Dynamic Factor Analysis for Short Panels: Estimating Performance Trajectories for Water Utilities

Nikolaos Zirogiannis, Indiana University Bloomington

nzirogia@indiana.edu

Yorghos Tripodis, Boston University

yorghos@bu.edu

Selected Paper prepared for presentation at the Agricultural and Applied Economics Association's 2014 AAEA Annual Meeting, Minneapolis, MN, July 27-29, 2014

Dynamic Factor Analysis for Short Panels: Estimating Performance Trajectories for Water Utilities*

Nikolaos Zirogiannis† and Yorghos Tripodis‡

Abstract

We develop a dynamic factor model for panel data with a short time dimension (i.e. n$\leq$ 15). Unlike most of the work in the DFM literature where one common factor is estimated for a group of cross sectional units, our interest lies in the estimation of a latent variable for each cross sectional unit at every point in time. This difference increases the computational challenges of the estimation process. To facilitate estimation we develop the "Two-Cycle Conditional Expectation-Maximization" (2CCEM) algorithm which is a variant of the EM algorithm and it's extensions (Dempster et al. 1977; Meng and Rubin 1993; Liu and Rubin 1994). Initially, the latent variable is estimated (first cycle)

*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, University of Massachusetts Amherst and Indiana University Bloomington 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.

and then the dynamic component 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. We provide simulation results demonstrating consistency of our 2CCEM estimator. One of the advantages of this work is that the estimation strategy can account for multiple cross sectional units with a short time dimension, and is flexible enough to be used in different types of applications. We apply our model to a dataset of 853 water and sanitation utilities from 45 countries and use the 2CCEM algorithm to estimate performance trajectories for each utility.
Keywords: Dynamic Factor Models, EM algorithm, Panel Data, State-Space models, Water utilities, IBNET

1 Introduction

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 (FA) models was limited to cross sectional datasets ignoring any dynamic analysis. Geweke (1977) as well as Sargent and Sims (1977) were the first to propose a new class of dynamic factor models (DFMs). Sargent and Sims (1977) introduced applications of both observable and unobservable index models estimated using their DFM. Stock and Watson (1989) built on those contributions using maximum likelihood to estimate a DFM for unobserved coincident and leading economic indices of the US economy.

Those initial DFMs were applied to macroeconomic data and focused on a specific country. Thus, they did not include a cross-sectional dimension. Forni and Reichlin (1996) as well as Forni et al. (2000) were the first to develop a DFM that could handle panel data (i.e. multiple observations, for multiple time periods, across multiple cross-sectional units). The focus of that work was to estimate unobserved indices in the form of common factors for groups cross sectional units. 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. 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. More recently Doz et al. (2012) addressed the issue of using principle components in DFMs of large dimensions. They argued 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. The inferential theory for DFMs with large dimensions (both time and cross section) has been examined by Bai (2003). 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.

However, prior research in the DFM literature ignores cases where the time dimension is short as well as cases where the interest lies in estimating one factor for every cross sectional unit as opposed to a group of cross sections. In the application we discuss in this paper the time dimension ranges between 5 and 15 years. Our goal is to estimate a dynamic index that assess the operational and financial performance of 853 water and sanitation utilities in 45 countries using data from the International Benchmarking Network (IBNET 2013). The index assumes the role of the latent variable in the DFM and is a trajectory summarizing information from several observable measures of performance.

2 Measuring performance of water utilities

The water utility industry is unique in several ways. More often than not, water utilities are government owned or managed by the state and are not subject to the same financial and operational constraints that firms in competitive markets are. Even when their financial performance is poor they will often receive some form of financial support by the state in order to avoid bankruptcy. This is primarily due to the public nature of the services they provide. Those characteristics make it increasingly difficult to externally evaluate the efficiency of a water utility (Van den Berg and Danilenko 2011).

Previous work in the field of performance measurement for water utilities and other non-profit institutions (i.e. hospitals, universities, etc.) has been conducted mainly through the use of Data Envelopment Analysis (Abbott and Cohen 2009). Data Envelopment Analysis (DEA) is a linear programming technique that measures efficiency by calculating the ratio of total inputs to total outputs for a given cross section (Charnes, Cooper, and Rhodes 1978). The main drawback of DEA is the lack of a random error term and the resulting omission of potential cross section level characteristics (Anwandter and Ozuna 2002). Furthermore, the choice of input and output variables in DEA models that assess the performance of water utilities is not consistent amongst different authors (Anwandter and Ozuna 2002; Abbott and Cohen 2009; Garcia-Valiñas and Muñiz 2007; Thanassoulis 2000). Our work considers the efficiency measure as an unobserved dynamic performance index that is estimated through the model that will be analyzed in section 4.2.

A critical issue in constructing performance 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 2011). The authors calculate a static index they call the "APGAR score" whose aim is to assess the health of a water utility based on a weighted sum of six indicators, namely 1) water coverage (percentage of the population within the utility's jurisdiction that has access to drinking water), 2) sewerage coverage (percentage of population within the utility's jurisdiction that has access access to sanitation services), 3) non revenue water (water provided to the network that is not being paid for), 4) affordability (percentage of the Gross National Income spent on water and sanitation services), 5) collection period (number of days it takes the water utility to get paid back by it's customers) and 6) operating cost coverage (ratio of operating revenues over operating costs). The "APGAR score" is a static index in that it evaluates water utilities based on performance at a given time period, without considering any information from the past.

Our dataset is comprised of 853 utilities from 45 countries. Even though the full IBNET database contains information on more than 2,000 water utilities, we keep only those utilities that have at least 5 years of data available. We consider the same observable indicators that Van den Berg and Danilenko use in their "APGAR score". Our goal is to estimate a dynamic performance index using the model that will be analyzed in section 4.2.

3 Contribution

Our work contributes to the DFM literature as well as the field of assessing performance for water utilities. Unlike previous work cited in section 1 we are interested in estimating a dynamic performance index (i.e. a latent variable) that is unique for each water utility and varies across time. We are able to assess performance in two ways: 1) first by summarizing information from the six observable indicators analyzed above into a single performance index, and 2) by estimating the trajectory of this index for each water utility and thus following the changes of performance over time. This sets our work apart from previous authors whose focus was on the estimation of a single common factor for a group of cross sections (Doz et al. 2012; Forni et al. 2000). In addition, the application that motivates our work has a very short time dimension ($n \leq 15$). As a result we face the computational challenge of having to estimate a large number of parameters using a short time dimension per panel. To address this challenge we develop a novel iterative estimation process, which we call the "Two-Cycle Conditional Expectation-Maximization" (2CCEM) algorithm and is a variant of the traditional EM algorithm developed by Dempster et al. (1977) and extended by Meng and Rubin (1993) and Liu and Rubin (1994). Initially, the unobserved performance index is estimated (first cycle) and then the dynamic component 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.

