# Hierarchical Markov–Switching Models for Multivariate Integer–valued Time–series

Leopoldo Catania[a], Roberto Di Mari[b]

[a]*Department of Economics and Business Economics, Aarhus University and CREATES, Denmark*
[b]*Department of Economics and Business, University of Catania, Italy*

**Abstract**

We propose a new flexible dynamic model for multivariate nonnegative integer–valued time–series. Observations are assumed to depend on the realization of two unobserved integer–valued stochastic variables which control for the time– and cross–dependence of the data. We provide conditional and unconditional (cross)–moments implied by the model, as well as the limiting distribution of the series. An Expectation–Maximization algorithm for maximum likelihood estimation of the model parameters is derived, and an extensive Monte Carlo experiment investigates the finite sample properties of the resulting maximum likelihood estimator. Constrained specifications of the model are also formulated by modifying the assumptions about the dependence structure of the latent variables, and model identification is discussed accordingly. An application by means of a crime data set from the New South Wales (NSW) Bureau Of Crime Statistics And Research with observations spanning beyond 20 years is reported to illustrate the methodology. Results indicate that the proposed approach provides a good description of the conditional distribution of crime records, outperforming the standard hidden Markov model.

*Keywords:* Markov–Switching Model, Hidden Markov Model, Mixture Model, Hierarchical Model, NSW crime data

## 1. Introduction

Many applied problems have recently involved the analysis of discrete–valued time–series data, requiring adequate methodology for modeling and prediction. Recent reviews are reported by Karlis (2015) and Scotto et al. (2015). Examples can be found in various fields such as social science, finance and economics, and epidemiology, just to mention a few. For instance, crime economists are often concerned with the modelling of crime records for different cities over time, see for instance Glaeser and Sacerdote (1999) and Duggan (2001). Traders can obtain a better description of the stock market activities by modelling the number of trades for a set of assets, see for example Jung et al. (2011) and Fokianos et al. (2019). Labour economists might design better policies to improve the well-being of workers by explaining and predicting immigrants arrivals and flows, see for instance Friedberg (2001) and Freeman (2006).

Although the literature on univariate discrete–valued time–series is well developed, there is a number of additional complications related to the multivariate case, especially to multivariate counts. Due to lack of flexible correlation structures for the variables and computational difficulties – like evaluation of multiple summation multiple times over all possible counts, which is cumbersome already in the bivariate case (Kocherlakota and Kocherlakota, 1992) – simple extensions from univariate distributions are, in most instances, hardly feasible.[1] Recently, INteger–valued AutoRegressive (INAR) models (Al-Osh and Alzaid, 1987) have been applied to bivariate counts by Pedeli and Karlis (2011). Extensions to more than two dimensions have also been considered in Pedeli and Karlis (2013a), Pedeli and Karlis (2013b) and Bulla et al. (2017), generalizing the bivariate INAR(1) to all possible couples of variables. Although estimation of these models can be carried out with composite likelihood methods, multivariate (joint) distributions for the innovations have to be assumed, limiting the range of possible applications to only few variables. Another extension of the INAR model, which includes the dependence of the process from an unobserved Markov chain, has been recently proposed by Bu and McCabe (2008), Olteanu and Rynkiewicz (2012), and Fernández-Fontelo et al. (2016). An alternative modelling framework is offered by the Poisson autoregressive model detailed in Rydberg and Shephard (2000) and Fokianos et al. (2009). However, maximum likelihood inference of multivariate extensions of this model poses substantial challenges, see Fokianos et al. (2019).

Our contribution consists of providing a new dynamic parameter–driven model for multivariate integer–valued time–series, which allows for arbitrarily flexible serial– and cross–correlation patterns, maintaining computational simplicity in a maximum likelihood context. This is achieved through a two discrete latent variable structure, conditional on which the count variables are assumed to be independent. That is, the latent structure is assumed to fully explain serial– and cross–dependencies between the variables. This assumption – conditional independence of the observed variables given the latent process – is common in longitudinal data modeling (see, for instance, Vermunt et al., 1999; Bartolucci and Farcomeni, 2009). Specifically, the latent variables are assumed to have the following hierarchical structure (Figure 1): given one of the possible $J$ realizations of the first latent variable, which is dynamic and has a first–order Markov structure, the second time–specific latent variable takes one of the $K$ values, and they together affect the distribution of the observables $\mathbf{Y}_t$. In other words, we are assuming that the underlying latent structure is a *compound* that can be decomposed into an unconditional Markovian state process $S_t$ that handles serial dependence in the data, and another process, $Z_t$, that conditionally on $S_t$ handles cross dependencies and within–state unobserved heterogeneity. For instance, an unobserved time–varying overall crime attitude can influence cross–

[1] As noted by Karlis (2015), modeling (possibly complex) associations among integer–valued variables exploiting copulas (Nelsen, 2006) is difficult for two main reasons. First, the dependence structure cannot be fully separated from the marginals. Second, even with mildly complex correlation structures, the copula approach might require performing daunting tasks like evaluation of multiple integrals multiple times.

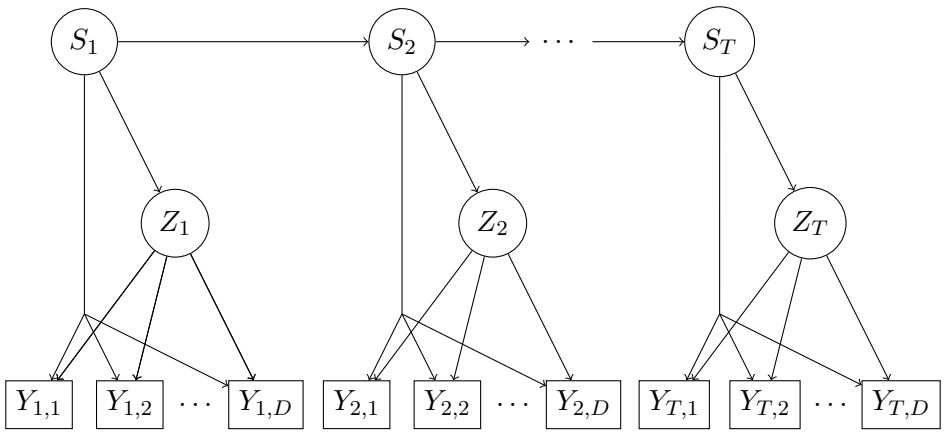



Figure 1: The model (Model 1, Equation (4)) path diagram. $S_t$ and $Z_t$ are two integer–valued unobserved stochastic variables. $S_t$ follows a first order Markov process with state space $\{1, \ldots, J\}$, while $Z_t$ is independently and identically distributed given $S_t$, with support $\{1, \ldots, K\}$. $Y_{i,t}$ for $i = 1, \ldots, D$ are the observed integer–valued random variables which are independently and identically distributed given $S_t$ and $Z_t$.

region criminal patterns (Carcach and Muscat, 2000), and they jointly determine crime rates in a given region/city. Alternative specifications obtained by modifying the dependence structure of the latent variables are also discussed in the paper. More in details, from the most general specification, we obtain two restricted specifications or sub-models. In the first sub-model the observed variable $\mathbf{Y}_t$ is affected by $S_t$ only through $Z_t$: for instance, the overall crime attitude does not directly affect the number of crimes in a given area, but only through cross–region criminal patterns. In the second sub-model the observed variable $\mathbf{Y}_t$ is affected by both $S_t$ and $Z_t$, which is now defined unconditionally with respect to $S_t$: for instance, the overall time–varying crime attitude and cross–region criminal patterns are independent, but they jointly determine crime rates in a given region/city. In our empirical application we will illustrate the three alternative modeling approaches, providing guidelines to the applied researcher about which approach to use under different circumstances.

A specification similar to ours was used by Geweke and Amisano (2011) to model univariate continuous time–series. Their hierarchical Markov normal mixture model adds flexibility to the standard hidden Markov model (HMM) to deal with possibly non–Gaussian components in the Markov states. The implementation is done within a Bayesian framework – with data augmentation for the latent variables and Markov chain Monte Carlo sampling of the posterior distribution. The difference with respect to our proposal goes beyond the chosen paradigm – *frequentist* or Bayesian. More specifically, although sharing a common Markov structure, our second latent variable is defined to describe the mutual association in the observed discrete variables, rather than adding flexibility to the standard HMM. Other similar modeling strategies have been proposed in the context of longitudinal data by Maruotti and Rydén (2008) and Altman (2007).[2] Model identifiability and

[2]Specifically, the mixed HMM they propose is a semiparametric HMM where the observed process is assumed to follow an inhomogeneous Poisson kernel. In their approach, the unobserved heterogeneity is modeled exploiting a generalized

other statistical features, i.e. conditional and unconditional (cross) moments implied by the model as well as the limiting distribution of the series, shall be presented. To compute the maximum likelihood estimates of the model parameters and make inference about the latent structure, we derive an Expectation-Maximization (EM) algorithm with closed form E and M steps by relying on the incomplete-data representation of the model. Finite sample properties of the ML estimator will be investigated – under several conditions – in an extended Monte Carlo experiment.

We estimate the model (and its constrained versions) on data from the NSW Bureau Of Crime Statistics And Research (BOCSAR). The BOCSAR data is a publicly available database[3] containing monthly counts of criminal incidents, reported to or detected by the NSW Police Force from January 1995 up to December 2016, for a total of 252 observations for each series.[4] These series exhibit a notable level of heterogeneity across cities and time: a stylized fact which can be conveniently incorporated by our specification and that it is common among crime records, see for example the discussion in Freeman (1999) and Levitt (2017). For instance, the series for some cities might show values in a relatively small range and modest time variation, whereas others might have both larger range of values and variability across time. Although our goal is not to establish any causal relation which might explain *why* crime records exhibit such a heterogeneous behavior, we find that the proposed model is able to well represent *how* these series evolve. Also, results indicate that the standard Markov Switching model is outperformed by our specification. The illustration concludes by investigating whether modifying the dependence structure among the latent and observed random variables leads to different results. Our analysis indicates that allowing for a flexible dependence structure is key for crime record modeling.

The paper proceeds as follows. Section 2 introduces the model, and its statistical properties are derived in Section 3. Section 4 describes two nested specification of the model. Details on model estimation are given in Section 5, and results from a Monte Carlo experiment assessing the finite sample properties of the estimator are showed in Section 6. Section 7 presents results on the NSW crime data, and Section 8 concludes. Proofs are gathered in the Appendix. The supplementary material accompanying this paper reports additional empirical results.

---

linear model structure by adding individual–specific continuous random effects in the link function. However, they exploit the finite mixture approach (Aitkin, 1996) to approximate these random effects ending up with a hierarchical structure similar to ours. Other examples of the use of mixed HMM for longitudinal data can be found in Lagona et al. (2014) and Marino and Alfó (2016). For a review on the topic, see Maruotti (2011).

[3] http://www.bocsar.nsw.gov.au/Pages/bocsar_crime_stats/bocsar_detailedspreadsheets.aspx

[4] Works on these data include, for instance, Jones et al. (2009) who studied the impact of alcohol availability on alcohol–related violence in the NSW state; Weatherburn et al. (2003), who examined the effects of supply–side drug law enforcement on the dynamics of the NSW heroin market and the harms associated with heroin, among others.

## 2. A Hierarchical Markov Switching Model (HMSM) for multivariate nonnegative integer–valued time–series

Let $\{S_t\}_{-\infty}^{\infty}$ be an unobserved first order stationary and ergodic Markov chain with state space $\mathcal{S} = \{1, \ldots, J\}$ and transition probability matrix $\mathbf{\Gamma} = [\gamma_{j,h}]$, where $\gamma_{j,h} = P(S_t = h|S_{t-1} = j, S_{t-u}\, u > 1) = P(S_t = h|S_{t-1} = j)$ and $\sum_{h\in\mathcal{S}} \gamma_{j,h} = 1$ for all $j, h \in \mathcal{S}$, and stationary distribution $\boldsymbol{\pi}_\infty$, which satisfies $\mathbf{\Gamma}\boldsymbol{\pi}_\infty = \boldsymbol{\pi}_\infty$. Let $\{Z_t\}_{-\infty}^{\infty}$ be an additional sequence of unobserved conditionally independent distributed integer–valued random variables given $S_t$, with support $\mathcal{K} = \{1, \ldots, K\}$. Given a realization $S_t = j$, we indicate by $\omega_{j,k}$ the probability that $Z_t = k$, that is $P(Z_t = k|S_t = j) = \omega_{j,k} > 0$, with $\sum_{k\in\mathcal{K}} \omega_{j,k} = 1$ for all $j \in \mathcal{S}$. All $\omega_{j,k}$ are collected in the $J \times K$ matrix $\mathbf{\Omega}$. The $D$–dimensional vector of random variables whose outcome is observed is denoted as $\mathbf{Y}_t \in \mathbb{N}_0^D$, where $\mathbb{N}_0$ is the set of natural numbers with 0 included. We denote the joint conditional probability mass function (*pmf*), indexed by the vector $\boldsymbol{\lambda}_{j,k}$, as $P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t|S_t = j, Z_t = k) > 0$ for all possible realizations $\mathbf{y}_t$ of $\mathbf{Y}_t$ and values of $\boldsymbol{\lambda}_{j,k} > \mathbf{0}$.[5] The notations $P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t|S_t = j, Z_t = k)$, $P(\mathbf{y}_t|S_t = j, Z_t = k)$ and $P(\mathbf{Y}_t = \mathbf{y}_t|S_t = j, Z_t = k)$ are used interchangeably through the paper. The following factorization of the conditional joint *pmf* of $\mathbf{Y}_t$ is assumed:

$$P_{\boldsymbol{\lambda}_j}(\mathbf{y}_t|S_t = j) \equiv \sum_{k=1}^{K} \omega_{j,k} P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t|S_t = j, Z_t = k), \tag{1}$$

where $\boldsymbol{\lambda}_j = (\boldsymbol{\lambda}'_{j,k}, k = 1, \ldots, K)'$, and

$$P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t|S_t = j, Z_t = k) \equiv \prod_{i=1}^{D} P_{\lambda_{j,k,i}}(y_{i,t}|S_t = j, Z_t = k), \tag{2}$$

where $y_{i,t}$ is a realization of $Y_{i,t}$, the $i$–th element of $\mathbf{Y}_t$, and $\boldsymbol{\lambda}_{j,k} = (\lambda_{j,k,i}, i = 1, \ldots, D)'$. Note that, according to Equation (2), conditionally on the realizations of the two random variables $S_t$ and $Z_t$, the univariate random variables $Y_{i,t}$, $i = 1, \ldots, D$, and $t = 1, \ldots, T$ are independent. That is, $Y_{i,t}|(S_t, Z_t) \perp\!\!\!\perp Y_{j,t}|(S_t, Z_t)$ for all $i \neq j$ and $Y_{i,t}|(S_t, Z_t) \perp\!\!\!\perp Y_{i,t-u}|(S_{t-u}, Z_{t-u})$ for all $u \neq 0$. Thus, the time– and cross–dependence structures which characterize $\mathbf{Y}_t$ are controlled by the two unobserved random variables $S_t$ and $Z_t$. For example, the random variable $Z_t$ can be thought of as an unobserved characteristic which is common to all the components of $\mathbf{Y}_t$ and, given a realization of $S_t$, determines their dependence structure.

