*University of Massachusetts Amherst*
**Department of Resource Economics**
**Working Paper No. 2013-1**
**http://www.umass.edu/resec/workingpapers**

# A Generalized Dynamic Factor Model for Panel Data: Estimation with a Two-Cycle Conditional Expectation-Maximization Algorithm*

**Nikolaos Zirogiannis[1], Yorghos Tripodis[2]**

January 15, 2013

Abstract:
We develop a generalized dynamic factor model for panel data with the goal of estimating an unobserved index. While similar models have been developed in the literature of dynamic factor analysis, our contribution is threefold. First, contrary to simple dynamic factor analysis where multiple attributes of the same subject are measured at each time period, our model also accounts for multiple subjects. It is therefore suitable to a panel data framework. Second, our model estimates a unique unobserved index for every subject for every time period, as opposed to previous work where a temporal index common to all subjects was used. Third, we develop a novel iterative estimation process which we call the Two-Cycle Conditional Expectation-Maximization (2CCEM) algorithm and is flexible enough to handle a variety of different types of datasets. The model is applied on a panel measuring attributes related to the operation of water and sanitation utilities.

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

JEL Classification: C32, C33, C51, Q25

---

* The authors would like to thank Alexander Danilenko for providing data from the International Benchmarking Network (www.ib-net.org). 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.

[1] Nikolaos Zirogiannis, Department of Resource Economics
University of Massachusetts, 80 Campus Center Way
Amherst, MA 01003
E: nzirogia@som.umass.edu P: 413-545-5736 F: 413-545-5853

[2] Yorghos Tripodis, Department of Biostatistics
801 Massachusetts Avenue, Boston University School of Public Health
Boston, MA 02118
E: yorghos@bu.edu P: 617-638-5844

# 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.

Methods involving estimation of latent variables have been gaining increasing attention in various fields of research, with factor analysis being one of the most important. 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 that contribution by estimating 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 was limited by the fact that it could not handle panel data, that is, multiple variables for multiple individuals spanning several years. Forni et al. (2000) extended DFMs by allowing for panel data estimation. They developed a generalized dynamic factor model that estimated one unobserved index for all individuals 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. A large body of literature has focused on the macroeconomic applications of such models (Stock and Watson, 2002, Forni et al., 2001). Bernanke and Boivin (2003) suggested that the model of Stock and Watson (2002) can assist the U.S. Federal Reserve in constructing macroeconomic indices using a larger number of indicators, compared to the relatively limited amount of information that the Fed has traditionally used, adding to the informative and predictive power of these indicators. 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. Jungbacker et al. (2011) use a similar maximum likelihood approach for DFMs and extend it to account for missing data.

Our work contributes to this literature by developing a generalized dynamic factor model for panel data. We develop a novel iterative estimation process, which we call "Two-Cycle Conditional Expectation-Maximization" (2CCEM) algorithm. Initially, the unobserved index is estimated and then the dynamic component of the index is incorporated into the estimation process. Our estimation strategy can account for multiple individuals, 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 individuals in their sample, we estimate one latent index for every individual.

The paper is organized as follows. In section 2, we present the theoretical framework, examine the correlation structure between the observed data and the unobserved index and discuss conditions for identifiability of the model. In addition, we illustrate various parameter formulations. Section 3 presents the 2CCEM algorithm and illustrates the estimation process for each of the two cycles. In section 4, we describe the data where we apply the model and discuss how we obtain initial values for the parameters. The section concludes with estimation results, diagnostic checking and simulations. In the final section, we draw conclusions based on the estimation results 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 cross correlations between individuals and is applicable to a panel data setting. In this section we present the theoretical foundation of our model, describe each of its components, address identifiability and illustrate various parameter formulations. 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 $j^{th}$ individual at time $t$ with:

- $i = 1, ..., p$ denoting the number of observed variables (indicators) in the model;
- $j = 1, ..., \theta$ denoting the number of individuals;
- $t = 1, ..., n$ 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 $\theta p \times 1$ vector with elements, $y_{ij,t}$, for $t$ fixed with $i = 1, ..., p$ and $j = 1, ..., \theta$;
- $\mathbf{Y}$ be a $n\theta p \times 1$ vector of all $p$ indicators for all $\theta$ individuals 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 $\theta$ 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:

$$\mathbf{Y}_t = \mathbf{B}\mathbf{U}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{D}), \tag{1}$$

$$\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}), \tag{2}$$

where $\mathbf{B}$ is the matrix of factor loadings with dimensions $\theta p \times \theta$, $\mathbf{U}_t$ is the $\theta \times 1$ unobserved state vector, $\mathbf{Y}_t$ is a $\theta p \times 1$ vector of observed variables, $\mathbf{T}$ is a $\theta \times \theta$ 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; Durbin and Koopman, 2001, p. 65). Equation (1) is known as the observation equation (or measurement equation) and equation (2) is called the state equation (or transition equation) and represents the first order autoregressive nature of the model (Harvey, 1991. p. 100). The state space formulation described in (1) and (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).

## 2.3 Correlation structure

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 $\boldsymbol{\Omega}$, is a $n\theta p \times n\theta p$ matrix with the following structure:

$$\text{Var}(\mathbf{Y}) = \underset{n\theta p \times n\theta p}{\boldsymbol{\Omega}} = \begin{bmatrix} \text{Var}(\mathbf{Y}_1) & \text{Cov}(\mathbf{Y}_1\mathbf{Y}_2) & ... & \text{Cov}(\mathbf{Y}_n\mathbf{Y}_1) \\ \text{Cov}(\mathbf{Y}_2\mathbf{Y}_1) & \text{Var}(\mathbf{Y}_2) & ... & ... \\ ... & ... & ... & ... \\ \text{Cov}(\mathbf{Y}_1\mathbf{Y}_n) & ... & ... & \text{Var}(\mathbf{Y}_n) \end{bmatrix}, \tag{3}$$

where $\text{Cov}(\mathbf{Y}_t, \mathbf{Y}_{t^*})$, with $t, t^* = 1, ..., n$ and $t \neq t^*$, is a $\theta p \times \theta p$ matrix. For ease of presentation, and without loss of generality, we assume that $\text{E}(\mathbf{Y}_t) = 0$. The unconditional covariance matrix of $\mathbf{Y}_t$, that is, the covariance matrix of all indicators for all individuals at a given time period $t$, is denoted by $\boldsymbol{\Sigma}$ and has the following structure:

$$\text{Var}(\mathbf{Y}_t) = \underset{\theta p \times \theta p}{\boldsymbol{\Sigma}} = \begin{bmatrix} \text{Var}(\mathbf{Y}_{1,t}) & \text{E}(\mathbf{Y}_{1,t}\mathbf{Y}_{2,t}) & ... & \text{E}(\mathbf{Y}_{1,t}\mathbf{Y}_{\theta,t}) \\ \text{E}(\mathbf{Y}_{2,t}\mathbf{Y}_{1,t}) & \text{Var}(\mathbf{Y}_{2,t}) & ... & ... \\ ... & ... & ... & ... \\ \text{E}(\mathbf{Y}_{\theta,t}\mathbf{Y}_{1,t}) & ... & ... & \text{Var}(\mathbf{Y}_{\theta,t}) \end{bmatrix}, \tag{4}$$

where $\text{Var}(\mathbf{Y}_{j,t})$ is a $p \times p$ covariance matrix of all $p$ indicators of the $j^{th}$ individual at time period $t$. It follows from (1) that:

$$\boldsymbol{\Sigma} = \text{Var}(\mathbf{Y}_t) = \text{E}(\mathbf{Y}_t\mathbf{Y}_t') = \text{E}([\mathbf{B}\mathbf{U}_t + \mathbf{e}_t]\,[\mathbf{B}\mathbf{U}_t + \mathbf{e}_t]') = \mathbf{B}\text{Var}(\mathbf{U}_t)\mathbf{B}' + \mathbf{D}. \tag{5}$$

The matrix $\boldsymbol{\Sigma}$ can be decomposed in two parts: 1) $\mathbf{B}\text{Var}(\mathbf{U}_t)\mathbf{B}^T$, known as the communality of the indicators, represents the variance of $\mathbf{Y}_t$ shared by all indicators via the unobserved state vector $\mathbf{U}_t$ and 2) $\mathbf{D}$ is the specific or unique variance and relates to the variability of $\mathbf{Y}_t$ that is not shared with other indicators (Everitt and Dunn, 1998). $\mathbf{D}$ is a $\theta p \times \theta p$ diagonal covariance matrix of the following form:

$$\mathbf{D} \quad = \quad \text{diag}(\sigma_{11}^2, ..., \sigma_{p\theta}^2). \tag{6}$$

Note that the diagonality of $\mathbf{D}$ is an important assumption of factor analysis. Given the unobserved state vector, the observed variables are independent (Rubin and Thayer, 1982; Ghahramani and Hinton, 1997), that is:

$$\text{Cov}(y_{ij,t}, y_{i*j,t}|\mathbf{U}_t) = \mathbf{0} \text{ and } \text{Var}(\mathbf{Y}_t|\mathbf{U}_t) = \mathbf{D},$$

where $i, i^* = 1, ..., p$ and $i \neq i^*$. In other words, in a factor analysis framework no correlation exists between the idiosyncratic errors $\mathbf{e}_t$. Any correlation between indicators is part of the unobserved common factor.

The off-diagonal elements of $\mathbf{\Omega}$ capture the covariance of $\mathbf{Y}_t$ across time. It follows from (1) and (2) that:

$$\begin{aligned} \text{E}(\mathbf{Y}_{t+1}\mathbf{Y}_t') &= \text{E}([\mathbf{B}\mathbf{U}_{t+1} + \mathbf{e}_{t+1}][\mathbf{B}\mathbf{U}_t + \mathbf{e}_t]') = \\ &= \text{E}([\mathbf{B}(\mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t) + \mathbf{e}_{t+1}][\mathbf{B}\mathbf{U}_t + \mathbf{e}_t]') = \\ &= \mathbf{B}\mathbf{T}\text{Var}(\mathbf{U}_t)\mathbf{B}', \end{aligned} \tag{7}$$

where (7) can be generalized as follows:

$$\text{E}(\mathbf{Y}_{t+h}\mathbf{Y}_t') = \mathbf{B}\mathbf{T}^h\text{Var}(\mathbf{U}_t)\mathbf{B}', \text{ for } h \geq 1.$$

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

$$\begin{aligned} \text{E}(\mathbf{U}_t\mathbf{U}_t') &= \text{E}\left[(\mathbf{T}\mathbf{U}_{t-1} + \boldsymbol{\eta}_{t-1})(\mathbf{T}\mathbf{U}_{t-1} + \boldsymbol{\eta}_{t-1})'\right] = \\ &= \mathbf{T}\text{Var}(\mathbf{U}_{t-1})\mathbf{T}' + \mathbf{Q}. \end{aligned} \tag{8}$$

Finally, $\text{E}(\mathbf{Y}_t\mathbf{U}_t)$ is:

$$\text{E}(\mathbf{Y}_t\mathbf{U}_t') = \text{E}\left[(\mathbf{B}\mathbf{U}_t + \mathbf{e}_t)\mathbf{U}_t'\right] = \mathbf{B}\text{Var}(\mathbf{U}_t). \tag{9}$$

From (5), (8) and (9) we determine the moments of the joint multivariate normal vector $(\mathbf{Y}_t^T, \mathbf{U}_t^T)^T$ with mean:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and a covariance matrix that can be calculated recursively, using the following equations:

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_t \mathbf{Y}_t') &= \mathbf{B}\mathrm{Var}(\mathbf{U}_t)\mathbf{B}' + \mathbf{D}, \\ \mathrm{E}(\mathbf{U}_t \mathbf{U}_t') &= \mathbf{T}\mathrm{Var}(\mathbf{U}_{t-1})\mathbf{T}' + \mathbf{Q}, \\ \mathrm{E}(\mathbf{Y}_t \mathbf{U}_t') &= \mathbf{B}\mathrm{Var}(\mathbf{U}_t). \end{aligned} \tag{10}$$

## 2.4 Identifiability

A central issue in the literature of unobserved component models is identifiability. Two parameters points (or structures) are observationally equivalent if they have the same joint density function. A structure is identifiable if there is no other structure which is observationally equivalent. A model, on the other hand, is identifiable if all its structures are identifiable (Rothenberg, 1971; Harvey, 1991, p. 205). Rather than invoking this general result it is preferable to 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 identifiability in the model specified in (1) and (2). In order to achieve that, we derive general formulas for the autocovariance function of our model.