To our knowledge this is the first paper that uses a DFM to estimate dynamic performance indices of water utilities thus allowing for comparisons both between utilities and across time. Previous work in the water utility assessment literature has relied on static DEA analysis (Thanassoulis 2000; Cubbin and Tzanidakis 1998) and more recently on dynamic versions of DEA models (Coelli and Walding 2006; Garcia-Valiñas and Muñiz 2007). Nevertheless, even the dynamic models of DEA suffer from the lack of the error term mentioned previously.

By summarizing information from several time series of observable indicators our estimated index can succinctly communicate whether the utility has been performing well or not. This feature makes our work directly relevant to policy makers, since the estimated performance indices can be used to dynamically rank utilities and provide critical information that can assist in determining policy interventions. The contributions of this index to the field of performance measurement of water utilities are the following:

- It is a dynamic measure of performance since at every time period it incorporates information from the entire sample.
- The use of the EM algorithm in estimation allows for the presence of missing data.

- It is a more transparent modeling strategy since there is no subjective weighting of the six observed indicators. As will be discussed in section 4.2 the weights of the indicators are obtained using the estimated factor loadings.

The paper is organized as follows. In Section 4, we present the theoretical framework, and examine the various components of the model. We also discuss necessary conditions for identifiability. Section 5 presents the 2CCEM algorithm and illustrates the estimation process for each of the two cycles. In order to examine the asymptotic behavior of our 2CCEM estimator we conduct a Monte Carlo study. The results are presented in section 6. In Section 7, we apply our model to a longitudinal dataset of water and sanitation utilities from 45 countries. We discuss how we obtain initial values for the parameters and present estimation results. The final section draws conclusions.

4 A Dynamic Factor Model for Short Panels

The main contribution of our work lies in the development of a DFM for panels with a short time dimension (i.e. n$\leq$ 15) that can estimate a performance trajectory unique to every water utility unit in the database. We begin this section by presenting the notation that will be used throughout the paper.

4.1 Notation

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

- $i = 1,...,p$ denoting the number of observable performance indicators in the mode. These are the six indicators analyzed in section 2;
- $j = 1,...,m$ denoting the number of water utilities where $m = 853$;

- $t = 1,...,n$ denoting the time point of an observation. Our dataset consists of an unbalanced pane where $n \in [5,15]$;

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 water utilities over all n years.

4.2 The theoretical framework of the model

State space models have been used extensively, particularly in the early literature of DFMs, since they allow for 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. As mentioned in section 3, one of the contributions of our work is that a unique latent variable $\mathbf{U}_t$ is estimated for each water utility. This comes in contrast to previous work in the literature that estimates one latent variable $\mathbf{U}_t$ for a group of cross sectional units. In our application the state variable $\mathbf{U}_t$ assumes the role of the dynamic performance index that is estimate for every water utility. 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:

$$\mathbf{Y}_t = \mathbf{B}\mathbf{U}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(0,\mathbf{D}), \tag{4.1}$$

$$\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \eta_t, \qquad \eta_t \sim N(0,\mathbf{Q}), \tag{4.2}$$

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. These are the six observable indicators of perfromance analyzed in section 2. $\mathbf{T}$ is an $m \times m$ transition matrix that describes the Markovian nature of the unobserved state vector, and $\mathbf{e}_t$ and η_t are error terms (Koopman 1993). Equation (4.1) is known as the observation equation (or measurement equation) and equation (4.2) is called the state equation (or transition equation) and represents the first order autoregressive nature of the model. The state space formulation described in (4.1) and (4.2) 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 η_t for all $t = 1, ..., n$. In addition, the error terms $\mathbf{e}_t$ and η_t are assumed to be independent, identically distributed (i.i.d.) and mutually uncorrelated (deJong 1991; Kohn and Ansley 1989; Koopman et al. 1999).

The matrix of factor loadings, is block diagonal and has the form: $\underset{mp \times m}{\mathbf{B}} = \text{diag}(\mathbf{b})$ where $\mathbf{b}$ is a $p \times 1$ vector of the factor loadings. We assume that all factor loadings are the same across water utilities. This is a plausible assumption, since our goal is to estimate a dynamic performance 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. In addition, the zero block off-diagonal vectors of $\mathbf{B}$ imply that the observable indicators of utility A do not load on the performance index of utility B. This is a reasonable assumption in the context of our application since observable indicators of utility A are not likely to affect the performance of utility B.

All other matrices that include the parameters of the model, namely $\mathbf{D}$, $\mathbf{T}$ and $\mathbf{Q}$ are also diagonal. The variance of the idiosyncratic errors in $\mathbf{D}$ is $\underset{mp \times mp}{\mathbf{D}} = \text{diag}(\mathbf{d})$ where $\mathbf{d}$ is a $p \times p$ diagonal matrix representing the variance of the error term for every cross section. The transition matrix $\mathbf{T}$ has the form $\underset{m \times m}{\mathbf{T}} = \text{diag}(\phi)$ where ϕ is the autoregressive parameter that determines the effect through time of a utility's own perfromance index. That is, it conveys information as to how performance in time period $t-1$ affects performance in period t. Finally, $\mathbf{Q}$, is an $m \times m$ diagonal matrix with elements σ^2, the variance of the error term of the state equation, along its diagonal.

