Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

DYNAMIC MIXTURES OF FACTOR ANALYZERS TO CHARACTERIZE MULTIVARIATE AIR POLLUTANT EXPOSURES

By Antonello Maruotti[*†], Jan Bulla[‡] Francesco Lagona[§] Marco Picone[¶] and Francesca Martella[‖]

University of Southampton[*], Libera Università Maria Ss. Assunta[†], University of Bergen[‡], Università di Roma Tre[§], The Institute for Environmental Protection and Research (ISPRA)[¶] and Sapienza Università di Roma[‖]

The assessment of pollution exposure is based on the analysis of multivariate time series that include the concentrations of several pollutants as well as the measurements of multiple atmospheric variables. It typically requires methods of dimensionality reduction that are capable to identify potentially dangerous combinations of pollutants and, simultaneously, to segment exposure periods according to air quality conditions. When the data are high-dimensional, however, efficient methods of dimensionality reduction are challenging because of the formidable structure of cross-correlations that arise from the dynamic interaction between weather conditions and natural/anthropogenic pollution sources. In order to assess pollution exposure in an urban area while taking the above mentioned difficulties into account, we develop a class of parsimonious hidden Markov models. In a multivariate time-series setting, this approach allows to simultaneously perform temporal segmentation and dimensionality reduction. We specifically approximate the distribution of multiple pollutant concentrations by mixtures of factor analysis models, whose parameters evolve according to a latent Markov chain. Covariates are included as predictors of the chain transition probabilities. Parameter constraints on the factorial component of the model are exploited to tune the flexibility of dimensionality reduction. In order to estimate the model parameters efficiently, we propose a novel three-step Alternating Expected Conditional Maximization (AECM) algorithm, which is also assessed in a simulation study. In the case study, the proposed methods were capable (1) to describe the exposure to pollution in terms of a few latent regimes, (2) to associate these regimes with specific combinations of pollutant concentration levels as well as distinct correlation structures between concentrations, and (3) to capture the influence of weather conditions on transitions between regimes.

MSC 2010 subject classifications: Primary 62-07, 62H25; secondary 62P12
Keywords and phrases: Hidden Markov models, AECM algorithm, Dimensionality reduction, three-step algorithm

1. Introduction. In urban areas, the study of pollutant concentrations is of high interest since negative effects of air pollution have been observed (Lee and Sahu, 2016; Lee et al., 2015; Sahu et al., 2014); this has in particular been reported for major pollutants such as carbon monoxide, nitrogen monoxide, particulate matter, and sulphur dioxide (Kim et al., 2013; Latza et al., 2009). Local and atmospheric sources of pollution may influence pollution concentrations, either directly, or indirectly, or both. Empirical and theoretical studies applied sophisticated strategies to link pollution with (observable and unobservable) atmospheric and other sources (see, e.g., Chattopadhyay et al., 2015; Cooley et al., 2012; Lagona et al., 2011; Paciorek et al., 2009; Park et al., 2001), modeling the fluctuations of concentrations over long periods, and accounting for uncertainty in the observed measurements. Although it is known that air pollution is a complex mixture of multiple pollutants, the pollution analyzes are in general conducted on each pollutant separately (see, e.g., Martinez-Zarzoso and Maruotti, 2013; Bornn et al., 2012; Greven et al., 2011; Shaddick et al., 2008). Hence, alternative multivariate approaches have been overlooked so far, although the investigation of exposures to multiple pollutants would be useful because it could provide a better understanding of potential combined effects of individual pollutants and their interactions.

We consider a multivariate setting, where pollution is understood as levels of the various pollutants, i.e., pollutant concentrations. Since these concentrations exhibit time-variability, multivariate hidden Markov models (HMMs) constitute a natural candidate for studying their joint temporal dynamics. More precisely, the latent process of the HMM approach allows to infer a finite number of concentration profiles, each of which is characterized by a specific combination of pollutant concentrations. These profiles are a direct result of clustering similarities occurring in both the variable space (i.e. between pollutants) and in a temporal neighborhood. Each profile is linked to an easily interpretable multivariate distribution, which provides an accurate representation of the respective concentration of air pollutants, illustrating the potential of this type of modeling for defining exposure characteristics (or conditions).

The proposed HMM accounts for three major dependency structures in multivariate time series data, including the correlation among multiple pollutants, temporal dependence, and heterogeneity. Following the HMMs literature, we assume that the hidden structure underlying the observed data is a first-order Markov chain, and that time-specific concentrations can be modeled as a multivariate process conditioning on the sequence of hidden states. The challenge of modeling multiple pollutants and their interactions

is fairly common to all analyzes of high dimensional data with many variable of interests. In the context of understanding exposure to multiple pollutants, dimensionality-related aspects present a challenge because these pollutants could be potentially highly correlated. Therefore, estimation and interpretation of the parameters of interest may become non-trivial. In order to examine the interrelationships between pollutant concentrations to perform dimensionality reduction in the variable space simultaneously (allowing for an easy interpretation of model parameters), we propose the use of a latent factor model (Field et al., 2016; Yao et al., 2005; Rosti et al., 2001). Accordingly, we define a general class of parsimonious HMMs by imposing a factor decomposition on state-specific covariance matrices. The loadings and noise terms of the covariance matrix may be constrained to be equal or unequal across latent states. In addition, the noise term may be subject to further restrictions, resulting in a set of eight parsimonious covariance structures (McNicholas and Murphy, 2010, 2008; McLachlan et al., 2003; Ghahramani and Hinton, 1997). This model structure allows to account for dependence between pollutants, and provides a clear interpretation of the (latent) association structure between pollutants. Moreover, unlike irreversible and progressive phenomena, pollution data often comprises periods of flat stretches (also known as background concentrations) and sudden bursts. Therefore, not only the identification of different levels of exposure, but also the investigation of their transitions patterns is of interest. In order to characterize transitions between hidden states, along with estimating the effects of observed atmospheric variables on these transitions, we use a multinomial logistic regression model, which is capable of revealing the heterogeneity in the transition process.

In this framework, model parameters can be estimated by a full maximum likelihood method based on the Alternating Expectation Conditional Maximization (AECM) algorithms (Yao et al., 2005; Field et al., 2016), and recursions widely used in the HMM literature (Baum et al., 1970). However, the inclusion of covariates in the hidden process often renders the estimation step more time-consuming. Furthermore, several other drawbacks such as slowness of convergence, instability, and multimodality in the likelihood function regularly arise. To overcome all the aforementioned obstacles, we extend the recent proposal of Bartolucci et al. (2015) and introduce a three-step AECM algorithm. Moreover, we consider a multiple random starting point strategy for enhancing the chance of reaching the global maximum of the likelihood.

After introducing the data sources of the analyzed pollutants (Section 2.1), we summarize various properties of pollutant concentrations and atmo-

spheric variables in Section 2.2. These include dynamics over time, marginal distributions, and bivariate correlation. The proposed modeling framework is introduced in Section 2.3: the non-homogeneous multivariate HMMs is specified, along with several possible constraints on the decomposition of covariance matrix that allows for dimensionality reduction. Identifiability conditions (Section 2.4) are also discussed. In Section 3 we outline an ad hoc version of the AECM algorithm to estimate model parameters to allow for the inclusion of covariates in the hidden part of the model. Computational details (Section 3.2) and hidden state decoding (Section 3.3) are also discussed. In Section 4, we illustrate the proposal by a simulation study to investigate the empirical behavior of the propose approach with respect to several factors, such as the number of times, variables and hidden states. At last, we discuss and interpret the results of applying the proposal to the data, and focus on several different aspects of the modelling, as e.g. clustering and model fit.

2. Data and Methods.

2.1. *Data source.* In Italy, the Institute for Environmental Protection and Research (ISPRA, http://www.isprambiente.gov.it) mainly provides information regarding the nationwide state of the environment and environmental trends. The activities of ISPRA generally focus on monitoring the status and dynamics of the environment in response to the directives of EU Commission. It is the natural reference for the European Environment Agency, and its research activity focuses on thematic issues, such as environmental pollution, waste, etc., as well as multidisciplinary issues (e.g. environmental assessment).

The data considered in this research originate from a regional monitoring network system developed by the Lazio Region. This system has been organized in order to respond to the increasing demand for environmental information, but also for providing reliable data suitable for policy-related aspects. The entire network produces large amounts of air pollution measurements, which are commonly summarized for public reporting. Indeed, when leaving the data in their raw form, it proves difficult to provide a complete overview of the overall exposure conditions. Therefore, quantitative methods (models/summary statistics) with the ability to condense the large volume data into a small number of categorical summaries are regularly introduced. However, the method selected often depends on personal preferences or political agendas, and statistical aspects may be neglected.

Therefore, the data present a good opportunity for investigating the advan-

tages of a purely quantitative approach, which is able to summarize overall exposure conditions by means of categorical variables. These are, in our case, identified by the hidden states and the corresponding conditional distributions, which provide a condensed overview of their characteristics via the respective concentration profile. Moreover, we link the dynamics of the hidden states and thus the evolution of the exposure conditions in time to observed atmospheric variables, which improves the characterization of the different levels of exposure.

2.2. *Data description.* In 2011, the ISPRA air quality network recorded concentrations of nine air pollutants on hourly basis at monitoring stations in the central area of the city of Rieti (Italy). We averaged the pollution data to daily frequency, and carried out a log-transformation. The original data series are freely available to download from the ISPRA website (`www.brace.sinanet.isprambiente.it`). The Rieti site has been chosen as it is classified as a traffic location, although it is at short distance from green areas and forests that facilitate the movement of air masses and removal of pollutants. The recorded pollutants are:

- NO_2, Nitrogen dioxide
- SO_2, Sulphur dioxide
- PM_{10}, Particulate matter (aerodynamic diameter less than 10 μm)
- $PM_{2.5}$, Particulate matter (aerodynamic diameter less than 2.5 μm)
- O_3, Ozone
- CO, Carbon monoxide
- Toluene
- Etilbenzene
- Oxylene

The concentrations are expressed in $\mu g/m^3$, except for CO, which is expressed in mg/m^3. This set of pollutants has often been used to build air quality indicators (see, e.g., Pollice and Jona-Lasinio, 2009; Fassò et al., 2007). Most of the pollutants show a seasonal pattern with high concentrations in winter due to the combination of car exhaust and heating emissions. These concentrations tend to decrease during summer (see Figure 1). However, ozone concentration is subject to the opposite cycle. The chemical reactions leading to ozone formation take place in conditions of high solar radiation and heat. Consequently, the highest ozone concentrations are observed during summer, while the lowest concentrations are recorded during cold seasons. Sulfur dioxide, generated almost exclusively by traffic exposition, shows an almost uniform behavior during the year, with few peaks and small bursts.

FIG 1. *Pollutant concentrations over time*

Log-transformed pollutant concentrations over the one year period considered. Vertical gray lines separate blocks of length three months.