Let $\boldsymbol{\phi}_d = \mathrm{vec}\left[\mathrm{diag}(\mathbf{T})\right]$ and $\boldsymbol{\phi}_{(\bullet)}$ be an operator that multiplies the $j^{th}$ element of the vector $\boldsymbol{\phi}_d$ with every element in rows $[(j-1)\,p+1]$ to $[jp]$ of a matrix where $j = 1, ..., \theta$. For example:

$$\boldsymbol{\phi}_{(\bullet)}\mathbf{Y}_t = \begin{pmatrix} \phi_1 \mathbf{Y}_{t[1:p][]} \\ \phi_2 \mathbf{Y}_{t[p+1:2p][]} \\ \vdots \\ \phi_\theta \mathbf{Y}_{t[(\theta-1)p+1:\theta p][]} \end{pmatrix} = \begin{pmatrix} \phi_1 y_{11,t} \\ \vdots \\ \phi_1 y_{p1,t} \\ \vdots \\ \phi_\theta y_{1\theta,t} \\ \vdots \\ \phi_\theta y_{p\theta,t} \end{pmatrix}.$$

Similarly:

$$\boldsymbol{\phi}_{(\bullet)}\mathbf{U}_t = \begin{pmatrix} \phi_1 u_{1,t} \\ \phi_2 u_{2,t} \\ \vdots \\ \phi_\theta u_{\theta,t} \end{pmatrix}. \tag{11}$$

Furthermore, letting $\mathbf{i}_p$ be a vector of ones with $p$ denoting the row dimension, we define $\boldsymbol{\phi}_{(\bullet)}(L)$ as:

$$\boldsymbol{\phi}_{(\bullet)}(L) = \mathbf{i}_p - \boldsymbol{\phi}_{(\bullet)}L,$$

where $L$ is the lag operator. Applying $\boldsymbol{\phi}_{(\bullet)}(L)$ to $\mathbf{Y}_t$ we have:

$$\boldsymbol{\phi}_{(\bullet)}(L)\mathbf{Y}_t = \left(\mathbf{i}_{\theta p} - \boldsymbol{\phi}_{(\bullet)}L\right)\mathbf{Y}_t = \mathbf{Y}_t - \boldsymbol{\phi}_{(\bullet)}L\mathbf{Y}_t = \mathbf{Y}_t - \boldsymbol{\phi}_{(\bullet)}\mathbf{Y}_{t-1}.$$

Similarly:

$$\boldsymbol{\phi}_{(\bullet)}(L)\mathbf{U}_t = \mathbf{U}_t - \boldsymbol{\phi}_{(\bullet)}L\mathbf{U}_t = \mathbf{U}_t - \boldsymbol{\phi}_{(\bullet)}\mathbf{U}_{t-1}.$$

In addition we define:

$$\mathbf{T}^* = \mathbf{T} - \text{diag}\{\mathbf{T}\}, \tag{12}$$

where $\mathbf{T}^*$ is a square matrix with the off-diagonal elements of $\mathbf{T}$ and zeros along the diagonal. Using $\boldsymbol{\phi}_{(\bullet)}$, $\boldsymbol{\phi}_{(\bullet)}(L)$ and $\mathbf{T}^*$, we rewrite the state equation in (2) as:

$$\begin{aligned} \mathbf{U}_t &= \boldsymbol{\phi}_{(\bullet)}\mathbf{U}_{t-1} + \mathbf{T}^*\mathbf{U}_{t-1} + \boldsymbol{\eta}_t \\ \mathbf{U}_t - \boldsymbol{\phi}_{(\bullet)}\mathbf{U}_{t-1} &= \mathbf{T}^*\mathbf{U}_{t-1} + \boldsymbol{\eta}_t \\ \boldsymbol{\phi}_{(\bullet)}(L)\mathbf{U}_t &= \mathbf{T}^*\mathbf{U}_{t-1} + \boldsymbol{\eta}_t. \end{aligned} \tag{13}$$

Furthermore, applying the $\boldsymbol{\phi}_{(\bullet)}(L)$ operator to the observation equation in (1) yields:

$$\boldsymbol{\phi}_{(\bullet)}(L)\mathbf{Y}_t = \mathbf{B}\boldsymbol{\phi}_{(\bullet)}(L)\mathbf{U}_t + \boldsymbol{\phi}_{(\bullet)}(L)\mathbf{e}_t. \tag{14}$$

Replacing (13) into (14) we have:

$$\begin{aligned} \boldsymbol{\phi}_{(\bullet)}(L)\mathbf{Y}_t &= \mathbf{B}\mathbf{T}^*\mathbf{U}_{t-1} + \mathbf{B}\boldsymbol{\eta}_t + \boldsymbol{\phi}_{(\bullet)}(L)\mathbf{e}_t \\ \mathbf{Y}_t &= \boldsymbol{\phi}_{(\bullet)}\mathbf{Y}_{t-1} + \mathbf{B}\mathbf{T}^*\mathbf{U}_{t-1} + \mathbf{B}\boldsymbol{\eta}_t + \mathbf{e}_t - \boldsymbol{\phi}_{(\bullet)}\mathbf{e}_{t-1}. \end{aligned} \tag{15}$$

Our next step is to calculate the covariances of the unobserved state vector $\mathbf{U}_t$. Assuming stationarity of the state vector such that:

$$\text{Var}(\mathbf{U}_t) = \text{Var}(\mathbf{U}_{t-1}) = \boldsymbol{\Gamma}_{\mathbf{U}}(0), \tag{16}$$

the variance of $\mathbf{U}_t$, becomes:

$$\boldsymbol{\Gamma}_{\mathbf{U}}(0) = \mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{T}' + \mathbf{Q}. \tag{17}$$

Furthermore, $\boldsymbol{\Gamma}_{\mathbf{U}}(1)$ is determined as follows:

$$\begin{aligned} \boldsymbol{\Gamma}_{\mathbf{U}}(1) &= \text{E}\left[\left(\mathbf{T}\mathbf{U}_{t-1} + \boldsymbol{\eta}_t\right)\mathbf{U}_{t-1}\right] \\ &= \mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0), \end{aligned} \tag{18}$$

with the general form of the autocovariance function of $\mathbf{U}_t$ being:

$$\boldsymbol{\Gamma}_{\mathbf{U}}(h) = \mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(h-1), \quad \text{for } h \geq 1. \tag{19}$$

A closed form solution for (17) can be obtained with the use of the vec operator as shown by Hamilton (1994; p. 265):

$$\begin{aligned} \text{vec}\left[\boldsymbol{\Gamma}_{\mathbf{U}}(0)\right] &= \text{vec}\left[\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{T}' + \mathbf{Q}\right] \\ &= (\mathbf{T}\otimes\mathbf{T})\text{vec}\left[\boldsymbol{\Gamma}(0)\right] + \text{vec}(\mathbf{Q}) \\ &= \left[\mathbf{I} - \mathbf{T}\otimes\mathbf{T}\right]^{-1}\text{vec}(\mathbf{Q}). \end{aligned} \tag{20}$$

The autocovariance function of (15) has the following form:

$$\boldsymbol{\Gamma}_{\mathbf{Y}}(0) = \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(1) + \mathbf{B}\mathbf{T}^{*}\text{Cov}(\mathbf{U}_{t-1}, \mathbf{Y}_t') + \mathbf{B}\mathbf{Q}\mathbf{B}' + \mathbf{D} + \boldsymbol{\phi}^2_{(\bullet)}\mathbf{D}, \tag{21}$$

$$\boldsymbol{\Gamma}_{\mathbf{Y}}(1) = \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(0) + \mathbf{B}\mathbf{T}^{*}\text{Cov}(\mathbf{U}_{t-1}, \mathbf{Y}_{t-1}')\mathbf{B}' + \boldsymbol{\phi}^2_{(\bullet)}\mathbf{D}, \tag{22}$$

$$\boldsymbol{\Gamma}_{\mathbf{Y}}(2) = \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(1) + \mathbf{B}\mathbf{T}^{*}\text{Cov}(\mathbf{U}_{t-1}, \mathbf{Y}_{t-2}'), \tag{23}$$

$$\vdots$$

$$\boldsymbol{\Gamma}_{\mathbf{Y}}(h) = \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(h-1) + \mathbf{B}\mathbf{T}^{*}\text{Cov}(\mathbf{U}_{t-1}, \mathbf{Y}_{t-h}'). \tag{24}$$

One of the components of the autocovariance function of (15) is the cross-covariance between $\mathbf{U}_{t-1}$ and $\mathbf{Y}_{t-h}$, which is given by:

$$\begin{aligned}
\text{Cov}(\mathbf{U}_{t-1}\mathbf{Y}_t') &= \text{Cov}\left[\mathbf{U}_{t-1},\, (\mathbf{B}\mathbf{U}_t + \mathbf{e}_t)'\right] = \boldsymbol{\Gamma}_{\mathbf{U}}(1)\mathbf{B}', && (25)\\
\text{Cov}(\mathbf{U}_{t-1}\mathbf{Y}_{t-1}') &= \text{Cov}\left[\mathbf{U}_{t-1},\, (\mathbf{B}\mathbf{U}_{t-1} + \mathbf{e}_{t-1})'\right] = \boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}', && (26)\\
&\vdots \qquad \vdots &&\\
\text{Cov}(\mathbf{U}_{t-1}\mathbf{Y}_{t-h}') &= \text{Cov}\left[\mathbf{U}_{t-1},\, (\mathbf{B}\mathbf{U}_{t-h} + \mathbf{e}_{t-h})'\right] = \boldsymbol{\Gamma}_{\mathbf{U}}(h-1)\mathbf{B}', && (27)
\end{aligned}$$

for $h \geq 1$. Using the result in (19) and replacing (25)-(27) in (21)-(24) we have:

$$\begin{aligned}
\boldsymbol{\Gamma}_{\mathbf{Y}}(0) &= \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(1) + \mathbf{B}\mathbf{T}^*\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{B}\mathbf{Q}\mathbf{B}' + \mathbf{D} + \boldsymbol{\phi}^2_{(\bullet)}\mathbf{D}, && (28)\\
\boldsymbol{\Gamma}_{\mathbf{Y}}(1) &= \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(0) + \mathbf{B}\mathbf{T}^*\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' - \boldsymbol{\phi}_{(\bullet)}\mathbf{D}, && (29)\\
\boldsymbol{\Gamma}_{\mathbf{Y}}(2) &= \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(1) + \mathbf{B}\mathbf{T}^*\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}', && (30)\\
&\vdots &&\\
\boldsymbol{\Gamma}_{\mathbf{Y}}(h) &= \boldsymbol{\phi}_{(\bullet)}\boldsymbol{\Gamma}_{\mathbf{Y}}(h-1) + \mathbf{B}\mathbf{T}^*\boldsymbol{\Gamma}_{\mathbf{U}}(h-1)\mathbf{B}', && (31)
\end{aligned}$$

where $h \geq 2$. Our final task is to calculate the closed form solution for (28)-(31). We derive the following results:

$$\begin{aligned}
\text{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] =& \{[\mathbf{i}_{\theta^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\,\text{vec}(\mathbf{Q})\left[\boldsymbol{\phi}_{(\bullet)}\left(\mathbf{B}\otimes\mathbf{B}\mathbf{T}^*\right) + \left(\mathbf{B}\otimes\mathbf{B}\mathbf{T}^*\mathbf{T}\right)\right]\\
&+ (\mathbf{B}\otimes\mathbf{B})\,\text{vec}\left(\mathbf{Q}\right) + \text{vec}(\mathbf{D})\}\left(\mathbf{i}_{\theta p} - \boldsymbol{\phi}^2_{(\bullet)}\mathbf{i}_{\theta p}\right)^{-1}, && (32)\\
\text{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(1)\right] =& \{[\mathbf{i}_{\theta^2} - \mathbf{T}\otimes\mathbf{T}]^{-1}\,\text{vec}(\mathbf{Q})\left[\boldsymbol{\phi}_{(\bullet)}\left(\mathbf{B}\otimes\mathbf{B}\mathbf{T}^*\mathbf{T}\right) + \left(\mathbf{B}\otimes\mathbf{B}\mathbf{T}^*\right)\right]\\
&+ \boldsymbol{\phi}_{(\bullet)}\text{vec}\left(\mathbf{B}\mathbf{Q}\mathbf{B}'\right) + \boldsymbol{\phi}^3_{(\bullet)}\text{vec}\left(\mathbf{D}\right)\}\left(\mathbf{i}_{\theta p} - \boldsymbol{\phi}^2_{(\bullet)}\mathbf{i}_{\theta p}\right)^{-1}. && (33)
\end{aligned}$$

