UNIVERSITA' COMMERCIALE "LUIGI BOCCONI" - MILANO

Dipartimento di Scienze delle Decisioni

Ph.D. in Statistics - XXI Cycle

# Bayesian hidden Markov models for performance-based analysis of electricity supply

Nadia Accoto

Matr. 1094846

Tutor: Sonia Petrone

“*Essentially, all models are wrong,*
*but some are useful*”
George Box

# Acknowledgements

...and now I’m here at the end of a route held over three years. Writing the thesis, even if it is a challenging experience, it is very interesting and stimulating. Like a little plant it needs the seed, the ground, but also to be grown and irradiated by warn sunbeams.

First of all I would like to thank professor Tobias Rydén, Lund University, and professor Piercesare Secchi, Politecnico di Milano. To them I say “Thank you” for helping, supporting and encouraging me, also when results were discouraging and over all I acknowledge them for saying me “Good job!!!”, that sometimes is worth more than technical or practical suggestions.

I acknowledge the Italian Regulatory Authority for Electricity and Gas for providing me the data, seed and row material of the whole work; in particular I would like to thank Ing. Luca Lo Schiavo.

Moreover, I thank who arranged the ground, all professors from whom I learned a lot during the first years; I thank professor Pietro Muliere and professor Sonia Petrone for the care and dedication used in their work.

I really thank who studied, struggled, but also smiled with me during this period: Emiliano, Lorenzo and Mattia. To them I say “Good luck”and I wish they will find their way.

I thank my family for the support and the constant presence; thank you for being there, as a rock to hang on, stop and recover the strength for a new adventure.

Finally I thank Ale, the inebriating light of my life.

# Ringraziamenti

...ed eccomi qua alla fine di questo percorso durato tre anni. L'esperienza della tesi, seppur impegnativa, è quella più interessante, stimolante. E come una piantina ha bisogno del seme, del terreno, ma anche e soprattutto di essere accudita e irradiata da caldi raggi solari.

Prima di tutto ringrazio il professor Tobias Rydén, della Lund University, e il professor Piercesare Secchi, del Politecnico di Milano. A loro dico grazie per avermi aiutato, sostenuto ed incoraggiato anche quando i risultati erano sconfortanti e soprattutto ringrazio loro per quel "Buon lavoro!!!", che a volte vale più di mille suggerimenti tecnici o pratici.

Ringrazio l'Autorità per l'Energia Elettrica e il Gas, per aver fornito i dati, seme e materia prima dell'intero lavoro. In particolar modo vorrei ringraziare l'Ing. Luca Lo Schiavo.

Ringrazio inoltre chi ha preparato il terreno, tutti i docenti da cui ho avuto modo di imparare tanto durante i primi anni di dottorato; ringrazio in particolar modo il professor Pietro Muliere e la professoressa Sonia Petrone per l'impegno e la dedizione con cui svolgono il loro lavoro.

Ringrazio di cuore chi con me ha studiato, lottato, ma anche sorriso nella stanzetta di Viale Isonzo: Emiliano, Lorenzo e Mattia. A loro va un sentito "in bocca al lupo" e a tutti loro auguro di trovare la propria strada.

Ringrazio la mia famiglia per il sostegno e la presenza costante; grazie per essere lì, come uno scoglio a cui aggrapparsi, sostare e riprendere le forze per un nuovo viaggio.

Ringrazio infine Ale, sole inebriante della mia vita.

# Abstract

In this work different hidden Markov models are proposed for identifying exceptional events in the electricity distribution. We propose a way for evaluating the utilities restoration schemes, by studying the distribution of the time needed to the electricity distribution utility to reestablish the normal operating status, given that an exceptional event occurs.

A Cluster analysis is performed and interpretation of the generated clusters is provided. With the same goal we introduce the hidden mixture Markov model, a model-based method for clustering utilities by means of the Markov chain's transitional dynamic.