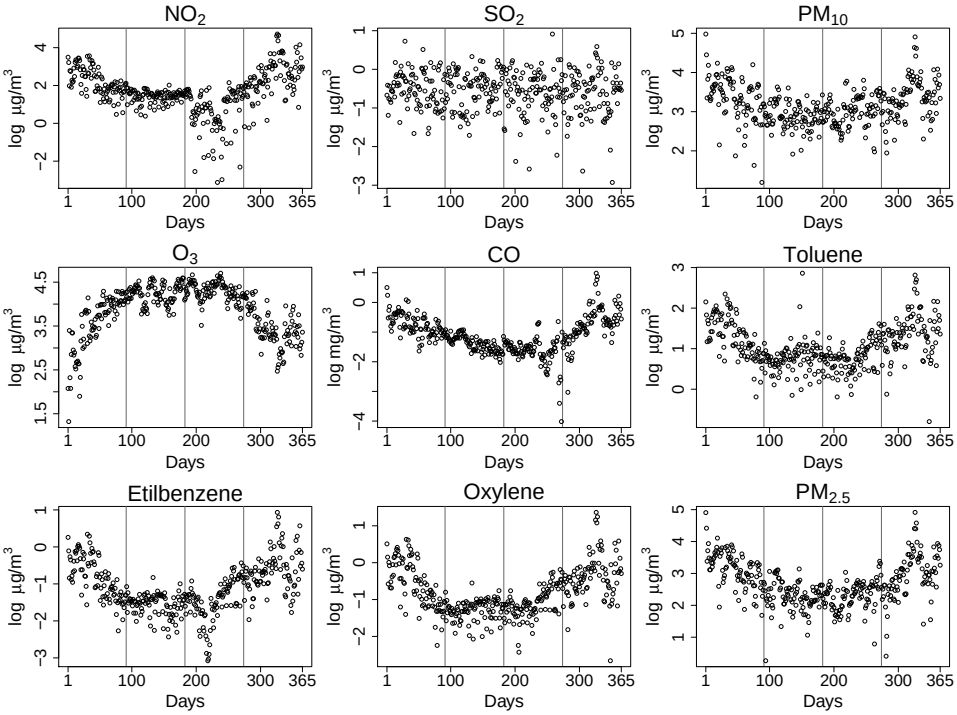
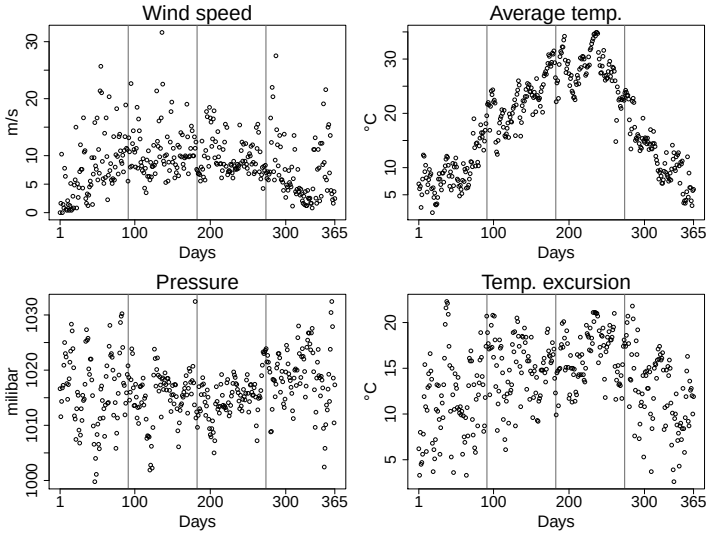


FIG 2. *Atmospheric variables over time*

Atmospheric variables over the one year period considered. Vertical gray lines separate blocks of length three months.



Therefore, atmospheric variables may play a major role in determining the level of exposure to particular pollutants on the one hand, but also for capturing time dependence as proxies for seasonal variations and characteristics on the other hand. Since pollution episodes are triggered by specific atmospheric factors, we also included the following atmospheric variables (see Figure 2) as covariates in the model, as they are important for deliquescence, formation, and conservation of pollutants characterizing the different levels of exposure:

- daily average wind speed
- daily average temperature
- daily average pressure
- temperature excursion.

Another noteworthy aspect is that the pollutants considered in our setting possess varying atmospheric lifetime. Thus, a strong daily dependence would be expected between pollutant concentrations (see e.g. Shaddick and Wakefield, 2002). Therefore, the association between different pollutants constitutes a crucial aspect of the analysis as well, and should not be neglected or treated as nuisance. Figure 3 shows scatter plots of the variables considered: the blue line results from simple linear correlation, the solid red line from a

local polynomial smoothing, and the red dashed line corresponds to confidence intervals of the smoothing line. Finally, the main diagonal contains the empirical densities. From the visual perspective, it is apparent that patterns of non-linear correlation can be detected, and that the correlation structure is rather heterogeneous since each pollutant seems to be related to others in different ways.

2.3. *Methodology.* Let $\{\mathbf{Y}_t, t = 1, \ldots, T\}$ denote a sequence of multivariate observations (i.e. pollutant concentrations), where $\mathbf{Y}_t = \{Y_{t1}, \ldots, Y_{tp}\} \in \mathcal{R}^p$, and $\{S_t, t = 1, \ldots, T\}$ denotes a Markov chain defined on the state space $\{1, 2, \ldots, K\}$. In the context of our application, all response variables are continuous, although HMMs may be also applied to variables having a different nature (see, e.g., Lagona et al., 2016, 2015; Bulla et al., 2012; Bartolucci and Farcomeni, 2009). For up to date reviews of HMMs under different settings see, e.g., Zucchini et al. (2016); Bartolucci et al. (2013); Maruotti (2011).

A HMM is a particular kind of dependent mixture. It is a stochastic process consisting of two parts: the underlying unobserved process $\{S_t\}$, fulfilling the Markov property, i.e.

$$\Pr(S_t \mid S_1, S_2, \ldots, S_{t-1}) = \Pr(S_t \mid S_{t-1}),$$

and the state-dependent observation process $\{\mathbf{Y}_t\}$ for which the conditional independence property holds, i.e.

$$f(\mathbf{Y}_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T, S_1, \ldots, S_T) = f(\mathbf{Y}_t \mid S_t),$$

where $f(\cdot)$ is a generic probability density function. Assuming that the hidden process follows a first-order Markov chain is equivalent to the assumption that any latent variable S_t given S_{t-1} is conditionally independent of $S_1, S_2, \ldots, S_{t-2}$. This dependence structure is seldom considered restrictive, and, due to its easy interpretation usually preferred to more complex structures of the latent variables.

The hidden Markov chain has K states, labeled from 1 to K, with initial probabilities

$$\pi_k = \Pr(S_1 = k), \quad k = 1, \ldots, K,$$

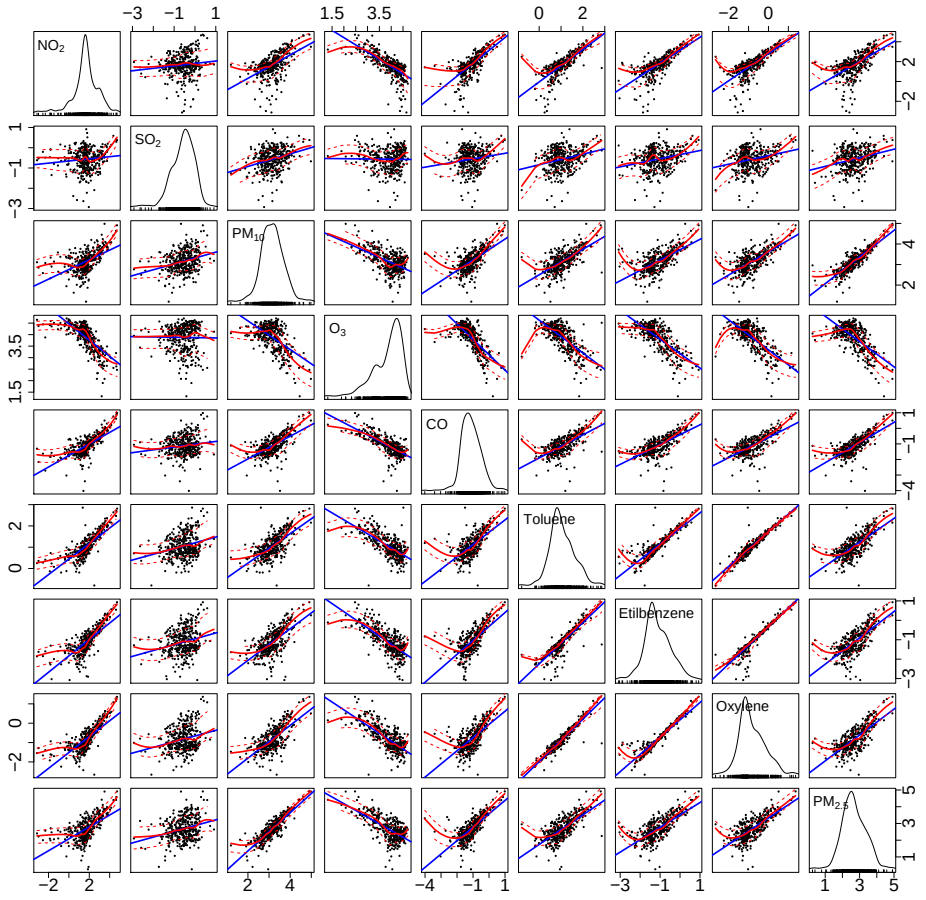
and transition probabilities

$${}_t\pi_{k|j} = \Pr(S_t = k \mid S_{t-1} = j), \quad t = 2, \ldots, T; j, k = 1, \ldots, K.$$

Note that k refers to the current state in the above definitions, whereas j refers to the one previously visited; this convention will be used throughout the paper. Moreover, the initial probabilities are collected in the K-dimensional vector $\boldsymbol{\pi}$, whereas the $K \times K$ transition probability matrix

FIG 3. *Pollutant concentrations: marginal distributions and bivariate correlations*

Marginal densities and scatter plots of log-transformed pollutant concentrations. The blue line results from robust linear regression, and the red line from local polynomial regression. The dashed red lines determine a 90% confidence interval of the smoother. The main diagonal shows marginal densities resulting from a kernel density estimator.



${}_t\mathbf{\Pi}$ contains the time-varying transition probabilities. The simplest model in this framework is the homogeneous HMM, which assumes time-homogeneous transition probabilities, i.e. independence of t and thus ${}_t\mathbf{\Pi} = \mathbf{\Pi}$. This specification fails to take into account how atmospheric observed conditions affect the evolution of unobserved exposure states and, in general, time heterogeneity of the transition probability matrix. In order to overcome this potential drawback, the transitions probabilities may be parametrized as a function of $\tilde{P}$ exogenous covariates $\mathbf{x}_t = \{x_{t1}, \ldots, x_{t\tilde{P}}\}$ by

$$ {}_t\pi_{k|j} = \frac{\exp(\mathbf{x}_t'\boldsymbol{\gamma}_{jk} + \gamma_{jk0})}{1 + \sum_{h=1}^{K} \exp(\mathbf{x}_t'\boldsymbol{\gamma}_{jh} + \gamma_{jh0})} \tag{2.1} $$

where $\boldsymbol{\gamma}_{jk} = \{\gamma_{jk1}, \ldots, \gamma_{jk\tilde{P}}\}$ represents a vector of fixed regressors and γ_{jk0} is an intercept term. To ensure identifiability, we impose $\boldsymbol{\gamma}_{jj} = 0$ and $\gamma_{jj0} = 0$ for $j = 1, ..., K$. Accordingly, the probability of no transition at time t is given by

$$ {}_t\pi_{j|j} = \frac{1}{1 + \sum_{h=1}^{K} \exp(\mathbf{x}_t'\boldsymbol{\gamma}_{jh} + \gamma_{jh0})}. $$

This model specification permits to investigate the dynamics of the hidden state sequence over time, and allows for a potential impact of covariates on its evolution (see also Maruotti and Rocci, 2012).