A detailed calculation of (32) and (33) is provided in section 6.1 of the Appendix. Equations (32) and (33) can be used to derive the restrictions necessary for the identifiability of a specific formulation of our generalized dynamic factor model. To illustrate this, we consider the case of one individual with multiple indicators in the following section.

### 2.4.1 The case of one individual with multiple indicators

In this formulation of our model, equations (32) and (33) have the following form:

$$\mathbf{\Gamma_Y}(0) = \left(\sigma_\eta^2 \mathbf{B}\mathbf{B}' + \mathbf{D}\right)\left(\mathbf{i}_p - \phi^2 \mathbf{i}_p\right)^{-1}, \tag{34}$$

$$\mathbf{\Gamma_Y}(1) = \left(\phi\sigma_\eta^2 \mathbf{B}\mathbf{B}' + \phi^3 \mathbf{D}\right)\left(\mathbf{i}_p - \phi^2 \mathbf{i}_p\right)^{-1}. \tag{35}$$

Note that all elements of (32) and (33) that involve $\mathbf{T}^*$ are absent from (34) and (35). The off-diagonal elements of (34) and (35) are represented as follows:

$$\mathbf{\Gamma_Y}_{[ii*]}(0) = \sigma_\eta^2 b_i b_{i*}\left(1-\phi^2\right)^{-1}, \tag{36}$$

$$\mathbf{\Gamma_Y}_{[ii*]}(1) = \phi\sigma_\eta^2 b_i b_{i*}\left(1-\phi^2\right)^{-1}. \tag{37}$$

Due to symmetry of $\mathbf{\Gamma_Y}(0)$ and $\mathbf{\Gamma_Y}(1)$ each of those two matrices has $\frac{p(p+1)}{2}$ elements. Therefore (34) and (35) form a system of $p^2+p$ equations with $2p+2$ unknown parameters ($p$ parameters in each of the two matrices $\mathbf{B}$ and $\mathbf{D}$ as well as parameters $\phi$ and $\sigma_\eta^2$). In order for this system to be fully identifiable we select the following restriction that makes the system linear in its parameters:

$$\sigma_\eta^2 = 1-\phi^2. \tag{38}$$

From (17), and assuming that $\mathrm{Var}(u_t) = \mathrm{Var}(u_{t-1})$, we determine that the variance of $u_t$ is:

$$\mathrm{Var}(u_t) = \phi^2 \mathrm{Var}(u_{t-1}) + \sigma_\eta^2 \Rightarrow \mathrm{Var}(u_t) = \frac{\sigma_\eta^2}{1-\phi^2}, \tag{39}$$

and replacing (38) into (39) we have:

$$\mathrm{Var}(u_t) = 1. \tag{40}$$

Moving to a more general case where multiple individuals are considered, while assuming no correlation among individuals, such that $\mathbf{T}^* = \mathbf{0}$, restriction (38) affects the moments of the joint multivariate normal vector $(\mathbf{Y}_t^T, \mathbf{U}_t^T)^T$ which are now:

$$\begin{pmatrix} \mathbf{B}\mathbf{B}^T + \mathbf{D} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{I} \end{pmatrix}. \tag{41}$$

McLachlan and Peel (2000, p. 243) use the same covariance matrix when discussing the case of a factor analytic model without a time dimension. Note that (41) is a simplified version of the covariance structure illustrated in (10).

When solving the system defined by $\mathbf{\Gamma_Y}(0)$ and $\mathbf{\Gamma_Y}(1)$ the elements of $\mathbf{B}$ can have both a positive and a negative root. We disregard the negative root of $\mathbf{B}$ by a simple transformation which makes all indicators positively correlated to each other.

## 2.5 Parameter formulation

In this section, we illustrate all possible formulations of the parameters of the model, namely $\mathbf{B}$, $\mathbf{D}$, $\mathbf{T}$ and $\mathbf{Q}$. The general form of the matrix of factor loadings $\mathbf{B}$ can be written as follows:

$$\underset{\theta p \times \theta}{\mathbf{B}} = \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} & \mathbf{b}_{13} & ... & \mathbf{b}_{1\theta} \\ \mathbf{b}_{21} & \mathbf{b}_{22} & \mathbf{b}_{23} & ... & \mathbf{b}_{2\theta} \\ \mathbf{b}_{31} & \mathbf{b}_{32} & \mathbf{b}_{33} & ... & \mathbf{b}_{3\theta} \\ ... & ... & ... & ... & .. \\ \mathbf{b}_{\theta 1} & \mathbf{b}_{\theta 2} & \mathbf{b}_{\theta 3} & ... & \mathbf{b}_{\theta\theta} \end{bmatrix},$$

where each $\mathbf{b}_{jj}$ $(j = 1, ..., \theta)$ along the diagonal of $\mathbf{B}$ is a $p \times 1$ vector of the factor loadings for the $j^{th}$ individual and each $\mathbf{b}_{jj^*}$ $(j,\ j^* = 1, ..., \theta$ and $j \neq j^*)$ on the off-diagonal of $\mathbf{B}$ is also a $p \times 1$ vector representing the loadings of the indicators of individual $j$ to the factor of individual $j^*$. For example, $\mathbf{b}_{11}$ contains the factor loadings of the first individual, while $\mathbf{b}_{12}$ loads the indicators of the first individual to the factor of the second individual. We distinguish four combinations for the diagonal and off-diagonal vectors of $\mathbf{B}$, illustrated in table 1:

| | | Off-diagonal vectors | | | |
|---|---|---|---|---|---|
| | | $\mathbf{b}_{jj^*} = 0$ | | $\mathbf{b}_{jj^*} \neq 0$ | |
| | | Notation | Parameters | Notation | Parameters |
| Diagonal | $\mathbf{b}_{jj} = c$ | $\mathbf{B}_1$ | $\theta$ | $\mathbf{B}_2$ | $\theta p \times [\theta - 1] + 1$ |
| vectors | $\mathbf{b}_{jj} \neq c$ | $\mathbf{B}_3$ | $\theta p$ | $\mathbf{B}_4$ | $\theta p \times \theta$ |

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

Formulations $\mathbf{B}_1$ and $\mathbf{B}_2$ represent the case where the factor loadings are the same for every individual. The difference between $\mathbf{B}_1$ and $\mathbf{B}_2$ lies in the assumptions regarding

the off-diagonal vectors. In $\mathbf{B}_1$ the indicators of individual $j$ do not load on the factor of individual $j^*$ since all the off-diagonal vectors are equal to zero. In case $\mathbf{B}_2$, $\mathbf{b}_{jj^*}$ is unconstrained. $\mathbf{B}_3$ also requires zero off-diagonal elements for $\mathbf{B}$, only this time, vectors $\mathbf{b}_{jj}$ are allowed to vary, hence every individual has a unique set of factor loadings. Finally, the most complex case is $\mathbf{B}_4$ where there are no constraints on the elements of $\mathbf{B}$. Following the classification in Rubin and Thayer (1982) cases $\mathbf{B}_1$ and $\mathbf{B}_3$ fall under the category of confirmatory factor analysis, where the researcher has a priori assumptions regarding the factor loadings, while $\mathbf{B}_4$ is considered an example of exploratory factor analysis, where no prior specification regarding the factor loadings is made (Kim and Mueller, 1978). Case $\mathbf{B}_2$ could be considered a hybrid.

Next, we consider the variance of the idiosyncratic errors in $\mathbf{D}$. The general form of $\mathbf{D}$ is:

$$\underset{\theta p \times \theta p}{\mathbf{D}} = \text{diag}(\mathbf{d}_j),$$

where $\mathbf{d}_j$ ($j = 1, ..., \theta$) is a $p \times p$ diagonal matrix representing the variance of the error term for every individual. As discussed in section 2.3, that diagonality of $\mathbf{D}$ is required due to the factor analytic nature of (1). The matrix form of each $\mathbf{d}_j$ is:

$$\underset{p \times p}{\mathbf{d}_j} = \text{diag}(\sigma_{ij}^2),$$

where $\sigma_{ij}^2$ is the variance of the error term of a specific individual. The following two cases, illustrated in table 2, are applicable for $\mathbf{D}$:

| | | Notation | Parameters |
|---|---|---|---|
| Diagonal | $\mathbf{d}_j = \mathbf{c}$ | $\mathbf{D}_1$ | $p$ |
| elements | $\mathbf{d}_j \neq \mathbf{c}$ | $\mathbf{D}_2$ | $\theta p$ |

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

Formulation $\mathbf{D}_1$ suggests that the variance of the error term is the same for every individual. On the other hand, $\mathbf{D}_2$ allows $\sigma_{ij}^2$ to vary for each individual.

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

$$\underset{\theta\times\theta}{\mathbf{T}} = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} & \cdots & \phi_{1\theta} \\ \phi_{21} & \phi_{22} & \phi_{23} & \cdots & \phi_{2\theta} \\ \phi_{31} & \phi_{32} & \phi_{33} & \cdots & \phi_{3\theta} \\ \cdots & \cdots & \cdots & \cdots & .. \\ \phi_{\theta 1} & \phi_{\theta 2} & \phi_{\theta 3} & \cdots & \phi_{\theta\theta} \end{bmatrix},$$

where $\phi_{jj}$ is the autoregressive parameter that determines the effect through time of an individual's own state variable. The off-diagonal elements $\phi_{jj^*}$ (where $j,\ j^* = 1,...,\theta$ and $j \neq j^*$), capture the correlation of the state variable between individuals across time. Note that we may have $\phi_{jj^*} \neq \phi_{j^*j}$. That is, the effect of $u_j$ to previous movements of $u_{j^*}$, could be different from the effect of $u_{j^*}$ to previous movements of $u_j$. We distinguish between six cases for $\mathbf{T}$, illustrated in table 3:

| | | Off-diagonal elements | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\phi_{jj^*} = 0$ | | $\phi_{jj^*} \neq 0$ | | Spatial correlation | |
| | | Notation | Parameters | Notation | Parameters | Notation | Parameters |
| Diagonal | $\phi_{jj} = c$ | $\mathbf{T}_1$ | 1 | $\mathbf{T}_2$ | $\theta[\theta-1]+1$ | $\mathbf{T}_5$ | $\theta[\theta-1]+1$ |
| elements | $\phi_{jj} \neq c$ | $\mathbf{T}_3$ | $\theta$ | $\mathbf{T}_4$ | $\theta^2$ | $\mathbf{T}_6$ | $\theta^2$ |

Table 3: Possible formulations of $\mathbf{T}$.

