A multivariate hidden semi-Markov model for the analysis of multiple air pollutants

Marco Mingione[a], Pierfrancesco Alaimo Di Loro[b], Francesco Lagona[b], and Antonello Maruotti[a]

[a]Dpt. of Political Sciences, Roma Tre University; marco.mingione@uniroma3.it, francesco.lagona@uniroma3.it
[b]Dpt. GEPLI, Libera Università Maria Ss. Assunta (LUMSA); p.alaimodiloro@lumsa.it, a.maruotti@lumsa.it

Abstract

The analysis of multivariate time series of pollutant concentrations is complicated by the complex interactions between pollutants, which may not be constant over over time. We propose to approximate the joint data distribution by a parsimonious finite mixture of vector auto-regressive models, whose parameters are driven by the evolution of a latent semi-Markov process. This results in a vector auto-regressive hidden semi-Markov model that is capable to (1) describe the exposure to pollution in terms of a few latent regimes, (2) associate these regimes with specific combinations of pollutant concentration levels as well as distinct correlation structures between concentrations, and (3) provide estimates of sojourn times in each regime and transition probabilities between regimes. We apply the proposed model to the daily time-series of nine pollutant concentrations recorded at the Marylebone Road station, in London, between 2012 to 2017.

Keywords: air quality, EM algorithm, hidden Markov model, semi-Markov process, penalized regression

1. Introduction

Environmental risks are defined as all the external physical, chemical, biological, and work related factors that affect a person's health and well-being. They include pollution, radiation, noise, land use patterns, work environment, and climate change, that are all well-recognised as important causes of disease burden for populations and National Health Systems. Among them, air pollution is classified as one of the greatest environmental risks to health by the WHO[1], causing millions premature deaths worldwide every year (7). Its extent can be measured by the concentrations of various pollutants in the air (10), making its nature intrinsically multi-dimensional.

Here, we propose a multivariate model-based approach, belonging to the class of multivariate vector auto-regressive Hidden Semi-Markov Models (HSMMs; (2)), that can capture the dynamic evolution of environmental risk factors determined by multiple time and cross-dependent components.

HSMMs are recognised as a promising tool driving the assessment of environmental risk building procedure. Specifically, they allow for policy makers to differentiate the overall air-pollution risk exposure depending on the state of the environment identified by the latent process. This synthesis resembles other existing composite metrics such as the air quality index proposed by the US EPA that provides six

[1] see https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

risk levels: “good”, “moderate”, “unhealthy for sensitive groups”, “unhealthy”, “very unhealthy”, “hazardous”. Furthermore, HSMMs are intrinsically dynamic and provide a conditional framework for the quantification of the environmental risk calculated on the resulting predictive distribution. HSMMs can be also viewed as a structured Hidden Markov Models (8; 11) and the predictive distribution is a finite mixture of the semi-Markovian emission distributions. Thus, they allow an analytical assessment of the overall current and future environmental risk conditional on the past values of the process, making them invaluable monitoring tools for regulatory agencies.

Nevertheless, the specification of our model poses two notable statistical challenges: (i) the typical model selection issue related to the order of the autoregression; (ii) the large number of parameters in the state-specific covariance matrices that generate unstable estimates and increasing computational burden. We propose a penalized likelihood approach to HSMMs to deal with these issues, following similar works in the field (3; 4). Here, we contribute to this literature by applying the LASSO regularization on both the hidden state covariance matrix and the vector autoregressive coefficients.

The remainder of the paper is organized as follows. Section 2. summarizes the modeling framework and the estimation process. Section 3. describes the application and the results on the air quality indicators (AQIs) recorded at the Marylebone Road station, in London (UK), from 2012 to 2017. Finally, Section 5. contains some final comments and further developments.