At each time, the hidden variable S_t corresponds to an unobserved environmental condition. The way in which these hidden conditions affect the corresponding multivariate response variable $\mathbf{Y}_t$ depends on the type of distribution assumed for the observed process. Nowadays, Gaussian HMMs are commonly used for clustering continuous data (see, e.g., Bartolucci and Farcomeni, 2010; Scott et al., 2005), although some robust (conditional) distributions have been recently proposed in the literature (Chatzis et al., 2009; Chatzis, 2010; Farcomeni and Greco, 2015; Punzo and Maruotti, 2016). Since we perform our analysis on log-transformed data, the selection of Gaussian conditional distributions for our HMM seems suitable. In addition, this specification permits to employ standard checks of model fit. Throughout the paper we consider the following conditional distribution of $\mathbf{Y}_t$ given S_t:

$$ \mathbf{Y}_t \mid S_t = k \sim \mathcal{N}_p\left(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right) \tag{2.2} $$

In the context of high dimensional data, parsimonious models should be used in order to reduce the general (heteroscedastic) model, and accordingly the number of parameters to be estimated. As pointed out by Bouveyron

TABLE 1
Covariance structure in a hidden Markov of factor analyzers framework

Model ID	Loading matrix $\mathbf{\Lambda}_k$	Error variance $\mathbf{\Psi}_k$	Isotropic $\mathbf{\Psi}_k = \psi_k \mathbf{I}_p$	Covariance structure	Covariance parameters
CCC	Constrained	Constrained	Constrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}\mathbf{\Lambda}' + \psi\mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}\mathbf{\Lambda}' + \psi_k\mathbf{I}_p$	$[pq - q(q-1)/2] + K$
CUU	Constrained	Unconstrained	Unconstrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}_k$	$[pq - q(q-1)/2] + Kp$
UCC	Unconstrained	Constrained	Constrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \psi\mathbf{I}_p$	$K[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \mathbf{\Psi}$	$K[pq - q(q-1)/2] + p$
UUC	Unconstrained	Constrained	Unconstrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \psi_k\mathbf{I}_p$	$K[pq - q(q-1)/2] + K$
UUU	Unconstrained	Unconstrained	Unconstrained	$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \mathbf{\Psi}_k$	$K[pq - q(q-1)/2] + Kp$

and Brunet (2014), a popular approach bases on mixtures of factor analyzers, mainly because the number of covariance parameters is linear in data-dimensionality. In its most general expression, the following decomposition of the covariance matrix is considered:

$$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \mathbf{\Psi}_k,$$

where $\mathbf{\Lambda}_k$ is a $p \times q$ non-sparse real matrix of state-specific factor loadings and $\mathbf{\Psi}_k = \text{diag}(\psi_{k1}, \ldots, \psi_{kp})$ represents a positive definite matrix containing the error variances.

In other words, an observation in state k follows a multivariate Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \mathbf{\Psi}_k$, i.e.

$$\mathbf{Y}_t \mid S_t = k \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k\mathbf{\Lambda}_k' + \mathbf{\Psi}_k).$$

In a factor analysis framework, the p-dimensional Gaussian random vector of state k can be decomposed in the following additive and independent parts. Formally,

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{\Lambda}_k\mathbf{u}_{tk} + \mathbf{e}_{tk},$$

where $\mathbf{u}_{tk}$ is a q-dimensional vector of state-specific latent factors drawn from $N(0, \mathbf{I}_q)$, and $\mathbf{e}_{tk}$ are Gaussian state-specific error terms with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Psi}_k$. Here and in the following, $\mathbf{I}_q$ denotes the q-dimensional identity matrix. By imposing constraints upon $\mathbf{\Lambda}_k$ and $\mathbf{\Psi}_k$, and on whether or not $\mathbf{\Psi}_k = \psi_k\mathbf{I}_p$ with positive-valued scalar ψ_k, and considering the special case $\psi_k = \psi$ ($\forall\, k = 1, \ldots, K$) as well, we provide a class of eight different parsimonious hidden Markov models (see Table 1).

2.4. *Identifiability.* When dealing with the proposed setting, an important aspect is to establish model identifiability. Identifiability is a necessary requirement, inter alia, for the common asymptotic theory in the context of

maximum likelihood estimation of the model parameters. As typically happens in connection with mixtures of factor analyzers and related models, we focus on local identifiability, i.e., identifiability in a neighborhood of a given parameter value.

For HMMs with finite observational space, generic identifiability of the parameters, i.e., identifiability except on a subset of the parameters space of Lebesgue measure zero, has been investigated by Allman et al. (2009). For general HMMs with state-dependent distributions belonging to some parametric family, identifiability has been proven up to label switching Leroux (1992). Accordingly, in order to ensure identifiability, we need to prove identifiability of the marginal mixtures (see, e.g., Dannmeann et al., 2014). In our case, the marginal finite mixtures are represented by the finite mixtures of Gaussian distributions. A sufficient condition for the identifiability of the mixture of multivariate Gaussian distributions can be summarized as follows (see e.g. Punzo and McNicholas, 2016): let $\|\cdot\|_F$ be the Frobenius norm. A finite mixture of Gaussian distributions is identifiable if $k_1 \neq k_2$ implies

$$\left\|\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_{k_2}\right\|_F^2 + \left\|\boldsymbol{\Sigma}_{k_1} - \boldsymbol{\Sigma}_{k_2}\right\|_F^2 \neq 0.$$

Based on Leroux (1992) and Dannmeann et al. (2014), the same sufficient condition for identifiability is inherited by our Gaussian HMM.

Further identifiability issues are related to the factor model used to parameterize the state-specific covariance matrix. Bearing in mind that unobserved factors are conditional independent given the hidden states, we observe that invariance under orthogonal transformations holds for the proposed model. In fact, the transformation $\mathbf{u}_{tk}^* = \boldsymbol{\Gamma}\mathbf{u}_{tk}$, where $\boldsymbol{\Gamma}$ is an orthogonal matrix leads to

$$Var(\boldsymbol{\Gamma}\mathbf{u}_{tk}) = \boldsymbol{\Gamma}\mathbf{I}_q\boldsymbol{\Gamma}' - \boldsymbol{\Gamma}\boldsymbol{\mu}_k\boldsymbol{\mu}_k'\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}Var(\mathbf{u}_{tk})\boldsymbol{\Gamma} = \mathbf{I}_q$$

since $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}_q$. Therefore, $\mathbf{y}_t = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k\mathbf{u}_{tk} + \mathbf{e}_{tk}$ is completely indistinguishable from the transformed model $\mathbf{y}_t = \boldsymbol{\mu}_k^* + \boldsymbol{\Lambda}_k^*\mathbf{u}_{tk}^* + \mathbf{e}_{tk}$, where $\boldsymbol{\Lambda}_k^* = \boldsymbol{\Gamma}\boldsymbol{\Lambda}_k$ and $\boldsymbol{\mu}_k = \boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_k$. The practical implication of this result is that several known rotations may be applied to the factors in order to improve the model interpretation. We implemented two common rotations, Varimax and Promax. However, we did not observe a major advantage of one or the other in terms of interpretability for our application. Nevertheless, since the factor loadings can be seen as correlations between single variables and factors for the orthogonal rotation, we chose Varimax for all models estimated and results presented in the following. A sufficient condition for local identifiability of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$ is that the Hadamard square of the matrix $\boldsymbol{\Psi}_k - \boldsymbol{\Lambda}_k(\boldsymbol{\Lambda}_k'\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k)^{-1}\boldsymbol{\Lambda}_k'$ is non-singular (Anderson and Rubin, 1956, Theorem 5.9).

3. Maximum likelihood estimation. Even in this relatively general framework, the parameters of the proposed parsimonious HMMs can be estimated using the method of maximum-likelihood. In order to perform maximum likelihood estimation of the above model on the basis of the multivariate response $\mathbf{y}_t = \{y_{t1}, \ldots, y_{tp}\}$, computation of the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\pi}' \mathbf{f}(\mathbf{y}_1)\, {}_2\boldsymbol{\Pi} \mathbf{f}(\mathbf{y}_2)\, {}_3\boldsymbol{\Pi} \ldots \mathbf{f}(\mathbf{y}_{T-1})\, {}_T\boldsymbol{\Pi} \mathbf{f}(\mathbf{y}_T)\mathbf{1} \tag{3.1}$$

is necessary. Here, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k, \boldsymbol{\gamma}_{jk}, \pi_k,\ k = 1, 2, \ldots, K\}$ is the set of all model parameters, $\mathbf{f}(\mathbf{y}_t)$ denotes a diagonal matrix with conditional probability densities $f(\mathbf{Y}_t = \mathbf{y}_t \mid S_t = k; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k)$ on the main diagonal (for further details, see, e.g., Zucchini et al., 2016) and $\mathbf{1}$ represents a unit vector of size K.

To maximize (3.1) with respect to $\boldsymbol{\theta}$, we introduce a three-step AECM based on the following steps:

Step 1. Fit a homogeneous HMM, i.e. without covariates, for the multivariate continuous outcomes. Maximum likelihood estimation is performed by maximizing (3.1) under the constraint ${}_t\boldsymbol{\Pi} = \boldsymbol{\Pi}$ using an AECM algorithm Meng and van Dyk (1997). The motivation beyond the use of the AECM algorithm lies in its ability to break the model into smaller models. On the basis of this preliminary fitting, we obtain the final estimates of the conditional distribution parameters.

Step 2. For each time $t = 1, \ldots, T$, we obtain the posterior expected values of state membership on the basis of the first step.

Step 3. Maximize the component of the (complete-data log-) likelihood involving the hidden structure parameters.

After an initial estimate of the latent parameters, the second and the third steps are iterated until convergence, while keeping fixed the estimates of the conditional distribution parameters from the first step.

3.1. *Likelihood inference and parameters estimation.* At the first step of the algorithm, we partition the set of unknown parameters $\boldsymbol{\theta}$ in two disjoint subsets ($\boldsymbol{\theta}_1$,$\boldsymbol{\theta}_2$): $\boldsymbol{\theta}_1$ contains the hidden chain parameters $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$ and the elements of the state-specific means $\boldsymbol{\mu}_k$, while $\boldsymbol{\theta}_2$ consists of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$. Then, the following steps are alternated until convergence in order to carry out the AECM algorithm:

- First stage

  E-step: compute the conditional expectation of the complete-data log-likelihood, given the observed data and the current estimate

of the parameter vector ($\boldsymbol{\mu}$, $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$), while keeping ($\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$) fixed at their values resulting from the previous iteration.

M-step: maximize the preceding expected complete-data log-likelihood function with respect to ($\boldsymbol{\mu}$, $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$).

- Second stage

E-step: compute the conditional expectation of the complete-data log-likelihood, in this step conditional on ($\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$), while considering ($\boldsymbol{\mu}$, $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$) fixed as given by the calculations in the first stage of the AECM algorithm.

CM-step: maximize the preceding expected complete-data log-likelihood function with respect to ($\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$). The (conditional) maximization step depends on the imposed model restrictions.