In this paper, we assume $Y_{i,t}|(S_t = j, Z_t = k)$ to be Poisson distributed, with intensity parameter $\lambda_{j,k,i} > 0$ and *pmf* given by:

$$P_{\lambda_{j,k,i}}(Y_{i,t} = q|S_t = j, Z_t = k) \equiv \frac{\lambda_{j,k,i}^q e^{-\lambda_{j,k,i}}}{q!}, \tag{3}$$

where $q = 0, 1, 2, \ldots$. We argue that the assumption of conditional independence is not restrictive at all. Rather, as will be discussed in Section 3, after marginalization of the latent states, it allows the

[5] The notation $\mathbf{x} > \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$ for two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimension is understood as element-wise.

distribution of $\mathbf{Y}_t$ to exhibit a wide range of time– and cross–dependencies.[6] The joint density of a sequence of $T$ variables, $\mathbf{Y}_{1:t}$, is obtained after marginalization of the latent variables as follow:

$$P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta}) = \sum_{S_1,\ldots,S_T \in \mathcal{S}} \delta_{S_1} \prod_{t=2}^{T} \gamma_{S_{t-1},S_t} \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_{j,k} P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t | S_t = j, Z_t = k), \tag{4}$$

or, in matrix notation,

$$P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta}) = \boldsymbol{\delta}' \mathbf{P}_1 \boldsymbol{\Gamma} \mathbf{P}_2 \cdots \boldsymbol{\Gamma} \mathbf{P}_T \mathbf{1}, \tag{5}$$

where $\mathbf{1}$ is a $J$-vector of ones and $\mathbf{P}_t$ is a $J \times J$ diagonal matrix with typical element $p_{j,j,t} = \boldsymbol{\omega}_j' \mathbf{p}_{j,t}$, where $\boldsymbol{\omega}_j = (\omega_{j,k}, \quad k = 1, \ldots, K)'$, and $\mathbf{p}_{j,t} = \left( \prod_{i=1}^{D} P(Y_{i,t} = y_{i,t} | S_t = j, Z_t = k), \quad k = 1, \ldots, K \right)'$. In this paper, we set the initial distribution of the chain, $\boldsymbol{\delta}$, equal to the stationary distribution, *i.e.*, $\boldsymbol{\delta} = \boldsymbol{\pi}_\infty$.

The model defined in Equation (4) can be extended to accommodate external auxiliary variables (covariates), that affect the unobserved as well as the observed variables of the model. Details on the parametrization and on how to compute the maximum likelihood estimator are given in Appendix A.

### *2.1. Model identifiability*

To establish identifiability (in a frequentist sense) of the HMSM we exploit Theorem 1 of Gassiat et al. (2016). Similar to other latent class models, identification is established conditional on a particular order of the state variables, that is up to label switching. The term "label switching" refers to the procedure of relabelling the latent variables and rearranging the state dependent distributions in a way to obtain an observationally equivalent model. In Section 5.1 we detail the ordering we induce on the HSMS latent variables to ensure identification.

Gassiat et al. (2016) consider the general class of non parametric HMMs with finite state space, like for instance HMMs where the emission distributions are estimated using kernels or using mixtures of distributions that belong to a parametric family - for example, the exponential one. Although the model reported in Equation (4) does not resemble a – strictly speaking – non parametric HMMs, Gassiat et al. (2016)'s results are general enough to be applied to our framework. Furthermore, as will be shown in Section 4, their results also apply to constrained versions of the HMSM. Identifiability of the model reported in Equation (4) is established in the following proposition.

[6] The conditional Poisson assumption is convenient for model tractability and estimation. However, it should be noted that if the aim is to model series with high (conditional to $\mathbf{Y}_{1:t-1}$) kurtosis coefficients, a large value of $K$ is required. To keep $K$ small, the Poisson assumption can be replaced by the Negative Binomial at the cost of a more involved estimation procedure. In this case, one can exploit the stochastic representation of a Negative Binomial random variable as a compound Poisson with a Logarithmic Series counting variable as in Quenouille (1949) and use results from Adamidis (1999). However, preliminary analysis has indicated that the resulting estimation algorithm is slow compared to the Poisson case and that precision, especially for the size parameter, can be low even for large samples.

**Proposition 1.** *Model identifiability*

*Consider the model defined by* (4)*, and assume that: i) $J$ and $K$ are known, ii) $S_t$ is irreducible, and iii) $\boldsymbol{\lambda}_{j_1,k_1} \neq \boldsymbol{\lambda}_{j_2,k_2}$ for all $j_1 \neq j_2, k_1 \neq k_2$. Then the parameters of the model of Equation* (4) *are identified up to label switching.*

Proposition 1 ensures that, conditional on a specific ordering of the latent states, $P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta}_0) = P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta})$ if $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}$, and $P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta}_0) \neq P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:t}; \boldsymbol{\Theta})$ when $\boldsymbol{\Theta}_0 \neq \boldsymbol{\Theta}$. Conditions *i*) – *iii*) are sufficient for model identifiability and commonly used in the HMM literature, see for instance Allman et al. (2009). Condition *i*) can be relaxed following the arguments of Alexandrovich et al. (2016). Note that, in *iii*) no conditions on $\boldsymbol{\Omega}$ are required (although we have imposed that mixtures probabilities are positive). This follows from the fact that the observable $\mathbf{Y}_t$ is jointly affected by $S_t$ and $Z_t$. In Section 4 we consider the case when $\mathbf{Y}_t$ is only affected by $Z_t$, and provide sufficient conditions for identifiability based on the rank of $\boldsymbol{\Omega}$.

## 3. Statistical properties of the model

In this section, we derive statistical properties of the HMSM model, additional results are reported in the supplementary material accompanying this paper. We first need to introduce the "*forward probabilities*" of $S_t$, $\boldsymbol{\alpha}_t' = \boldsymbol{\alpha}_{t-1}'\boldsymbol{\Gamma}\mathbf{P}_t$, with $\boldsymbol{\alpha}_1' = \boldsymbol{\delta}'\mathbf{P}_1$, such that $\boldsymbol{\alpha}_t = (\alpha_{j,t}, j = 1, \ldots, J)'$, where $\alpha_{j,t} = P(\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}, S_t = j)$ and $\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}$ indicates $\mathbf{Y}_s = \mathbf{y}_s$ for $s = 1, \ldots, t$. Exploiting the forward probabilities, the predictive distributions of the latent chain and the observables at time $t + h$, for $h > 0$, given observations up to time $t$ are simply given by:

$$\boldsymbol{\pi}_{t+h|t} = P(S_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) = \boldsymbol{\alpha}_t'\boldsymbol{\Gamma}^h/(\boldsymbol{\alpha}_t'\mathbf{1}), \tag{6}$$

and

$$p_{\boldsymbol{\lambda}}(\mathbf{Y}_{t+h} = \mathbf{y}_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) = \sum_{j=1}^{J} \pi_{j,t+h|t} \sum_{k=1}^{K} \omega_{j,k} \prod_{i=1}^{D} \frac{\lambda_{j,k,i}^{y_{i,t+h}} e^{-\lambda_{j,k,i}}}{y_{i,t+h}!}, \tag{7}$$

respectively. Note that, Equation (7) is the *pmf* of a mixture of mixuters of (conditionally independent) Poisson distributions. The first layer of the mixture is obtained after marginalization of $S_t$, while the second layer is obtained after marginalization of $Z_t$. The *pmf* of the univariate random variable $Y_{i,t}$ is recovered from (7) by marginalization of the other variables:

$$\begin{aligned} p(Y_{i,t+h} = y_{i,t+h}|\mathbf{Y}_{1:t}) &= \sum_{y_{1,t+h}} \cdots \sum_{y_{i-1,t+h}} \sum_{y_{i+1,t+h}} \cdots \sum_{y_{D,t+h}} p(\mathbf{Y}_{t+h} = \mathbf{y}_{t+h}|\mathbf{Y}_{1:t}) \\ &= \sum_{j=1}^{J} \pi_{j,t+h|t} \sum_{k=1}^{K} \omega_{j,k} \frac{\lambda_{j,k,i}^{y_{i,t+h}} e^{-\lambda_{j,k,i}}}{y_{i,t+h}!}. \end{aligned} \tag{8}$$

Equation (8) shows that the univariate conditional distribution in the HMSM model is a mixture of mixtures of univariate Poisson distributions. Knowing the exact formulation of these distributions allows us to easily derive all the conditional and unconditional moments of $\mathbf{Y}_t$. Furthermore, by

letting $h \to \infty$, we recover the stationary distribution of $S_t$, $\boldsymbol{\pi}_{t+h|t} \to \boldsymbol{\pi}_\infty$. The limiting distribution of $\mathbf{Y}_t$ is obtained by replacing $\pi_{j,t+h|t}$ with $\pi_{j,\infty}$, the $j$–th element of $\boldsymbol{\pi}_\infty$, in Equations (7) and (8). Moments of the conditional and unconditional distributions of $\mathbf{Y}_t$ are readily available exploiting the mixture structure – given by either the predicted probabilities $\boldsymbol{\pi}_{t+h|t}$, or the limiting distribution of the Markov chain $\boldsymbol{\pi}_\infty$ as first layer of probabilities. In the HMSM model, all moments of $\mathbf{Y}_{t+h}|\mathbf{Y}_{1:t}$ exist and can be recovered from the moment generating function, $M_{\mathbf{Y}_{t+h|t}}(\mathbf{u})$, given by:

$$M_{\mathbf{Y}_{t+h|t}}(\mathbf{u}) \equiv \sum_{j=1}^{J} \pi_{j,t+h|t} \sum_{k=1}^{K} \omega_{j,k} e^{\sum_{i=1}^{D} \lambda_{j,k,i}(e^{u_i}-1)}, \tag{9}$$

where $\mathbf{u} = (u_i, i = 1, \ldots, D)'$ is a vector of constants. From (9) we easily recover the first two moments around the origin. The expected value $\boldsymbol{\mu}_{t+h|t} = \mathbb{E}[\mathbf{Y}_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}]$ is given by:

$$\boldsymbol{\mu}_{t+h|t} = \sum_{j=1}^{J} \pi_{j,t+h|t} \sum_{k=1}^{K} \omega_{j,k} \boldsymbol{\lambda}_{j,k}, \tag{10}$$

while the $(h, b)$–th element of the second moment $\mathbb{E}[\mathbf{Y}_{t+h}\mathbf{Y}'_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}] = [m_{hb,t+h|t}]_{h,b=1}^{D}$ is given by:

$$m_{hb,t+h|t} = \sum_{j=1}^{J} \pi_{j,t+h|t} \sum_{k=1}^{K} \omega_{j,k} \tilde{\lambda}_{j,k}^{(hb)}, \tag{11}$$

where

$$\tilde{\lambda}_{j,k}^{(hb)} = \begin{cases} \lambda_{j,k,h}(1 + \lambda_{j,k,h}), & \text{if } h = b \\ \lambda_{j,k,h}\lambda_{j,k,b}, & \text{otherwise.} \end{cases} \tag{12}$$

The covariance matrix of $\mathbf{Y}_{t+h}$ can easily be evaluated as $Cov(\mathbf{Y}_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) = \mathbb{E}[\mathbf{Y}'_{t+h}\mathbf{Y}_{t+h}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}] - \boldsymbol{\mu}_{t+h|t}\boldsymbol{\mu}'_{t+h|t}$, showing how overdispersion and cross sectional correlation is induced by the two latent variables $S_t$ and $Z_t$. Higher order moments easily follow.