4.3 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 (4.1) and (4.2). 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 Ω, has the following structure:

$$\mathrm{Var}(\mathbf{Y}) = \underset{nmp\times nmp}{\Omega} = \begin{bmatrix} \mathrm{Var}(\mathbf{Y}_1) & \mathrm{Cov}(\mathbf{Y}_1\mathbf{Y}_2) & ... & \mathrm{Cov}(\mathbf{Y}_1\mathbf{Y}_n) \\ \mathrm{Cov}(\mathbf{Y}_2\mathbf{Y}_1) & \mathrm{Var}(\mathbf{Y}_2) & ... & ... \\ ... & ... & ... & ... \\ \mathrm{Cov}(\mathbf{Y}_n\mathbf{Y}_1) & ... & ... & \mathrm{Var}(\mathbf{Y}_n) \end{bmatrix}, \tag{4.3}$$

where $\mathrm{Cov}(\mathbf{Y}_t, \mathbf{Y}_t^*)$ is an $mp\times mp$ matrix, for $t, t^* = 1, ..., n$ and $t \neq t^*$. The off-diagonal elements of Ω capture the covariance of $\mathbf{Y}_t$ across time. For ease of presentation, and without loss of generality, we assume that $\mathrm{E}(\mathbf{Y}_t) = \mathrm{E}(\mathbf{U}_t) = 0$. 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 Σ. It follows from (4.1) and (4.2) that:

$$\Sigma = \mathrm{Var}(\mathbf{Y}_t) = \mathbf{B}\mathrm{Var}(\mathbf{U}_t)\mathbf{B}' + \mathbf{D}, \tag{4.4}$$

and

$$\mathrm{E}(\mathbf{Y}_{t+1}\mathbf{Y}_t') = \mathbf{BT}\mathrm{Var}(\mathbf{U}_t)\mathbf{B}'. \tag{4.5}$$

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

$$\mathrm{E}(\mathbf{U}_t\mathbf{U}_t') = \mathbf{T}\mathrm{Var}(\mathbf{U}_{t-1})\mathbf{T}' + \mathbf{Q}, \tag{4.6}$$

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

$$\mathrm{E}(\mathbf{Y}_t\mathbf{U}_t') = \mathrm{E}\left[(\mathbf{B}\mathbf{U}_t + \mathbf{e}_t)\mathbf{U}_t'\right] = \mathbf{B}\mathrm{Var}(\mathbf{U}_t). \tag{4.7}$$

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 (4.4)-(4.7). In order to obtain the necessary conditions for indentifiability, we first derive the autocovariance function of $\mathbf{Y}_t$ in the following lemma.

Lemma 4.1. *The autocovariance function of* $\mathbf{Y}_t$ *is:*

$$\begin{aligned}
\mathrm{vec}[\Gamma_{\mathbf{Y}}(0)] &= \mathbf{B}\otimes\mathbf{B}\{[\mathbf{I}_{m^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\mathrm{vec}(\mathbf{Q})\} + \mathrm{vec}(\mathbf{D}) && (4.8)\\
\mathrm{vec}[\Gamma_{\mathbf{Y}}(1)] &= \mathbf{B}\otimes(\mathbf{B}\mathbf{T})\{[\mathbf{I}_{m^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\mathrm{vec}(\mathbf{Q})\} && (4.9)\\
\mathrm{vec}[\Gamma_{\mathbf{Y}}(h)] &= \mathbf{B}\otimes(\mathbf{B}\mathbf{T})\{[\mathbf{I}_{m^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}, \textit{ for } h > 1 && (4.10)
\end{aligned}$$

Proof. The proof is provided in Appendix A. □

Theorem 4.2 provides the necessary conditions for the model to be identifiable.

Theorem 4.2. *The necessary conditions for the model in* (4.1) *and* (4.2) *to be identifiable are:*

1.

$$\Gamma_{\mathbf{U}}(0) = \mathbf{C}, \tag{4.11}$$

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

2.

$$m > \frac{1}{3p - 2 - \frac{2}{p}} \tag{4.12}$$

Proof. The proof is provided in Appendix B. □

The choice of **C** is arbitrary as long as the conditions for a symmetric positive definite matrix are satisfied.

Remark 4.1. 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 4.2 that, when $\mathbf{C} = \mathbf{I}$, the necessary conditions for identifiability imply that $\mathbf{Q} = \mathbf{I} - \mathbf{T}\mathbf{T}'$.

5 The 2CCEM algorithm

One of the contributions of our work is the development of the 2CCEM algorithm, which is a novel approach to the estimation of DFMs. The high dimensionality of the data vector $\mathbf{Y}_t$ ($m = 853$), the short time dimension per panel ($t \in [5, 15]$) as well as the fact that our goal is to estimate one latent variable for each cross section at every point in time, 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 (4.1) and (4.2) is:

$$L(\mathbf{B}, \mathbf{D}, \mathbf{T}, \mathbf{Q}; \mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{t=2}^{n} f(\mathbf{Y}_1) f_{\mathbf{Y}}(\mathbf{Y}_t; [\mathbf{B}, \mathbf{D}, \mathbf{T}, \mathbf{Q}] \, | \mathbb{Y}_{t-1}), \tag{5.1}$$

where $\mathbb{Y}_{t-1}$ represents the set of past observations $\mathbf{Y}_1, ..., \mathbf{Y}_{t-1}$ and the model parameters to be estimated are **B**, **D**, **T** and **Q**. We showed in Theorem 4.2 that the parameterization of **Q** depends on **T** for identifiability of the model. Therefore, although **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 (4.1) and (4.2) feasible through an iterative two-cycle process. The 2CCEM algorithm is an extension of the Expectation/Conditional Maximization Either (ECME) algorithm introduced by Liu and Rubin (1994) which is itself an extension of the EM algorithm (Dempster et al. 1977) and the ECM algorithm (Meng and Rubin 1993). 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 (4.1) and (4.2). 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 Ψ, if $\mathbf{Y}_t$ and $\mathbf{U}_t$ were fully observable, be:

$$\log L_c(\Psi) = \log f_c(\mathbf{Y}_t, \mathbf{U}_t; \Psi), \tag{5.2}$$

where the subscript c denotes the complete-data likelihood.

The 2CCEM algorithm starts by partitioning the vector of unknown parameters Ψ into (Ψ_1, Ψ_2) where Ψ_1 contains the elements of $\mathbf{B}$ and $\mathbf{D}$ that need to be estimated, while Ψ_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 $\Psi_1^{(k-1)}$ and $\Psi_2^{(k-1)}$ denote the initial values of Ψ where k denotes the number of iterations in the estimation process with $k = 1, ..., l$. 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. Each cycle includes two E-steps and 2 M-steps.