To illustrate the principle modus operandi of this algorithm, we need to specify the sources of incompleteness in our setting. The common source arises from the fact that we do not know the state membership and its evolution over time. This source of incompleteness is introduced in the formulation of the model via the definition of the unobserved state membership, i.e., $\mathbf{z}_t = (z_{t1}, z_{t2}, \ldots, z_{tK})$ and $\mathbf{w}_t = (w_{t11}, w_{t12}, \ldots, w_{tjk}, \ldots, w_{tKK})$ as missing data with

$$z_{tk} = \begin{cases} 1 & S_t = k \\ 0 & otherwise \end{cases}$$

$$w_{tjk} = \begin{cases} 1 & S_{t-1} = j, S_t = k \\ 0 & otherwise \end{cases}$$

Then, the complete-data log-likelihood function has the following form

$$\begin{aligned} \mathcal{L}_{c_1}(\mathbf{y}, \mathbf{z}, \mathbf{w}) &= \sum_{k=1}^{K} z_{1k} \log(\pi_k) \\ &+ \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=2}^{T} w_{tjk} \log\left(\pi_{k|j}\right) \\ &+ \sum_{k=1}^{K}\sum_{t=1}^{T} z_{tk}\left[-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left(|\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Psi}_k|\right)\right. \\ &- \left.\frac{1}{2}\mathrm{tr}\left\{(\mathbf{y}_t - \boldsymbol{\mu}_k)(\mathbf{y}_t - \boldsymbol{\mu}_k)'(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Psi}_k)^{-1}\right\}\right] \end{aligned}$$

The expected value of the complete-data log-likelihood is therefore

$$\begin{aligned}\mathcal{H}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^{(m)}\right) &= \sum_{k=1}^{K} \hat{z}_{1k} \log\left(\pi_k\right) \\ &+ \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=2}^{T} \widehat{w}_{tjk} \log\left(\pi_{k|j}\right) \\ &- \frac{1}{2}\sum_{k=1}^{K}\sum_{t=1}^{T} \hat{z}_{tk} \log\left(|\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Psi}_k|\right) \\ &- \frac{1}{2}\sum_{k=1}^{K}\sum_{t=1}^{T} \hat{z}_{tk} \mathrm{tr}\left\{(\tilde{\boldsymbol{\Sigma}}_k(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Psi}_k)^{-1}\right\}\end{aligned}$$

where

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{\sum_{t=1}^{T} \hat{z}_{tk}(\mathbf{y}_t - \boldsymbol{\mu}_k)(\mathbf{y}_t - \boldsymbol{\mu}_k)'}{\sum_{t=1}^{T} \hat{z}_{tk}}$$

is the state-specific sample covariance matrix for all $k = 1, \ldots, K$, and

$$\hat{z}_{tk} = \mathtt{E}(z_{tk} \mid \mathbf{y}_1, \ldots, \mathbf{y}_T),$$

$$\widehat{w}_{tjk} = \mathtt{E}(w_{tjk} \mid \mathbf{y}_1, \ldots, \mathbf{y}_T).$$

At the first-stage M-step, maximizing with respect to $\boldsymbol{\pi}_k$, $\boldsymbol{\Pi}$ and $\boldsymbol{\mu}_k$ yields

$$\pi_k = \hat{z}_{1k}$$

$$\pi_{k|j} = \frac{\sum_{t=2}^{T} \widehat{w}_{tjk}}{\sum_{k=1}^{K}\sum_{t=2}^{T} \widehat{w}_{tjk}}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{t=1}^{T} \hat{z}_{tk}\boldsymbol{y}_t}{\sum_{t=1}^{T} \hat{z}_{tk}}$$

With regards to the maximization with respect to π_k, it is in general not reasonable to try to estimate the initial distribution from just one observation at time 1, especially as the state of the Markov chain itself is not observed. However, it should be noted that $\boldsymbol{\pi}$ is one of the K possible unit vectors at a maximum of the likelihood, and, accordingly, its components do not require all to be estimated. An alternative would be to assume that $\boldsymbol{\pi}$ corresponds to the stationary distribution of the Markov chain (Bulla and Berzel, 2008). This would, after all, complicate the M-step additionally, with only little additional gain in settings with a comparably high number of observations, which is why we did not investigate this approach.

At the second stage of the algorithm, we consider the state labels and the latent factor to be missing data when estimating $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Lambda}_k$. Therefore, here the complete-data log-likelihood is of the form

$$\begin{aligned}\mathcal{L}_{c_2}(\mathbf{y},\mathbf{z},\mathbf{u}) &= C+\sum_{k=1}^{K}\left[-\frac{\sum_{t=1}^{T} z_{tk}}{2}\log\left(|\boldsymbol{\Psi}_k^{-1}|\right)-\frac{\sum_{t=1}^{T} z_{tk}}{2}\mathrm{tr}\left(\boldsymbol{\Psi}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k\right)\right.\\ &+ \left.\sum_{t=1}^{T} z_{tk}(\mathbf{y}_t-\boldsymbol{\mu}_k)'\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k\mathbf{u}_k-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}_k'\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k\sum_{t=1}^{T} z_{tk}\mathbf{u}_k\mathbf{u}_k'\right)\right],\end{aligned}$$

where C is a constant. The expected complete-data log-likelihood is then given by

$$\begin{aligned}\mathcal{H}_2\left(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2^{(m)}\right) &= C+\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\hat{z}_{tk}}{2}\log\left(|\boldsymbol{\Psi}_k^{-1}|\right)\\ &- \sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\hat{z}_{tk}}{2}\mathrm{tr}\left(\boldsymbol{\Psi}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k\right)\\ &+ \sum_{t=1}^{T}\hat{z}_{tk}(\boldsymbol{y}_t-\boldsymbol{\mu}_k)\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k\mathrm{E}(\mathbf{u}_{tk}\mid\mathbf{y}_t,\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k,\boldsymbol{\Psi}_k)\\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Lambda}_k'\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_k\sum_{t=1}^{T}\hat{z}_{tk}\mathrm{E}\left(\mathbf{u}_{tk}\mathbf{u}_{tk}'\mid\mathbf{y}_t,\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k,\boldsymbol{\Psi}_k\right)\right\},\end{aligned}$$

where the $\hat{z}_{tk}$ are computed by considering the updated values of $\boldsymbol{\mu}_k$ $\boldsymbol{\pi}_k$ and $\boldsymbol{\Pi}_k$. The resulting estimates of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$ matrices can be easily derived from the expression for $\mathcal{H}_2(\cdot)$. In the most general case UUU, we have

$$\begin{aligned}\boldsymbol{\Lambda}_k^{new} &= \tilde{\boldsymbol{\Sigma}}_k\boldsymbol{\beta}_k'\boldsymbol{\Delta}_k^{-1},\\ \Psi_k^{new} &= \mathrm{diag}\left(\tilde{\boldsymbol{\Sigma}}_k-\boldsymbol{\Lambda}_k^{new}\boldsymbol{\beta}_k\tilde{\boldsymbol{\Sigma}}_k\right),\end{aligned}$$

where

$$\boldsymbol{\beta}_k=\boldsymbol{\Lambda}_k'\left(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k'+\boldsymbol{\Psi}_k\right)^{-1},\quad \boldsymbol{\Delta}_k=\mathbf{I}_q-\boldsymbol{\beta}_k\boldsymbol{\Lambda}_k+\boldsymbol{\beta}_k\tilde{\boldsymbol{\Sigma}}_k\boldsymbol{\beta}_k'.$$

Estimates for other models in Table 1 can be easily obtained by imposing that $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$, and/or $\tilde{\boldsymbol{\Sigma}}$ are state-invariant.

Once the AECM achieves convergence at the first step, we obtain $\widehat{w}_{tjk}$ and use these to get estimates of ${}_t\pi_{k|j}$. The estimated parameters for the hidden process are the solutions of

$$\sum_{t=2}^{T}\sum_{k=1}^{K}\widehat{w}_{tjk}\frac{\partial_t\pi_{k|j}}{\partial(\boldsymbol{\gamma}_{jk},\gamma_{jk0})}=0 \tag{3.2}$$

which are weighted sums of K multinomial regressions with weights $\widehat{w}_{tjk}$. We then update $\widehat{w}_{tjk}$ and iterate Step 2 and Step 3, plugging in the estimated transition probabilities into the log-likelihood function, till further convergence.

3.2. *Computational details.* The quantities $\hat{z}_{tk}$ and $\widehat{w}_{tjk}$ can be computed recursively (Baum et al., 1970, Welch, 2003). Let us define the forward variable

$$\alpha_{tk} = f(\mathbf{y}_1, \ldots, \mathbf{y}_t, S_t = k),$$

which represents the probability of seeing the partial sequence ending up in state k at time t, and the corresponding backward variable

$$\beta_{tk} = f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T \mid S_t = k).$$

It is worth noting that the computation of the forward and backward probabilities is susceptible to errors resulting from numerical under- or overflow. In applying the algorithm as described here, a scaling procedure is adopted in order to prevent, or at least reduce, the risk of such errors. For convenience, we work on the log-scale; the forward recursion is then given by

$$\log(\alpha_{1k}) = \log\left[f\left(\mathbf{y}_1 \mid S_1 = k\right)\right] + \log(\pi_k),$$

and for $t = 2, \ldots, T$ we compute

$$\log(\alpha_{tk}) = \log\left[f\left(\mathbf{y}_t \mid S_t = k\right)\right] + \sum_{j=1}^{K} \log(\alpha_{t-1,j}) + \log({}_t\pi_{k|j}).$$

Similarly, it is possible to implement the following backward recursion:

$$\log(\beta_{Tk}) = 0,$$

and for $t = T-1, \ldots, 1$ we have

$$\log(\beta_{tj}) = \sum_{k=1}^{K} \log\left[f\left(\mathbf{y}_{t+1} \mid S_{t+1} = k\right)\right] + \log(\beta_{t+1,k}) + \log({}_t\pi_{k|j}).$$

Then, the expected values of the quantities involved in the E-step can be computed as

$$\hat{z}_{tk} = \frac{\alpha_{tk}\beta_{tk}}{\sum_{k=1}^{K} \alpha_{tk}\beta_{tk}}$$

and

$$\widehat{w}_{tjk} = \frac{{}_t\pi_{k|j}\alpha_{t-1,k} f(\mathbf{y}_t \mid S_t = k)\beta_{tk}}{\sum_{k=1}^{K} \alpha_{Tk}}.$$

3.3. *Path prediction.* A major issue of interest in a HMM framework is the prediction of the hidden state sequence, which can be determined by maximizing the posterior probability $\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T \mid \mathbf{y}_1, \ldots, \mathbf{y}_T)$ with respect to $(s_1, s_2, \ldots, s_T)$. To avoid inconsistent sequences and to account for the joint probability of the entire latent sequence, the Viterbi algorithm (Viterbi, 1967) is the most suitable approach for this prediction, a procedure often also termed 'global decoding'. With $\rho_t(\mathbf{s}) = \max_{s_1,\ldots,s_T} \Pr(s_1, \ldots, s_{t-1}, s_t, \mathbf{y}_1, \ldots, \mathbf{y}_t)$, the algorithm performs the following steps:

1. Compute $\rho_1(k) = \pi_k f(\mathbf{y} \mid S_1 = k)$ for $k \in \{1, 2, \ldots, K\}$.
2. Calculate $\rho_t(k) = f(\mathbf{y}_t \mid S_{t+1} = k) \max_j [\rho_t(j)_t \pi_{k|j}]$ for $t = 2, \ldots, T$, $k = 1, \ldots, K$.
3. Find the optimal $\tilde{s}_T = \arg\max_k \rho_T(k)$.
4. Determine $\tilde{s}_t$ by $\tilde{s}_t = \arg\max_j \rho_t(j)_{t+1} \pi_{\tilde{s}_{t+1}|j}$ for $t = T-1, T-2, \ldots, 1$.