As for standard HMMs, in our case generally $\mathbb{E}[Y_{i,t}Y_{l,t-\tau}] \neq 0$ for $i, l = 1, \ldots, D$ and $\tau > 0$. That is, the model allows for any type of (cross) serial autocorrelation. Bivariate INAR models, for instance, do not allow for negative cross–correlation in their baseline version, where either bivariate Poisson or negative binomial distributions are assumed for the innovation joint distribution. Let $B(\tau) = Cov(\mathbf{Y}_t, \mathbf{Y}_{t-\tau})$ be the $D \times D$ matrix of cross–covariances with typical element $Cov(Y_{i,t}Y_{l,t-\tau})$. In the HMSM model we have that,

$$B(\tau) = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{h=1}^{J}\sum_{b=1}^{K} \pi_{h,\infty}[\boldsymbol{\Gamma}^\tau]_{h,j}\omega_{j,k}\omega_{h,b}\boldsymbol{\lambda}_{j,k}\boldsymbol{\lambda}'_{h,b} - \boldsymbol{\mu}_\infty\boldsymbol{\mu}'_\infty, \tag{13}$$

where $\boldsymbol{\mu}_\infty$ is obtained from (10) by replacing $\pi_{j,t+h|t}$ with $\pi_{j,\infty}$, the $j$–th element of $\boldsymbol{\pi}_\infty$. From (13) we see that the $i, j$–th element of $B(\tau)$ decreases with $\tau$. However, we note that, differently from standard HMMs, also the second layer of probabilities, $\omega_{l,m}, \quad l = 1, \ldots, J, \quad m = 1, \ldots, K$, enters the autocovariance structure of $\mathbf{Y}_t$. Furthermore, from (13) we conclude that the process $\{\mathbf{Y}_t\}_{-\infty}^{\infty}$ is covariance stationary.

*3.1. Constrained MS representation*

As detailed by Bartolucci (2006) and Geweke and Amisano (2011), an interesting feature of the Hierarchical Markov–Switching Model (Section 2) is that it can always be represented as a constrained Markov–Switching Model with a particular structure imposed to the transition probability matrix. Specifically, the Hierarchical Markov–Switching Model with $J$ regimes and $K$ mixture components has an equivalent representation in terms of a standard Markov–Switching Model with $JK$ regimes. However, the reverse is not true. The transition probability matrix associated to the equivalent MS representation is given by:

$$\mathbf{\Gamma}^* = \mathbf{u}\text{vec}(\mathbf{\Omega}')' \cdot (\mathbf{\Gamma} \otimes \mathbf{U}), \tag{14}$$

where $\mathbf{U}$ and $\mathbf{u}$ are a $K \times K$ matrix and a $JK$–vector of ones, and $\cdot$ and $\otimes$ are the element–wise multiplication and Kronecker operators, respectively. The distributions defined in the regimes of the new system coincide with the mixture components of the original HMSM ordered such that the first $K$ are the distributions of the first regime in the original model, from $K+1$ to $2K$ of the second, and so on.

## 4. Special cases of the HMSM model

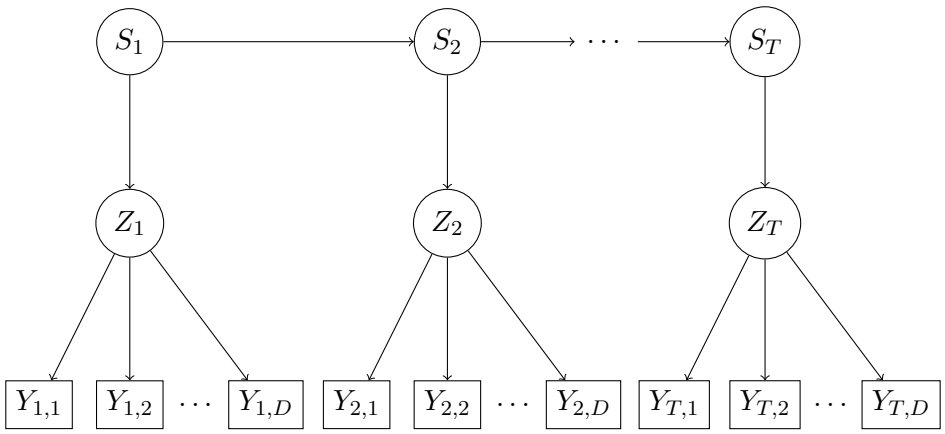
The HMSM model of Equation (4) imposes a general dependence structure between $S_t$, $Z_t$, and $\mathbf{Y}_t$, which is depicted in Figure 1. From this general specification, which we refer to as "Model 1", it is possible to obtain two restricted specifications or sub-models. In the first sub-model, Model 2, the observed process $\mathbf{Y}_t$ is affected by $S_t$ only through $Z_t$: for instance, the overall crime attitude does not directly affect the number of crimes in a given area, but only through cross–region criminal patterns. The joint *pmf* of a sequence of $T$ variables in Model 2 is given by:

$$P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \mathbf{\Theta}) = \sum_{S_1,\ldots,S_T \in \mathcal{S}} \delta_{S_1} \prod_{t=2}^{T} \gamma_{S_{t-1},S_t} \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_{j,k} P_{\boldsymbol{\lambda}_{\cdot,k}}(\mathbf{y}_t | S_t = j, Z_t = k), \tag{15}$$

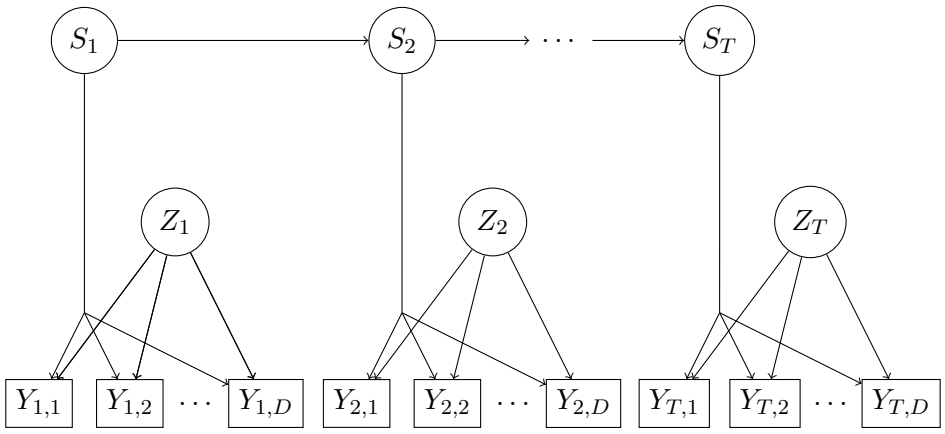
where we use the notation $\boldsymbol{\lambda}_{\cdot,k}$ to stress that, relative to the general specification of Equation (4), here the distribution of the observables depends only on $S_t$ through $Z_t$. The path diagram of Model 2 is reported in Figure 2a. Identification of Model 2 is established in the following Proposition 2.

**Proposition 2.** *Identifiability of Model 2 ($S_t$ does not affect $\mathbf{Y}_t$)*
*Consider the Model 2 defined by* (15)*, and assume that: i) $J$ and $K$ are known and $J \geq K$, ii) $\boldsymbol{\lambda}_{\cdot,k_1} \neq \boldsymbol{\lambda}_{\cdot,k_2}$ for all $k_1 \neq k_2$, iii) $S_t$ is irreducible, and iv) $\mathbf{\Omega}$ has rank $J$. Then, the parameters of Model 2 are identified up to label switching.*

The proof of Proposition 2 is made exploiting Proposition 2 of Gassiat et al. (2016) where a specification which resembles our Model 2 is considered. Conditions $i$) and $iv$) can be relaxed following the arguments of Alexandrovich et al. (2016).




(a) Model 2, Equation (15)




(b) Model 3, Equation (16)

Figure 2: $S_t$ and $Z_t$ are two integer–valued unobserved stochastic variables. $S_t$ follows a first order Markov process with state space $\{1, \ldots, J\}$, and $Y_{i,t}$ for $i = 1, \ldots, D$ are the observed integer–valued random variable. In Model 2, $\mathbf{Y}_t$ depends from $S_t$ only through $Z_t$. In Model 3, $\mathbf{Y}_t$ directly depends from $S_t$ and $Z_t$, and $Z_t$ does not depend on $S_t$. The support of $Z_t$ is $\{1, \ldots, K\}$.

In the second sub-model, Model 3, the observed variable $\mathbf{Y}_t$ is affected by both $S_t$ and $Z_t$, where $Z_t$ is now defined unconditionally with respect to $S_t$: for instance, the overall time–varying crime attitude and cross–region criminal patterns are independent, but they jointly determine crime rates in a given region/city. Notice that, in Model 3, the process $Z_t$ has no serial correlation. The *pmf* of $\mathbf{Y}_{1:T}$ in Model 3 is given by

$$P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \boldsymbol{\Theta}) = \sum_{S_1,\ldots,S_T \in \mathcal{S}} \delta_{S_1} \prod_{t=2}^{T} \gamma_{S_{t-1},S_t} \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_{\cdot,k} P_{\boldsymbol{\lambda}_{j,k}}(\mathbf{y}_t | S_t = j, Z_t = k), \tag{16}$$

where now the notation $\omega_{\cdot,k}$ is used to indicate that $Z_t$ does not depend on the chain $S_t$. The path diagram of Model 3 is displayed in Figure 2b. Identification of Model 3 is reported in the following Proposition 3.

**Proposition 3.** *Identifiability of Model 3 ($S_t$ does not affect $Z_t$)*
*Consider the Model 3 defined by* (16)*, and assume that the conditions of Proposition 1 hold. Then, the parameters of Model 3 are identified up to label switching.*

## 5. Maximum Likelihood parameters estimation

Estimation of the HMSM can be done by maximum likelihood. We collect all model parameters in $\boldsymbol{\Theta}$ and write the likelihood function as:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{y}_{1:T}) = P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \boldsymbol{\Theta}), \tag{17}$$

where $P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \boldsymbol{\Theta})$ is reported in Equation (5). In principle, direct (constrained) maximization of Equation (17) with respect to $\boldsymbol{\Theta}$ is feasible using a gradient descent method via, for example, the well–known Broyden–Fletcher–Goldfarb–Shanno algorithm. However, this solution can be very costly in terms of computational time due to the possible high number of parameters when $D$ is relatively large. The Expectation–Maximization (EM) algorithm of Dempster et al. (1977) provides an elegant alternative to the numerical likelihood optimization. We now present the EM algorithm for the HMSM in its more general specification (Model 1) reported in Equation (4) and represented in Figure 1. The modifications required to fit Models 2 and 3 are reported in Appendix C. Details about the modifications of the algorithm due to the inclusion of exogenous variables are reported in Appendix A. We start by introducing the following additional variables:

$$u_{j,t} = \begin{cases} 1, & \text{if } S_t = j \\ 0, & \text{otherwise.} \end{cases}, \quad v_{j,l,t} = \begin{cases} 1, & \text{if } S_{t-1} = j, \quad S_t = l, \\ 0, & \text{otherwise.} \end{cases}, \quad z_{j,k,t} = \begin{cases} 1, & \text{if } Z_t = k \text{ given } S_t = j \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The first two sets of variables, $u_{j,t}$ and $v_{j,l,t}$ for $j, l = 1, \ldots, J$, follow from the standard implementation of the algorithm for Markov–Switching Models (McLachlan and Peel, 2000), whereas the third set, $z_{j,k,t}$ (for $j = 1, \ldots, J$, and $k = 1, \ldots, K$), is specific to our model and is related to the additional latent variables $Z_t$, for $t = 1, \ldots, T$. Thus, augmenting the sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ with the new variables $\mathbf{u}_t = \{u_{j,t}, \quad j = 1, \ldots, J\}$, $\mathbf{v}_t = \{v_{j,l,t}, \quad j, l = 1, \ldots, J\}$ and $\mathbf{z}_t = \{z_{j,k,t}, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K\}$ allows us to write the so–called Complete–Data Log–Likelihood (CDLL):

$$\begin{aligned} \log \mathcal{L}^{\mathsf{c}}(\boldsymbol{\Theta}|\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \mathbf{v}_{2:T}, \mathbf{z}_{1:T}) = &\sum_{j=1}^{J} u_{j,1} \log(\delta_j) + \sum_{t=2}^{T} \sum_{j=1}^{J} \sum_{l=1}^{J} v_{j,l,t} \log(\gamma_{j,l}) + \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{k=1}^{K} u_{j,t} z_{j,k,t} \log(\omega_{j,k}) \\ &+ \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{D} u_{j,t} z_{j,k,t} (y_{i,t} \log(\lambda_{j,k,i}) - \lambda_{j,k,i} - \log(y_{i,t}!)), \end{aligned} \tag{19}$$

where $\mathbf{u}_{1:T} = (\mathbf{u}_1', \ldots, \mathbf{u}_T')'$, $\mathbf{v}_{2:T} = (\mathbf{v}_2', \ldots, \mathbf{v}_T')'$ and $\mathbf{z}_{1:T} = (\mathbf{z}_1', \ldots, \mathbf{z}_T')'$. The EM algorithm iterates between the Expectation–step (E–step) and Maximization–step (M–step) until convergence. Given a value of the model parameters at iteration $m$, $\boldsymbol{\Theta}^{(m)}$, the E–step consists in the evaluation of the so–called $\mathcal{Q}$ function defined as:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(m)}) = \mathbb{E}\left[\log \mathcal{L}^{\mathsf{c}}(\boldsymbol{\Theta}|\mathbf{y}_{1:T})\right], \tag{20}$$

where the expectation is taken with respect to the joint distribution of the augmenting variables – $\mathbf{u}_{1:T}$, $\mathbf{v}_{2:T}$ and $\mathbf{z}_{1:T}$ – conditional on $\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}$. Exploiting the formulation of the CDLL in (19),