Finally, a prior specification in Bayesian hidden Markov model, based on the Reinforced Urn Processes, is considered.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

## 1.1 Presentation of the problem

Quality regulation in electricity distribution has received significant attention in recent years, since, in 1999, a reliability performance regulation was introduced for the first time. Analyzes of continuity of supply indicators are fundamental for setting regulatory targets, monitoring utility performance and disseminating information to the public (CEER, 2005; Fumagalli *et al.* 2007, 2008).

One of the main problems in these analyzes is how to identify events that are exceptional with respect to the normal performance. The exclusion of these extreme cases from the dataset enables utilities, regulators and the public to observe more meaningful trends in 'normal operation' performance, that would be, otherwise, hard to capture. In addition regulators usually design incentive mechanisms specifically targeted at quality supply, assuming the form of financial penalties and rewards to the distribution utility based on the expected performance. It is therefore crucial to understand when failure in meeting the regulatory targets is due to the utility behavior or to events that are outside the utility's control and that could be considered exceptional. Moreover, even if some events, such as extreme weather conditions, are unavoidable, regulators have become more interested in controlling the efficiency and effectiveness of utility restoration schemes.

Traditional criteria for identifying data due to exceptional situations are based on definitions of exceptional events, given in terms of number of customers interrupted, duration of the interruptions, weather conditions, extent of the mechanical damage to the distribution system and combinations of these factors. The application of this criterion, however, resulted in some practical cases quite difficult and ambiguous. For these reasons the introduction of statistical methodologies was suggested; a statistical approach, in fact, is expected to present significant advantages because of a reduction in ambiguities and an increase in fairness. Nevertheless, statistical analyzes of exceptional events can be performed in very different ways, depending on the choice of the quality indicator, on the spatial and temporal units of such measure, and, of course, on the statistical methods employed.

Hidden Markov Models (HMMs) describe the relationship between an observed process $\{Y_t\}_{t>0}$ and an underlying and unobserved process $\{X_t\}_{t\geq 0}$; the hidden process is assumed to follow a Markov chain whose realization $X_k$ governs the distribution of the correspondent $Y_k$. For example, when the Business Cycle is under analysis it is possible to assume that the gross national product depends on whether the economy is expanding or contracting; therefore an HMM with a two state Markov chain can be considered: $X_k = 1$ for the expansion status and $X_k = 0$ for the regression status (Hamilton, 1989).

The interpretation of the electrical faults as a signal of an underlying and not observable process naturally leads to analyze the problem by means of an HMM. In other words, in order to attain the regulators goals, we propose a statistical method based on the idea that the number of interruptions depends on the latent status of the "global" system controlling the electricity distribution; the latent system operating status along time is modeled by a Markov chain, whose realization controls the distribution of the corresponding number of observed faults.

In Section 1.2 we will present the method adopted by the Italian Regulatory Authority (Autorità per l'energia elettrica e il gas - AEEG) in order to identify exceptional events, for the third regulatory period 2008-2011. In Sections 1.3 and 1.4 we will introduce the HMM and how to make inference in a Bayesian framework. The choice of the number of possible states of the Markov chain will be the objective of Section 1.5. In Section 1.6 we will analyze the time needed to the system to reestablish the normal situation, given that an exceptional event occurred. Finally, in Section 1.7 the organization of the work will be explained.

## 1.2 The method adopted by the Italian Regulatory Authority

Since in 2000 the Italian Regulatory Authority introduced the first continuity of supply regulation, different methods have been proposed, applied and in some cases criticized and improved (see Fumagalli *et al.*, 2008 for a review).

Before presenting the method adopted by the AEEG for the third Regulatory period and in order to better understand it, we briefly describe the methods proposed during the years.

In the first regulatory period (2000-2003), the regulation required companies to classify interruptions according to three categories: *Force Majeure*, external causes (*i.e.* third party responsibilities and interruptions originated on the transmission grid) and utility responsibility. AEEG accepted a *Force Majeure* attribution only if the exceptional nature of the event could be proven by technical or administrative evidence. For instance, a formal declaration of calamity made by the government or measures of wind speed made by an independent weather center. In practical terms, this procedure turned out to be onerous both for the companies, that were collecting the data, and for the regulatory authority, that was controlling the documentation provided. In addition, a few controversial cases, where the exceptional nature of the event was claimed by the companies, but could not be formally proven, generated a large amount of disputes.