In other words, the algorithm performs a forward recursion to compute the above quantities, and subsequently finds the most likely latent sequence with a backward recursion. All of the above quantities are computed on the basis of the maximum likelihood parameter estimates.

4. Simulation study. In this section, we illustrate and discuss the results of a simulation study aimed at assessing the properties of the maximum likelihood estimator outlined above. Various settings are investigated, such as different lengths of time T, numbers of analyzed variables p, hidden states K, and latent factors q. We treat only the UUU case in the following, since the others can be seen as special cases. We obtained maximum likelihood estimates by following the three-step AECM algorithm described in Section 3.1. To initialize state memberships, we randomly generate state assignments from a multinomial distribution. The hidden chain parameters are initialized according to the obtained partition. Similarly, all $\tilde{\Sigma}_k$ are computed based on the initial partition. Then, we obtain the initial values for the elements of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$ following the procedure described by McNicholas and Murphy (2008).

4.1. *Simulation design.* For each of the scenarios described below, we repeated the procedure $B = 100$ times, and averaged the resulting measures of performance over the replications. We considered the following experimental designs:

- number of observed times $T = 100, 365, 1000$;
- number of states $K = 2, 3$;

- number of latent factors $q = 2, 4$;
- number of observed variables $p = 10, 50, 100$;
- hidden chain parameters and state-specific means defined as follows
  - $K = 2$: we generate transition probabilities from the model

$$ {}_t\pi_{k|j} = \frac{\exp(x'_{it}\gamma_{jk} + \gamma_{jk0})}{1 + \sum_{h=1:h\neq j}^{2} \exp(x'_{it}\gamma_{jh} + \gamma_{jh0})}, $$

    where x_{it} is a covariate independently drawn from a $N(0, 1)$ distribution, $\gamma_{12} = -0.75$, $\gamma_{21} = -0.25$, and $\gamma_{11} = \gamma_{22} = 0$ due to identification constraints. The intercepts take the values $\gamma_{120} = -1$ and $\gamma_{210} = -0.7$, and we set $\pi_1 = 1$. Then,
    - for $p = 10$ we further specify the state-specific mean vector by

$$ \boldsymbol{\mu}_1 = \{-2, 0, 1, 0.5, -1, 2, 0, -1, 0, -1\} $$
$$ \boldsymbol{\mu}_2 = \{2, 0, -1, 0, -1, 0, 0, 0, -0.4, 1.5\} $$

    - for $p = 50, 100$ we draw the state-specific mean vectors from uniform distributions defined by

$$ \boldsymbol{\mu}_1 \sim U(-1, 2) $$
$$ \boldsymbol{\mu}_2 \sim U(-2, 0) $$

  - $K = 3$: we generate transitions probabilities from the following set of parameters

$$ [\gamma_{110}, \gamma_{120}, \gamma_{130}] = [0.00; -0.50; -0.25] $$
$$ [\gamma_{210}, \gamma_{220}, \gamma_{230}] = [-0.70; 0.00; -0.50] $$
$$ [\gamma_{310}, \gamma_{320}, \gamma_{330}] = [-0.70; -0.20; 0.00] $$
$$ [\gamma_{11}, \gamma_{12}, \gamma_{13}] = [0.00; -0.75; -0.20] $$
$$ [\gamma_{21}, \gamma_{22}, \gamma_{23}] = [-0.25; 0.00; -0.25] $$
$$ [\gamma_{31}, \gamma_{32}, \gamma_{33}] = [-0.75; -0.20; 0.00] $$

    and

    - for $p = 10$ we define the state-specific mean vector by

$$ \boldsymbol{\mu}_3 = \{0, 2, -1, .5, 0, -2, 0, 1, 0.4, 0\} $$

∗ for $p = 50, 100$ the stated-specific mean vectors result from

$$\boldsymbol{\mu}_1 \sim U(-1, 2)$$
$$\boldsymbol{\mu}_2 \sim U(-2, 0)$$

as before.

At last, we generated the state-specific factor loading matrices $\boldsymbol{\Lambda}_k$ randomly from a uniform distribution on the interval $[-1, 1]$, and a reasonable level of noise was added by generating $\boldsymbol{\Psi}_k = \text{diag}(\text{Unif}(0, 1), p)$. In total, we generated 1800 data sets. Figures 4 and 5 show the scatter plot matrices for samples drawn for the settings commonly characterized by $p = 10$, $q = 2$, $T = 365$, but with varying $K = 2, 3$.

We analyzed the performance of the proposed approach, obtained when employing the three-step AECM algorithm, in terms of mean and standard deviation values of

- the sum of the squared differences between the true and the estimated state-specific means

$$S_{obs} = \sum_{k=1}^{K} \|\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k\|_2^2$$

  and the sum of the squared distances between the true and the estimated initial and transition probabilities, respectively,

$$S_{hidden} = \|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|_2^2 + \sum_{t=1}^{T} \|_t\boldsymbol{\Pi} - {}_t\hat{\boldsymbol{\Pi}}\|_F^2,$$

  where $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the Euclidean norm and Frobenius norm, respectively. Both indices serve for evaluating the accuracy of the estimators.
- the Adjusted Rand index (ARand, see Hubert and Arabie, 1985), which is a measure of agreement between the true and the estimated cluster memberships. It takes the value one in case of perfect agreement.

We halted the estimation process and assumed convergence to the maximum when the log-likelihood obtained from two successive iteration steps increased by less than 10^{-5}.

FIG 4. *Example for a simulation setting with* $p = 10$, $K = 2$, $q = 2$, *and* $T = 365$

Marginal densities and scatter plots of a simulated data set. Each color identifies observations belonging to a different state. The main diagonal shows marginal densities resulting from a kernel density estimator.

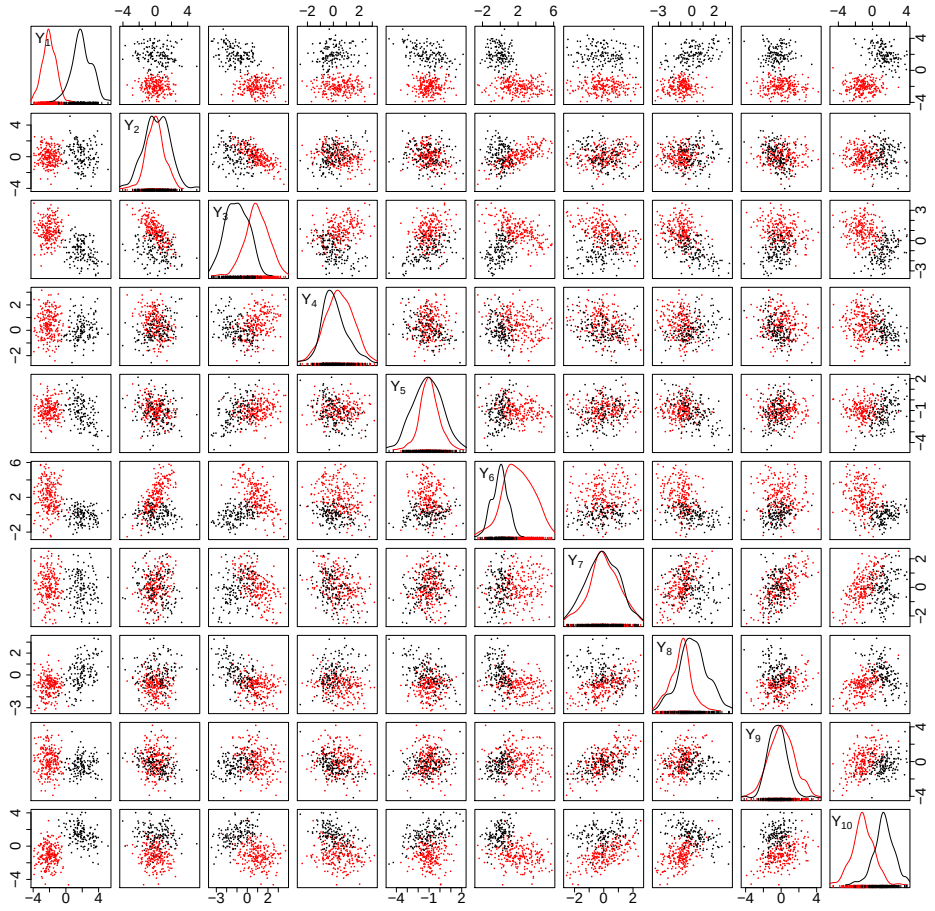


FIG 5. *Example for a simulation setting with* $p = 10$, $K = 3$, $q = 2$, *and* $T = 365$

Marginal densities and scatter plots of a simulated data set. Each color identifies observations belonging to a different state. The main diagonal shows marginal densities resulting from a kernel density estimator.

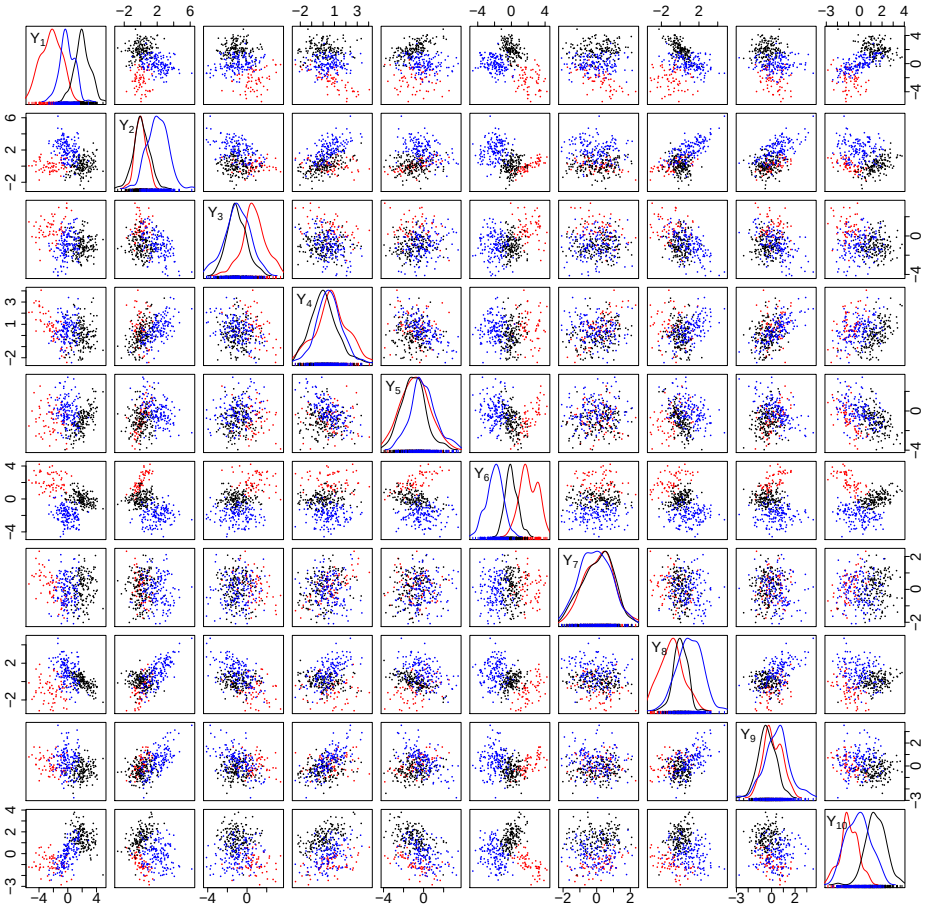


TABLE 2

Mean and standard deviation (in brackets) values of S_{obs}, S_{hidden}, and the Adjusted Rand index (ARand) obtained by the AECM algorithms over 100 samples in the different settings investigated.

