&= (\mathbf{T}\otimes\mathbf{T})\text{vec}[\Gamma_{\mathbf{U}}(0)] + \text{vec}(\mathbf{Q})\\
&= [\mathbf{I}_{m^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\text{vec}(\mathbf{Q}). && \text{(A.5)}
\end{aligned}$$

Using assumption (A.1) and applying the vec operator to (A.2) we have:

$$\begin{aligned}
\text{vec}[\Gamma_{\mathbf{Y}}(0)] &= \text{vec}[\mathbf{B}\Gamma_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{D}]\\
&= \text{vec}[\mathbf{B}\Gamma_{\mathbf{U}}(0)\mathbf{B}'] + \text{vec}[\mathbf{D}]\\
&= \mathbf{B}\otimes\mathbf{B}\text{vec}[\Gamma_{\mathbf{U}}(0)] + \text{vec}(\mathbf{D}) && \text{(A.6)}
\end{aligned}$$

Replacing (A.5) into (A.6) we have:

$$\text{vec}[\Gamma_{\mathbf{Y}}(0)] = \mathbf{B}\otimes\mathbf{B}\{[\mathbf{I}_{m^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\text{vec}(\mathbf{Q})\} + \text{vec}(\mathbf{D}) \qquad \text{(A.7)}$$

Similarly for (A.4) we have:

$$\begin{aligned} \text{vec}[\Gamma_{\mathbf{Y}}(1)] &= \text{vec}[\mathbf{BT}\Gamma_{\mathbf{U}}(0)\mathbf{B}'] \\ &= \mathbf{B} \otimes (\mathbf{BT})\text{vec}[\Gamma_{\mathbf{U}}(0)] \\ &= \mathbf{B} \otimes (\mathbf{BT})\{[\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1}\text{vec}(\mathbf{Q})\} \end{aligned} \tag{A.8}$$

Finally the general form of the autocovariance function of $\mathbf{Y}$ is:

$$\Gamma_{\mathbf{Y}}(h) = \mathbf{BT}\Gamma_{\mathbf{U}}(h-1)\mathbf{B}' \text{ for } h > 1, \tag{A.9}$$

where:

$$\Gamma_{\mathbf{U}}(h-1) = \mathbf{T}\Gamma_{\mathbf{U}}(h-1)\mathbf{T}' \Rightarrow \tag{A.10}$$

$$\text{vec}[\Gamma_{\mathbf{U}}(h-1)] = [\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1} \tag{A.11}$$

Replacing (A.11) into (A.9) and applying the vec operator we have:

$$\Gamma_{\mathbf{Y}}(h) = \mathbf{B} \otimes (\mathbf{BT})\{[\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1} \tag{A.12}$$

B Proof of Theorem 2.2

Identifiability of the model requires that in the system defined by (A.7) and (A.8) we have more equations than unknowns and that those equations are linear in their parameters. The

latter is accomplished by setting the following restriction:

$$\Gamma_{\mathbf{U}}(0) = \mathbf{C} \tag{B.1}$$

Applying the vec operator to (B.1) we have:

$$\text{vec}\Gamma_{\mathbf{U}}(0) = \text{vec}(\mathbf{C}) \tag{B.2}$$

Replacing (A.5) into (B.2) we have:

$$\begin{aligned} \text{vec}(\mathbf{C}) &= [\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1}\text{vec}(\mathbf{Q}) \Rightarrow \\ \text{vec}(\mathbf{Q}) &= [\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]\text{vec}(\mathbf{C}) \\ &= \mathbf{I}_{m^2}\text{vec}(\mathbf{C}) - \mathbf{T} \otimes \mathbf{T}\text{vec}(\mathbf{C}) \Rightarrow \\ \mathbf{Q} &= \mathbf{C} - \mathbf{T}\mathbf{C}\mathbf{T}' \end{aligned} \tag{B.3}$$

In the most general case of the model we have the following number of parameters: $mp \times m$ parameters in $\mathbf{B}$, mp parameters in $\mathbf{D}$ and m^2 parameters in $\mathbf{T}$.