Formulation $\mathbf{T}_1$ suggests that there is no correlation between the values of the state variable of different individuals across time. Additionally, all individuals share the same autoregressive parameter. $\mathbf{T}_3$ retains the same assumption regarding the off-diagonal elements of $\mathbf{T}$. However, in this case the autoregressive parameter is allowed to vary by individual. Formulations $\mathbf{T}_2$ and $\mathbf{T}_4$ have unconstrained $\phi_{jj^*}$, hence accounting for cross-temporal correlations between the state variables of different individuals. This correlation is further expanded in formulations $\mathbf{T}_5$ and $\mathbf{T}_6$ where spatial effects are considered. Under those two cases $\mathbf{T}$ will have the following form:

$$\underset{\theta\times\theta}{\mathbf{T}} = \begin{bmatrix} \phi_{11} & \phi_{12}(s_1,s_2) & \phi_{13}(s_1,s_3) & ... & \phi_{1\theta}(s_1,s_\theta) \\ \phi_{21}(s_2,s_1) & \phi_{22} & \phi_{23}(s_2,s_3) & ... & \phi_{2\theta}(s_2,s_\theta) \\ \phi_{31}(s_3,s_1) & \phi_{32}(s_3,s_2) & \phi_{33} & ... & \phi_{3\theta}(s_3,s_\theta) \\ ... & ... & ... & ... & .. \\ \phi_{\theta 1}(s_\theta,s_1) & \phi_{\theta 2}(s_\theta,s_2) & \phi_{\theta 3}(s_\theta,s_3) & ... & \phi_{\theta\theta} \end{bmatrix},$$

where $\phi_{jj^*}(s_j,s_{j^*})$ (with $j$, $j^* = 1,...,\theta$ and $j \neq j^*$) denotes the cross-temporal correlation of the state variable between individuals which is a function of locations $s_j$ and $s_{j^*}$ of individuals $j$ and $j^*$. Alternatively, we could think of parameter $\phi_{jj^*}(s_j,s_{j^*})$ as a spatial component determined by the distance between $s_j$ and $s_{j^*}$. The closer two individuals are, the higher the value of their spatial correlation is.

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:

$$\underset{\theta\times\theta}{\mathbf{Q}} = \begin{bmatrix} \sigma_1^2 & \mathrm{E}(\eta_1\eta_2) & \mathrm{E}(\eta_1\eta_3) & ... & \mathrm{E}(\eta_1\eta_\theta) \\ \mathrm{E}(\eta_2\eta_1) & \sigma_2^2 & \mathrm{E}(\eta_2\eta_3) & ... & \mathrm{E}(\eta_2\eta_\theta) \\ \mathrm{E}(\eta_3\eta_1) & \mathrm{E}(\eta_3\eta_2) & \sigma_3^2 & ... & \mathrm{E}(\eta_3\eta_\theta) \\ ... & ... & ... & ... & .. \\ \mathrm{E}(\eta_\theta\eta_1) & \mathrm{E}(\eta_\theta\eta_2) & \mathrm{E}(\eta_\theta\eta_3) & ... & \sigma_\theta^2 \end{bmatrix},$$

where the diagonal elements $\sigma_j^2$ are the variances of error term of the state equation. The off-diagonal elements $\mathrm{E}(\eta_j\eta_{j^*})$ (where $j$, $j^* = 1,...,\theta$ and $j \neq j^*$) represent covariances, with $\mathrm{E}(\eta_j\eta_{j^*}) = \mathrm{E}(\eta_{j^*}\eta_j)$ by symmetry of $\mathbf{Q}$. There are several alternatives for the formulation of $\mathbf{Q}$ presented in Table 4:

| | | Off-diagonal elements | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\mathrm{E}(\eta_j\eta_{j^*}) = 0$ | | $\mathrm{E}(\eta_j\eta_{j^*}) \neq 0$ | | Spatial correlation | |
| | | Notation | Parameters | Notation | Parameters | Notation | Parameters |
| Diagonal | $\sigma_j^2 = c$ | $\mathbf{Q}_1$ | 1 | $\mathbf{Q}_2$ | $\frac{\theta(\theta+1)}{2} - (\theta-1)$ | $\mathbf{Q}_5$ | $\frac{\theta(\theta+1)}{2} - (\theta-1)$ |
| elements | $\sigma_j^2 \neq c$ | $\mathbf{Q}_3$ | $\theta$ | $\mathbf{Q}_4$ | $\frac{\theta(\theta+1)}{2}$ | $\mathbf{Q}_6$ | $\frac{\theta(\theta+1)}{2}$ |

Table 4: Possible formulations of $\mathbf{Q}$.

The difference between non-zero $\mathrm{E}(\eta_j\eta_{j^*})$ and $\phi_{jj^*}$ is that the former captures the effect of a contemporaneous shock, as opposed to the latter which pertains to the effect of a

relationship that persists through time. Consider the case of a group of water utilities affected by a severe drought at one particular time period. The effect of that drought would enter the model through the off-diagonal elements of $\mathbf{Q}$. However, if two neighboring utilities are competing over the same water resources, a conflict that is likely to persist through time, then this spatial correlation would affect their performance every year and hence would enter the model through the off-diagonal elements of $\mathbf{T}$ in formulations $\mathbf{T}_5$ or $\mathbf{T}_6$.

# 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. Section 3.1 explains the need for the development of the algorithm. Each of the two cycles is analyzed in detail in sections 3.2 and 3.3.

## 3.1 Challenges and a new approach

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 (1) and (2) is:

$$L(\mathbf{B}, \mathbf{D}, \mathbf{T}, \mathbf{Q}; \mathbf{Y}_1, ..., \mathbf{Y}_n) = \prod_{t=2}^{n} f(\mathbf{Y}_1) f_{\mathbf{Y}}(\mathbf{Y}_t; [\mathbf{B}, \mathbf{D}, \mathbf{T}, \mathbf{Q}] \,|\, \tilde{\mathbf{Y}}_{t-1}), \tag{42}$$

where $\tilde{\mathbf{Y}}_{t-1}$ 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 introduce the 2CCEM algorithm that makes estimation of the model specified in (1) and (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. 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). Shumway and Stoffer (1982) were the first to use the EM

algorithm to estimate state space models, similar to the one specified in (1) and (2).

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

$$\log L_c(\mathbf{\Psi}) = \log f_c(\mathbf{Y}_t, \mathbf{U}_t; \mathbf{\Psi}), \tag{43}$$

where the subscript $c$ denotes the complete-data likelihood. In a conventional Maximum Likelihood Estimation (MLE) setting, maximization of (43) would occur by differentiating the function with respect to its parameters and setting the derivative equal to zero:

$$\frac{\partial \log f_c(\mathbf{Y}_t, \mathbf{U}_t; \mathbf{\Psi})}{\partial \mathbf{\Psi}} = 0.$$

However, since we only observed $\mathbf{Y}_t$ and as a result the observed data are incomplete such a maximization cannot be performed. At this point, a traditional EM algorithm proceeds by replacing the complete-data log likelihood with the conditional expectation of the incomplete-data given $\mathbf{Y}$.

## 3.2 First cycle of the 2CCEM

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 $\mathbf{\Psi}_2$. During the $k^{th}$ iteration of the first cycle, the E-step of the 2CCEM algorithm requires the following calculation:

$$\mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) = \mathrm{E}_{\mathbf{\Psi}_1}\left\{ \log \mathrm{L}_c(\mathbf{\Psi}_1) \,|\mathbf{Y}, \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)} \right\}. \tag{44}$$

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

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\boldsymbol{\Psi}_1^{(k/2)};\boldsymbol{\Psi}_1^{(k-1)},\boldsymbol{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_1}(\boldsymbol{\Psi}_1;\boldsymbol{\Psi}_1^{(k-1)},\boldsymbol{\Psi}_2^{(k-1)}), \tag{45}$$

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

$$\mathbf{Z}_{\boldsymbol{\Psi}_1}(\boldsymbol{\Psi}_1^{(k)};\boldsymbol{\Psi}_1^{(k-1)},\boldsymbol{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_1}(\boldsymbol{\Psi}_1^{(k/2)};\boldsymbol{\Psi}_1^{(k-1)},\boldsymbol{\Psi}_2^{(k-1)}) \tag{46}$$

### 3.2.1 Estimation of 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 $\mathbf{Y}$, $\boldsymbol{\Sigma}$, is denoted by $\mathbf{C}_{yy}$, where:

$$\mathbf{C}_{yy} = \mathbf{Y}\mathbf{Y}'. \tag{47}$$

$\mathbf{C}_{yy}$ is the main building block in the E-step of the first cycle of the 2CCEM algorithm described in (44). Equation (44), also appearing in a traditional EM algorithm, 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:

$$\begin{aligned}
&\mathrm{E}(\mathbf{Y}\mathbf{Y}^T|\mathbf{Y}) = \mathbf{C}_{yy},\\
&\mathrm{E}(\mathbf{Y}^T\mathbf{U}|\mathbf{Y}) = \mathbf{C}_{yy}\boldsymbol{\gamma},\\
&\mathrm{E}(\mathbf{U}^T\mathbf{U}|\mathbf{Y}) = \boldsymbol{\gamma}^T\mathbf{C}_{yy}\boldsymbol{\gamma} + n\boldsymbol{\omega},
\end{aligned} \tag{48}$$

where:

$$\boldsymbol{\gamma} = \left(\mathbf{B}\mathbf{B}^T + \mathbf{D}\right)^{-1}\mathbf{B}, \tag{49}$$

and:

$$\boldsymbol{\omega} = \mathbf{I} - \boldsymbol{\gamma}^T\mathbf{B}. \tag{50}$$

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

$$\mathbf{U}_t|\mathbf{Y}_t \sim N(\boldsymbol{\gamma}^T\mathbf{Y}_t, \mathbf{I} - \boldsymbol{\gamma}^T\mathbf{B}). \tag{51}$$

Note that (48)-(50) constitute the E-step of the first cycle of the 2CCEM algorithm illustrated in (44). The subsequent first M-step, illustrated in (45), is identical to the M-step of the traditional EM algorithm which involves plugging the sufficient statistics (48) into $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ and differentiating with respect to $\boldsymbol{\Psi}_1$. The functional form of $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ is:

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}_1) = &\frac{n}{2}\log\left\{|\mathbf{D}^{-1}| + \log|\mathbf{Q}^{-1}|\right\} \\ &-\frac{1}{2}\sum_{t=1}^{n}\left\{(\mathbf{y}_t - \mathbf{B}\hat{\mathbf{u}}_t)^T\mathbf{D}^{-1}(\mathbf{y}_t - \mathbf{B}\hat{\mathbf{u}}_t) - (\hat{\mathbf{u}}_{t+1} - \mathbf{T}\hat{\mathbf{u}}_t)^T\mathbf{Q}^{-1}(\hat{\mathbf{u}}_{t+1} - \mathbf{T}\hat{\mathbf{u}}_t)\right\}. \end{aligned} \tag{52}$$

Equating the first derivatives of $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ to zero yields:

$$\mathbf{B}^{(k/2)} = \mathbf{C}_{yy}\boldsymbol{\gamma}\left\{\boldsymbol{\gamma}^T\mathbf{C}_{yy}\boldsymbol{\gamma} + n\boldsymbol{\omega}\right\}^{-1}, \tag{53}$$

and

$$\mathbf{D}^{(k/2)} = n^{-1}\mathrm{diag}\left\{\mathbf{C}_{yy} - \mathbf{C}_{yy}\boldsymbol{\gamma}\mathbf{B}^T\right\}, \tag{54}$$

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 (52) is maximized, through a Newton-Raphson algorithm, with respect to $\boldsymbol{\Psi}_1$, using (53) and (54) 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.