	$p = 10, q = 2, K = 2$				$p = 10, q = 2, K = 3$		
	$T = 100$	$T = 365$	$T = 1000$		$T = 100$	$T = 365$	$T = 1000$
S_{obs}	0.106 (0.246)	0.019 (0.011)	0.007 (0.003)	S_{obs}	0.225 (0.632)	0.021 (0.011)	0.007 (0.004)
S_{hidden}	1.746 (2.576)	0.311 (0.294)	0.120 (0.096)	S_{hidden}	12.471 (17.056)	1.919 (1.124)	1.106 (0.377)
ARand	0.966 (0.148)	0.999 (0.003)	0.999 (0.002)	ARand	0.903 (0.158)	0.990 (0.009)	0.992 (0.005)
	$p = 50, q = 2, K = 2$				$p = 50, q = 2, K = 3$		
	$T = 100$	$T = 365$	$T = 1000$		$T = 100$	$T = 365$	$T = 1000$
S_{obs}	0.101 (0.150)	0.014 (0.006)	0.005 (0.002)	S_{obs}	0.934 (1.202)	0.044 (0.007)	0.008 (0.009)
S_{hidden}	2.307 (2.812)	0.346 (0.329)	0.114 (0.102)	S_{hidden}	13.510 (45.523)	1.960 (0.977)	1.098 (0.998)
ARand	0.916 (0.223)	0.981 (0.032)	0.998 (0.002)	ARand	0.734 (0.195)	0.972 (0.072)	0.998 (0.019)
	$p = 100, q = 2, K = 2$				$p = 100, q = 2, K = 3$		
	$T = 100$	$T = 365$	$T = 1000$		$T = 100$	$T = 365$	$T = 1000$
S_{obs}	0.230 (0.294)	0.086 (0.315)	0.022 (0.129)	S_{obs}	2.273 (1.850)	0.610 (1.555)	0.108 (0.276)
S_{hidden}	2.880 (2.816)	0.479 (0.508)	0.191 (0.238)	S_{hidden}	22.451 (75.743)	2.019 (0.982)	1.078 (0.995)
ARand	0.743 (0.329)	0.937 (0.221)	0.983 (0.123)	ARand	0.681 (0.200)	0.957 (0.118)	0.998 (0.019)

4.2. *Simulation results.* Table 2 presents results concerning the accuracy of the estimates (S_{obs} and S_{hidden}), as well as the ability of recovering the true partition (ARand). For all settings, the length of time T influences the accuracy of the estimates and the goodness of classification: when T increases, the average values of S_{obs}/S_{hidden} and the Adjusted Rand Index decrease and increase, respectively (keeping p, K, and q fixed). Considering only small values of T ($T = 100$), the algorithms better recover the partition in settings with simpler data structure and lower number of variables. To a much minor degree, the same holds true for $T = 365$. On the other hand, the recovery of the true partition remains stable at high values close to one over all settings for $T = 1000$. Overall, the results obtained indicate a satisfactory performance of the AECM algorithm in the simulated settings. In addition, one may observe relatively high values for the standard deviation of almost all quantities estimated in all settings. This is most likely a consequence from a relatively strong influence of choice for initial values.

5. Results. We fitted all eight models (see Table 1) to the data for $K = 2, \ldots, 4$ and $q = 1, \ldots, 3$, and computed the Akaike Information Criterion (AIC), the Bayesian Information Criterion, and Integrated Completed Likelihood (ICL) values for each model. The ICL essentially penalizes the

BIC for estimated mean entropy, and it is given given by

$$ICL = BIC + \sum_{t=1}^{T}\sum_{k=1}^{K} \hat{z}_{tk} \log \hat{z}_{tk}.$$

All criteria involve penalization terms depending on the number of redundant parameters, which is equal to the sum of the following:

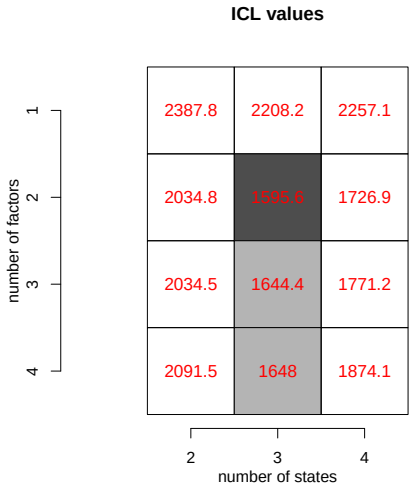
- $(K-1)$ initial probabilities,
- $K(K-1)(\tilde{P}+1)$, corresponding to the number of regression parameters $(\boldsymbol{\gamma}_{jk}, j, k = 1, \ldots, K, j \neq k)$ needed to compute transition probabilities collected in ${}_t\boldsymbol{\Pi}$,
- KP state-specific means, collected in $\boldsymbol{\mu}_k, k = 1, \ldots, K$,
- covariance parameters, as defined in Table 1.

Figure 6 shows the minimum values over the eight models of each pair (q, K) for the ICL. We would like to remark that both BIC and ICL select the UUU model with $K = 3$ and $q = 2$, whilst AIC selects the UUU model with $K = 4$ and $q = 4$. We would further remark that the non-homogeneous HMMs lead to significant improvements in model fitting, as measured by log-likelihoods and penalized criteria, with respect to their homogeneous counterparts. Model selection results for the homogeneous HMMs are not shown for the sake of brevity, but are available upon request.

Since our main intention lies in the detection of meaningful clusters that characterize particularly different exposure regimes, we choose to work with the model selected by ICL (and BIC). One should, however, be taken into consideration that a minor drawback of this rather restrictive model selection could be a reduced fit to the density, in particular in the tails. In order to investigate this potential shortcoming, Figure 7 displays the Q-Q plots of the pseudo-residuals (see Zucchini et al., 2016, Chapter 6.2) for each pollutant belonging to the observed process. The plots show, on the one hand, an overall satisfactory fit for most pollutants, with NO_2 and O_3 exhibiting the largest deviations from normality due to reduced fit in the tails and skewness, respectively. This could be improved, e.g. by a less parsimonious model or generalizations of the Gaussian distribution. Furthermore, the model selected by AIC constitutes a straightforward alternative if the focus lies on a more appropriate density approximation. On the other hand, by considering the ICL-based solution we obtain a more certain and less fuzzy classification. It is possible to quantify the uncertainty surrounding the obtained classification by looking at histograms or rootograms of the posterior state probabilities. In fact, both are suitable to usually assess the hidden structure. As a more informative alternative, a scatter plot of the

FIG 6. *Heat map of model selection criteria*

Heat maps of ICL values for each (K, q). For fixed (K, q), the minimum value is determined from all eight models.



state-specific posterior probabilities may be used: a good classification results in the distribution to be concentrated near the vertices (0,0), (0,1) and (1,0). In our simple example the states are well separated (see Figure 8), the obtained classification is therefore close to the so-called 'hard case'.

To properly provide a physical meaning of the hidden states, we inferred the latent process by applying the Viterbi algorithm. Figure 9 shows the resulting most likely sequence of hidden states and suggests that the inferred states can be interpreted as latent exposures strongly characterized by seasonal weather conditions. More specifically, State 1 tends to cluster days in late autumn, winter, and early spring, while State 3 is generally inferred during mid- to late summer, but also occurs sporadically during autumn. Finally, State 2 mainly arises during late spring/early summer, however, it can also be observed during various periods of the rest of the year. Table 3 displays the estimated effects of the standardized meteorological variables on the transition probabilities and their standard errors. The estimation procedure outlined in Section 3.1 does not produce standard errors of the estimates. Therefore, following the suggestion of Visser et al. (2000), we implemented a parametric bootstrap approach to obtain the standard errors: we re-fitted the model to 200 bootstrap samples, which were simulated from

FIG 7. *Normality plots*

Q-Q plots of the pseudo residuals for each pollutant.

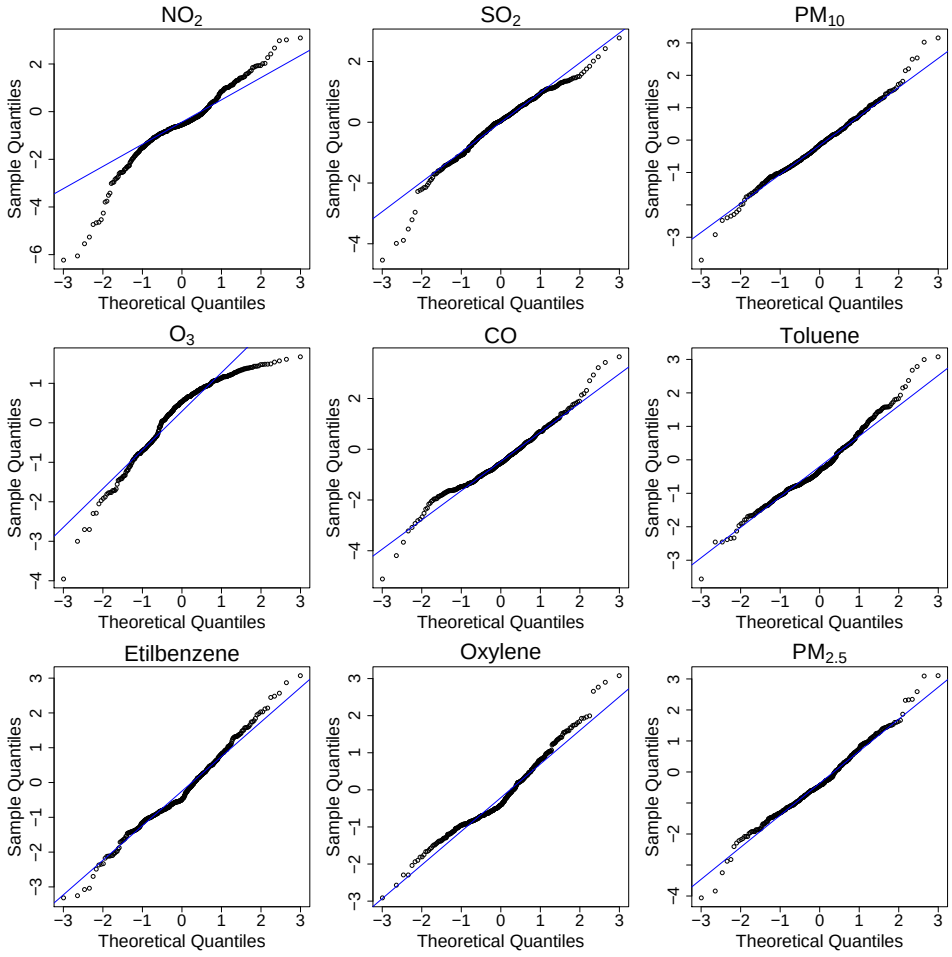
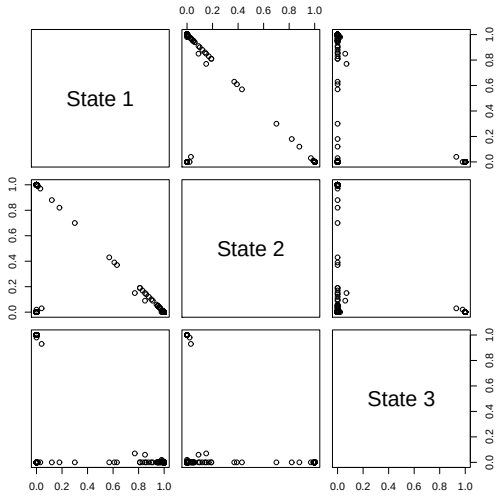


FIG 8. *Uncertainty surrounding the classification*

State-specific posterior probabilities for the three states model with two factors, selected by the BIC (and ICL) criterion.



the estimated model parameters. On average, convergence was achieved in 5.04 seconds (s.d. 5.87) on average, and required 418 iterations (s.d. 177).