For these reasons, the AEEG began to consider a simpler procedure for identifying an exceptional event on the basis of the nature of the interruption it caused, compared to the characteristics of the interruptions caused by "normal events". Then, in 2004, for the regulatory period 2004 - 2007, AEEG introduced a statistical methodology for identifying exceptional events, based on the idea that such events are characterized by a longer-than-average restoration time (Christie (2003), Warren *et al.* (2003), Fumagalli *et al.* (2006)). Even if a significant reduction in administrative work resulted, on both the regulator's and the utilities' side, the method showed some empirical problems. In particular, consumers found confusing that the same event, when affecting different districts within smaller geographical proximity, might result in exclusion of minutes lost in some, but not in other districts; moreover, for a small number of districts the methodology was found unable to identify events that the companies would have classified as exceptional, on the basis of their knowledge and experience.

During the consultation process for the third regulatory period additional elements emerged that lead to a new statistical analysis of exceptional events; the new proposed method was incorporated by the AEEG, in the Regulatory Order 333/07 (AEEG, 2007).

In this method a 6 hour interval is deemed to be exceptional when the number of faults registered in that interval is larger than an *exceptionality threshold*; the procedure for the computation of these thresholds is briefly presented in the following.

First of all, for each electricity distribution utility, the distribution of the number $X$ of faults in 6 hours time-interval is clustered in two groups: that for Ordinary Interval (OI) and that for Exceptional Interval (EI). The threshold separating the two clusters is computed using a $k$-means algorithm with $k = 2$.

Hence the distribution for the OI number of faults is modeled with a geometric distribution with parameter $(1 - e^{-\lambda})$

$$P(X = h) = (1 - e^{-\lambda})e^{-\lambda h}$$

where $h \in \{0, 1, \ldots\}$. The parameter $\lambda$ is estimated with the maximum likelihood estimator

$$\hat{\lambda} = -\log\left(\frac{\bar{\mu}_1}{1 + \bar{\mu}_1}\right),$$

where $\bar{\mu}_1$ is the mean of the observations of $X$ smaller than the threshold separating the two groups (*i.e.* it is the mean of the observations in the first OI-cluster).

Finally a quantile $q_\alpha$ of the fitted geometric distribution for the OI number of faults is computed and an interval is declared exceptional for the utility under exam when the number of faults observed in the period is larger than $q_\alpha$, which is therefore called the *exceptionality threshold*. In particular $\alpha$ is set so that, according to the fitted distribution for the OI number of faults, a value of $X$ greater than $q_\alpha$ would be seen once every $t$ years, with $t$ large; for example suppose that an event is considered exceptional if it happens every 20 years (*i.e.* $t = 20$), then $\alpha = 1 - \frac{1}{365*4*20} = 0.9999658$. Therefore, by the definition of quantile, the *exceptionality threshold* $q_\alpha$ is the minimum real value such that $P(X \leq q_\alpha) = \alpha$: for each utility

$$q_\alpha \quad = \quad \left[ -1 - \frac{\log(1-\alpha)}{\hat{\lambda}} \right]$$

where $[x]$ indicates the integer part of $x$.

Moreover, for ease of elicitation by the AEEG $q_\alpha$ is expressed as a linear function of the average number $m$ of faults per 6 hours period computed over the available years

$$q_\alpha = \beta_0 + \beta_1 * m$$

with $\beta_0 \neq 0$.

Given the EIs (where the number of faults occurred is greater than $q_\alpha$) an Exceptional Period (EP) is defined considering 3 hours before the beginning of an EI and 3 hours after the end of the same interval. Then the methodology labels as an EP a larger time span and in the simplest case (where there are no contiguous EIs) the EP covers a period of 12 hours.