5.1 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:

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1;\Psi_1^{(k-1)},\Psi_2^{(k-1)}) = \mathrm{E}_{\boldsymbol{\Psi}_1}\left\{\sum_{t=1}^{n}\log\mathrm{L}_c\left(\Psi_1;\mathbf{U}_t|\mathbf{Y}_t,\Psi_1^{(k-1)},\Psi_2^{(k-1)}\right)\right\}. \tag{5.3}$$

The first M-step involves differentiating $\mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1;\Psi_1^{(k-1)},\Psi_2^{(k-1)})$ with respect to Ψ_1 in order to obtain $\Psi_1^{(k/2)}$:

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1^{(k/2)};\Psi_1^{(k-1)},\Psi_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1;\Psi_1^{(k-1)},\Psi_2^{(k-1)}), \tag{5.4}$$

The second E-step replaces $\mathbf{U}_t$ with it's filtered estimate. Here we calculate:

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1^{(k/2)};\Psi_1^{(k-1)},\Psi_2^{(k-1)}) = \mathrm{E}_{\boldsymbol{\Psi}_1}\left\{\sum_{t=1}^{n}\log\mathrm{L}_c\left(\Psi_1;\mathbf{U}_t|\mathbb{Y}_{t-1},\Psi_1^{(k-1)},\Psi_2^{(k-1)}\right)\right\}. \tag{5.5}$$

The second M-step maximizes $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ with respect to $\mathbf{B}$ using $\Psi_1^{(k/2)}$ as the initial value of the parameters. Our goal, in this step, is to obtain $\Psi_1^{(k)}$ such that:

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1^{(k)};\Psi_1^{(k-1)},\Psi_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_1}(\Psi_1^{(k/2)};\Psi_1^{(k-1)},\Psi_2^{(k-1)}) \tag{5.6}$$

5.1.1 Estimation of the first cycle

As mentioned in section 4.3 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 $\mathbf{Y}$, Σ, 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 (5.3) and treats the unobserved state vector $\mathbf{U}_t$ as missing data while iteratively maximizing $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ 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:

$$
\begin{aligned}
\mathrm{E}(\mathbf{Y}\mathbf{Y}^T|\mathbf{Y}) &= \mathbf{C}_{yy}, \\
\mathrm{E}(\mathbf{Y}^T\mathbf{U}|\mathbf{Y}) &= \mathbf{C}_{yy}\gamma, \\
\mathrm{E}(\mathbf{U}^T\mathbf{U}|\mathbf{Y}) &= \gamma^T\mathbf{C}_{yy}\gamma + n\omega, \qquad (5.7)
\end{aligned}
$$

where:

$$\gamma = \left(\mathbf{B}\mathbf{B}^T + \mathbf{D}\right)^{-1}\mathbf{B} \text{ and } \omega = \mathbf{I} - \gamma^T\mathbf{B}. \qquad (5.8)$$

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

$$\mathbf{U}_t|\mathbf{Y}_t \sim N(\gamma^T\mathbf{Y}_t, \mathbf{I} - \gamma^T\mathbf{B}). \qquad (5.9)$$

Equations (5.7) and (5.8) constitute the E-step of the first cycle of the 2CCEM algorithm illustrated in (5.3). The subsequent first M-step, illustrated in equation (5.4), is identical to the M-step of the traditional EM algorithm which involves replacing the sufficient statistics in (5.7) into $\mathbf{Z}_{\Psi_1}$ and differentiating with respect to Ψ_1. The functional form of $\mathbf{Z}_{\Psi_1}$ is:

$$
\begin{aligned}
\log L_c(\Psi_1) &= \frac{n}{2}\log\{|\mathbf{D}^{-1}| + \log|\mathbf{Q}^{-1}|\} - \frac{1}{2}\sum_{t=1}^{n}\{(\mathbf{y}_t - \mathbf{B}\hat{u}_t)^T\mathbf{D}^{-1}(\mathbf{y}_t - \mathbf{B}\hat{u}_t) \\
&\quad -(\hat{u}_{t+1} - \mathbf{T}\hat{u}_t)^T\mathbf{Q}^{-1}(\hat{u}_{t+1} - \mathbf{T}\hat{u}_t)\}. \qquad (5.10)
\end{aligned}
$$

Equation (5.10) is the complete data log likelihood; complete both in terms of data and parameters. Setting the first derivatives of $\mathbf{Z}_{\Psi_1}$ equal to zero yields the following first order conditions:

$$\mathbf{B}^{(k/2)} = \mathbf{C}_{yy}\gamma\left\{\gamma^T\mathbf{C}_{yy}\gamma+n\omega\right\}^{-1}, \tag{5.11}$$

$$\mathbf{D}^{(k/2)} = n^{-1}\mathrm{diag}\left\{\mathbf{C}_{yy}-\mathbf{C}_{yy}\gamma\mathbf{B}^T\right\}, \tag{5.12}$$

where $\mathbf{B}^{(k/2)}$ and $\mathbf{D}^{(k/2)}$ represent the updated values $\Psi_1^{(k/2)}$. The second E-step replaces the latent variable in (5.10) with it's filtered estimate. The subsequent second M-step, maximizes (5.10) through a Newton-Raphson algorithm, with respect to **B**, using (5.11) and (5.12) as initial values. Upon convergence of this maximization we obtain the final updated values for $\Psi_1^{(k)}$.

As mentioned in the beginning of section 5 our approach is an extension of the Expectation / Conditional Maximization Either (ECME) algorithm introduced by Liu and Rubin (1994) which is itself an extension of the EM algorithm (Dempster et al. 1977) and the ECM algorithm (Meng and Rubin 1993). The ECME algorithm uses the same first E-step and M-step as we do, but does not include a second E-step. In the second M-step Liu and Rubin (1994) maximize the log likelihood with respect to **D**, holding **B** fixed at $\mathbf{B}^{(k/2)}$ (McLachlan and Peel 2000).