2. A model for multivariate times series of pollutant concentrations

The data that motivated our proposal are in the form of a multivariate time series $(\boldsymbol{y}_t, t \geq 0)$, where the vector $\boldsymbol{y}_t = [y_{t1}, \ldots, y_{tp}]$ includes the concentrations of p pollutants at time t, $t = 1, \ldots, T$. A popular modelling approach in this setting is provided by the class of vector auto-regressive (VAR) models with lag H, where the conditional distribution of the process at time t given the past $\mathcal{H}_t = \{\boldsymbol{y}_\tau,\ \tau = 1, \ldots, t-1\}$:

$$f_{\boldsymbol{Y}_t}(\boldsymbol{y}_t|\mathcal{H}_t) = f_{\boldsymbol{Y}_t}\left(\boldsymbol{y}_t \mid \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{Y}_{t-H} = \boldsymbol{y}_{t-H}\right), \quad \boldsymbol{y}_t \in \mathbb{R}^p,\ t = 1, \ldots, T.$$

is a multivariate normal distribution with mean

$$\boldsymbol{\mu}_t = \boldsymbol{\beta}_0' + \sum_{h=1}^{H} \boldsymbol{\alpha}_h' \boldsymbol{y}_{t-h}, \tag{1}$$

and covariance matrix Σ. Under a traditional VAR model, the effect of past values on the process is time-constant and measured by the autoregressive coefficients $\boldsymbol{\alpha}$. It is however well known that the dynamic of air quality is affected by unobserved, time-varying weather conditions. This source of latent heterogeneity can be captured by assuming that the conditional distribution of the process given the past is a mixture of K VARs (5). A mixed VAR can be seen as a hierarchical VAR, whose parameters evolve according to the values taken by a sequence of independent multinomial distributions. Such an independence assumption is an obvious limitation of the model, which can be relaxed by allowing latent multinomial distributions to be temporally correlated. Such extension can be easily obtained by specifying the latent multinomial process as a homogeneous Markov chain. A mixture of VARs whose parameters evolve according to a latent Markov chain are known in the literature as multivariate hidden Markov models, and they have been already exploited for the analysis of multivariate environmental time series (9). Under a multivariate hidden Markov model, the sojourn times that the latent process spends in a specific state follow a Geometric distribution. Although this assumption can be realistic in some settings, it can be a serious shortcoming in the analysis of urban pollution, where sojourn times of specific air quality regimes may well show more complicated patterns. To avoid this source of misspecification, we propose a finite mixture of VARs whose parameters vary according to the states of a semi-Markov process. A discrete-time semi-Markov process $(S_t, t \geq 0)$ taking values in a finite state space, say $\{1, 2, \ldots K\}$, can be defined as follows. First, let $(U_t, t \geq 0)$ be a homogeneous Markov chain with a special transition probabilities matrix having all diagonal entries equal to zero (i.e the chain is not allowed to remain in the same state in two subsequent times). A realization of a semi-Markov process is obtained by replacing the state k of the Markov chain at time t by a run of d copies of state k, where d is drawn from a state-specific dwell-time distribution d_k. When the dwell time-distributions are all equal to a geometric distribution, then a semi-Markov process reduces to a Markov chain. Otherwise, it extends the Markov chain framework by allowing for flexible dwell times.

Let $(S_t, t \geq 0)$ be a latent semi-Markov process, taking values in the set $\{1 \dots K\}$, whose distribution is known up to a $K \times K$ matrix of transition probabilities with zero diagonal entries and a battery of K dwell time distributions that depend on a finite set of parameters. Conditionally on the value k taken by the latent process at time t, we assume that the conditional distribution of the observed process given the past is a multivariate normal distribution with mean

$$\boldsymbol{\mu}_t = \boldsymbol{\beta}'_{k0} + \sum_{h=1}^{H} \boldsymbol{\alpha}'_{kh} \boldsymbol{y}_{t-h}, \tag{2}$$

and covariance matrix Σ_k. We remark that the resulting mixture VAR is stable, hence stationary, if all the companion matrices:

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{\alpha}_{k1} & \boldsymbol{\alpha}_{k2} & \cdots & \boldsymbol{\alpha}_{kH} \\ \boldsymbol{I}_p & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_p & \mathbf{0} \end{bmatrix}, \quad k = 1, \dots, K,$$

have all the eigen-values with modulus < 1 (5).