the $\mathcal{Q}$ function can be factorized as:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(m)}) = \sum_{j=1}^{J} \widehat{u}_{j,1} \log(\delta_j) + \sum_{t=2}^{T}\sum_{j=1}^{J}\sum_{l=1}^{J} \widehat{v}_{j,l,t} \log(\gamma_{j,l}) + \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K} \widehat{u}_{j,t}\widehat{z}_{j,k,t} \log(\omega_{j,k}) + \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{D} \widehat{u}_{j,t}\widehat{z}_{j,k,t}(y_{i,t}\log(\lambda_{j,k,i}) - \lambda_{j,k,i} - \log(y_{i,t}!)), \tag{21}$$

where $\widehat{u}_{j,t} = P(S_t = j|\mathbf{y}_{1:T}, \boldsymbol{\Theta}^{(m)}) = \alpha_{j,t}\beta_{j,t}/(\boldsymbol{\alpha}_T'\mathbf{1})$,

$$\widehat{v}_{j,l,t} = P(S_{t-1} = j, S_t = l|\mathbf{y}_{1:T}, \boldsymbol{\Theta}^{(m)}) \tag{22}$$

$$= \alpha_{j,t-1}\gamma_{j,l}P(\mathbf{y}_t|S_t = l)\beta_{j,t}/(\boldsymbol{\alpha}_T'\mathbf{1}) \tag{23}$$

and $\beta_{j,t} = P(\mathbf{Y}_{t+1:T} = \mathbf{y}_{t+1:T}|S_t = j)$ is the $j$–th element of the so–called "backward probabilities" vector $\boldsymbol{\beta}_t = (\beta_{j,t}, j = 1, \ldots, J)'$, evaluated as $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\mathbf{P}_{t+1}\boldsymbol{\beta}_{t+1}$, where $\boldsymbol{\beta}_T = \mathbf{1}$ (see e.g. Frühwirth-Schnatter, 2006). The last quantities $\widehat{z}_{j,k,t}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $t = 1, \ldots, T$, are the posterior probabilities of sampling from the $k$–th mixture component of regime $j$ at time $t$, and are evaluated as follows:

$$\begin{aligned}\widehat{z}_{j,k,t} &= P(Z_t = k|S_t = j, \mathbf{y}_{1:T}, \boldsymbol{\Theta}^{(m)}) \\ &= \frac{\omega_{j,k}P(\mathbf{Y}_t = \mathbf{y}_t|S_t = j, Z_t = k)}{\sum_{l=1}^{K}\omega_{j,l}P(\mathbf{Y}_t = \mathbf{y}_t|S_t = j, Z_t = l)}.\end{aligned} \tag{24}$$

During the M–step of the algorithm, the $\mathcal{Q}$ function is maximized with respect to the model parameters $\boldsymbol{\Theta}$. Solving the Lagrangian associated to this (constrained) optimization we obtain the following solution to the maximization problem:

$$\gamma_{j,l}^{(m+1)} = \frac{\sum_{t=2}^{T}\widehat{v}_{j,l,t}}{\sum_{l=1}^{J}\sum_{t=2}^{T}\widehat{v}_{j,l,t}}, \quad \omega_{j,k}^{(m+1)} = \frac{\sum_{t=1}^{T}\widehat{u}_{j,t}\widehat{z}_{j,k,t}}{\sum_{t=1}^{T}\sum_{l=1}^{J}\widehat{u}_{l,t}\widehat{z}_{l,k,t}}, \quad \lambda_{j,k,i}^{(m+1)} = \frac{\sum_{t=1}^{T}\widehat{u}_{j,t}\widehat{z}_{j,k,t}y_{i,t}}{\sum_{t=1}^{T}\sum_{l=1}^{J}\widehat{u}_{l,t}\widehat{z}_{l,k,t}}. \tag{25}$$

Note that, in the derivation of $\gamma_{j,l}^{(m+1)}$, we have omitted from the maximization the first term of the $\mathcal{Q}$ function in order to remain with a closed form optimization step. This approach is a popular alternative to the numerical optimization step:

$$\max_{\gamma_{j,l}}\Big\{\sum_{j=1}^{J}\widehat{u}_{j,1}\log(\delta_j) + \sum_{t=2}^{T}\sum_{j=1}^{J}\sum_{l=1}^{J}\widehat{v}_{j,l,t}\log(\gamma_{j,l})\Big\}; \tag{26}$$

see Bulla and Berzel (2008). However, the solutions of the two optimizations are numerically equivalent for moderate time–series length.

Given an initial guess $\boldsymbol{\Theta}^{(0)}$, the algorithm iterates between the E– and the M–steps until convergence.[7] Convergence to a local optimum is guaranteed since the M–step increases the likelihood value at each iteration. As for standard HMMs, the likelihood function can present several local optima and there is no guarantee that convergence to the global optimum is achieved. To this

[7] For instance, when $|\mathcal{L}(\boldsymbol{\Theta}^{(m)}|\mathbf{y}_{1:T}) - \mathcal{L}(\boldsymbol{\Theta}^{(m-1)}|\mathbf{y}_{1:T})| < 10^{-8}$.

end, running the algorithm several times with different starting values is a standard procedure to better explore the likelihood surface. In our application, we run the algorithm 500 times, initializing the model parameters with random values. Then, the solution with the highest likelihood is selected. This is then compared, in terms of likelihood, with the following (more refined) initialization strategy, which follows a stepwise logic not new in Markov–Switching modeling (see, for instance, Bartolucci et al., 2015; Di Mari et al., 2016; Di Mari and Bakk, 2017), and works as follows:

1. Generate $J$ random Poisson parameters, initial state and transition probability matrix and fit a standard Markov–Switching Model with $J$ states, in order to obtain a starting point for $\mathbf{\Gamma}$;
2. Assign observations to latent states with global decoding (using the Viterbi algorithm);
3. Estimate $J$ mixtures, each with $K$ Poisson components, using the decoded observation, in order to get starting points for $\boldsymbol{\lambda}_j$ and $\omega_{j,k}$, with $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Then proceed alternating E and M steps until convergence, as described above. Finally, the solution we consider between the random and the *refined* initialization strategies is the one with the highest likelihood value. Standard errors and confidence intervals of the estimated parameters can be computed by parametric bootstrap (Efron and Tibshirani, 1993) as detailed by Zucchini et al. (2017). Note that parametric bootstrap is particularly convenient in our context since simulation from the model is straightforward and estimation via the EM algorithm is fast.

### *5.1. Handling label switching*

Mixture distributions are known to be identifiable up to label switching due to invariance under relabeling of the components (Redner and Walker, 1984). Although in classical ML inference this is not an issue[8] this practically means that, for instance, different local maximizers of the likelihood cannot directly be compared in terms of estimated parameters, unless some reparameterization is done. The model presented is Section 2 has one additional complication: label switching occurs both in the states of the hidden Markov chain and the components within each state. That is, there are $J!$ possible permutations of the states and, for each state, $K!$ permutations of the components defining the same parametric family. We handle this unidentifiability problem by relabeling the states according to an ordering defined on the $\ell_1$ norms of the Poisson parameters for each state and component. The new first state is defined as the $\arg\min_j\{\min_k\{\|\boldsymbol{\lambda}_{j,k}\|_{\ell_1}\}\}$, and the relabeling for the remaining states follow similarly. In the $j$–th state, the new first component is the one which $\arg\min_k\{\|\boldsymbol{\lambda}_{j,k}\|_{\ell_1}\}$, and consequently follows the relabeling for the subsequent components. Intuitively, we let the first state be relabeled as the one with the $K$–plet of Poisson parameters yielding the smallest norm, the second state be relabeled as the one with the $K$–plet of Poisson parameters yielding the second smallest norm, and so on. Within, say, state $j$, components are then relabeled such that the first component

[8] In Bayesian inference it can be a serious problem (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006)

is the one with the smallest norm, the second component is the one with the second smallest norm, and so on.

## 6. Finite sample properties of the Maximum Likelihood estimator

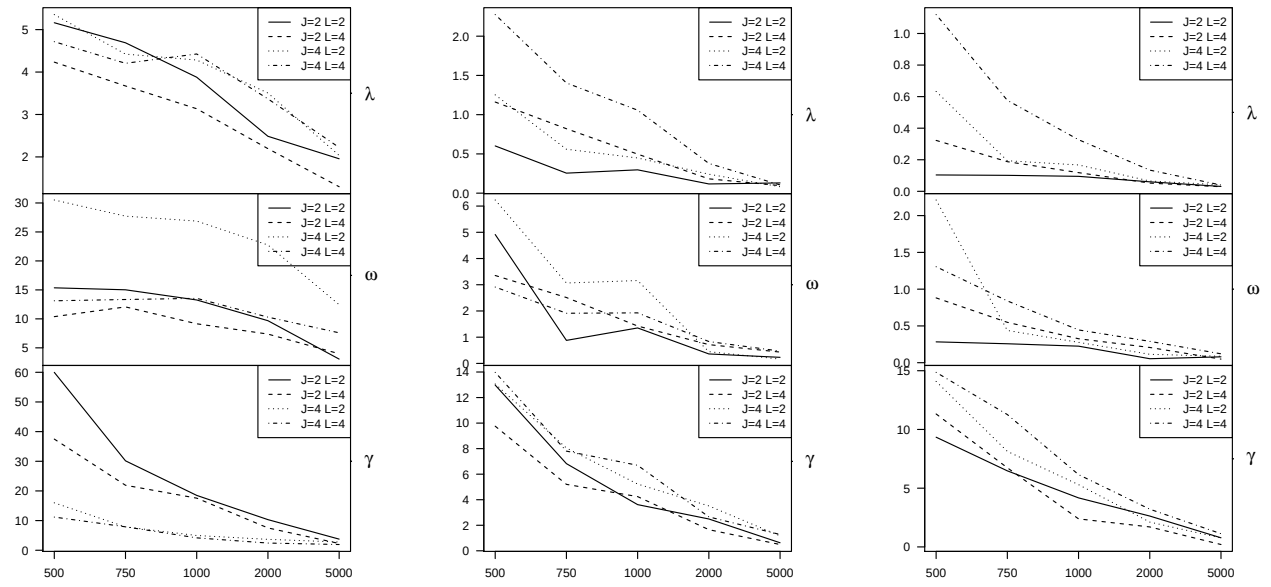
In this section we report an extensive Monte Carlo experiment in order to investigate the finite sample properties of the ML estimator for the HMSM detailed in Section 2. We select different options for: *i*) the number of Markov regimes, $J$, *ii*) the number of mixture components, $K$, *iii*) the cross–section dimension, $D$ and, *iv*) the length of the time–series, $T$. Specifically, we select $J \in (2, 4)$, $K \in (2, 4)$, $D \in (1, 4, 6)$ and $T \in (500, 750, 1000, 2000, 5000)$ such that we can investigate the properties of the estimator among different scenario of model complexity and data availability. Furthermore, since we expect class separation to have an impact on parameters estimation (see, for instance, Vermunt, 2010), we also consider a setting of low and medium separation, which we label as Sep = "Low" and Sep = "Medium", respectively. The experiment proceeds as follow. For each triplet $(J, K, D)$, we generate $T$ observations from the true Data Generating Process for selected parameter values. Subsequently, parameters are estimated from the simulated data using the EM algorithm detailed in Section 5. The procedure is iterated $B = 1000$ times, and for each iteration the difference across the estimated and the true parameters is stored.

Model parameters are chosen in order to replicate the dynamic properties of usual empirical data, such as a persistent evolution of the Markov chain and heterogeneity across the mixture components. Precisely, we set $p_{j,j} = 0.99$ and $p_{j,j} = 0.97$ when $J = 2$ and $J = 4$, respectively. The transition probabilities between regimes are fixed at $p_{i,j} = 0.01$ for $i, j = 1, \ldots, J$ and $i \neq j$. Mixing probabilities are randomly chosen during each iteration between $\boldsymbol{\omega}_j = (0.2, 0.8)$ and $\boldsymbol{\omega}_j = (0.8, 0.2)$ when $K = 2$ and between $\boldsymbol{\omega}_j = (0.1, 0.2, 0.3, 0.4)$ and $\boldsymbol{\omega}_j = (0.4, 0.3, 0.2, 0.1)$ when $K = 4$. Values for the location parameters $\lambda_{j,k,i}$ for $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $i = 1, \ldots, D$ are randomly generated from a uniform distribution with support $(1, \ldots, JKD)$ when Sep = "Low" and $(1, \ldots, JKD \times 10)$ when Sep = "Medium".