5.2 Second cycle of the 2CCEM

In the first E-step of the second cycle we estimate $\Psi_2^{(k)}$. We proceed by calculating:

$$\mathbf{Z}_{\Psi_2}(\Psi_2;\Psi_1^{(k)},\Psi_2^{(k-1)}) = \mathrm{E}_{\Psi_2}\left\{\sum_{t=1}^{n}\log\mathrm{L}_c\left(\Psi_2;\mathbf{U}_t|\mathbf{Y}_t,\Psi_1^{(k)},\Psi_2^{(k-1)}\right)\right\}. \tag{5.13}$$

The first E-step involves forming the expected complete-data log likelihood by conditioning $\mathbf{Z}_{\Psi_2}$ on the estimates $\Psi_1^{(k)}$. The subsequent M-step involves differentiating $\mathbf{Z}_{\Psi_2}(\Psi_2;\Psi_1^{(k)},\Psi_2^{(k-1)})$ with respect to Ψ_2. We choose $\Psi_2^{(k)}$ such that:

$$\mathbf{Z}_{\Psi_2}(\Psi_2^{(k/2)};\Psi_1^{(k)},\Psi_2^{(k-1)}) \geq \mathbf{Z}_{\Psi_2}(\Psi_2;\Psi_1^{(k)},\Psi_2^{(k-1)}). \tag{5.14}$$

The second E-step is the equivalent of that in the first cycle. We replace $\mathbf{U}_t$ with it's filtered estimate to calculate:

$$\mathbf{Z}_{\Psi_2}(\Psi_2^{(k/2)};\Psi_1^{(k)},\Psi_2^{(k-1)}) = \mathrm{E}_{\Psi_2}\left\{\sum_{t=1}^{n}\log\mathrm{L}_c\left(\Psi_1;\mathbf{U}_t|\mathbb{Y}_{t-1},\Psi_1^{(k)},\Psi_2^{(k-1)}\right)\right\}. \tag{5.15}$$

Finally the second M-step maximizes $\mathbf{Z}_{\Psi_2}$ with respect to $\mathbf{T}$ using $\Psi_2^{(k/2)}$ as the initial value of the parameters. Here we obtain $\Psi_2{}^{(k)}$ such that:

$$\mathbf{Z}_{\Psi_2}(\Psi_2^{(k)};\Psi_1^{(k)},\Psi_2^{(k-1)}) \geq \mathbf{Z}_{\Psi_2}(\Psi_1^{(k/2)};\Psi_1^{(k-1)},\Psi_2^{(k-1)}) \tag{5.16}$$

Upon maximization of $\mathbf{Z}_{\Psi_2}$, the estimate $\Psi_2^{(k)}$ is used in the E-step of the first cycle. This iterative maximization process will continue until convergence of both likelihood functions $\mathbf{Z}_{\Psi_1}$ and $\mathbf{Z}_{\Psi_2}$ is achieved.

5.2.1 Estimation in the second cycle

In the second cycle we utilize the prediction error decomposition form which is the observational equivalent of the likelihood function in (5.10) (Harvey 1989) :

$$\log\mathbf{L}_c(\Psi_2) = n - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\left[\log|\mathbf{F}_t| + \mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t\right], \tag{5.17}$$

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, (5.17) is maximized with respect to Ψ_2 , as illustrated in (5.16).

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).

6 Monte Carlo study

In order to examine the performance of our 2CCEM estimator we conduct a Monte Carlo study. The basis for our simulations is the model proposed by Doz et al. (2012). We begin by defining $\mathbf{B} = \text{diag}(\mathbf{f})$, where $\mathbf{f}$ is a $p \times 1$ vector of factor loadings with $\mathbf{f}_{[k]} \sim \mathscr{U}(0,1)$ subject to $\sum_{k=1}^{p} \mathbf{f}_{[k]} = 1$. Furthermore, we set $\mathbf{D} = \text{diag}(\mathbf{d})$, where $\mathbf{d}$ a $p \times 1$ vector of variances for the idiosyncratic elements, with $\mathbf{d}_{[k]} = \mathbf{f}_{[k]} \frac{\beta_k}{1-\beta_k}$ with $\beta_k \sim \mathscr{U}(0.1, 0.9)$. Both $\mathbf{f}$ and $\mathbf{d}$ are held constant for each cross sectional unit. Finally we let $\mathbf{T} = \text{diag}(\phi)$ and set $\phi = 0.9$.

Our simulations are based on a series of combinations for the three dimensions of our dataset (i.e. cross-sectional, time and number of indicators). For each of the three dimensions we use different specifications. For the number of observable indicators, p, we use 5 and 10, for the number of cross-sectional units we use $10, 50, 100, 200$ and 300 and finally for the number of time periods n we use $3, 5, 7, 10$ and 15. Therefore we have a total of 50 sets of combinations of the three dimensions. For each of the 50 combinations we run 1,000 simulations based on the model described in (4.1) and (4.2) using the parameter specifications for $\mathbf{B}, \mathbf{D}$ and $\mathbf{T}$ described above.

Using the 2CCEM algorithm we estimate the latent variable $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2, \ldots, \hat{\mathbf{U}}_T)'$ for each combination of the three dimensions analyzed above. In order to evaluate the performance of the algorithm we calculate the following trace statistic which is a goodness of fit measure used by Doz et al. (2012) in their simulations. This statistic measures the accuracy of the estimation of the factors for each cross-section, versus the true simulated factors, as sample size increases.

$$\frac{\mathrm{tr}(\mathbf{U}'\hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{U})}{\mathrm{tr}(\mathbf{U}'\mathbf{U})}$$

We report the results of our Monte Carlo study in Table 1. The reported trace statistics are the averages across all 1000 simulations per combination. It is interesting to note that as the sample size increases in all dimensions, the accuracy of the estimated factors improves. For example, holding the number of cross sectional units constant, the trace statistic increases both along the time dimension and as the number of observable indicators increases. However, the increase in the accuracy of the estimators as the cross sectional dimension increases, diminishes after m reaches 100. This finding suggests that there are no significant consistency gains by greatly increasing the cross sectional dimension.

7 Application

We apply our model to a dataset of 853 water and sanitation utilities in 45 countries. Our goal is to estimate a dynamic performance index for every water utility that will summarize information from the six observables indicators outlined in section 2. Table 2 lists all 45 countries along with their respective number of utilities in our dataset. Table 3 presents descriptive statistics for each of the six observable indicators by continent. The data are obtained from the International Benchmarking

Network (IBNET) for Water and Sanitation Utilities (IBNET 2013). IBNET was launched in 1996 with the goal of facilitating a standardized comparison amongst water utilities with respect to their financial and operational performance.

7.1 Initial values

In this section we discuss the selection of initial values for each of the parameters. The initial value of $\mathbf{B}$ is denoted by $\mathbf{B}^0$. Every block diagonal vector $\mathbf{b}$ is denoted by $\mathbf{b}^0$ where $\mathbf{b}^0 = \left(\frac{1}{p}\right)\mathbf{i}_p$. The formulation of $\mathbf{D}$ discussed in section 4.2 assumes 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:

$$\mathbf{D}^0 = \text{diag}\left\{\mathbf{C}_{yy} - \left(\mathbf{B}^0 \times \left(\mathbf{B}^0\right)^T\right)\right\}. \tag{7.1}$$

Following the recommendations of Little (2009), we begin by transforming the indicators so that they are positively correlated to the dynamic performance index. We accomplish this by multiplying non-revenue water, affordability and collection period by -1, since those three indicators were negatively correlated to the index. Furthermore, to enable comparisons between the factor loadings, all indicators are standardized.