2.1 Estimation

The proposed hidden semi-Markov VAR model can be represented as a multivariate HMM by appropriately augmenting the state sequence, as proposed by Langrock and Zucchini (8). Such representation is exact if all dwell time distributions have finite support. It otherwise provides an approximation with the desired accuracy. An HMM representation of our proposal allows us to exploit the efficient likelihood maximization algorithms that have been developed for HMMs. These algorithms rely on a EM approach, where the maximization of a weighted (data-augmented) log-likelihood function is alternated with weights updating, up to convergence.

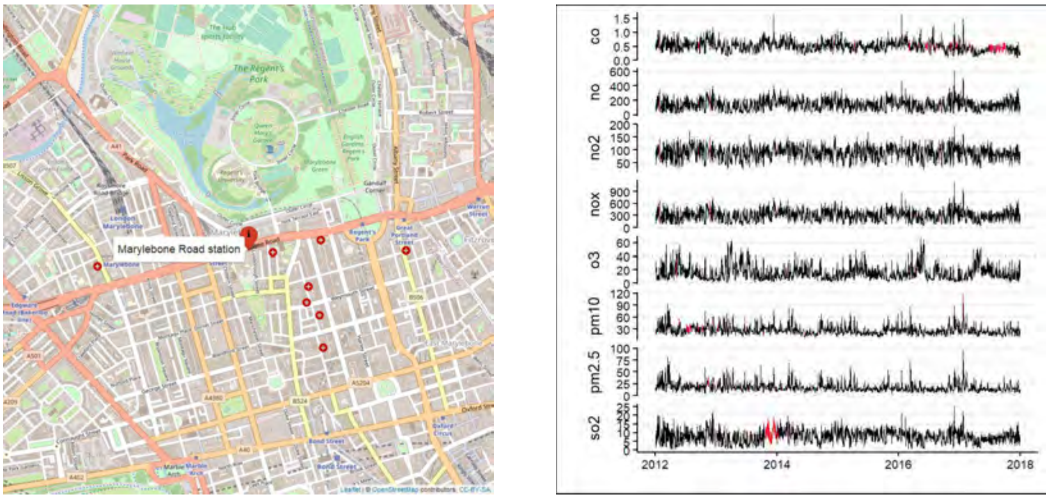
The maximization step is further integrated by considering a penalized approach for the estimation of the covariance matrix. This regularizes the estimation process, favoring the model identifiability and highlighting the most evident correlation patterns across the p outcomes. Following (4), we use a convex combination of the maximum likelihood estimator and a scaled identity matrix with the same trace:

$$\boldsymbol{\Sigma}_k = \frac{1}{1+\lambda_\Sigma} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{\boldsymbol{ML}} + \frac{\lambda_\Sigma}{1+\lambda_\Sigma} c\mathbf{I}, \quad \text{with} \quad \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{\boldsymbol{ML}}) = tr(c\mathbf{I}).$$

The coefficient $\lambda_\Sigma \geq 0$ is a shrinking parameter that controls for the strength of the penalization. Similarly, we consider a L1-norm penalty on the VAR coefficients and introduce a LASSO shrinking parameter λ_α in order to regularize their estimation and simultaneously select a smaller subset of lags that exhibits the strongest effects.

3. Application to urban pollution

We apply the proposed model to the daily time-series of $p = 9$ air quality indicators (AQIs) that have been recorded at the Marylebone Road station from 2012 to 2017 in London, for a total of $T = 2,192$ days of observation. This station is placed in the north-western part of the city on a very busy urban road (part of the London ring) and is close to Regent's park (see Figure 1a). Data have been downloaded from the `openair R` package (1), which provides several tools for the analysis of air pollution data, and include the concentrations of the following: carbon monoxide (co), nitrogen dioxide (no2), nitrogen



(a) (b)

Figure 1: (a) Location of MY1 (Marylebone Road station); (b) daily time-series of the AQIs, including imputed missing values (red line); (c) marginal distribution of the AQIs.