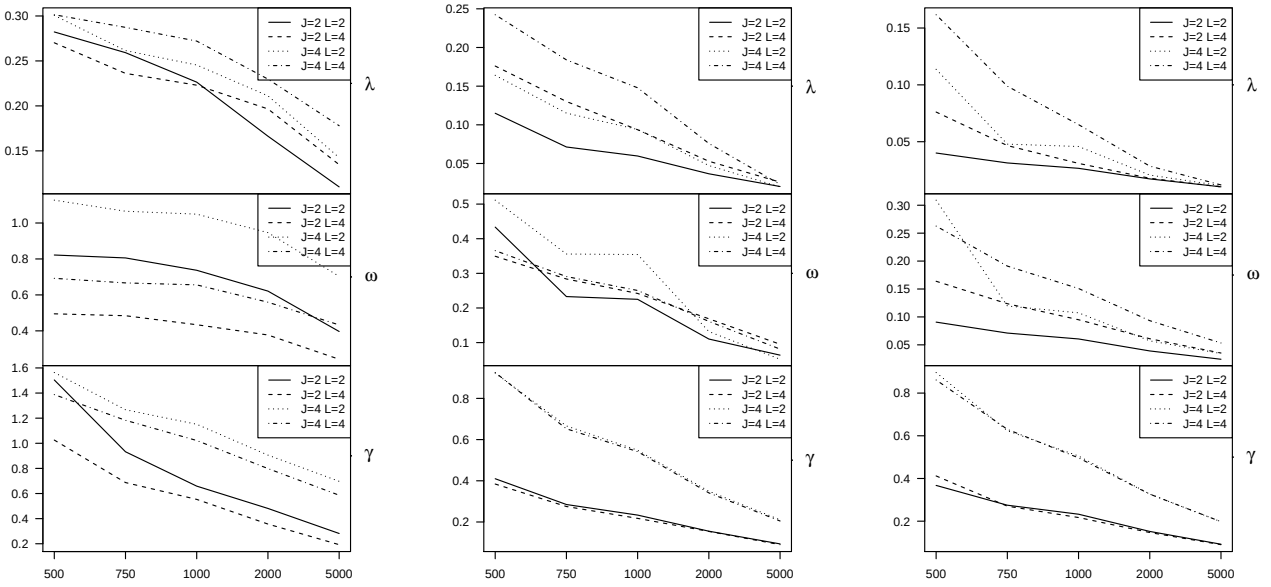
Figures 3 and 4 report the average absolute relative bias in percentage points (AARB) and average root mean squared error (ARMSE) between true and estimated parameters when the class separation is low and medium, respectively.[9] Results are reported for each type of parameters labeled as "$\lambda$" for the Poisson locations, "$\omega$" for the mixture probabilities and "$\gamma$" for the Markov chain transition probabilities. For instance, the AARB and ARMSE relative to the Poisson location parameters $\lambda_{j,k,i}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $i = 1, \ldots, D$ are evaluated as:

$$\text{AARB}(\lambda) = \frac{1}{JDKB} \sum_{b=1}^{B} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{D} \frac{|\widehat{\lambda}_{j,k,i}^{(b)} - \lambda_{j,k,i}^{(b)}|}{\lambda_{j,k,i}^{(b)}} \times 100, \tag{27}$$

[9] 5% of the total number of simulations have been discarded due to convergence to unreasonable values.




(A) $D = 1$ (B) $D = 3$ (C) $D = 6$




(D) $D = 1$ (E) $D = 3$ (F) $D = 6$

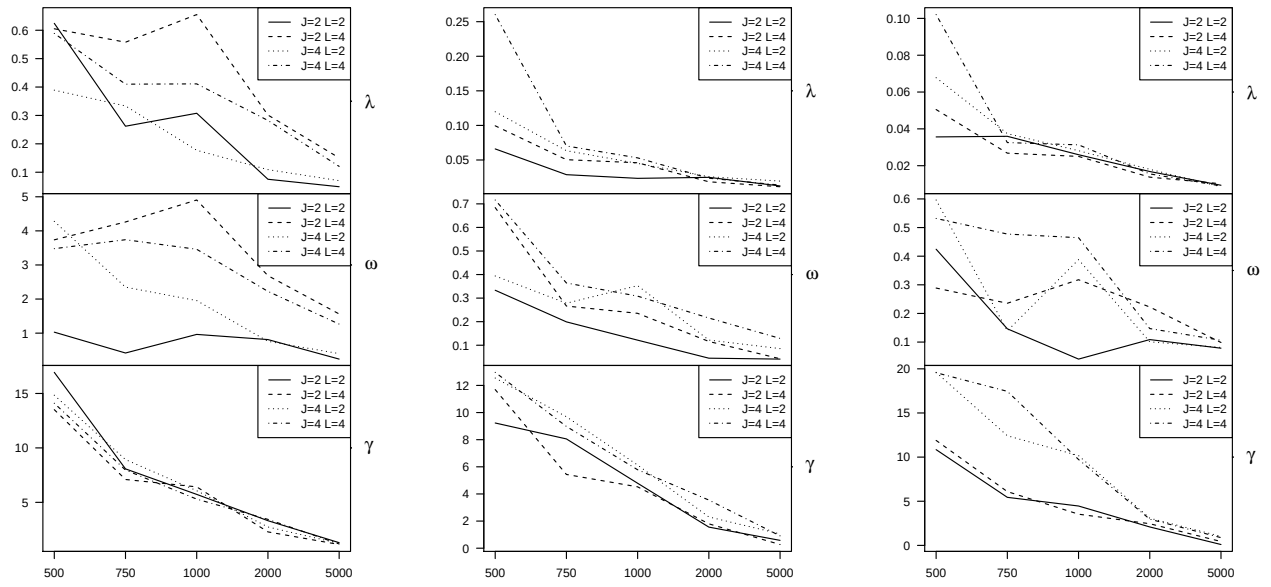
FIGURE 3: Panels (A), (B), and (C) report the AARB in percentage points of the HMSM parameters for different choices of $J$ and $L$. The horizontal axis indicates the lengths of the sample, $T$. Panels (D), (E), and (F) report the ARMSE. Results are for low separation between classes Sep = "Low".

and:

$$\text{ARMSE}(\lambda) = \frac{1}{JDK} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{D} \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left( \widehat{\lambda}_{j,k,i}^{(b)} - \lambda_{j,k,i}^{(b)} \right)^2}, \tag{28}$$

where $\widehat{\lambda}_{j,k,i}^{(b)}$ is the estimate for $\lambda_{j,k,i}^{(b)}$ at iteration $b = 1, \ldots, B$, respectively.[10] Results are very

[10]Recall that also the true parameters depend on the iteration since these are randomly generated.




(A) $D = 1$ (B) $D = 3$ (C) $D = 6$

(D) $D = 1$ (E) $D = 3$ (F) $D = 6$

FIGURE 4: Panels (A), (B), and (C) report the AARB in percentage points of the HMSM parameters for different choices of $J$ and $L$. The horizontal axis indicates the lengths of the sample, $T$. Panels (D), (E), and (F) report the ARMSE. Results are for medium separation between classes Sep = "Medium".

encouraging and indicate good performance of our estimator in finite samples. The AARB is very low for all parameters, the ARMSE is generally small and decreases with the increase of the sample size. We find that AARB is lower when separation is medium. We also note that the increase of the cross–section dimension $D$ increases the precision of the estimates and reduces their ARMSE. This result is probably related to the fact that the number of parameters only increases at rate $JK$, which is generally lower than, for example, $2D + 1$, *i.e.* the rate for a (static) multivariate Poisson model.

Results indicates that, in the particular case of $D = 1$ and low separation, biases in the Poisson parameters can be relatively large in the range (2% - 5%) depending on $T$. In addition, we note that the bias in the estimation of the transition probability matrix $\Gamma$ when the sample size is not large and $D$ is small. This result is not surprising and follows from our choice of the diagonal elements of $\mathbf{\Gamma}$. Specifically, since we have specified a persistent dynamic for $S_t$, in small sample the number of regimes changes will be on average very low, implying bad estimates of $\mathbf{\Gamma}$. Biases vanish for large sample sizes. Results from Figures 3 and 4 can be summarized as follows: *i*) a larger cross-section dimension ($D$) delivers more accurate parameter estimates (smaller ARMSE); *ii*) results (both in terms of AARB and ARMSE) tend to be overall better for larger levels of separation; *iii*) in general, we get reliable estimates of model parameters (AARB $< 10\%$) for $T > 750$.

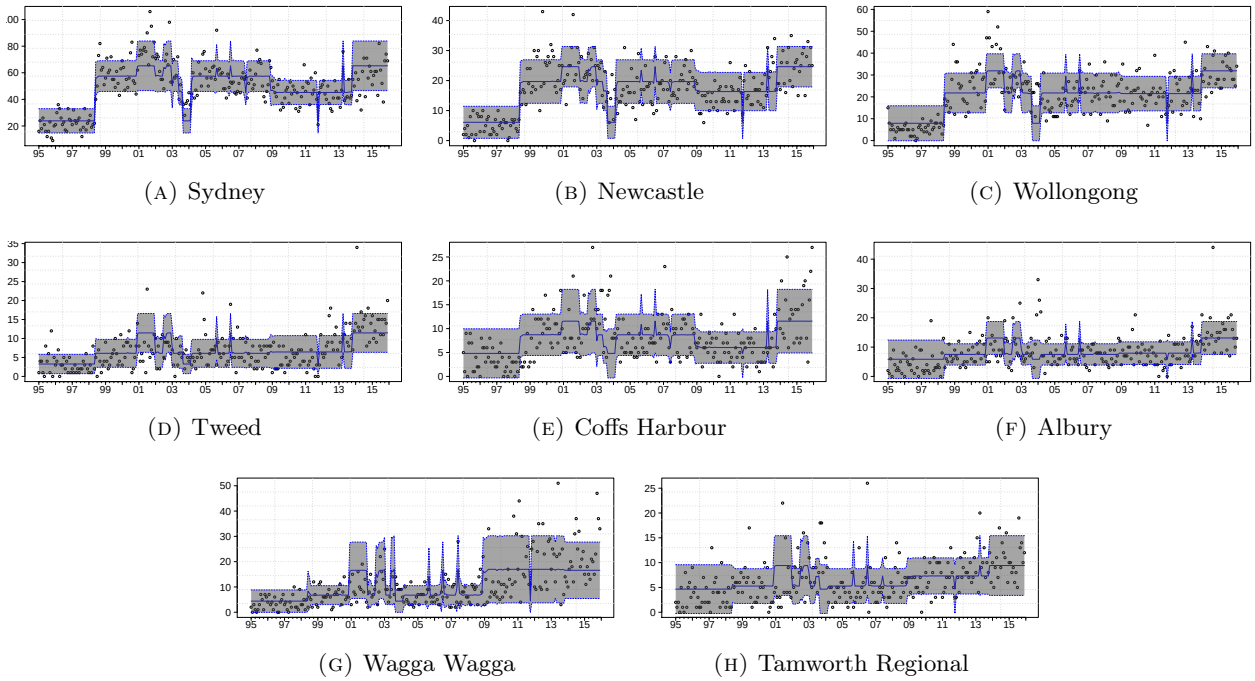
## 7. An Empirical Illustration Using Crime Data in the NSW State (Australia)

The goal of this application is twofold. First, to illustrate how the HMSM can be employed in an multivariate framework where individual series are characterised by different levels of heterogeneity. Second, to provide a technical framework to model cross section and time variability of crime records, which in the seminal paper of Glaeser et al. (1996) is referred to as the "most intriguing aspect" in the economics of crime - and still is (Levitt, 2017). Indeed, a better description of crime records might help in the development of policies aimed at improving the well-being of individuals and societies (Sah, 1991). Crime also has a large impact on the economy. For instance, the 2001's overall estimated crime social costs in Australia - medical costs, lost output, intangible costs caused by pain, suffering and lost quality of life - come to nearly $32 billion per year, including the additional costs of policing, prisons and security (Mayhew, 2003). Such a substantial social cost (about 10% of Australian GDP) makes crime a relevant issue for policy makers (to see how individual decision problems link to aggregate crime data analysis we refer to Kadane, 1985). In this spirit, Sickles and Williams (2008) studied the impact of both deterrent and preventive policies on potential criminal behavior.

The offenses category reported in the BOCSAR data are 21, including the most serious personal violence and property offenses – like assault, robbery, burglary, and malicious damage. Our application focuses on prohibited and regulated weapon offences, which have experienced a dramatic increase in the last twenty years. The series are collected, with no missing records, for the eight cities: Sydney, Newcastle, Wollongong, Tweed, Coffs Harbour, Albury, Wagga Wagga and, Tamworth Regional. The total number of data points available for model estimation is $2,016$. More details on both the crime definitions and the cities included in the sample are available at BOCSAR's website.

### *7.1. Model selection and signal extraction*

All pairs of models with $J \in (1, \ldots, 8)$ and $K \in (1, \ldots, 8)$ are estimated via the EM algorithm detailed in Section 5 on the whole sample of data. Model selection is performed using the Bayesian



(A) Sydney (B) Newcastle (C) Wollongong

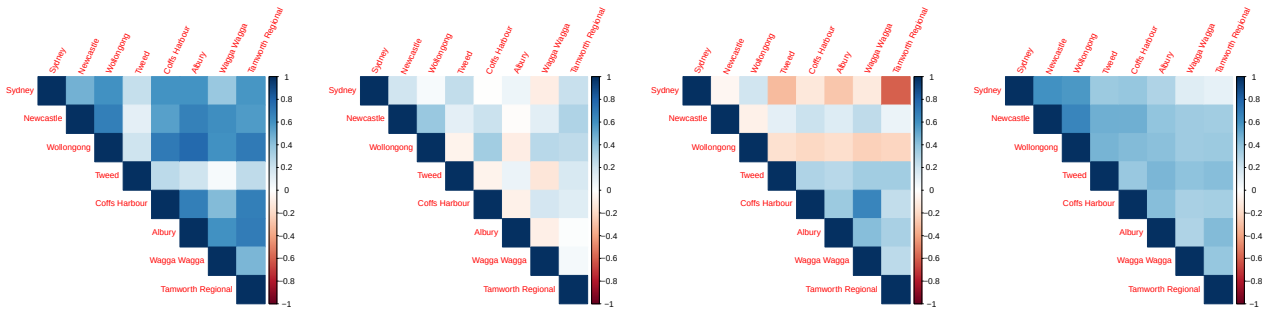
(D) Tweed (E) Coffs Harbour (F) Albury

(G) Wagga Wagga (H) Tamworth Regional

FIGURE 5: This figure reports the smoothed mean of the number of monthly weapon offences in eight cities of the NSW state (Australia) during the sample January, 1995 – January, 2016. Results are reported relative to the Hierarchical Markov–Switching model with $J = 4$ and $K = 3$, estimated during the period January, 1995 – January, 2016. One standard deviation confidence bounds are reported.