Given the specification of $\mathbf{B}^0$ and $\mathbf{D}^0$, the first cycle of the 2CCEM algorithm outlined in equation (5.3)-(5.6) 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 $\mathbf{Q} = 0$. 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): $\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \eta_t$. 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 $\mathbf{U}_1 \sim N(\check{\mathbf{U}}_1, \mathbf{P}_1)$ with $\check{\mathbf{U}}_1$ 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 (5.9), is equal to $\gamma^T\mathbf{Y}_1$. Finally to ensure that the model is identifiable we make use of Remark 2.1 and set $\Gamma_{\mathbf{U}}(0) = \mathbf{I}$.

7.2 Results

We initially run a model that estimates performance trajectories for each of the 853 utilities in our sample. The resulting factor loadings and error variances are illustrated in table 4. These results suggest that the two most important indicators of performance are water and sewerage coverage, followed by the remaining four indicators. This is plausible given the high number of low and middle income countries in our dataset. For those countries the primary driver of performance is provision of water and sanitation services with less of a focus on financial indicators like collection period or operating cost coverage.

We then run separate models for each of the 45 countries. For ease of presentation we illustrate the results from countries that have at least 10 utilities in tables 5 and 6. It is not surprising that the importance of the factor loadings (based on the rankings of their magnitude) change from country to country. This result is plausible and relies on the different conditions under which utilities operate in different countries. Croatian utilities for example demonstrate high factor loadings for water and sewerage coverage. The median values for those indicators in the country are 82% and 48% respectively, far below the median values for Europe (92% and 68%). It is therefore intuitive that greater importance is placed on those indicators. On the other hand, Bosnia and Herzegovina has a high factor loading for affordabililty. Looking more closely at this indicator we realize that the country lags it's European counter parts given that 50% more of Bosnia's GNI is spent paying for water and sewerage services compared to the relevant number in Europe (i.e. median affordability is 1.8% in Bosnia and Herzegovina compared to a European median of 1.2%).

In figure 1 we plot the estimated dynamic index in relation to the standardized observable indicators of performance as well as the static APGAR score. Similar graphs can be produced for each of the 853 water utilities in our sample. The top panel of figure 1 shows the performance trajectory of a utility in Moldova. A higher value of the dynamic index implies improved performance. As a result the Moldovan utility has been consistently improving it's operation with the exception of year 2010. The utility depicted in the bottom panel of figure 1 on the other hand demonstrates deteriorating performance until 2004.

At every point in time the dynamic index is estimated using information from both before and after a particular time period. It is therefore able to more accurately capture the performance trajectory of a utility, given that it is less sensitive to big jumps in the value of the observable indicators. This property of the dynamic index can be best exemplified when compared with the trajectory of the static APGAR score that is also depicted in figure 1. The latter demonstrates significant variation precisely because it is heavily affected by coincident changes in the observable indicators.

8 Conclusion

Our paper contributes to the literature of DFMs by introducing a dynamic factor model for panels with a short time dimension. Most of the DFM literature has so far not considered panels with this attribute (Stock and Watson 1989; Doz et al. 2012; Forni et al. 2000). To address the computational complexities that such an estimation process entails, we introduce the 2CCEM algorithm.

Previous DFMs have used similar estimation algorithms that relied on two separate cycles. In the first cycle of those models, the parameters are 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 have illustrated the conditions that are required for the model to be identifiable as well as provided the results of a Monte Carlo study that demonstrates consistency of the estimator. Future work will focus on estimating different specifications of the model with a larger number of parameters.

Our application utilizes a dataset of 853 utilities from 45 countries and estimates a unique performance trajectory for every water utility. The estimation algorithm can account for the short time dimension, the missing observations as well as the unbalanced nature of the panel. The performance index that we estimate is a superior benchmarking tool compared to previous work in the literature since it incorporates information from the entire sample at very point in time, thus estimating a performance measure that is less susceptible to variability caused by external shocks.

Appendix A: Proof of Lemma 4.1

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 (4.4)-(4.6) as follows:

$$\Gamma_{\mathbf{Y}}(0) = \mathbf{B}\Gamma_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{D}, \tag{A.2}$$

$$\Gamma_{\mathbf{Y}}(1) = \mathbf{B}\mathbf{T}\Gamma_{\mathbf{U}}(0)\mathbf{B}'. \tag{A.3}$$

$$\Gamma_{\mathbf{U}}(0) = \mathbf{T}\Gamma_{\mathbf{U}}(0)\mathbf{T}' + \mathbf{Q}, \tag{A.4}$$

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

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

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

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

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

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

Similarly for (A.4) we have:

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

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

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

where:

$$\begin{aligned}
\Gamma_{\mathbf{U}}(h-1) &= \mathbf{T}\Gamma_{\mathbf{U}}(h-1)\mathbf{T}' \Rightarrow \qquad (A.10) \\
\mathrm{vec}[\Gamma_{\mathbf{U}}(h-1)] &= [\mathbf{I}_{m^2} - \mathbf{T} \otimes \mathbf{T}]^{-1} \qquad (A.11)
\end{aligned}$$

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

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

Appendix B: Proof of Theorem 4.2

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:

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

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:

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

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 $m^2 p^2$ unique elements. Therefore, identifiability of the model requires that:

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

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

Abbott, M. and B. Cohen (2009). Productivity and efficiency in the water industry. *Utilities Policy 17*(3), 233–244.

Anwandter, L. and T. Ozuna (2002). Can public sector reforms improve the efficiency of public water utilities? *Environment and Development Economics 7*(4), 687–700.

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.

Charnes, A., W. W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European journal of operational research 2*(6), 429–444.

Coelli, T. and S. Walding (2006). Performance measurement in the australian water supply industry: A preliminary analysis. *Performance measurement and regulation of network utilities*, 29–62.

Cubbin, J. and G. Tzanidakis (1998). Regression versus data envelopment analysis for efficiency measurement: an application to the england and wales regulated water industry. *Utilities Policy 7*(2), 75–85.

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 (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics 94*(4), 1014–1024.

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.

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.