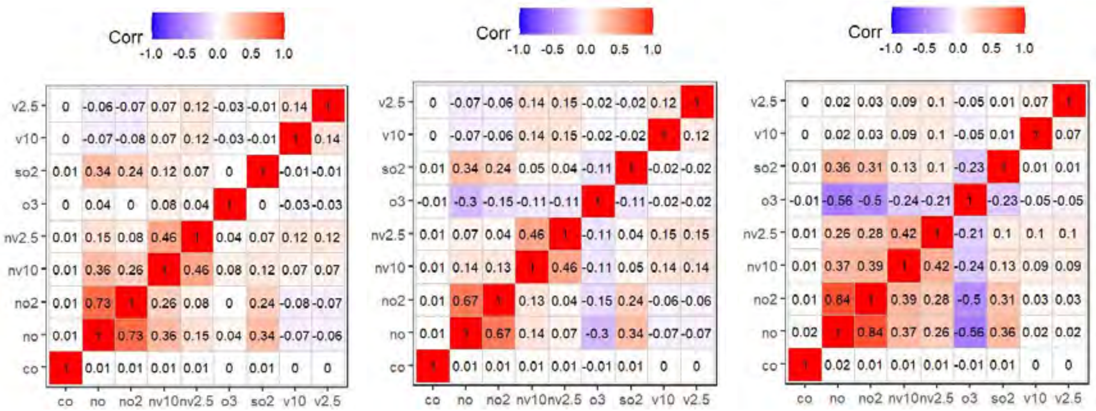
oxide (no), ozone (o3), sulfur dioxide (so2), volatile and non-volatile matter at $2.5\mu m$ (v2.5, nv2.5) and at $10\mu m$ (v10, nv10). As it is common in such applications, the raw data however contained some missing values, perhaps arising from the malfunctioning of the monitoring device. Note that missing values are not necessarily concomitant and the missing data patterns may vary across the AQIs. In our application, we assume that the missing data pattern is Missing at Random (MAR) for each AQI, and there is independence in the missing data mechanism across AQI (i.e. the probability of recording a missing value for one indicator does not depend on concomitant or previous missing values of another indicator). Overall, we counted a total of 804 missing out of $T \times p = 17,536$ values ($< 5\%$), with a percentage of missingness going from a minimum of $\approx 2\%$ for no and no2, up to $\approx 10\%$ for co. Given these low percentages, we used the `Amelia R` package (6) that combines bootstrapping and EM algorithms to provide efficient multiple imputation of incomplete set of data, including time-series data. Here, we use 50 multiple imputations and averaged all the results to get the final estimate of the missing values. A more rigorous approach would rely on integrating the missing value imputation within the proposed EM algorithm, at the price of an additional computational burden. The obtained time-series are shown in Figure 1b, where the imputed values are coloured in red. The series appear stationary through the whole considered time-window and show substantial correlation, particularly when spikes of high concentrations are recorded.

4. Results

We run a preliminary analysis on the data described in Section 3. We set the shrinking parameter of the covariance matrix $\lambda_\Sigma = 0.1$ and use cross-validation within the EM algorithm to adaptively select the best LASSO shrinking parameter λ_α at each iteration. We set $K = 3$ as to test the hypothesis of observing three possible patterns: pollutants below, in the mid-range, or above average. We consider only the effect of one lag, i.e. $H = 1$. The estimated conditional averages of pollutant concentrations in each state are reported in Table 1. Looking at the point estimates, state 2 can be interpreted as the high-pollution state, showing the largest pollutant average concentrations, except for ozone, which is however uncorrelated with all other pollutants (see Figure 2b). On the other hand, state 3 is the low-pollution state, state 1 is the intermediate state. We estimate 931 days in state 1, 39 days in state 2 and 1222 days in state 3. As expected, the dwell time distribution of state n. 3 is more skewed to the right than the other two distributions, as shown in Figure 3. The LASSO shuts down the effect of carbon oxide on all other pollutants.

state	co	no	no2	nv10	nv2.5	o3	so2	v10	v2.5
2	0.79	289.83	123.56	42.43	30.36	6.34	14.57	6.63	5.93
1	0.6	186.92	105.78	27.77	18.49	10.52	9.94	3.89	3.16
3	0.42	95.44	75.38	18.94	12.04	19.08	5.97	2.91	2.19

Table 1: Estimated conditional averages in each of the considered latent states.