The negative signs of the estimated intercepts confirm that, overall, state persistence is the norm. Furthermore, comparison of these values indicates that most of the transitions can be expected to happen from State 1 to State 2 and from State 3 to State 1. Taking the effect of covariates into account, one can notice that transitions from State 1 to State 2 are more likely when the temperature increases. The contrary holds for reverse transitions (from State 2 to State 1), which can be associated with low temperature values, combined with low pressure and high values of wind speed and temperature excursion. Furthermore, transitions from State 1 to State 3 are more likely when the average temperature increases and temperature excursion decreases, while the likelihood of a reverse transition increases during autumn, when small temperature excursions are accompanied by low wind speeds and average temperatures. The remaining transitions between State 2 and State 3 can be mainly associated with typical summer days with low pressure and large temperature excursions.

Table 4 displays the estimated, state-specific parameters (and their standard errors), obtained under a UUU model with $K = 3$ states and $q = 2$ fac-

FIG 9. *Hidden state sequence*
The most likely sequence of states inferred by the Viterbi algorithm.

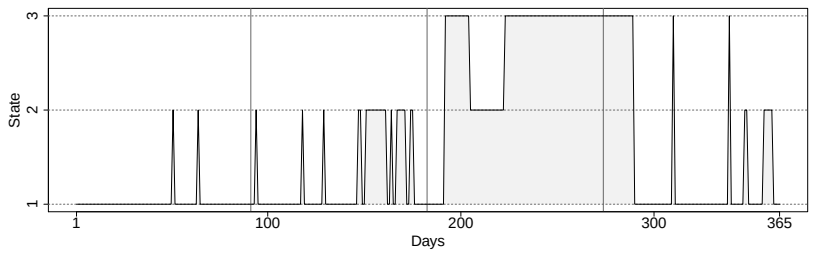


TABLE 3
Estimated regression coefficients and standard errors (within parentheses).

	γ_{12}	γ_{13}	γ_{21}	γ_{23}	γ_{31}	γ_{32}
Intercept (γ_{jk0})	-2.71	-5.92	-0.73	-6.90	-2.13	-8.59
	(0.24)	(1.08)	(0.14)	(1.17)	(0.38)	(2.31)
Wind Speed	0.28	0.31	0.63	-2.65	-1.21	1.26
	(0.21)	(0.38)	(0.15)	(0.77)	(0.26)	(0.83)
Average Temperature	0.82	3.09	-1.75	-2.24	-4.18	-1.05
	(0.33)	(1.02)	(0.23)	(0.54)	(0.56)	(0.94)
Pressure	-0.05	-0.48	-0.34	-1.23	0.91	-3.46
	(0.26)	(0.54)	(0.14)	(0.45)	(0.25)	(1.07)
Temperature Excursion	-0.36	-2.41	1.33	6.48	-0.65	2.35
	(0.33)	(0.83)	(0.23)	(1.13)	(0.30)	(1.20)

TABLE 4
Estimated parameters and standard errors (within parentheses)

	Variable	Mean		Factor 1		Factor 2		Residual Variance	
State 1	NO_2	2.12	(0.38)	0.84	(0.18)	0.43	(0.35)	0.12	(0.16)
	SO_2	-0.54	(0.13)	0.15	(0.22)	0.20	(0.27)	0.94	(0.19)
	PM_{10}	3.24	(0.16)	0.41	(0.24)	0.81	(0.33)	0.17	(0.21)
	O_3	3.70	(0.05)	-0.66	(0.21)	-0.50	(0.35)	0.32	(0.14)
	CO	-0.88	(0.23)	0.63	(0.32)	0.63	(0.26)	0.20	(0.26)
	Toluene	1.16	(0.18)	0.87	(0.31)	0.47	(0.33)	0.02	(0.13)
	Ethylbenzene	-0.95	(0.19)	0.88	(0.14)	0.47	(0.21)	0.01	(0.13)
	Oxylene	-0.70	(0.19)	0.89	(0.20)	0.46	(0.26)	0.01	(0.11)
	$PM_{2.5}$	2.90	(0.29)	0.43	(0.23)	0.90	(0.32)	0.01	(0.21)
State 2	NO_2	1.36	(0.30)	0.65	(0.18)	0.37	(0.34)	0.44	(0.15)
	SO_2	-0.71	(0.13)	0.02	(0.21)	0.50	(0.22)	0.75	(0.17)
	PM_{10}	2.98	(0.04)	0.62	(0.27)	0.18	(0.37)	0.59	(0.23)
	O_3	4.07	(0.09)	-0.84	(0.28)	0.08	(0.33)	0.29	(0.18)
	CO	-1.36	(0.14)	0.97	(0.42)	0.20	(0.22)	0.01	(0.43)
	Toluene	0.88	(0.05)	0.21	(0.35)	0.93	(0.33)	0.09	(0.11)
	Ethylbenzene	-1.72	(0.13)	0.67	(0.13)	0.49	(0.16)	0.31	(0.15)
	Oxylene	-1.18	(0.07)	0.47	(0.23)	0.75	(0.24)	0.21	(0.11)
	$PM_{2.5}$	2.27	(0.06)	0.59	(0.23)	0.23	(0.32)	0.60	(0.23)
State 3	NO_2	0.63	(0.19)	0.84	(0.17)	-0.11	(0.33)	0.27	(0.10)
	SO_2	-0.56	(0.11)	-0.03	(0.16)	0.48	(0.20)	0.77	(0.13)
	PM_{10}	3.03	(0.08)	0.17	(0.22)	0.90	(0.32)	0.16	(0.14)
	O_3	4.26	(0.19)	-0.54	(0.20)	0.44	(0.35)	0.51	(0.15)
	CO	-1.68	(0.08)	0.05	(0.35)	0.16	(0.20)	0.97	(0.35)
	Toluene	0.85	(0.08)	0.93	(0.22)	0.32	(0.25)	0.04	(0.06)
	Ethylbenzene	-1.21	(0.16)	0.94	(0.11)	0.32	(0.16)	0.01	(0.08)
	Oxylene	-0.95	(0.10)	0.93	(0.14)	0.35	(0.19)	0.02	(0.07)
	$PM_{2.5}$	2.31	(0.12)	0.19	(0.21)	0.90	(0.25)	0.16	(0.14)

tors. Conditionally on each hidden state, these estimates can be interpreted as if they were the output of a factor model, which is a distinct advantage of our approach. State-specific means allow to better characterize air conditions observed during different times of the year. High pollution is estimated in State 1, where all the pollutants but the O_3 achieve their highest values. This is not surprising as State 1 is observed during winter and autumn, mainly, and when the weather gets cold enough, pollution from vehicles and homes becomes a very visible mist in the air. State 3 is characterized by high values of O_3, while all the other pollutants do not show particular behaviors. At last, State 2 is the one with the best air conditions. Low values for both primary and secondary pollutants are observed. Columns three and four of the table indicate the estimated factor loadings for each state, i.e. the conditional correlation between each variable and the factor. Furthermore, factors flexibly accommodate different correlation structures within states. Consequently, the interpretation of each factor is state-dependent, which is an additional advantage of our approach. By focusing on the loadings that are significant at a 95% level, the first factor is positively associated with nitrogen dioxide (NO_2) and aromatic compounds (Toluene, Ethylbenzene and Oxylene) in State 1. These pollutants form mostly in cases of incomplete combustion, occurring when fuels are burned at high temperatures. In urban areas, incomplete combustion is typically associated with transportation vehicles and heating devices. Conditionally on State 1, Factor 2 is positively associated with harmful secondary air pollutants, such as particulate matter in both fine ($PM_{2.5}$) and coarse form (PM_{10}), which are mainly produced when precursor gases (captured by Factor 1) condense in the atmosphere. A similar factorial interpretation holds under State 3, where Factors 1 and 2 are positively associated with primary and secondary pollutants, respectively. However, in State 1 the second factor is negatively associated with Ozone, while this correlation is positive in State 3. This pattern can be expected, because the low temperature levels in winter and autumn do not allow for the formation of this secondary pollutant, which is instead typically produced during summer. In State 2, Factor 1 is positively correlated with carbon monoxide (CO), and Factor 2 with some aromatic compounds, hence capturing a specific pattern of pollution emissions that occur during spring.

The communality of each variable for each state can be obtained as the complement of estimated residual variance (last column of Table 4). It indicates the proportion of variability captured by the factorial space in each state. The heterogeneity of these communalities across states (Figure 10) reflects a data correlation structure that significantly varies with the states.

FIG 10. *Communalities*
Box plots of the communalities for each state.

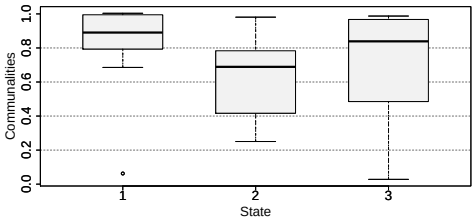


TABLE 5
Proportion of total variance associated with each factor.

State	Factor 1	Factor 2	Residual variance
1	0.47	0.33	0.20
2	0.39	0.24	0.37
3	0.41	0.27	0.32

For State 1 and 3, high correlation values allow most of the data variability to be represented in a two-dimensional factorial space. However, in State 2 lower correlations are responsible for a larger information loss due to the dimensional reduction. These results are confirmed by Table 5, which displays the proportion of total data variability explained by each factor, thus providing a measure of the importance of this factor in determining the variability of the data for each state.

6. Discussion. The assessment of pollution exposure is not only complicated by the complex structure of cross-correlations between multiple pollutants, but also by the non-stationary evolution of these correlations in time. In this paper we propose to assess pollution exposure via a dynamic method of dimension reduction, the core of which consists in the extension of mixtures of factor analyzers to a temporal setting. More specifically, we assume that the joint distribution of a multivariate time series of pollutant concentrations can be approximated by a dynamic mixture of lower-dimensional Gaussian densities in the factor space, and that the mixture parameters evolve according to a latent non-homogeneous Markov chain.