Forni, M. and L. Reichlin (1996). Dynamic common factors in large cross-sections. *Empirical economics 21*(1), 27–42.

Garcia-Valiñas, M. A. and M. A. Muñiz (2007). Is dea useful in the regulation of water utilities? a dynamic efficiency evaluation (a dynamic efficiency evaluation of water utilities). *Applied Economics 39*(2), 245–252.

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.

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

Harvey, A. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

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

IBNET (2013). 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.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space using ssfpack 2.2. *The Econometrics Journal 2*(1), 107–160.

Little, T. D. (2009). *Longitudinal structural equation modeling*. Guilford Press.

Liu, C. and D. B. Rubin (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika 81*(4), 633–648.

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.

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*.

Thanassoulis, E. (2000). The use of data envelopment analysis in the regulation of uk water utilities: water distribution. *European Journal of Operational Research 126*(2), 436–453.

Van den Berg, C. and A. Danilenko (2011). *The IBNET Water Supply and Sanitation Performance Blue Book: The International Benchmarking Network of Water and Sanitation Utilities Databook*. World Bank Publications.

Table 1: Performance of factor estimators from 1000 simulations

		T=3	T=5	T=7	T=10	T=15
	p=5	0.586	0.615	0.645	0.652	0.674
n=10	p=10	0.673	0.679	0.693	0.705	0.720
	p=5	0.617	0.672	0.681	0.696	0.708
n=50	p=10	0.744	0.750	0.745	0.752	0.764
	p=5	0.657	0.676	0.690	0.708	0.718
n=100	p=10	0.759	0.752	0.754	0.755	0.774
	p=5	0.656	0.686	0.697	0.708	0.720
n=200	p=10	0.757	0.762	0.764	0.766	0.776
	p=5	0.661	0.691	0.700	0.704	0.719
n=300	p=10	0.758	0.760	0.761	0.764	0.777

Table 2: Countries and number of utilities by continent

Continent	Country	# of utilities	Continent	Country	# of utilities
Africa	Benin	(1)	Europe	Albania	(55)
	Congo. Dem. Rep.	(1)		Belarus	(21)
	Cote d'Ivoire	(1)		Bosnia and Herzegovina	(19)
	Ghana	(1)		Bulgaria	(9)
	Namibia	(1)		Croatia	(12)
	South Africa	(12)		Czech Republic	(20)
	Sudan	(3)		Hungary	(21)
	Tanzania	(9)		Macedonia. FYR	(2)
	Togo	(1)		Moldova	(39)
	Uganda	(1)		Poland	(35)
	Zambia	(5)		Romania	(26)
Asia	Armenia	(3)		Russia	(81)
	Azerbaijan	(1)		Slovakia	(2)
	Georgia	(24)		Turkey	(19)
	Kazakhstan	(15)		Ukraine	(81)
	Kyrgyz Republic	(4)	South America	Mexico	(12)
	Mongolia	(1)		Panama	(1)
	Pakistan	(4)		Argentina	(4)
	Philippines	(1)		Bolivia	(1)
	Tajikistan	(9)		Brazil	(230)
	Uzbekistan	(8)		Chile	(1)
	Vietnam	(8)		Peru	(47)
				Uruguay	(1)

Table 3: Descriptive statistics by continent

Africa

	First quantile	Median	Third quantile
water coverage	60%	85%	98%
sewerage coverage	4%	24%	71%
non revenue water (m^3/km/day)	9.2	20.2	41
affordability	1.60%	2.40%	4%
collection period (days)	62	105	226
operating cost coverage ratio	0.86	1.05	1.31

Europe

	First quantile	Median	Third quantile
water coverage	71%	92%	100%
sewerage coverage	39%	68%	88%
non revenue water (m^3/km/day)	6.4	18	47.8
affordability	0.90%	1.20%	2%
collection period (days)	69	126	245
operating cost coverage ratio	0.87	1.05	1.26

Asia

	First quantile	Median	Third quantile
water coverage	50%	74%	100%
sewerage coverage	20%	43%	66%
non revenue water (m^3/km/day)	10.7	23.6	79.2
affordability	0.50%	0.90%	1%
collection period (days)	119	316	823
operating cost coverage ratio	0.79	1	1.34

South America

	First quantile	Median	Third quantile
water coverage	80%	92%	99%
sewerage coverage	23%	65%	87%
non revenue water (m^3/km/day)	13.7	29.8	51.3
affordability	0.60%	0.80%	1%
collection period (days)	52	97	176
operating cost coverage ratio	0.93	1.08	1.32

Table 4: Factor loadings, idiosyncratic variance and AR(1) coefficient estimates for all 853 utilities

Indicators	**B**	**D**
Water Coverage	0.850	0.826
Sewerage Coverage	0.860	0.849
Non Revenue Water	0.731	0.840
Affordability	0.712	0.818
Collection period	0.715	0.812
Operating Cost Coverage	0.650	1.017

Table 5: Factor loadings, idiosyncratic variance and AR(1) coefficient estimates by country

	Country					
	Albania		Belarus		Bosnia Herzegovina	
Indicators	**B**	**D**	**B**	**D**	**B**	**D**
Water Coverage	0.183	0.586	0.202	0.275	0.162	0.492
Sewerage Coverage	0.224	0.379	0.233	0.315	0.162	0.422
Non Revenue Water	0.177	0.532	0.239	0.237	0.110	0.356
Affordability	0.080	0.666	0.120	0.362	0.331	0.405
Collection period	0.024	0.557	0.000	0.346	0.169	0.711
Operating Cost Coverage	0.313	0.474	0.205	0.409	0.065	0.493
ϕ	0.955		0.617		0.765	
	Country					
	Brazil		Croatia		Czech Republic	
Indicators	**B**	**D**	**B**	**D**	**B**	**D**
Water Coverage	0.1458	0.4497	0.2922	0.5700	0.2262	0.6335
Sewerage Coverage	0.1780	0.4117	0.3081	0.5430	0.3538	0.4185
Non Revenue Water	0.2174	0.3899	0.0765	0.9293	0.1622	0.5347
Affordability	0.2040	0.4368	0.0903	0.6698	0.1866	0.6036
Collection period	0.1300	0.4155	0.0000	0.6549	0.0712	0.6875
Operating Cost Coverage	0.1248	0.4794	0.2329	0.6421	0.0000	0.5592
ϕ	0.8637		0.8998		0.9168	
	Country					
	Georgia		Hungary		Kazakhstan	
Indicators	**B**	**D**	**B**	**D**	**B**	**D**
Water Coverage	0.341	0.481	0.065	0.635	0.279	0.392
Sewerage Coverage	0.195	0.449	0.221	0.542	0.329	0.385
Non Revenue Water	0.255	0.302	0.239	0.507	0.209	0.509
Affordability	0.187	0.503	0.142	0.581	0.043	0.576
Collection period	0.018	0.450	0.141	0.494	0.114	0.507
Operating Cost Coverage	0.005	0.766	0.193	0.548	0.027	0.639
ϕ	0.940		0.648		0.899	