Information Criterion (BIC).[11] Specifications with $K = 1$ correspond to standard $J$–state Markov–Switching Models (MSM) thus allowing us to investigate the benefits of HMSM ($K > 1$) versus MSM (for an example of MSM used for detecting patterns of criminal activities, see Bartolucci et al., 2007). We find that BIC selects a model with four regimes, $J = 4$, and three mixture components, $K = 3$ which corresponds to a BIC of 12971 (BIC values for all combinations of $J$ and $K$ are reported in Appendix D). The number of (free) parameters for this specification is $J(J-1) + J(K-1) + JKD = 116$. Interestingly, the standard $J$–state MSM is always outperformed by our hierarchical model. Figure 5 plots the smoothed mean along with one standard deviation confidence bounds for the eight cities in our data set. We note that, for all cities, the monthly number of weapon offences exhibits considerable time variation in its levels. The first two remarkable increases in weapon offences level occur around 1998 and during the early 2000's. After a decreasing period around the middle of the sample, almost all cities have again gone up since 2014. At the end of our sample, most cities exhibit the highest weapon offence levels of the past 20 years. Interestingly, we note that the dependence structure also changed over the sampling period. Indeed, while at the

[11] Order estimation of HMMs is not trivial and we are not aware of any consistency result of BIC in our context of HMSM. Here we follow the common practice and use BIC to select $J$ and $K$. We note that, if $K$ is kept fixed and known, results from Csiszár and Shields (2000) can be applied as discussed in Cappé et al. (2005, Chapter 15).




(A) January, 1995 (B) June, 2005 (C) January, 2016 (D) Limiting Distribution

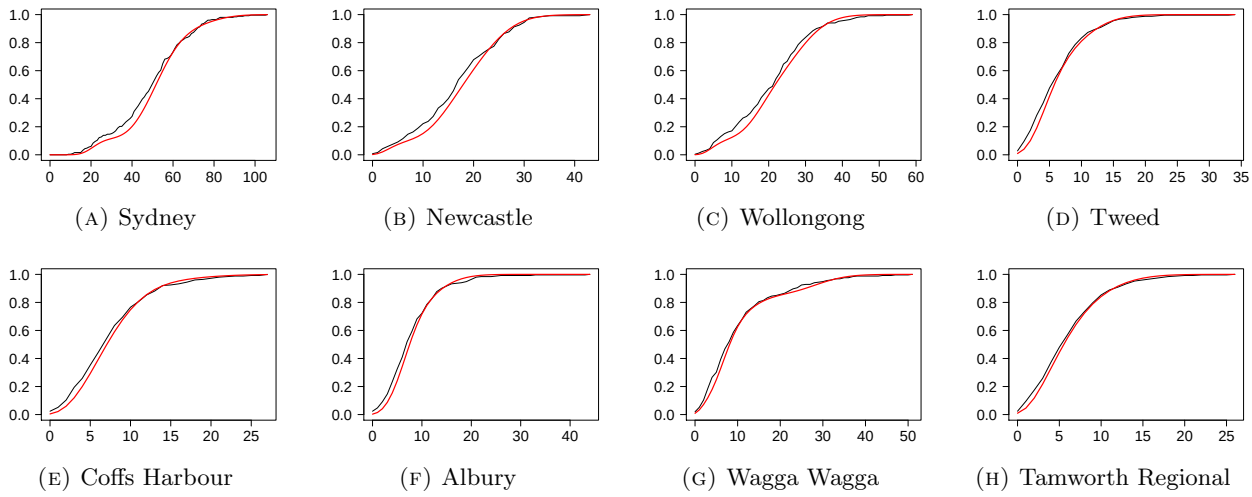
FIGURE 6: Figures (A), (B) and (C) report the smoothed correlation matrices between the monthly number of weapon offences across eight cities of the NSW state (Australia) at dates: *i*) January, 1995, *ii*) June, 2005, and *iii*) January, 2016, respectively. Figure (D) reports the correlation matrix implied by the limiting distribution. Results are reported for to the HMSM with $J = 4$ and $K = 3$, estimated during the period January, 1995 – January, 2016.

beginning of the sample all series evolve following a similar path (see for instance at the begin of 1998), in the middle of the sample their movements become less correlated. To further investigate this aspect, we report in Figure 6 the smoothed correlation matrix at the begin (January 1995), the middle (June 2005), and end (January 2016) of the sample. We observe that the dependence structure across the series has changed during the observation time, as previously noted. The last Figure 6d reports the correlation matrix implied by the limiting distribution of the series. We see that, in the limit, all cities exhibit positive correlation in the rage 0.2 – 0.7.[12]

### *7.2. Model goodness of fit*

To study the performance of the HMSM we report a goodness of fit analysis. Figure 7 reports a comparison between the limiting cumulative distribution of each city and the corresponding empirical one. We note that these distributions are heterogeneous among them, yet the estimated model is able to accommodate this feature. Univariate limiting distributions are computed starting from the joint limiting distribution - estimated by the HMSM with $J = 4$ and $K = 3$ components - as detailed in Section 3. In order to statistically assess the model's fit, we compute the *nonrandomized* yet uniform probability integral transform for discrete variables defined in Czado et al. (2009) for each univariate time series. These are subsequently compared with the theoretical Uniform distribution divided in $B$ bins exploiting a standard $\chi^2$ test, where the null hypothesis is the equivalence between the empirical

[12] In the supplementary material accompanying this paper, we derive an index of criminal offenses for the NSW State by combining each univariate series. We show that the value of this index has increased considerably over the last years and we propose a policy intervention to reduce its value in the future by exploiting the dependence structure among the NSW cities. The supplementary material also reports a graphical comparison between the empirical and theoretical pmfs, similar to that of Figure 7. Additional results for the comparison of HMSM and the two sub models are also reported.



(A) Sydney (B) Newcastle (C) Wollongong (D) Tweed

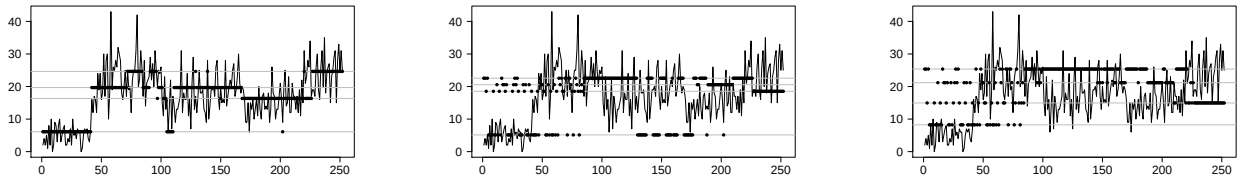
(E) Coffs Harbour (F) Albury (G) Wagga Wagga (H) Tamworth Regional

FIGURE 7: This figure reports a comparison between the empirical and limiting cumulative distribution of the number of monthly weapon offences in eight cities of the NSW State (Australia) over the sample January, 1995 – January, 2016. Results are reported relative to the Hierarchical Markov–Switching model with $J = 4$ and $K = 3$ components. The black lines indicate the empirical cumulative distribution, while the red lines the estimated ones.

and estimated distributions. Associated $p$–values are higher than 95% when $B = 195$, and higher than 99% for a larger number of bins.

### *7.3. Model comparison*

We conclude our empirical illustration by comparing the HMSM model with its two nested specifications reported in Section 4. Model 2 (Equation (15) and Figure 2a) and Model 3 (Equation (16) and Figure 2b) are estimated by setting $J = 4$ and $K = 3$. BIC values are 13733 and 12977 respectively for Model 2 and Model 3, indicating that Model 1 provides the best fit to the data. The difference between the three specifications in terms of state compositions can be studied by means of the Adjusted Rand Index (ARI, Hubert and Arabie, 1985). Overall, the reported ARI values indicate that there is a certain agreement between the state profiles: in particular, we observe a relatively higher agreement between state decodings of Models 1 and 2 (ARI $\approx 0.49$), whereby Model 3 is relatively further apart with an ARI of approximately 0.3 (with Model 1) and 0.27 (with Model 2). In fact, to show that differences are greater than similarities, in Figure 8 we report the decoded means according to each of the three models for the series of Newcastle. The state means estimated by Model 2 are closer with each other providing less separation across the states, whereby both Model 1 and 3's means are more separated with each other. Overall, we observe that Model 1 delivers the most plausible state profiles; Model 2's profiles are relatively better than Model 3's in terms of plausibility, but they adapt more rigidly to changes in the observed series. Indeed, Model 2 predicts much less state switches compared to Model 1.



(A) Model 1 (B) Model 2 (C) Model 3

FIGURE 8: This figure reports a comparison between the hidden state profiles for Model 1 (the HMSM model), Model 2 and Model 3 for the city of Newcastle, with $J = 4$ and $K = 3$. In the background we report the observed series; light gray lines indicate state means, obtained by summing over the second hidden layer, whereby black dots indicate the state (global) decodings at each time point.

## 8. Conclusion

This article showed that flexible dynamic modeling of multivariate nonnegative integer–valued time–series is possible with the HMSM model. Whereas the Markov latent variable captures time dependencies, the static latent variable models cross–dependencies, as well as additional features of the data within each Markov regime. The number of states and the number of components, if no prior knowledge is available, can be chosen based on well–known information criteria (like AIC or BIC). We have worked out the statistical properties of the model, and provided an EM algorithm for ML estimation. The latter, as results from an extended Monte Carlo experiment indicate, has showed good performances in finite samples. Two restricted nested specifications have been derived by modifying our assumptions on the dependence structure between the unobserved and observed random variables.

In the empirical application we analyzed the monthly number of weapon offences in eight cities in the NWS state, Australia. Our results can be summarized as follow: *i*) the HMSM reports a good description of the conditional distribution of crime records, *ii*) it outperforms the standard HMM, and *iii*) it is able to capture the correlation between NSW cities in the number of convictions for weapon offence. We have also compared the main specification with two nested models: results have shown that the HMSM model delivers the most plausible state crime profiles compared to the constrained alternative specifications. In general, we believe the choice of either of the three specifications should be driven by *i*) substantive interpretation of the joint latent process $P(S_t, Z_t)$ and how this should be decomposed, and *ii*) (for Models 1 and 2) considerations related to the within state heterogeneity of the series - as Model 1 can handle substantive within state heterogeneity compared to Model 2 which, in contrast, is much less parametrized.

Our empirical analysis was concerned with modeling variation in weapon offences within cities and heterogeneity across them. As such, we contribute to the literature of crime economics by providing a

sound modeling approach to handle crime time and *spatial* patterns. Although we do not address any causal relation - i.e. *why* such patterns occur - future research might bring our modeling approach into causal inference, as is commonly done in the psychometric literature (see, for instance Marcoulides and Moustaki, 2014).

Several extensions of the model we have proposed are possible. For instance, the performance of the model in forecasting applications can be investigated, whereas possible extensions of the current model might *i*) accommodate for zero–inflated nonnegative integer–valued time–series, *ii*) allow for variables of mixed type, or *iii*) include time–constant and time–varying exogenous regressors. We also note that, for longitudinal categorical data, recently Vidotto et al. (2019) have proposed a mixed latent Markov model for multiple imputation. The underlying modeling idea, similarly to what we do in our work, is that of decomposing the joint distribution of $S_t$ and $Z_t$. However, instead of considering our factorization $P(S_t, Z_t) = P(Z_t|S_t)P(S_t)$, they consider $P(S_t, Z_t) = P(S_t|Z_t)P(Z_t)$. Since their specification turns out to be more parameterized that ours, a comparison between the two seems in place.

**Acknowledgments**

We would like to thank the Editor, the Associate Editor, and three anonymous Reviewers for their helpful and constructive comments, which have helped us improving the manuscript. We would also like to thank Jeroen K. Vermunt and Antonello Maruotti for insightful comments on the paper and interesting suggestions for future works. Our gratitude goes also to Elisabeth Gassiat for helpful discussion on model identification, as well as to Luca Brugnolini, Rob Hyndman, Salvatore Ingrassia, Nicola Loperfido, Robert Jung, Antonio Punzo, Tommaso Proietti, and the participants of the fourth Model-Based Clustering and Classification (MBC$^2$) conference held in Catania (Italy) on September 5-7, 2018.

## Appendix A. HMSM with covariates

Let us assume a vector of $P$ external variables (covariates) $\mathbf{x}_t$ is observed alongside with $\mathbf{y}_t$, for $t = 1, \ldots, T$. The hierarchical Markov switching model of Equation (4) can accommodate covariates by allowing the Poisson rates, the initial, transition and mixture probabilities to have, for instance, the following parametrizations

$$
\begin{aligned}
\log \lambda_{t,j,k,i} &= \eta_{0,i,j,k} + \boldsymbol{\eta}'_{i,j,k}\mathbf{x}_t \\
\log\left[\frac{P(S_1 = j \mid \mathbf{x}_1)}{P(S_1 = 1 \mid \mathbf{x}_1)}\right] &= \log\left[\frac{\delta_j}{\delta_1}\right] = \kappa_{0,j} + \boldsymbol{\kappa}'_j\mathbf{x}_1 && j > 1 \\
\log\left[\frac{P(S_t = j \mid S_{t-1} = h, \mathbf{x}_t)}{P(S_t = h \mid S_{t-1} = h, \mathbf{x}_t)}\right] &= \log\left[\frac{\gamma_{t,h,j}}{\gamma_{t,h,h}}\right] = \xi_{0,h,j} + \boldsymbol{\xi}'_{h,j}\mathbf{x}_t && h \neq j \\
\log\left[\frac{P(Z_t = k \mid S_t = j, \mathbf{x}_t)}{P(Z_t = 1 \mid S_t = j, \mathbf{x}_t)}\right] &= \log\left[\frac{\omega_{t,j,k}}{\omega_{t,j,1}}\right] = \rho_{0,j,k} + \boldsymbol{\rho}'_{j,k}\mathbf{x}_t && k > 1, \qquad \text{(A.1)}
\end{aligned}
$$

where $\delta_1 = 1 - \sum_{j=2}^{J}\delta_j$, $\gamma_{t,h,h} = 1 - \sum_{l\neq h}\gamma_{t,h,l}$ for all $h = 1,\ldots,J$, and $\omega_{t,j,1} = 1 - \sum_{h=2}^{K}\omega_{t,j,h}$, for all $j = 1,\ldots,J$.