Taking a factor analysis approach for dimensionality reduction has advantages, but limitations as well. On the one hand, the requirement of mul-

tivariate normality presents a major limitation. In our case study, this issue was successfully addressed by taking the logarithm of the observed concentrations. However, different applications may well require less obvious transformations. On the other hand, the linear structure of factor analyzers simplifies the interpretation of the low-dimensional Gaussian densities as regimes of air quality: the factor loadings indicate which pollutants mainly contribute to each regime, while the residual variances describe the quality of the a dimensionality reduction within each regime. Furthermore, a factor analysis approach provides a simple framework for parameter constraints, which can be specified via structured covariance matrices. From a practical viewpoint, these different structures are not only important for tuning the flexibility of dimensionality reduction. They also allow the model to be fitted in high-dimensional settings, because the number of covariance parameters is linear in data dimensionality. Therefore, our approach is potentially well suited to the analysis of high-dimensional data. While our application is motivated by questions arising ecological analyzes, it can be easily adapted to a wide range of real-world cases (e.g., financial time series).

We use a non-homogeneous Markov chain to capture time-varying interactions between pollution and weather conditions, while accounting for temporal dependence simultaneously. In our case study, this model adequately represented the alternating periods of flat stretches and sudden bursts of pollution data. However, other case studies might require more complex models that relax the Markovian assumption, which results in a geometric distribution of the dwell times as implicit consequence. This limitation could, e.g., be overcome by considering a hidden semi-Markov chain, which allows for more general dwell time distributions.

We propose an algorithm which allows to estimate the models considered in a stable manner, which is also a direct consequence of assuming constraints on the covariance structure. The proposed algorithm bases on the AECM algorithm, introducing a further estimation step to estimate the effect of covariate on transitions between hidden states. The resulting method is a valid alternative to the full maximum likelihood approach, which is typically based on the AECM algorithm only, when the latter becomes difficult to use due to the large number of response variables and latent states. The further step introduced in the estimation process modifies the algorithm in a suitable manner in order to obtain more precise estimates of the latent parameters. This last step is iterated until convergence, while the estimates of the conditional distributions are kept fixed. An aspect that seems worth investigating in the future might be the dependence of the algorithm on 'good' initial values, i.e., values leading to a convergence to the global max-

imum of the log-likelihood. While carrying out additional simulation runs, we noticed that the algorithm seems robust to poor initial values, but it could be of interest to investigate different initialization strategies to speed up the convergence of the algorithm.

We assumed that the number of factors is constant across latent states. This simplifies model selection procedures on the one hand, but also imposes limitations in terms of model flexibility on the other hand. An alternative could be a model specification allowing for state-varying numbers of factors. Such an approach does, however, not necessarily guarantee a simple interpretation of the results and is certainly more demanding from a computational viewpoint than our model. Although not explicitly shown, we noted that a model with state-varying number of factors may be implemented by extending our algorithm, taking into consideration few (relevant) adjustments.

Our approach could be further extended to directly deal and identify atypical observations (outliers, spurious points or noise), which may affect the parameter estimates. However, despite of the wide literature on robust estimation of mixture models, the development of robust approaches in a HMM framework is still in its infancy, and algorithms related to possible model extensions in this direction could be even more cumbersome.

Last, the assessment of pollution is in this paper based on the analysis of a multivariate time series. This approach is sensible in small urban areas, such as the one considered in our case study. However, the spatial dimension of pollution cannot be ignored when the study involves larger areas. In principle, an extension of our approach to the spatial setting is possible by considering mixtures of factor analyzers with parameters that vary across space according to a latent Markov random field, hence obtaining a hidden Markov random field. Markov random fields, also known as Gibbs fields, are spatial multinomial processes that extend the Markov property to the spatial setting and naturally provide the spatial counterpart of Markov chains. However, the distribution of a Markov field is typically only known up to a computationally intractable normalizing constant, which complicates the estimation step. Further research is therefore needed to extend our approach to the spatial setting.

7. Acknowledgments. The work of J. Bulla and A. Maruotti has been partly supported by The Finance Market Fund (project number 261570). The work of F. Lagona is developed under the PRIN2015 supported-project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)', funded by MIUR (Italian Ministry of Education, University and Scientific Research).

References.

Allman, E.S., Matias, C. and Rhodes, J.A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37: 3099–3132.

Anderson, T.W. and Rubin, H. (1956). Statistical interference in factor analysis, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed., Univ. of California Press, 5: 111–150.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, 104: 816–831.

Bartolucci, F. and Farcomeni, A. (2010). A note on the mixture transition distribution and hidden Markov models. *Journal of Time Series Analysis*, 31: 132–138.

Bartolucci, F., Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Boca-Raton, FL: CRC Press.

Bartolucci, F., Montanari, G. E. and Pandolfi, S. (2015). Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis*, 83: 287–301.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41: 164–171.

Bornn, L., Shaddick, G. and Zidek, J.V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107: 281–289.

Bouveyron, C. and Brunet, C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71: 52–78.

Bulla, J. Berzel, A. (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23: 1–18.

Bulla, J., Lagona, F., Maruotti, A. and Picone, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological and Environmental Statistics*, 17: 544–567.

Chatzis, S.P., Kosmopoulos, D.I. and Varvarigou, T.A. (2009). Robust sequential data modelling using an outlier tolerant hidden Markov model. *IEEE Transactions on Pattern Analysis and Machine Learning*, 31: 1657–1669.

Chatzis, S.P. (2010). Hidden Markov models with nonelliptically contoured state densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32: 2297–2304.

Chattopadhyay, A.K., Mondal, S. and Biswas, A. (2015). Independent component analysis and clustering for pollution data. *Environmental and Ecological Statistics*, 22: 33–43.

Cooley, D., Davis, R.A. and Naveau, P. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *Annals of Applied Statistics*, 6: 1406–1429.

Dannemann, J., Holzmann, H. and Lesiter, A. (2014) Semiparameteric hidden Markov models: identifiability and estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6: 418–425.

Farcomeni, A. and Greco, L. (2015). S-estimation of hidden Markov models. *Computational Statistics*, 30: 57–80.

Fassó, A., Cameletti, M. and Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. *Environmentrics*, 18: 245–264.

Field, M., Stirling, D., Pan, Z. and Naghdy, F. (2016). Learning trajectories for robot programming by demonstration using a coordinated mixture of factor analyzers. *IEEE Transactions on Cybernetics*, 46: 706–717.

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

Greven, S., Dominici, F. and Zeger, S. (2011). An approach to the estimation of chronic air pollution effects using spatio-temporal information. *Journal of the American Statistical Association*, 106: 396–406.

Kim, K.H., Jahan, S.A. and Kabir, E. (2013). A review on human health perspective of air pollution with respect to allergies and asthma. *Environmental International*, 59: 41–52.

Lagona, F., Maruotti, A. and Picone, M. (2011). A Non-Homogeneous Hidden Markov Model for the Analysis of Multi-Pollutant Exceedances Data. In *Hidden Markov Models: Theory and Applications*. P. Dymarski (ed.), p. 207–222. InTech Publisher.

Lagona, F. , Maruotti, A. and Padovano, F. (2015). Multilevel multivariate modelling of legislative count data, with hidden Markov chain. *Journal of the Royal Statistical Society - Series A*, 178: 705–723.

Lagona, F., Picone, M. and Maruotti, A. (2016). A hidden Markov model for the analysis of cylindrical time series. *Environmetrics*,26: 534–544.

Latza, U., Gerdes, S. and Baur, X. (2009). Effects of nitrogen dioxide on human health: Systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *International Journal of Hygiene and Environmental Health* 212: 271–287.

Lee, D., Rushworth, A. and Sahu, S. (2014). A Bayesian localised conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70: 419–429.

Lee, D. and Sahu, S. (2016). Estimating the health impact of environmental pollution fields. In *Handbook of Spatial Epidemiology*. Lawson, A., Banerjee, S., Haining, R. and Ugarte, L. (eds.), in press.

Leroux, G.B. (1992). Maximum-likelihood estimation for hidden Markov models.*Stochastic Processes ad their Applications*, 40: 127–143.

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review*,79: 427–454.

Maruotti, A. and Rocci, R. (2012). A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine*, 31: 871–886.

McLachlan, G.J., Peel, D. and Bean, R.W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* 41: 379–388.

McNicholas, P. and Murphy, B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18: 285–296.

McNicholas, P. and Murphy, B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, 26: 2705–2712.

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

Martinez-Zarzoso, I. and Maruotti, A. (2013). The Environmental Kuznets Curve: functional form, time-dependent heterogeneity and outliers in a panel setting. *Environmetrics*, 24: 461–475.

Paciorek, C.J., Yanosky, J.D., Puett, R.C., Laden, F. and Suh, H.H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Annals of Applied Statistics*, 1: 370–397.

Park, E.S., Guttorp, P. and Henry, R.C. (2001). Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association*, 96: 1171–1183.

Pollice, A. and Jona Lasinio, G. (2009). Two approaches to imputation and adjustment of air quality data from a composite monitoring network. *Journal of Data Science*, 7: 43–59.

Punzo, A. and Maruotti, A. (2016). Clustering multivariate longitudinal observations: The

contaminated Gaussian hidden Markov model. *Jorunal of Graphical and Computational Statistics*, 25: 1097–1116.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58: 1506–1537.

Rosti, A.V.I. and Gales, M.J.F. (2002). Factor analysed hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 949–952.

Sahu, S., Baffour, B., Harper, P.R., Minty, J.H. and Sarran, C. (2014).A Hierarchical Bayesian Model for Improving Short-Term Forecasting of Hospital Demand by Including Meteorological Information. *Journal of the Royal Statistical Society, Series A*, 177: 39–61.

Scott, S.L., James, G.M. and Sugar, C.A. (2005). Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association*, 100: 359–369.

Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society - Series C*, 51: 351–372.

Shaddick, G., Lee, D., Zidek, J.V. and Salway, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Annals of Applied Statistics* 2: 1249–1270.

Visser, I., Raijmakers, M., and Molenaar, P. (2000). Confidence Intervals for HiddenMarkovModel Parameters. *British Journal of Mathematical & Statistical Psychology* 53: 317–327.

Yao, K., Paliwal, K.K. and Lee, T.W. (2005). Generative factor analyzed HMM for automatic speech recognition. *Speech Communication*, 45: 435–454.

Zucchini, W., MacDonald I.L. and Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R, 2nd Edition*, London: Chapman & Hall.

Antonello Maruotti
Centre for Innovation and Leadership in Health Sciences
University of Southampton
E-mail: a.maruotti@soton.ac.uk

Antonello Maruotti
Dipartimento di Scienze Economiche, Politiche e delle Lingue Moderne
Libera Università Maria Ss. Assunta
E-mail: a.maruotti@lumsa.it

Jan Bulla
Department of Mathematics
University of Bergen
E-mail: Jan.Bulla@uib.no

Francesco Lagona
Dipartimento di Scienze Politiche
Università Di Roma Tre
E-mail: lagona@uniroma3.it

Marco Picone
Dipartimento di Tutela delle Acque Interne e Marine
The Institute for Environmental Protection and Research (ISPRA)
E-mail: marco.picone@isprambiente.it

Francesca Martella
Dipartimento di Scienze Statistiche
Sapienza Università di Roma
E-mail: francesca.martella@uniroma1.it