Table 6: Factor loadings, idiosyncratic variance and AR(1) coefficient estimates by country

	Country					
	Mexico		Moldova		Peru	
Indicators	**B**	**D**	**B**	**D**	**B**	**D**
Water Coverage	0.251	0.212	0.342	0.563	0.216	0.678
Sewerage Coverage	0.251	0.213	0.313	0.524	0.228	0.663
Non Revenue Water	0.255	0.234	0.194	0.450	0.069	0.649
Affordability	0.052	0.925	0.103	0.740	0.155	0.749
Collection period	0.000	0.873	0.034	0.836	0.134	0.695
Operating Cost Coverage	0.190	0.488	0.014	0.898	0.197	0.752
ϕ	0.491		0.940		0.709	

	Country					
	Poland		Romania		Russia	
Indicators	**B**	**D**	**B**	**D**	**B**	**D**
Water Coverage	0.110	0.323	0.158	0.622	0.259	0.629
Sewerage Coverage	0.277	0.526	0.122	0.538	0.283	0.577
Non Revenue Water	0.202	0.639	0.262	0.545	0.163	0.798
Affordability	0.134	0.665	0.126	0.670	0.129	0.681
Collection period	0.114	0.717	0.113	0.388	0.026	0.731
Operating Cost Coverage	0.164	0.599	0.218	0.655	0.139	0.855
ϕ	0.928		0.869		0.700	

	Country			
	Turkey		Ukraine	
Indicators	**B**	**D**	**B**	**D**
Water Coverage	0.333	0.580	0.245	0.230
Sewerage Coverage	0.267	0.718	0.241	0.257
Non Revenue Water	0.273	0.301	0.220	0.240
Affordability	0.127	0.663	0.117	0.343
Collection period	0.000	0.181	0.072	0.357
Operating Cost Coverage	0.000	0.712	0.104	0.290
ϕ	0.836		0.730	

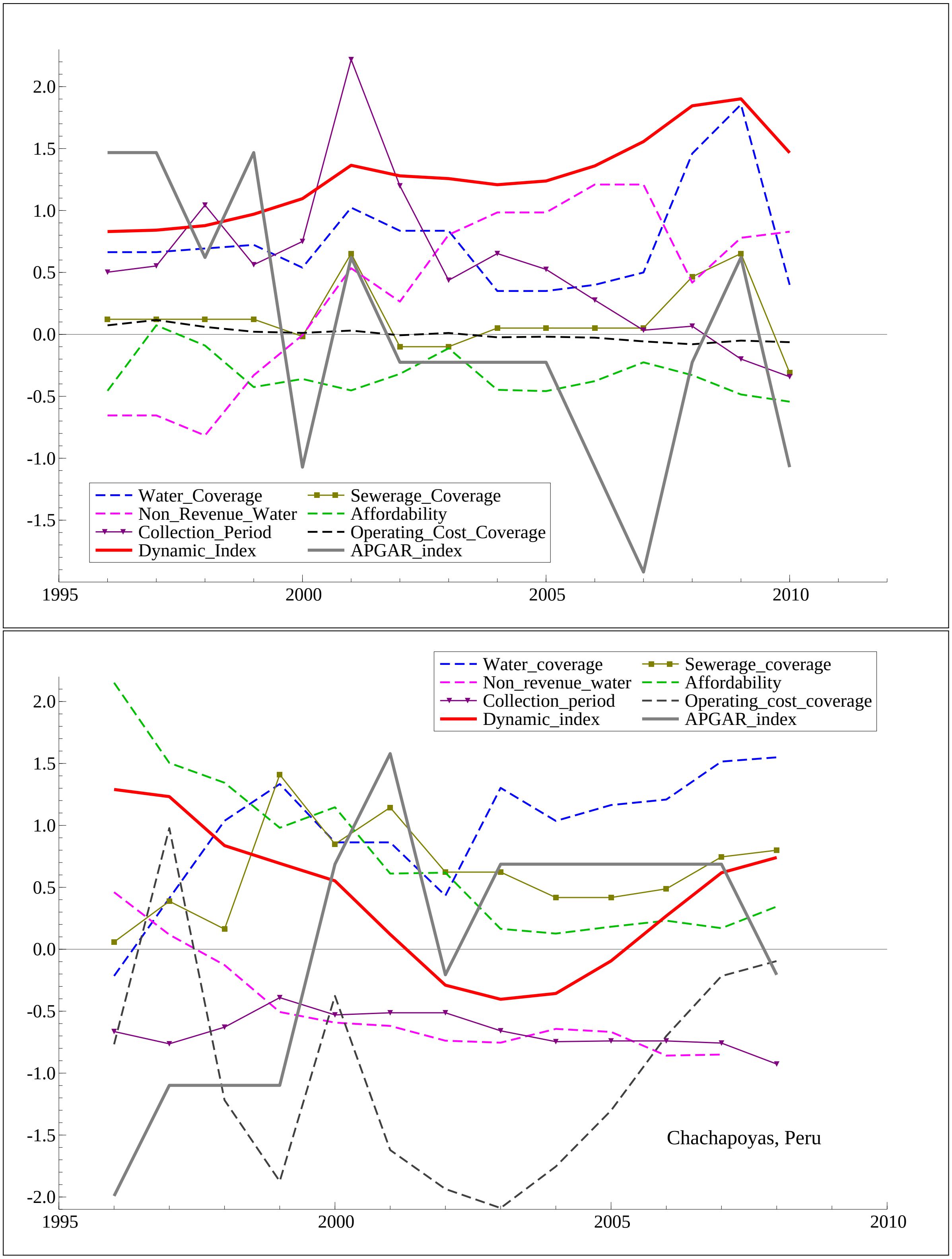

Figure 1: The estimated dynamic index with relation to the observable indicators and the static APGAR index