## 3.3 Second cycle of the 2CCEM

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

$$\mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}) = \mathrm{E}_{\boldsymbol{\Psi}_2}\left\{\mathrm{logL}_c(\boldsymbol{\Psi}_2)\,|\mathbf{Y}, \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}\right\}. \tag{55}$$

In other words, 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:

$$\mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2^{(k)}; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\boldsymbol{\Psi}_2}(\boldsymbol{\Psi}_2; \boldsymbol{\Psi}_1^{(k)}, \boldsymbol{\Psi}_2^{(k-1)}). \tag{56}$$

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

### 3.3.1 Estimation of the second cycle

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

$$\log \mathbf{L}_c(\boldsymbol{\Psi}_2) = n - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\left[\log|\mathbf{F}_t| + \mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t\right], \tag{57}$$

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. The log likelihood in (57) is also known as the prediction error decomposition (Harvey, 1991, p. 126). 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 our knowledge of the system to be updated every time an additional observation $\mathbf{Y}_t$ is added to the model (Kalman, 1960; Durbin and Koopman, 2001, p. 11). Let $\tilde{\mathbf{Y}}_{t-1}$ be the set of past observations $\mathbf{Y}_1, ..., \mathbf{Y}_{t-1}$ and assume that $\mathbf{U}_t|\tilde{\mathbf{Y}}_{t-1} \sim N(\hat{\mathbf{U}}_t, \mathbf{P}_t)$, 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. The set of filtering equations that are required for the calculation of $\hat{\mathbf{U}}_{t+1}$ and $\mathbf{P}_{t+1}$ is illustrated below:

$$\begin{array}{rclcrcl}
\mathbf{v}_t & = & \mathbf{y}_t - \mathbf{B}\hat{\mathbf{U}}_\mathbf{t}, & & \mathbf{F}_t & = & \mathbf{B}\mathbf{P}_t\mathbf{B}' + \mathbf{D}, \\
\mathbf{K}_t & = & \mathbf{T}\mathbf{P}_t\mathbf{B}'\mathbf{F}_t^{-1}, & & \mathbf{L}_t & = & \mathbf{T} - \mathbf{K}_t\mathbf{B}, \\
\hat{\mathbf{U}}_{t+1} & = & \mathbf{T}\hat{\mathbf{U}}_t + \mathbf{K}_t\mathbf{v}_t, & & \mathbf{P}_{t+1} & = & \mathbf{T}\mathbf{P}_t\mathbf{L}_t' + \mathbf{Q},
\end{array} \tag{58}$$

where $\hat{\mathbf{U}}_t$ is the filtered estimate of the unobserved state vector conditional on $\tilde{\mathbf{Y}}_{t-1}$. Note that step (55) involves plugging the ML estimates $\mathbf{B}^{(k)}$ and $\mathbf{D}^{(k)}$, obtained in the first cycle of the 2CCEM algorithm, into the filtering equations (58). Once $\mathbf{v}_t$ and $\mathbf{F}_t$ are calculated, (57) is maximized with respect to $\boldsymbol{\Psi}_2$ , as illustrated in (56).

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, $\mathbf{Y}_1, ..., \mathbf{Y}_n$. The set of smoothing equations for the state vector is:

$$\begin{aligned} \mathbf{r}_{t-1} &= \mathbf{B}'\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{L}'\mathbf{r}_t, \qquad \tilde{\mathbf{U}}_t &= \hat{\mathbf{U}}_t + \mathbf{P}_t\mathbf{r}_{t-1}, \\ \mathbf{N}_{t-1} &= \mathbf{B}'\mathbf{F}_t'\mathbf{B} + \mathbf{L}_t\mathbf{N}_t\mathbf{L}_t, \qquad \mathbf{V}_t &= \mathbf{P}_t - \mathbf{P}_t\mathbf{N}_{t-1}\mathbf{P}_t, \end{aligned} \tag{59}$$

where $\tilde{\mathbf{U}}_t$ represents the smoothed estimate of the unobserved index and $\mathbf{V}_t$ is the variance of the smoothed estimate (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. Section 4.1 provides background information regarding the data and explains the role of water utilities in developing countries. In section 4.2 we present previous relevant work with regards to water utility index formation and discuss our contributions. Section 4.3 outlines the process of initialization of the 2CCEM algorithm and presents estimation results. Finally, we conduct post-estimation tests and simulations that ensure the uniqueness of the estimated smoothed index.

## 4.1 Data

Our 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. It is a publicly available on-line database where utility executives can upload information on various indicators regarding the utility they manage. The information is available to a variety of stakeholders and policy makers. The IBNET database includes 105 indicators that can be grouped in financial, operational and quality of service indicators. The data pertain to utilities from more than 100 countries with the time dimension spanning from 1994 to 2012 (IBNET, 2005). 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).

Water utilities in low and middle income countries are organizations that deliver drinking water and sanitation services to the public and are either government owned or managed by the state. This provides a financial safety net that shields utilities from competitive pres-

sures. For example, a utility would never be left to go bankrupt given the importance of the services it delivers. In fact, it is very common for governments to provide "bailout packages" in order to improve the financial stability of utilities. Moreover, utilities do not operate as for-profit companies. Their goal is to provide affordable services while maintaining and operating a large infrastructure network. Those characteristics make water utilities unique in many ways, one of which is the difficulty in assessing their financial and operational health. Unlike companies that operate in competitive markets, there exists no standardized index that measures the performance of water utilities. Nevertheless, there are individual indicators, such as cost recovery, population coverage, quality of water sold or tariff structure, that could be considered when assessing a utility's effectiveness and financial viability. However, the variety of important indicators that are relevant in this effort makes the development of a comprehensive performance index rather challenging (van den Berg and Danilenko, 2010).

## 4.2 The APGAR score vs. a dynamic smoothed index

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 van den Berg and Danilenko (2010). The authors develop a static index whose aim is to assess the health of a utility based on a weighted sum of six indicators. Van den Berg and Danilenko call that index the "APGAR score" after Virginia Apgar who in 1953 introduced a similar measure, formed as the weighted sum of several indicators, to assess the health of newborn babies (Apgar, 1953). The APGAR score developed by van den Berg and Danilenko considers six continuous indicators, presented in table 5.

| Indicator | Formula |
|---|---|
| Water coverage | $\frac{\text{Population with easy access to water services}}{\text{Total population under utility's nominal responsibility}} \times 100$ |
| Sewerage coverage | $\frac{\text{Population with sewerage services}}{\text{Total population under utility's nominal responsibility}} \times 100$ |
| Non revenue water | $\frac{\text{Volume of water produced in } m^3\text{-Volume of water sold in } m^3}{\frac{\text{Length of water distribution network in km}}{365}} \times 1,000,000$ |
| Affordability | $\frac{\frac{\text{Total operating revenue}}{\text{Exchange rate}}}{\text{GNI per capita} \times \text{Population with access to water} \times 1{,}000}$ |
| Collection period | $\frac{\text{Year end accounts receivable}}{\text{Total Operating Revenue} \times 365}$ |
| Operating Cost Coverage | $\frac{\text{Total Operating Revenue}}{\text{Total Operating Expenses}}$ |

Table 5: The six APGAR score indicators. Source: van den Berg and Danilenko, 2010.

Our sample of eight utilities from the IBNET database considers the same six indicators that van den Berg and Danilenko use in their APGAR score. Table 6 presents the descriptive statistics for our sample.

| **Indicator** | **Q1** | **Median** | **Q3** |
|---|---|---|---|
| Water Coverage | 54% | 69% | 94% |
| Sewerage Coverage | 37% | 47% | 73% |
| Non Revenue Water | 18.24 $m^3$/km/day | 32.97 $m^3$/km/day | 149.82 $m^3$/km/day |
| Affordability | .76% | 1.03% | 2.25% |
| Collection period | 90 days | 198 days | 297 days |
| Operating Cost Coverage ratio | 0.81 | 0.93 | 1.03 |

Table 6: Descriptive statistics of indicators for the eight utilities in our sample. Source: IBNET database.

It is worth noting the disparity between water and sewerage coverage in our sample. Overall, 69% of the urban population within the jurisdiction of the utilities in the sample

has access to water. On the other hand, the relevant figure for access to sanitation services is only 47%. Table 6 suggests that the typical utility loses 33 $m^3$ per kilometer of water network per day either due to physical or commercial reasons. The latter could include theft, illegal connections and/or inefficiencies in billing (van den Berg and Danilenko, 2010). Furthermore, median expenditure for water and wastewater services accounts for approximately 1% of national GNI per capita in the countries included in our sample. Finally, the two indicators that capture financial sustainability are collection period and operating cost coverage. The former, with a median value of 198 days, suggests that the typical utility needs 198 days to collect payments from customers. Operating cost coverage, with a median of 0.93, illustrates that the typical utility cannot cover its costs through its operating revenues.

In order to calculate the APGAR score van den Berg and Danilenko (2010) convert the indicators presented in table 5 into discrete variables using a series of thresholds based on expert knowledge. Table 11 in the Appendix presents the details of the transformation. The discrete indicators are then summed to create the final APGAR score of a utility. While the methodology developed by van den Berg and Danilenko is very important in identifying utilities that are at risk, the results of their APGAR score are constrained by the subjectivity of the thresholds used. In addition, the APGAR score is a static index that rates every utility based on its performance at a specific time period, without considering previous performance.

Our goal is to extend the APGAR score of van den Berg and Danilenko by estimating a performance index using the generalized dynamic factor model for panel data presented in section 2. The advantage of our index is twofold: 1) It is dynamic as performance in every time period is measured by a smoothed index that includes 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 à vis other water providers at the national, regional or even international level. Finally, the trend of the index can identify inefficient practices and reinforce successful policies.

## 4.3 Initial values

In section 2.5 we specified several scenarios with respect to parameter formulation. In this application we have chosen to use scenarios $\mathbf{B}_1$, $\mathbf{D}_1$, $\mathbf{T}_1$ and $\mathbf{Q}_1$. This parameter formulation is equivalent to the case illustrated in section 2.4.1, the only difference being that here we are considering multiple utilities as opposed to just one. Since no correlation exists between utilities the only restriction required for the identifiability of the model is that presented in (38). 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.

We specify the following initial value for $\mathbf{B}$, denoted by $\mathbf{B}^0$:

$$\underset{(\theta\times p)\times\theta}{\mathbf{B}^0} = \begin{bmatrix} \mathbf{b}^0 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{b}^0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{b}^0 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{b}^0 \end{bmatrix}, \tag{60}$$

where $\mathbf{b}^0 = \left(\frac{1}{p}\right)\mathbf{i}_p$ . In section 2.4.1 we discussed the need to consider only the positive roots of the factor loadings. This restriction will not affect the estimation of the smoothed index so long as all of the indicators are positively correlated to the unobserved state vector. 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.

For the covariance matrix of the idiosyncratic errors, we choose formulation $\mathbf{D}_1$ such that:

$$\underset{(\theta\times p)\times(\theta\times p)}{\mathbf{D}} = \text{diag}(\mathbf{d}_j), \tag{61}$$

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:

$$\mathbf{D}^0 = \text{diag}\left\{\mathbf{C}_{yy} - \left(\mathbf{B}^0 \times \left(\mathbf{B}^0\right)^T\right)\right\}, \tag{62}$$

where $\mathbf{C}_{yy}$ will have the following form:

$$\underset{(\theta\times p)\times(\theta\times p)}{\mathbf{C}_{yy}} = \begin{bmatrix} Var(\mathbf{Y}_{1,t}) & 0 & \dots & 0 \\ 0 & Var(\mathbf{Y}_{2,t}) & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & Var(\mathbf{Y}_{\theta,t}) \end{bmatrix}, \tag{63}$$

with every element along the diagonal of $\mathbf{C}_{yy}$, $\text{Var}(\mathbf{Y}_{j,t})$, being a $p \times p$ covariance matrix. Note that (62) was determined after solving (5) for $\mathbf{D}$ and incorporating the result in (40). Given the specification of $\mathbf{B}^0$ and $\mathbf{D}^0$, the first cycle of the 2CCEM algorithm outlined in (44)-(46) will yield ML estimates of $\mathbf{B}$ and $\mathbf{D}$. Note that during the first iteration of the first cycle of the 2CCEM algorithm we set $\mathbf{T} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$. In addition, we use formulations $\mathbf{T}_1$ and $\mathbf{Q}_1$. 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):

$$\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t. \tag{64}$$

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 (51), is equal to $\boldsymbol{\gamma}^T\mathbf{Y}_1$.