We define $\dot{\boldsymbol{\eta}}_{i,j,k} = (\eta_{0,i,j,k}, \boldsymbol{\eta}'_{i,j,k})'$ for all $i = 1,\ldots,D$, $j = 1,\ldots,J$, and $k = 1,\ldots,K$, $\dot{\boldsymbol{\kappa}}_j = (\kappa_{0,j}, \boldsymbol{\kappa}'_j)'$ for all $j = 2,\ldots,J$, $\dot{\boldsymbol{\xi}}_{h,j} = (\xi_{0,h,j}, \boldsymbol{\xi}'_{h,j})'$, for all $j, h = 1,\ldots,J$ and $h \neq j$, and $\dot{\boldsymbol{\rho}}_{j,k} = (\rho_{0,j,k}, \boldsymbol{\rho}'_{j,k})'$ for all $j = 1,\ldots,J$, and $k = 1,\ldots,K$, the model parameters that we collect in a vector $\boldsymbol{\Psi}$ of dimension $[DJK + (J-1) + J(J-1) + J(K-1)](1+P)$.

Given the parametrizations in (A.1) and a sample of observations $\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^{T}$, we specify the log likelihood function as follows

$$
\log \mathcal{L}(\boldsymbol{\Psi}) = \log P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T} | \mathbf{x}_{1:T}; \boldsymbol{\Psi}), \qquad \text{(A.2)}
$$

where the model joint density is now expressed conditioning on the available covariates. In order to find the ML estimate of $\boldsymbol{\Psi}$, Equation (A.2) can be maximized by means of the EM algorithm. With a similar data augmentation as in Section 5, the EM algorithm iteratively maximizes the more tractable Complete–Data Log–Likelihood (CDLL)

$$
\begin{aligned}
\log \mathcal{L}^{\mathsf{c}}(\boldsymbol{\Psi} | \mathbf{y}_{1:T}, \mathbf{x}_{1:T}, \mathbf{u}_{1:T}, \mathbf{v}_{2:T}, \mathbf{z}_{1:T}) = &\sum_{j=1}^{J} u_{j,1}\log(\delta_j) + \sum_{t=2}^{T}\sum_{j=1}^{J}\sum_{h=1}^{J} v_{j,h,t}\log(\gamma_{t,j,h}) \\
&+ \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K} u_{j,t}z_{j,k,t}\log(\omega_{t,j,k}) \\
&+ \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{D} u_{j,t}z_{j,k,t}(y_{i,t}\log(\lambda_{t,j,k,i}) - \lambda_{t,j,k,i} - \log(y_{i,t}!)).
\end{aligned}
\qquad \text{(A.3)}
$$

The E-step consists in computing the expectations of those functions of the missing data that appear in the complete-data log-likelihood given the observations and given the current estimate of $\boldsymbol{\Psi}$. Such expectations are computed similarly to those in Section 5.

Let us consider partitioning the M step into four sub-problems, where the expected complete log-likelihood is maximized with respect to a subset of parameters given the current values of the others. This leads to a (local) maximum because each sub-problem is mathematically independent from the other. The current updates for the model parameters are found by solving the following sets of equations

$$\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{D}\widehat{u}_{j,t}\widehat{z}_{j,k,t}\frac{\partial\log(P(y_{i,t}|S_t=j,Z_t=k,\mathbf{x}_t))}{\partial\dot{\boldsymbol{\eta}}_{j,k,i}}=\mathbf{0}, \tag{A.4}$$

which are $JKD$ weighted Poisson regressions;

$$\sum_{j=1}^{J}\widehat{u}_{j,1}\frac{\partial\log(\delta_j)}{\partial\dot{\boldsymbol{\kappa}}_j}=\mathbf{0}, \tag{A.5}$$

$$\sum_{t=2}^{T}\sum_{j=1}^{J}\sum_{h=1}^{J}\widehat{v}_{j,h,t}\frac{\partial\log(\gamma_{t,j,h})}{\partial\dot{\boldsymbol{\xi}}_h}=\mathbf{0}, \tag{A.6}$$

and

$$\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\widehat{u}_{j,t}\widehat{z}_{j,k,t}\frac{\partial\log(\omega_{t,j,k})}{\partial\dot{\boldsymbol{\rho}}_j}=\mathbf{0} \tag{A.7}$$

which are, respectively, 1, $J$ and $J$ weighted multinomial logit regressions. Each of the Equations (A.4)-(A.7) can be solved, for instance, using a Fisher scoring algorithm, for which the score vector and the expected information matrix are both available in closed form, see for instance Agresti (2003).

E and M steps are alternated until some pre-specified convergence criterion is met.

## Appendix B. Proofs

*Proof.* Proposition 1

The proof of Proposition 1 is made exploiting Theorem 1 of Gassiat et al. (2016). In Theorem 1 of Gassiat et al. (2016) (GCR) identifiability of a general HMM with $J$ discrete states is considered under conditions on the emission distributions and the transition probability matrix of the Markov chain. Their conditions are: GCR1) $J$ is known, GCR2) the emission distributions are linearly independent, and GCR3) the transition probability matrix, $\mathbf{\Gamma}$, has full rank. The proof proceeds by simply checking that conditions GCR1)–GCR3) are satisfied in our context. Our condition *i*) is analogous to GCR1) where also $K$ has been included. Our condition *ii*), $\boldsymbol{\lambda}_{j_1,k_1}\neq\boldsymbol{\lambda}_{j_2,k_2}$, is sufficient for linear independence of our emission distributions (mixtures of products of independent Poisson variates), which is condition GCR2). Finally, our condition *iii*), $S_t$ is irreducible, implies that $\mathbf{\Gamma}$ has full rank, which is condition CGR3). Thus, Theorem 1 of Gassiat et al. (2016) can be applied and identifiability of the model (4) is established. □

**Derivation of $M_{\mathbf{Y}_t}(\mathbf{u})$**

First note that:

$$M_{\mathbf{Y}_t}(\mathbf{u})=\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|\mathbf{Y}_{1:t-1}=\mathbf{y}_{1:t-1}] \tag{B.1}$$

$$=\mathbb{E}[\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t]|\mathbf{Y}_{1:t-1}=\mathbf{y}_{1:t-1}] \tag{B.2}$$

$$=\sum_{j=1}^{J}\pi_{j,t|t-1}\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t=j] \tag{B.3}$$

then by again applying the law of total expectation on $\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t = j]$:

$$\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t] = \mathbb{E}[\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t, Z_t]] \tag{B.4}$$

$$= \sum_{k=1}^{K} \omega_{j,k} \mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t = j, Z_t = k] \tag{B.5}$$

exploiting the conditional independence of the components of $\mathbf{Y}_t$:

$$\mathbb{E}[e^{\mathbf{u}'\mathbf{Y}_t}|S_t = j, Z_t = k] = \prod_{i=1}^{D} \mathbb{E}[e^{Y_{i,t}u_i}|S_t = j, Z_t = k] \tag{B.6}$$

$$= e^{\sum_{i=1}^{D} \lambda_{j,k,i}(e^{u_i}-1)} \tag{B.7}$$

since $\mathbb{E}[e^{Y_{i,t}u_i}|S_t = j, Z_t = k] = e^{\lambda_{j,k,i}(e^{u_i}-1)}$ is the moment generating function of a Poisson distributed random variable. When substituting, we recover:

$$M_{\mathbf{Y}_t}(\mathbf{u}) = \sum_{j=1}^{J} \pi_{j,t|t-1} \sum_{k=1}^{K} \omega_{j,k} e^{\sum_{i=1}^{D} \lambda_{j,k,i}(e^{u_i}-1)}, \tag{B.8}$$

which completes the derivation.

**Derivation of** $Cov(\mathbf{Y}_t, \mathbf{Y}_{t-\tau})$

We first write:

$$Cov(\mathbf{Y}_t, \mathbf{Y}_{t-\tau}) = \mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}] - \mathbb{E}[\mathbf{Y}_t]\mathbb{E}[\mathbf{Y}'_{t-\tau}], \tag{B.9}$$

we indicate by $\boldsymbol{\mu}_\infty$ the expected value of $\mathbf{Y}_t$ which does not depend on $t$, hence $\mathbb{E}[\mathbf{Y}_t] = \mathbb{E}[\mathbf{Y}_{t-\tau}]$. The formula for $\mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}]$ is evaluated by iteratively applying the law of total expectations and the conditional independence of $\mathbf{Y}_t|S_t$ from $\mathbf{Y}_{t-\tau}|S_{t-\tau}$ as follows:

$$\mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}] = \mathbb{E}[\mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}|S_t, S_{t-\tau}]] \tag{B.10}$$

$$= \sum_{j=1}^{J} \sum_{h=1}^{J} \mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}|S_t = j, S_{t-\tau} = h] P(S_t = j, S_{t-\tau} = h) \tag{B.11}$$

$$= \sum_{j=1}^{J} \sum_{h=1}^{J} \mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}|S_t = j, S_{t-\tau} = h] P(S_t = j|S_{t-\tau} = h) P(S_{t-\tau} = h) \tag{B.12}$$

$$= \sum_{j=1}^{J} \sum_{h=1}^{J} \mathbb{E}[\mathbf{Y}_t \mathbf{Y}'_{t-\tau}|S_t = j, S_{t-\tau} = h] [\boldsymbol{\Gamma}^\tau]_{h,j} \pi_{h,\infty} \tag{B.13}$$

$$= \sum_{j=1}^{J} \sum_{h=1}^{J} \mathbb{E}[\mathbf{Y}_t|S_t = j] \mathbb{E}[\mathbf{Y}'_{t-\tau}|S_{t-\tau} = h] [\boldsymbol{\Gamma}^\tau]_{h,j} \pi_{h,\infty} \tag{B.14}$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{h=1}^{J} \sum_{b=1}^{K} \pi_{h,\infty} [\boldsymbol{\Gamma}^\tau]_{h,j} \omega_{j,k} \omega_{h,b} \boldsymbol{\lambda}_{j,k} \boldsymbol{\lambda}'_{h,b}. \tag{B.15}$$

*Proof.* Proposition 2

The proof of Proposition 2 is made exploiting Proposition 2 and Theorem 1 of Gassiat et al. (2016). Proposition 2 of Gassiat et al. (2016) extends their Theorem 1 by considering emission distributions which are mixtures that only depend on the second hidden layer. Their formulation thus resembles our Model 2. Their conditions are: GCR1b) $J$ and $K$ are known and $K \geq J$, GCR2b) the mixture component distributions are linearly independent, GCR3b) the transition probability matrix, $\mathbf{\Gamma}$, has full rank, and CGR4b) $\mathbf{\Omega}$ has rank $J$. The proof proceeds by simply checking that conditions GCR1b)–GCR4b) are satisfied in our context for Model 2. Our condition $i$) is analogous to GCR1b). Our condition $ii$), $\boldsymbol{\lambda}_{\cdot,k_1} \neq \boldsymbol{\lambda}_{\cdot,k_2}$, is sufficient for linear independence of the mixture components (products of Poisson variates), which is condition GCR2b). Our condition $iii$), $S_t$ is irreducible, implies that $\mathbf{\Gamma}$ has full rank, which is condition CGR3b). Our conditions $iv$) is analogous to CGR4b). Thus, Proposition 2 of Gassiat et al. (2016) can be applied and identifiability of Model 2 is established. □

*Proof.* Proposition 3

The proof is analogous to that of Proposition 1 and is thus omitted. □

## Appendix C. E and M steps for ML estimation of the parameters of the two submodels

### *Appendix C.1. Model 2*

The augmenting variables are defined as in Equation (18). Let $\boldsymbol{\alpha}_t' = \boldsymbol{\alpha}_{t-1}'\mathbf{\Gamma}\mathbf{P}_t$, with $\boldsymbol{\alpha}_1' = \boldsymbol{\delta}'\mathbf{P}_1$, be the "*forward probabilities*" of $S_t$ such that $\boldsymbol{\alpha}_t = (\alpha_{j,t}, j = 1, \ldots, J)'$. $\mathbf{P}_t$ is a $J \times J$ diagonal matrix with typical element $p_{j,j,t} = \boldsymbol{\omega}_j'\mathbf{p}_{\cdot,t}$, where $\boldsymbol{\omega}_j = (\omega_{j,k}, \quad k = 1, \ldots, K)'$, and $\mathbf{p}_{\cdot,t} = \left(\prod_{i=1}^{D} P(y_{i,t}|Z_t = k), \quad k = 1, \ldots, K\right)'$. Let the "backward probabilities" vector $\boldsymbol{\beta}_t = (\beta_{j,t}, j = 1, \ldots, J)'$ be such that $\boldsymbol{\beta}_t = \mathbf{\Gamma}\mathbf{P}_{t+1}\boldsymbol{\beta}_{t+1}$, where $\boldsymbol{\beta}_T = \mathbf{1}$. As before, we let the initial distribution be equal to the stationary distribution, that is $\boldsymbol{\delta} = \boldsymbol{\pi}_\infty$. The E and M steps of the algorithm are as follows

Step E.