There are as many equations as there are elements of $\Gamma_{\mathbf{Y}}(0)$ and $\Gamma_{\mathbf{Y}}(1)$. $\Gamma_{\mathbf{Y}}(0)$ is symmetric with $\frac{mp(mp+1)}{2}$ unique elements, while $\Gamma_{\mathbf{Y}}(1)$ is non-symmetric with m^2p^2 unique elements. Therefore, identifiability of the model requires that:

$$\begin{aligned} \frac{mp(mp+1)}{2} + m^2p^2 &> m^2p + mp + m^2 \\ m &> \frac{1}{3p - 2 - \frac{2}{p}} \end{aligned} \tag{B.4}$$

The denominator of (B.4) has two real roots, namely -0.15 and 1.48. Therefore, the necessary

condition for theoretical identifiability of the model requires that $m, p > 1$.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Boivin, J. and S. Ng (2006). Are more data always better for factor analysis? *Journal of Econometrics 132*(1), 169–194.

deJong, P. (1989). Smoothing and interpolation with the state-space model. *Journal of the American Statistical Association 84*(408), 1085–1088.

deJong, P. (1991). The diffuse kalman filter. *The Annals of Statistics 19*(2), 1073–1083.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Doz, C., D. Giannone, and L. Reichlin (2011). A quasi maximum likelihood approach for large approximate dynamic factor models. *Review of Economics and Statistics*.

Durbin, J. and S. Koopman (2001). *Time Series Analysis by State Space Methods*. Number 24 in Oxford Statistical Science Series. Oxford, U.K.: Oxford University Press.

FEEM (2011). FEEM sustainability index: Methodological report 2011. Technical report, Fondazione Eni Enrico Mattei.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics 82*(4), 540–554.

Geweke, J. (1977). The dynamic factor analysis of economic Time-Series model. In *Latent*

Variables in Socio-Economic Models, Contributions to Economic Analysis. Amsterdam, The Netherlands: North-Holland.

Grosskopf, S. and V. Valdmanis (1987). Measuring hospital performance: A non-parametric approach. *Journal of Health Economics 6*(2), 89–107.

Hamilton, J. D. (1994). *Time Series Analysis* (1 ed.). Princeton University Press.

Hotta, L. K. (1989). Identification of unobserved components models. *Journal of Time Series Analysis 10*(3), 25–270.

IBNET (2005). International benchmarking network for water and sanitation utilities. http://www.ib-net.org.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 35–45.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*(1), 65–79.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*(1), 117–126.

Marginson, S. (2007). Global university rankings: Implications in general and for australia. *Journal of Higher Education Policy and Management 29*(2), 131–142.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models* (first ed.). Wiley-Interscience.

McLachlan, G. J. and T. Krishnan (1996). *The EM Algorithm and Extensions* (1 ed.). Wiley-Interscience.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*(2), 267–278.

Meng, X.-L. and D. Van Dyk (1997). The EM algorithm-an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(3), 511–567.

Rubin, D. and D. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika 47*(1), 69–76.

Sargent, T. J. and C. Sims (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Paper 55, Federal Reserve Bank of Minneapolis.

Sen, A. and S. Anand (1994). Human development index: Methodology and measurement. Human Development Occasional Papers (1992-2007) HDOCPA-1994-02, Human Development Report Office (HDRO), United Nations Development Programme (UNDP).

Shumway, R. H. and D. S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis 3*(4), 253–264.

Stock, J. and M. Watson (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual 4*, 351–394.

Stock, J. and M. Watson (2010). Dynamic factor models. In *Oxford Handbook of Economic Forecasting*.

van den Berg, C. and A. Danilenko (2010). The IBNET water supply and sanitation performance blue book. Technical report, The World Bank, Washington DC.