## 4.4 Results

The estimated parameters are presented in Table 7. The transformation discussed in section 4.3, has resulted in positive values for the estimated factor loadings. In addition, the fact that the indicators are standardized, makes comparisons between factors loadings meaningful. Our results indicate that water coverage, affordability and collection period are the

| Indicator | **B** | **D** |
|---|---|---|
| Water Coverage | .5779 | .8515 |
| Sewerage Coverage | .2383 | .9747 |
| Non Revenue Water | .38447 | .9343 |
| Affordability | .5372 | .8717 |
| Collection period | .4835 | .89604 |
| Operating Cost Coverage | .2376 | .9749 |
| **T** | .78415 | |

Table 7: Factor loadings, variance of the error term and AR(1) coefficient estimates.

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, suggesting that being able to promptly collect payments from customers is a very important indicator for a utility's performance. 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. Sewerage coverage ranks fifth with a low factor loading suggesting that provision of sanitation services is not critical in judging a utility's performance. Operating cost coverage is the least important out of the six indicators. This result underlines the fact that due to the nature of the industry, public water utilities can be expected to operate at a loss.

Figure 1 illustrates the smoothed estimate of the performance index for each of the eight utilities in the sample. The index is denoted by a bold red line while the six standardized indicators are denoted by the dotted lines. 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.

Figure 2 compares the smoothed index, estimated by our model, to the APGAR score of van den Berg and Danilenko analyzed in section 4.2. It is clear that the smoothed index exhibits less variability than the APGAR score. The latter evaluates a utility only on the

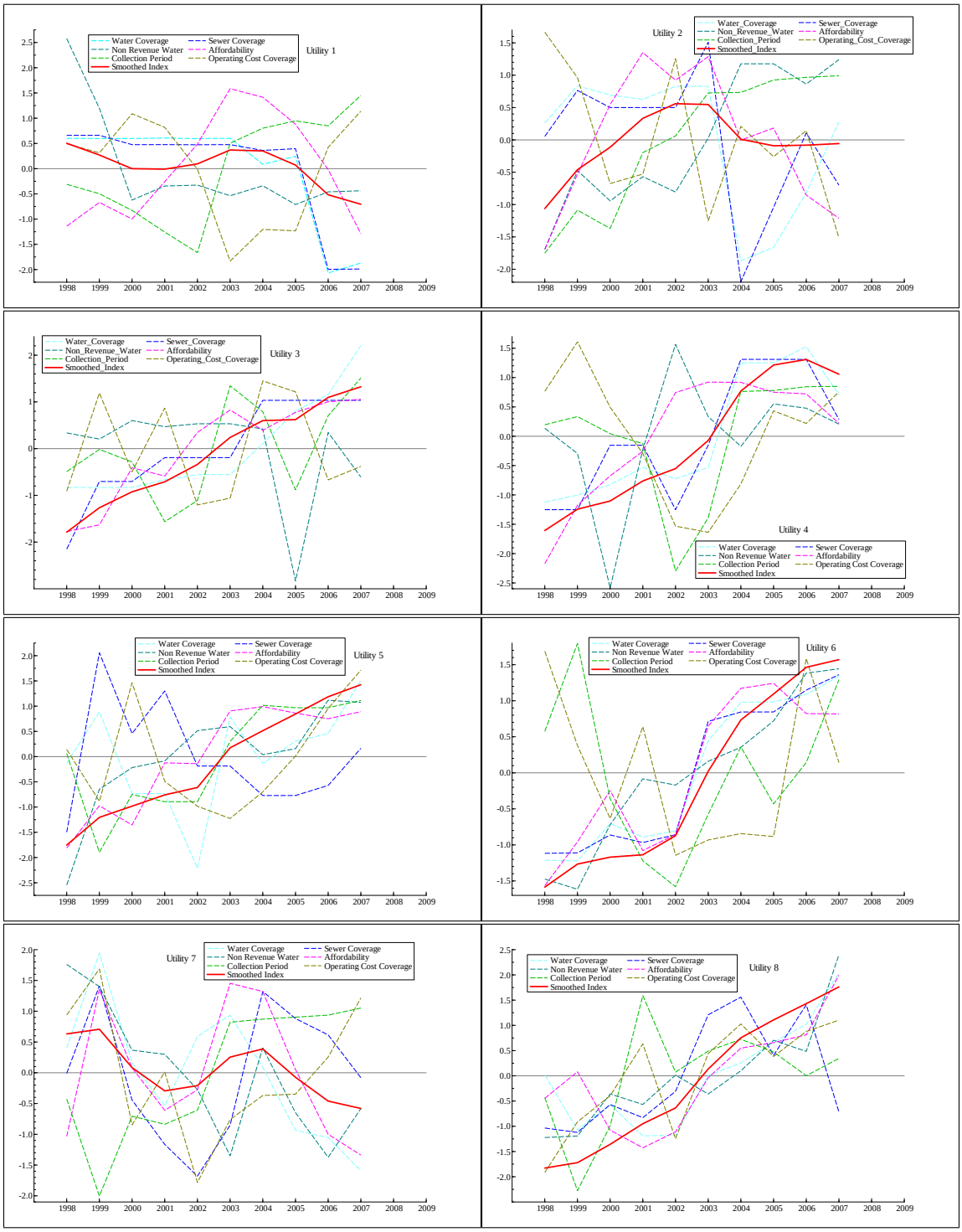


Figure 1: Standardized indicators and smoothed index for the 8 water utilities.

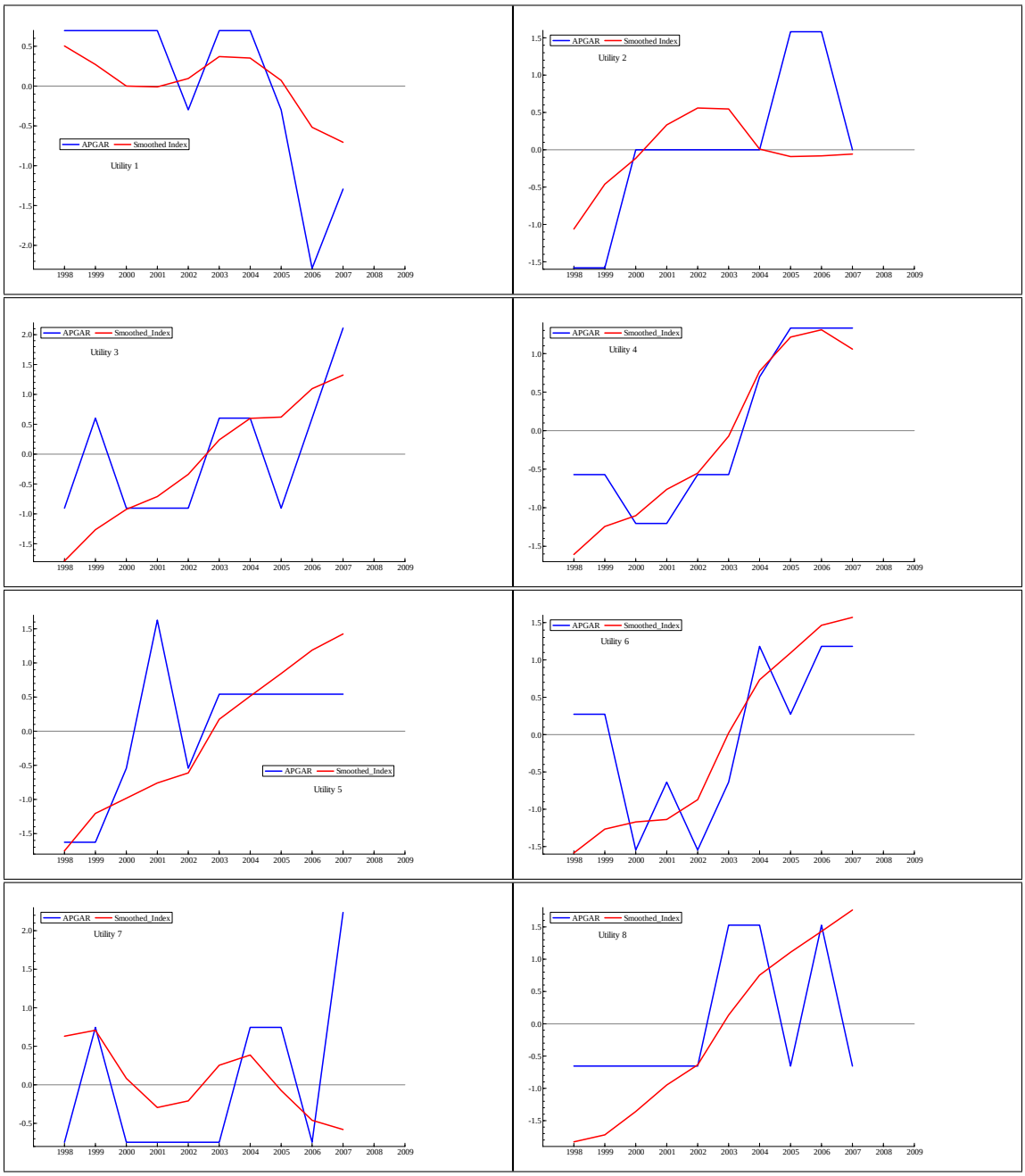


Figure 2: Smoothed index vis a vis the APGAR score for the 8 water utilities.

basis of its performance in a specific year and is hence highly volatile. In all panels of Figure 2 we detect significant jumps in the value of the APGAR score from year to year. On the other hand, the smoothed index evaluates a utility's performance using information from the entire sample. Thus, poor performance in a specific year is not penalized as much. For example, the APGAR score of utility 1 drops drastically between 2004-2006. However, the value of the smoothed index does not decrease as much as the APGAR score given the previous performance of the utility. As a result, the smoothed index is a more effective planning and assessment tool for utility managers.

Convergence of the 2CCEM algorithm is achieved after 25 iterations. Figure 3 illustrates the value of the log-likelihood for both cycles of the 2CCEM algorithm. The top panel, labeled $\mathbf{Z}_{\boldsymbol{\Psi}_1}$, pertains to equation (52) while the bottom panel, labeled $\mathbf{Z}_{\boldsymbol{\Psi}_2}$, pertains to equation (57).

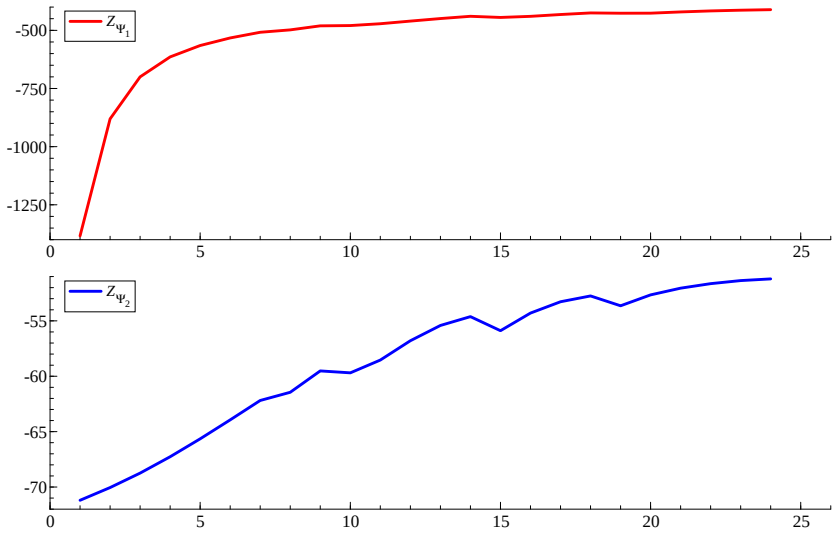


Figure 3: Log Likelihood values of $\mathbf{Z}_{\boldsymbol{\Psi}_1}$ and $\mathbf{Z}_{\boldsymbol{\Psi}_2}$.