- $\widehat{u}_{j,t} = P(S_t = j|\mathbf{y}_{1:T}, \mathbf{\Theta}^{(m)}) = \alpha_{j,t}\beta_{j,t}/(\boldsymbol{\alpha}_T'\mathbf{1})$
- $\widehat{v}_{j,l,t} = P(S_{t-1} = j, S_t = l|\mathbf{y}_{1:T}, \mathbf{\Theta}^{(m)}) = \alpha_{j,t-1}\gamma_{j,l}p_{l,l,t}\beta_{j,t}/(\boldsymbol{\alpha}_T'\mathbf{1})$
- $\widehat{z}_{j,k,t} = P(Z_t = k|S_t = j, \mathbf{y}_{1:T}, \mathbf{\Theta}^{(m)}) = \frac{\omega_{j,k}[\mathbf{p}_{\cdot,t}]_k}{\sum_{l=1}^{K}\omega_{j,l}[\mathbf{p}_{\cdot,t}]_l}$.

Step M

- $\gamma_{j,l}^{(m+1)} = \frac{\sum_{t=2}^{T}\widehat{v}_{j,l,t}}{\sum_{l=1}^{J}\sum_{t=2}^{T}\widehat{v}_{j,l,t}}$
- $\omega_{j,k}^{(m+1)} = \frac{\sum_{t=1}^{T}\widehat{u}_{j,t}\widehat{z}_{j,k,t}}{\sum_{t=1}^{T}\sum_{l=1}^{J}\widehat{u}_{l,t}\widehat{z}_{l,k,t}}$
- $\lambda_{\cdot,k,i}^{(m+1)} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\widehat{u}_{j,t}\widehat{z}_{j,k,t}y_{i,t}}{\sum_{t=1}^{T}\sum_{l=1}^{J}\widehat{u}_{l,t}\widehat{z}_{l,k,t}}$.

*Appendix C.2. Model 3*

Let us define

$$z_{k,t} = \begin{cases} 1, & \text{if } Z_t = k \\ 0, & \text{otherwise.} \end{cases} \tag{C.1}$$

The other augmenting variables are defined as in Equations (18). Let $\boldsymbol{\alpha}'_t = \boldsymbol{\alpha}'_{t-1}\boldsymbol{\Gamma}\mathbf{P}_t$, with $\boldsymbol{\alpha}'_1 = \boldsymbol{\delta}'\mathbf{P}_1$, be the "*forward probabilities*" of $S_t$ such that $\boldsymbol{\alpha}_t = (\alpha_{j,t}, j = 1, \ldots, J)'$, where $\alpha_{j,t} = P(\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}, S_t = j)$ and $\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}$ indicates $\mathbf{Y}_s = \mathbf{y}_s$ for $s = 1, \ldots, t$. $\mathbf{P}_t$ is a $J \times J$ diagonal matrix with typical element $p_{j,j,t} = \boldsymbol{\omega}'\mathbf{p}_{j,t}$, where $\boldsymbol{\omega} = (\omega_{.,k}, \quad k = 1, \ldots, K)'$, and $\mathbf{p}_{j,t} = \left(\prod_{i=1}^{D} P(y_{i,t}|S_t = j, Z_t = k), \quad k = 1, \ldots, K\right)'$. As before, we let the initial distribution be equal to the stationary distribution, that is $\boldsymbol{\delta} = \boldsymbol{\pi}_\infty$. The E and M steps of the algorithm are as follows

Step E:

- $\widehat{u}_{j,t} = P(S_t = j|\mathbf{y}_{1:T}) = \alpha_{j,t}\beta_{j,t}/(\boldsymbol{\alpha}'_T\mathbf{1})$
- $\widehat{v}_{j,l,t} = P(S_{t-1} = j, S_t = l|\mathbf{y}_{1:T}, \boldsymbol{\Theta}^{(m)}) = \alpha_{j,t-1}\gamma_{j,l}P(\mathbf{y}_t|S_t = l)\beta_{j,t}/(\boldsymbol{\alpha}'_T\mathbf{1})$
- $\widehat{z}_{k,t} = \frac{\sum_{j=1}^{J}\delta_j\sum_{h=1}^{J}[\boldsymbol{\Gamma}^t]_{h,j}\omega_{.,k}[\mathbf{p}_{j,t}]_k}{\sum_{j=1}^{J}\delta_j\sum_{h=1}^{J}[\boldsymbol{\Gamma}^t]_{h,j}\sum_{m=1}^{K}\omega_{.,m}[\mathbf{p}_{j,t}]_m}$

Step M:

$$\gamma_{j,l}^{(m+1)} = \frac{\sum_{t=2}^{T}\widehat{v}_{j,l,t}}{\sum_{l=1}^{J}\sum_{t=2}^{T}\widehat{v}_{j,l,t}} \tag{C.2}$$

$$\omega_k^{(m+1)} = \frac{1}{T}\sum_{t=1}^{T}\widehat{z}_{k,t} \tag{C.3}$$

$$\lambda_{j,k,i}^{(m+1)} = \frac{\sum_{t=1}^{T}\widehat{u}_{j,t}\widehat{z}_{j,k,t}y_{i,t}}{\sum_{t=1}^{T}\sum_{l=1}^{J}\widehat{u}_{l,t}\widehat{z}_{l,k,t}}. \tag{C.4}$$

## Appendix D. Model selection with BIC

| | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ | $K=7$ | $K=8$ |
|---|---|---|---|---|---|---|---|---|
| $J=1$ | 17508 | 14764 | 13920 | 13654 | 13452 | 13359 | 13299 | 13272 |
| $J=2$ | 14588 | 13495 | 13205 | 13117 | 13083 | 13075 | 13070 | 13101 |
| $J=3$ | 13702 | 13121 | 13012 | 13008 | 13039 | 13062 | 13123 | 13172 |
| $J=4$ | 13400 | 13005 | 12971 | 13029 | 13094 | 13182 | 13306 | 13403 |
| $J=5$ | 13198 | 12976 | 12973 | 13101 | 13211 | 13373 | 13540 | 13701 |
| $J=6$ | 13102 | 13000 | 13082 | 13245 | 13386 | 13580 | 13787 | 14023 |
| $J=7$ | 13102 | 13058 | 13181 | 13371 | 13562 | 13816 | 14096 | 14314 |
| $J=8$ | 13129 | 13128 | 13301 | 13517 | 13815 | 14117 | 14428 | 14745 |

Table D.1: Bayesian information criterion (BIC) for the Hierarchical Markov Switching model with $J \in (1, \ldots, 8)$ and $K \in (1, \ldots, 8)$ estimated using the monthly number of weapon offences in eight cities of the NSW state (Australia) in the period spanning from January, 1995 to January, 2016, for a total of 256 observations for each series. Model estimation is performed via the EM algorithm detailed in Section 5 of the accompanying paper. The model specification with the lowest BIC value is indicated in gray. BIC values associated to the standard $J$–states Hidden Markov Model are reported in the first column ($K = 1$).

## References


Adamidis, K. (1999). Theory & methods: An em algorithm for estimating negative binomial parameters. *Australian & New Zealand Journal of Statistics*, 41(2):213–221.

Agresti, A. (2003). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6(3):251–262.

Al-Osh, M. and Alzaid, A. (1987). Firstorder integervalued autoregressive process. *Journal of Time Series Analysis*, 8(3):261–275.

Alexandrovich, G., Holzmann, H., and Leister, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden markov models. *Biometrika*, 103(2):423–434.

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Altman, R. M. (2007). Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

Bartolucci, F. (2006). Likelihood inference for a class of latent markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):155–178.

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

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):115–132.

Bu, R. and McCabe, B. (2008). Model selection, estimation and forecasting in inar (p) models: a likelihood-based markov chain approach. *International Journal of Forecasting*, 24(1):151–162.

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

Bulla, J., Chesneau, C., and Kachour, M. (2017). A bivariate first-order signed integer-valued autoregressive process. *Communications in Statistics-Theory and Methods*, 46(13):6590–6604.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York.

Carcach, C. and Muscat, G. (2000). An analysis of regional variations in crime using crime concentration indexes. *Proceedings of Crime mapping: Adding value to crime prevention and control*.

Csiszár, I. and Shields, P. C. (2000). The consistency of the bic markov order estimator. *The Annals of Statistics*, 28(6):1601–1619.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Di Mari, R. and Bakk, Z. (2017). Mostly harmless direct effects: a comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Di Mari, R., Oberski, D. L., and Vermunt, J. K. (2016). Bias–adjusted three–step latent markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5):649–660.

Duggan, M. (2001). More guns, more crime. *Journal of political Economy*, 109(5):1086–1114.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. CRC press.

Fernández-Fontelo, A., Cabaña, A., Puig, P., and Moriña, D. (2016). Under–reported data analysis with inar–hidden markov chains. *Statistics in Medicine*, 35(26):4875–4890.

Fokianos, K., Rahbek, A., and Tjstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Fokianos, K., Stø ve, B., Tjø stheim, D., and Doukhan, P. (2019). Multivariate count autoregression. *Bernoulli (Forthcoming)*.

Freeman, R. B. (1999). The economics of crime. volume 3 of *Handbook of Labor Economics*, pages 3529 – 3571. Elsevier.

Freeman, R. B. (2006). People flows in globalization. *Journal of Economic Perspectives*, 20(2):145–170.

Friedberg, R. M. (2001). The impact of mass migration on the israeli labor market. *The Quarterly Journal of Economics*, 116(4):1373–1408.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Science & Business Media.

Gassiat, É., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1-2):61–71.

Geweke, J. and Amisano, G. (2011). Hierarchical markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26(1):1–29.

Glaeser, E. L. and Sacerdote, B. (1999). Why is there more crime in cities? *Journal of political economy*, 107(S6):S225–S258.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jones, C., Kypri, K., Moffatt, S., Borzycki, C., and Price, B. (2009). The impact of restricted alcohol availability on alcohol–related violence in newcastle, nsw. *NSW Bureau of Crime Statistics and Research*.

Jung, R. C., Liesenfeld, R., and Richard, J.-F. (2011). Dynamic factor models for multivariate count data: An application to stock-market trading activity. *Journal of Business and Economic Statistics*, 29(1):73–85.

Kadane, J. B. (1985). Is victimization chronic? a bayesian analysis of multinomial missing data. *Journal of Econometrics*, 29(1-2):47–67.

Karlis, D. (2015). *Models for multivariate count time series*, pages 407–424. Chapman & Hall: Boca Raton, FL.

Kocherlakota, S. and Kocherlakota, K. (1992). Bivariate discrete distributions. volume 132 of *Statistics: Textbooks and Monographs*. Markel Dekker, New York.

Lagona, F., Jdanov, D., and Shkolnikova, M. (2014). Latent time-varying factors in longitudinal analysis: a linear mixed hidden markov model for heart rates. *Statistics in Medicine*, 33(23):4116–4134.

Levitt, S. D. (2017). The economics of crime. *Journal of Political Economy*, 125(6):1920–1925.

Marcoulides, G. A. and Moustaki, I. (2014). *Latent variable and latent structure models*. Psychology Press.

Marino, M. F. and Alfó, M. (2016). Gaussian quadrature approximations in mixed hidden markov models for longitudinal data: A simulation study. *Computational Statistics & Data Analysis*, 94:193–209.

Maruotti, A. (2011). Mixed hidden markov models for longitudinal data: an overview. *International Statistical Review*, 79(3):427–454.

Maruotti, A. and Rydén, T. (2008). A semiparametric approach to hidden markov models under longitudinal observations. *Statistics and Computing*, 19(4):381.

Mayhew, P. (2003). *Counting the Costs of Crime in Australia*. Australian Institute of Criminology.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Nelsen, R. B. (2006). *An Introduction to Copulas, 2nd edn.* SpringerVerlag, New York.

Olteanu, M. and Rynkiewicz, J. (2012). Asymptotic properties of autoregressive regime-switching models. *ESAIM: PS*, 16:25–47.

Pedeli, X. and Karlis, D. (2011). A bivariate INAR (1) process with application. *Statistical Modelling*, 11(4):325–349.

Pedeli, X. and Karlis, D. (2013a). On composite likelihood estimation of a multivariate INAR (1) model. *Journal of Time Series Analysis*, 34(2):206–220.

Pedeli, X. and Karlis, D. (2013b). Some properties of multivariate INAR (1) processes. *Computational Statistics and Data Analysis*, 67:213–225.

Quenouille, M. H. (1949). A relation between the logarithmic, poisson, and negative binomial series. *Biometrics*, 5(2):162–164.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Rydberg, T. H. and Shephard, N. (2000). A modelling framework for the prices and times of trades made on the new york stock exchange. *Nonlinear and Nonstationary Signal Processing*, pages 217–246.

Sah, R. K. (1991). Social osmosis and patterns of crime. *Journal of political Economy*, 99(6):1272–1295.

Scotto, M. G., Weiss, C., and Gouveia, S. (2015). Thinning–based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, 15(6):590–618.

Sickles, R. C. and Williams, J. (2008). Turning from crime: A dynamic perspective. *Journal of Econometrics*, 145(1):158–173.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.

Vermunt, J. K., Langeheine, R., and B ockenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24:179–207.

Vidotto, D., Vermunt, J. K., and Deun, K. V. (2019). Multiple imputation of longitudinal categorical data through bayesian mixture latent markov models. *Journal of Applied Statistics*.

Weatherburn, D., Jones, C., Freeman, K., and Makkai, T. (2003). Supply control and harm reduction: lessons from the australian heroin drought. *Addiction*, 98(1):83–91.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.