(a) (b) (c)

Figure 2: Estimated correlation matrices of the pollutants when they are in state 1 (a), 2 (b) and 3 (c).

5. Conclusions and further developments

The proposed model is able to describe complex multivariate patterns that present temporal and cross-variable correlation. The non-parametric HSMM specification makes it able to detect and estimate the lower-dimensional structure controlling the different regimes of the multivariate process. The shrinking factors control for the substantial complexity of its specification. Indeed, the full model envisions a faceted dependence structure which depends on a great number of parameters. The adoption of regularized estimation methods becomes a necessity more than an opportunity. An interesting byproduct of the LASSO penalty is its ability to perform selection together with the shrinking. This allows highlighting the most relevant auto-regressive patterns across different components of the multivariate process.

The preliminary analysis on the time-series of pollutants recorded at the Marylebone Road station in London returns promising results. In the near future we aim at implementing a specific selection proce-

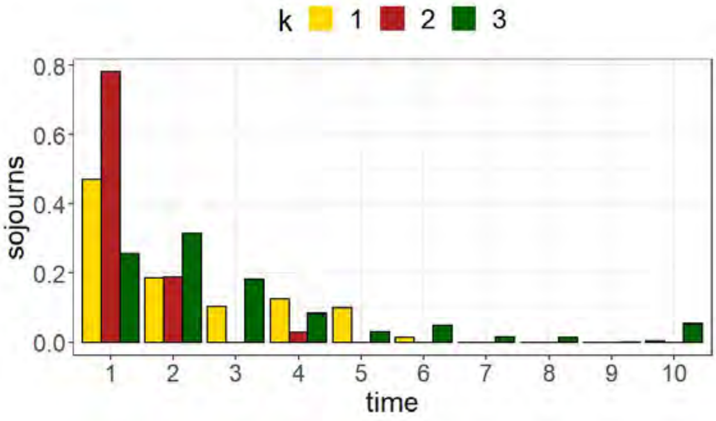


Figure 3: Estimated sojourn probabilities.

dure for the best number of states K. Furthermore, we can enrich the results analysis by accompanying them with risk and conditional risk measures, other than associating to each state the probabilities to exceed the risk thresholds fixed by the WHO.

References

[1] D. C. Carslaw and K. Ropkins. Openair—an r package for air quality data analysis. *Environmental Modelling & Software*, 27:52–61, 2012.

[2] M. Ehrmann, M. Ellison, and N. Valla. Regime-dependent impulse response functions in a markov-switching vector autoregression model. *Economics Letters*, 78(3):295–299, 2003.

[3] A. Farcomeni. Penalized estimation in latent markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency. *Biometrical Journal*, 59(5):1035–1046, 2017.

[4] M. Fiecas, J. Franke, R. von Sachs, and J. Tadjuidje Kamgaing. Shrinkage estimation for multivariate hidden markov models. *Journal of the American Statistical Association*, 112(517):424–435, 2017.

[5] P. W. Fong, W. K. Li, C. Yau, and C. S. Wong. On a mixture vector autoregressive model. *Canadian Journal of Statistics*, 35(1):135–150, 2007.

[6] J. Honaker, G. King, M. Blackwell, and M. M. Blackwell. Package 'amelia'. *Version. View Article*, 2010.

[7] P. J. Landrigan. Air pollution and health. *The Lancet Public Health*, 2(1):e4–e5, 2017.

[8] R. Langrock and W. Zucchini. Hidden markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1):715–724, 2011.

[9] A. Maruotti, J. Bulla, F. Lagona, M. Picone, and F. Martella. Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, 11(3):1617 – 1648, 2017.

[10] W. H. Organization et al. Who global air quality guidelines: particulate matter (pm2. 5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary. 2021.

[11] J. Pohle, T. Adam, and L. T. Beumer. Flexible estimation of the state dwell-time distribution in hidden semi-markov models. *Computational Statistics & Data Analysis*, 172:107479, 2022.