## 4.5 Diagnostic checking

One of the main assumptions of our model is that error terms of both the observation equation and the state equation are normally distributed. This assumption can be verified by considering the one step ahead forecast errors presented in (58). Durbin and Koopman

(2001, p. 33) consider the standardized one-step ahead forecast errors $\boldsymbol{\xi}_t$ given by:

$$\boldsymbol{\xi}_t = \frac{\mathbf{v}_t}{\sqrt{\mathbf{F}_t}}, \qquad \text{where } t = 1, ..., n. \tag{65}$$

The computation of $\boldsymbol{\xi}_t$ involves calculating the Cholesky decomposition of $\mathbf{F}_t$ for every time period $t$. The sampling distribution of $\boldsymbol{\xi}_t$ is presented in figure 4. Normality of the sampling distribution can be assessed by the skewness and kurtosis statistics, denoted by $\sqrt{b_1}$ and $b_2$ respectively (D'Agostino and Pearson, 1973). The statistics are defined as:

$$\sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}}, \qquad b_2 = \frac{m_4}{m_2^2},$$

where $m_2$, $m_3$ and $m_4$ are the second, third and forth moments of the standardized forecast errors. Bowman and Shenton (1975) suggest that the skewness and kurtosis statistics are asymptotically normally distributed as:

$$\sqrt{b_1} \sim N(0, \frac{6}{n}), \qquad b_2 \sim N(3, \frac{24}{n}).$$

Furthermore, the two statistics can be combined in an omnibus test using:

$$X^2(\sqrt{b_1}) + X^2(b_2),$$

where $X(\sqrt{b_1})$ and $X(b_2)$ are standardized values of the skewness and kurtosis statistics respectively. The statistic of the omnibus test is distributed as a $\chi^2$ with two degrees of freedom (Bowman and Shenton, 1975). The estimated values for the skewness, kurtosis and omnibus test statistics are presented in table 8. The P-value of $b_2$ suggests that we fail to reject the null hypothesis of normal kurtosis. In addition, at the 1% level of significance the omnibus test further supports the normality of $\boldsymbol{\xi}_t$ . However, the null hypothesis of normal skewness is rejected.

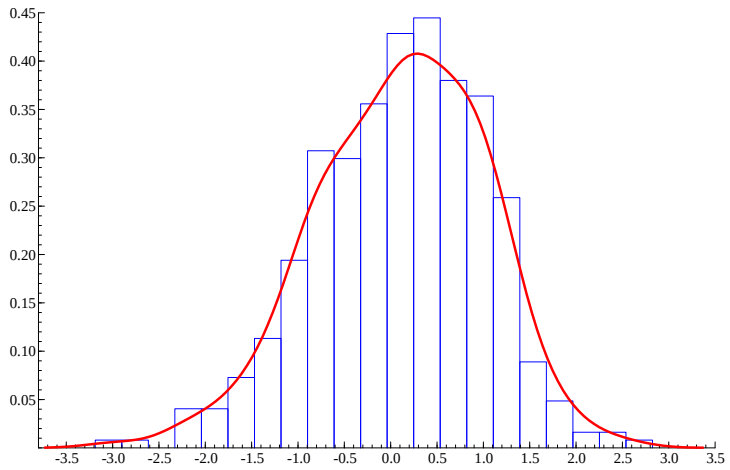


Figure 4: Standardized one step ahead forecast errors.

| | Parameter value | P-value |
|---|---|---|
| Skewness | -0.327 | 0.005 |
| Kurtosis | 3.072 | 0.761 |
| Omnibus test | 7.793 | 0.020 |

Table 8: Estimated skewness, kurtosis and omnibus test statistics.

## 4.6 Confidence Intervals of factor loadings

In order to make any inference regarding the factor loadings, we estimate the asymptotic variance of those parameters. The asymptotic covariance matrix of the ML estimator can be computed by evaluating the inverse of the matrix of second derivatives at the ML estimates (Cramer, 1946, p. 489; Amemiya, 1985, p. 123; Greene, 2008, p. 481). Following this methodology, we calculate the second derivative of $\mathbf{Z}_{\mathbf{\Psi}_1}$ with respect to $\mathbf{B}$ as follows:

$$\text{Var}(\mathbf{B}) = \left( -\frac{\partial^2 \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi})}{\partial \mathbf{B} \partial \mathbf{B}^T} \right)^{-1} = \left( 2 \sum_{t=1}^{n} \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t \right)^{-1}.$$

The complete derivation of the asymptotic variance can be found in section 6.3 of the Appendix. Table 9, shows the variance of the factor loadings and the corresponding 95% confidence intervals.

| Indicator | **B** | Var(**B**) | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower bound | Upper bound |
| Water Coverage | .5779 | .00319 | 0.4679 | 0.6878 |
| Sewerage Coverage | .2383 | .00366 | 0.1206 | 0.3559 |
| Non Revenue Water | .3845 | .00352 | 0.2693 | 0.499 |
| Affordability | .5372 | .00327 | 0.4250 | 0.648 |
| Collection period | .4835 | .00337 | 0.3707 | 0.596 |
| Operating Cost Coverage | .2376 | .00366 | 0.1199 | 0.3552 |

Table 9: Confidence intervals of the estimated factor loadings.

### 4.7 Uniqueness of the smoothed index

To verify uniqueness of the smoothed index we conduct a series of simulations. We create five thousand bootstrapped samples from our sample of eight utilities. The parameters of each bootstrapped sample are then estimated using the 2CCEM algorithm and a set of smoothed indices is obtained from every estimation. Those bootstrapped indices are then compared, on a utility by utility basis, with the indices estimated from the original sample. If the parameters of our model are uniquely estimated this would suggest that both the bootstrapped and the original smoothed indices are equal, or else, their difference is equal to zero. Figure 5 illustrates the sampling distributions of the difference between each of the five thousand bootstrapped samples and the original sample. There are ten panels in Figure 5 each representing the distribution of a particular year. Finally, Table 10 shows the t-statistics of those differences. At the 5% level of significance we fail to reject the null hypothesis of zero difference between the bootstrapped and the original smoothed indices.

## 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 individual, firm or economy (Stock and Watson, 1989). Even when multiple individuals are considered (Forni et al. 2000) only a single unobserved index, common for all individuals, is estimated for every time period. We develop

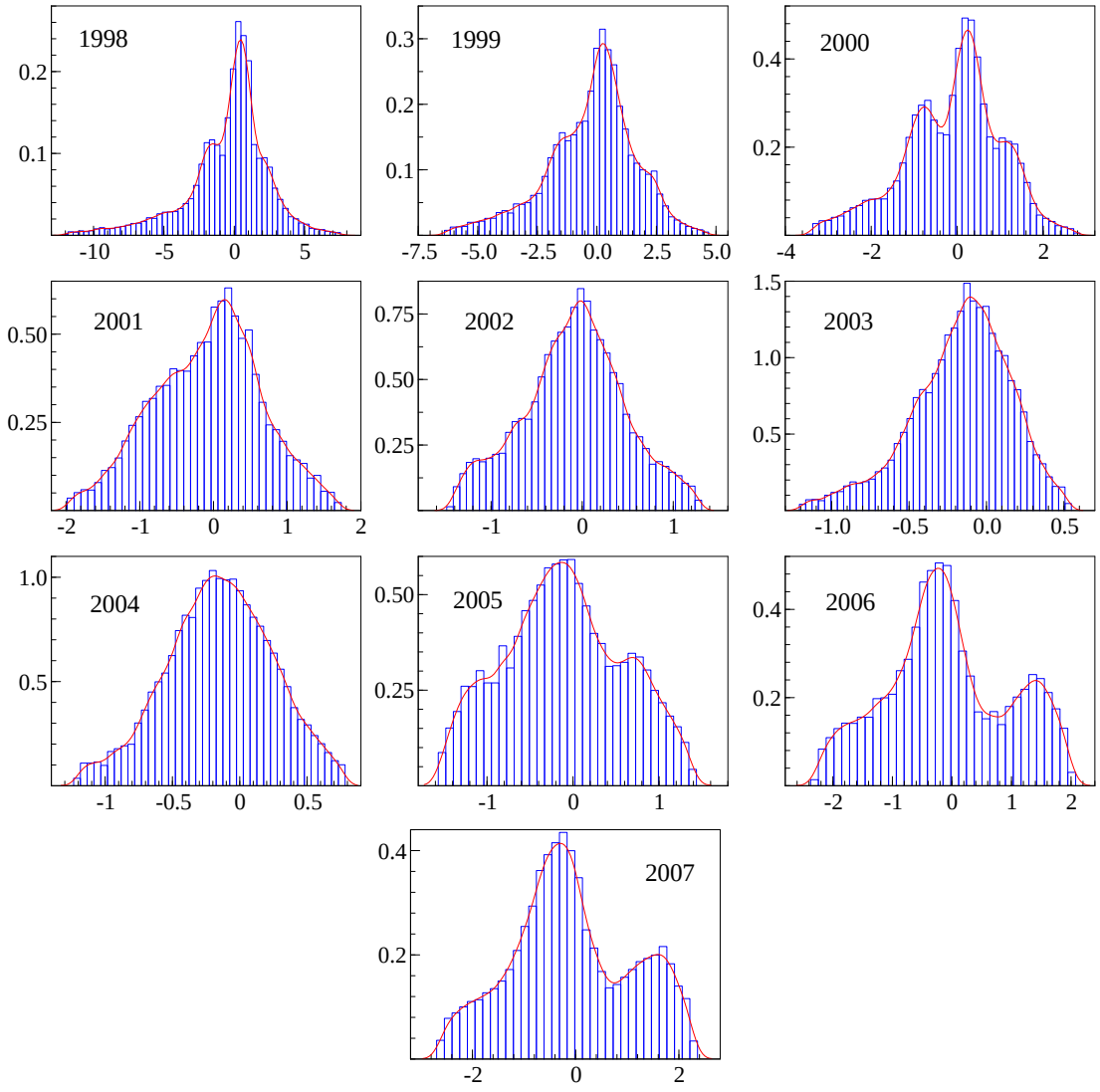


Figure 5: Distributions of the difference between the bootstrapped and the original smoothed indices.

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|---|
| Utility | -0.032 | -0.170 | -0.377 | -0.598 | -0.849 | -0.861 | 0.015 | 0.788 | 1.581 | 1.721 |
| 1 | **0.974** | **0.865** | **0.706** | **0.550** | **0.397** | **0.390** | **0.988** | **0.431** | **0.115** | **0.086** |
| Utility | 0.014 | -0.045 | -0.180 | -0.466 | -0.798 | -0.683 | 0.673 | 1.029 | 0.946 | 0.877 |
| 2 | **0.989** | **0.965** | **0.857** | **0.642** | **0.425** | **0.495** | **0.502** | **0.304** | **0.345** | **0.381** |
| Utility | 0.033 | 0.060 | 0.079 | 0.106 | -0.121 | -0.527 | -0.631 | -0.264 | -0.729 | -0.904 |
| 3 | **0.973** | **0.952** | **0.937** | **0.915** | **0.904** | **0.599** | **0.528** | **0.792** | **0.466** | **0.367** |
| Utility | 0.031 | 0.059 | 0.110 | 0.092 | 0.134 | 0.162 | -0.762 | -1.288 | -1.000 | -0.541 |
| 4 | **0.975** | **0.953** | **0.913** | **0.927** | **0.893** | **0.871** | **0.446** | **0.199** | **0.318** | **0.589** |
| Utility | 0.038 | 0.054 | 0.077 | 0.093 | 0.190 | -0.175 | -0.271 | -0.630 | -0.834 | -1.025 |
| 5 | **0.970** | **0.957** | **0.939** | **0.926** | **0.849** | **0.861** | **0.787** | **0.529** | **0.405** | **0.306** |
| Utility | 0.028 | 0.052 | 0.187 | 0.545 | 0.798 | 0.027 | -0.984 | -1.181 | -1.244 | -1.209 |
| 6 | **0.978** | **0.958** | **0.852** | **0.586** | **0.426** | **0.978** | **0.326** | **0.238** | **0.214** | **0.228** |
| Utility | -0.046 | -0.207 | -0.237 | -0.144 | -0.150 | -0.277 | -0.024 | 1.019 | 1.497 | 1.567 |
| 7 | **0.963** | **0.836** | **0.813** | **0.886** | **0.881** | **0.782** | **0.981** | **0.309** | **0.135** | **0.118** |
| Utility | 0.040 | 0.117 | 0.171 | 0.173 | 0.191 | -0.106 | -0.675 | -1.072 | -1.174 | -1.461 |
| 8 | **0.968** | **0.907** | **0.864** | **0.863** | **0.849** | **0.915** | **0.500** | **0.284** | **0.241** | **0.145** |

Table 10: T-statistics of differences between original and bootstrapped smoothed indices. P-values appear in bold.

a model that estimates one index for every individual 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 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 and sanitation 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 the Federazione Eni Enrico Mattei (FEEM, 2011).

# 6 Appendix

## 6.1 Closed form solution for $\boldsymbol{\Gamma}_{\mathbf{Y}}(0)$ and $\boldsymbol{\Gamma}_{\mathbf{Y}}(1)$

Obtaining a closed form solution for $\boldsymbol{\Gamma}_{\mathbf{Y}}(0)$ requires replacing (25) into (28) which yields:

$$\boldsymbol{\Gamma}_{\mathbf{Y}}(0) = \boldsymbol{\phi}_{(\bullet)}^2 \boldsymbol{\Gamma}_{\mathbf{Y}}(0) + \boldsymbol{\phi}_{(\bullet)} \mathbf{B}\mathbf{T}^* \mathbf{T}^2 \boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' - \boldsymbol{\phi}_{(\bullet)}^2 \mathbf{D} + \mathbf{B}\mathbf{T}^*\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{B}\mathbf{Q}\mathbf{B}' + \mathbf{D} + \boldsymbol{\phi}_{(\bullet)}^2 \mathbf{D}. \tag{66}$$

Applying the vec operator to (66) we get:

$$\begin{aligned}
\operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] =& \operatorname{vec}\left[\boldsymbol{\phi}_{(\bullet)}^2 \boldsymbol{\Gamma}_{\mathbf{Y}}(0) + \boldsymbol{\phi}_{(\bullet)} \mathbf{B}\mathbf{T}^* \boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{B}\mathbf{T}^*\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}' + \mathbf{B}\mathbf{Q}\mathbf{B}' + \mathbf{D}\right] \\
=& \boldsymbol{\phi}_{(\bullet)}^2 \operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] + \boldsymbol{\phi}_{(\bullet)} \operatorname{vec}\left[\mathbf{B}\mathbf{T}^* \boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}'\right] + \operatorname{vec}\left[\mathbf{B}\mathbf{T}^*\mathbf{T}\boldsymbol{\Gamma}_{\mathbf{U}}(0)\mathbf{B}'\right] + \operatorname{vec}\left(\mathbf{B}\mathbf{Q}\mathbf{B}'\right) + \operatorname{vec}(\mathbf{D}) \\
=& \boldsymbol{\phi}_{(\bullet)}^2 \operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] + \boldsymbol{\phi}_{(\bullet)} \left\{(\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*) \operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{U}}(0)\right]\right\} \\
& + \left\{(\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*\mathbf{T}) \operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{U}}(0)\right]\right\} + (\mathbf{B} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{Q}) + \operatorname{vec}(\mathbf{D})
\end{aligned} \tag{67}$$

Replacing (20) into (67) we have:

$$\begin{aligned}
\operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] =& \boldsymbol{\phi}_{(\bullet)}^2 \operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] + \boldsymbol{\phi}_{(\bullet)} \left\{(\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*) \left[\mathbf{I} - \mathbf{T} \otimes \mathbf{T}\right]^{-1} \operatorname{vec}(\mathbf{Q})\right\} \\
& + \left\{(\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*\mathbf{T}) \left[\mathbf{I} - \mathbf{T} \otimes \mathbf{T}\right]^{-1} \operatorname{vec}(\mathbf{Q})\right\} + (\mathbf{B} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{Q}) + \operatorname{vec}(\mathbf{D}).
\end{aligned} \tag{68}$$

Finally, solving 68 for $\operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right]$ yields:

$$\begin{aligned}
\operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(0)\right] =& \left\{\left[\mathbf{I} - \mathbf{T} \otimes \mathbf{T}\right]^{-1} \operatorname{vec}(\mathbf{Q}) \left[\boldsymbol{\phi}_{(\bullet)} (\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*) + (\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*\mathbf{T})\right]\right. \\
& \left. + (\mathbf{B} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{Q}) + \operatorname{vec}(\mathbf{D})\right\} \left(\mathbf{i}_{\theta p} - \boldsymbol{\phi}_{(\bullet)}^2 \mathbf{i}_{\theta p}\right)
\end{aligned}$$

The closed form of $\boldsymbol{\Gamma}_{\mathbf{Y}}(1)$, obtained in a similar way, is:

$$\begin{aligned}
\operatorname{vec}\left[\boldsymbol{\Gamma}_{\mathbf{Y}}(1)\right] =& \left\{\left[\mathbf{I} - \mathbf{T} \otimes \mathbf{T}\right]^{-1} \operatorname{vec}(\mathbf{Q}) \left[\boldsymbol{\phi}_{(\bullet)} (\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*\mathbf{T}) + (\mathbf{B} \otimes \mathbf{B}\mathbf{T}^*)\right]\right. \\
& \left. + \boldsymbol{\phi}_{(\bullet)} \operatorname{vec}\left(\mathbf{B}\mathbf{Q}\mathbf{B}'\right) + \boldsymbol{\phi}_{(\bullet)}^3 \operatorname{vec}(\mathbf{D})\right\} \left(\mathbf{i}_{\theta p} - \boldsymbol{\phi}_{(\bullet)}^2 \mathbf{i}_{\theta p}\right)
\end{aligned}$$

## 6.2 APGAR thresholds

Table 11 illustrates the thresholds based on which each of the six continuous indicators are transformed into discrete variables to form the APGAR score of van den Berg and Danilenko (2010).

| Indicator code in IBNET | Indicator name | Value |
|---|---|---|
| 1.1 | Water Coverage | 0 if $\leq$ 75% |
| | | 1 if between 75% and 90% |
| | | 2 if $>$ 90% |
| 2.1 | Sewerage Coverage | 0 if $\leq$ 50% |
| | | 1 if between 50% and 80% |
| | | 2 if $>$ 80% |
| 6.2 | Non Revenue Water | 0 if $\geq$ 40 |
| | | 1 if between 10 and 40 |
| | | 2 if $<$ 10 |
| 19.1 | Affordability | 0 if $>$ 2.5% |
| | | 1 if between 1% and 2% |
| | | 2 if $<$ 1% |
| 23.1 | Collection period | 0 if $\geq$ 180 days |
| | | 1 if between 90 and 180 days |
| | | 2 if < 90 days |
| 24.1 | Operating Cost Coverage | 0 if < 1 |
| | | 1 if between 1 and 1.4 |
| | | 2 if $\geq$ 1.4 |
| | Overall APGAR score | Critically low $\leq$ 3.6 |
| | | 3.6 <Fairly low$\leq$ 7.2 |
| | | Normal > 7.2 |

Table 11: APGAR score thresholds. Source: van den Berg and Danilenko (2010).

## 6.3 Variance of the estimated factor loadings

In this part of the Appendix we present in detail the calculations required to derive the asymptotic variance of the factor loadings. Our goal is to show that:

$$\text{Asymptotic Var}(\mathbf{B}) = \left( -\frac{\partial^2 \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi})}{\partial \mathbf{B} \partial \mathbf{B}^T} \right)^{-1} = \left( 2 \sum_{t=1}^{n} \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t \right)^{-1}.$$

We start by differentiating (52) with respect to $\mathbf{B}$.

$$\begin{aligned} \frac{\partial \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi})}{\partial \mathbf{B}} &= -\sum_{t=1}^{n} -\hat{\mathbf{u}}_t^T \mathbf{D}^{-1} (\mathbf{Y}_t - \mathbf{B}\hat{\mathbf{u}}_t) - (\mathbf{Y}_t - \mathbf{B}\hat{\mathbf{u}}_t)^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t = \\ &= \sum_{t=1}^{n} \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} (\mathbf{Y}_t - \mathbf{B}\hat{\mathbf{u}}_t) + (\mathbf{Y}_t - \mathbf{B}\hat{\mathbf{u}}_t)^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t. \end{aligned}$$

The second derivative of (52) with respect to $\mathbf{B}$ is:

$$\frac{\partial^2 \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi})}{\partial \mathbf{B} \partial \mathbf{B}^T} = \sum_{t=1}^{n} -\hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t - \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t = -2 \sum_{t=1}^{n} \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t.$$

Therefore:

$$\left( -\frac{\partial^2 \mathbf{Z}_{\mathbf{\Psi}_1}(\mathbf{\Psi}_1; \mathbf{\Psi})}{\partial \mathbf{B} \partial \mathbf{B}^T} \right)^{-1} = \left( 2 \sum_{t=1}^{n} \hat{\mathbf{u}}_t^T \mathbf{D}^{-1} \hat{\mathbf{u}}_t \right)^{-1} \quad \square$$

# References


Amemiya, T. (1985). *Advanced Econometrics* (1 ed.). Harvard University Press.

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Anesthesia & Analgesia 32*(4), 260–267.

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

Bernanke, B. S. and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics 50*(3), 525–546.

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

Bowman, K. O. and L. R. Shenton (1975). Omnibus test contours for departures from normality based on b1 and b2. *Biometrika 62*(2), 243–250.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

D'Agostino, R. and E. S. Pearson (1973). Tests for departure from normality. empirical results for the distributions of b2 and b1. *Biometrika 60*(3), 613–622.

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.

Everitt, B. and G. Dunn (1998). *Applied Multivariate data analysis* (sixth ed.). New York: John Wiley.

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.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2001). Coincident and leading indicators for the euro area. *The Economic Journal 111*(471), 62–85.

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.

Ghahramani, Z. and G. Hinton (1997). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Toronto.

Greene, W. (2008). *Econometric Analysis* (6th ed.). Prentice Hall.

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.

Harvey, A. C. (1991). *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 (2005). International benchmarking network for water and sanitation utilities. http://www.ib-net.org.

Jungbacker, B., S. Koopman, and M. van der Wel (2011). Maximum likelihood estimation for dynamic factor models with missing data. *Journal of Economic Dynamics and Control 35*(8), 1358–1368.

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

Kim, J.-O. and C. W. Mueller (1978). *Introduction to Factor Analysis: What It Is and How To Do It*. Sage Publications, Inc.

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.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica 39*(3), 577–591.

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 (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.

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.