The application of a statistical method, even if it is simple or very sophisticated, needs the understanding of some technical aspect of the field in which the method will be applied. For this reason before introducing the HMM we focus our attention on the method applied by the AEEG for gathering considerations that are useful to the implementation of the analysis we will perform. Then we underline that the method adopted by the Authority analyzes each utility separately from the others and using only the observed performance, without considering any other type of information. Moreover this method incorporates the idea that an exceptional event causes several faults protracting in time; in fact it considers the number of interruptions greater than a threshold and the analysis are based on the interruptions occurred in a 6 hours time interval. Finally enlarging an exceptional interval to the 3 hours before the beginning and 3 hours after the end of the same interval, the AEEG method incorporates the idea that the exceptional event is preceded and followed by some instability conditions.

THE AEEG method in general and these considerations in particular will be the reference frame of the analysis we will perform.

## 1.3 The model

We now introduce the general properties of the model; the book by Cappé *et al.* (2005) represents a complete and clear reference on HMMs.

An Hidden Markov Model (HMM) is a bivariate discrete time process $\{X_t, Y_t\}$, where $\{X_t\}_{t \geq 0}$ is a Markov chain and, conditional on $\{X_t\}$, $\{Y_t\}_{t>0}$ is a sequence of independent random variable such that the conditional distribution of $Y_k$ only depends on $X_k$. The dependence structure of an HMM can be represented by a graphical model as in Figure 1.1, in which nodes are the random variables, and the missing edges between the nodes represent conditional independencies.

$$\begin{tikzcd}
\cdots \arrow[r] & X_{k-1} \arrow[r] \arrow[d] & X_k \arrow[r] \arrow[d] & X_{k+1} \arrow[r] \arrow[d] & \cdots \\
\cdots & Y_{k-1} & Y_k & Y_{k+1} & \cdots
\end{tikzcd}$$

**Figure 1.1:** Graphical representation of the dependence structure of an HMM, where $\{Y_t\}$ is the observable process and $\{X_t\}$ is the hidden Markov chain.

In other words, the distribution of a variable $X_{k+1}$ conditionally on the history of the process, $X_0, \dots, X_k$, is determined by the value taken by the preceding one, $X_k$ (Markov property); likewise, the distribution of $Y_k$, conditionally on the past observations $Y_1, \dots, Y_{k-1}$ and the past value of the state $X_0, \dots, X_k$, is determined by $X_k$ only.

Denote the state space of the Markov chain $\{X_t\}$ by $\mathsf{X}$ and the set in which $\{Y_t\}$ takes its values by $\mathsf{Y}$.

When $\mathsf{X}$ and $\mathsf{Y}$ are finite sets we have a *finite hidden Markov model*, that can be characterized by the initial state distribution $\pi = \{\pi_i\}$, with $\pi_i = P(X_0 = i)$, $i \in \mathsf{X}$, the transition matrix $A = \{a_{i,j}\}$, where $a_{i,j} = P(X_{k+1} = j | X_k = i)$,

$i, j \in \mathsf{X}$ and the emission matrix $B = \{b_i(y)\}$, with the conditional probabilities $b_i(y) = P(Y_k = y|X_k = i)$, $i \in \mathsf{X}$, $y \in \mathsf{Y}$.

A *parametric hidden Markov model* assumes that the conditional distributions of $Y_k$ given $X_k$ all belong to a single parametric family, with parameters indexed by $X_k$, *i.e.* $Y_k|X_k = i \sim f(y_k|\,\xi_i)$. In this case the HMM is characterized by the initial state distribution $\pi$, the transition matrix $A$ and the emission parameters $\boldsymbol{\xi}$; for example, in the Poisson HMM with a four state Markov chain, $Y_k|X_k = i \sim \text{Pois}(y_t|\lambda_i)$ and $\boldsymbol{\xi} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$.

Usually it is assumed that the state space of the hidden Markov chain, $\mathsf{X}$, is finite; Beal *et. al.* (2005) hypothesizes that $\mathsf{X}$ is countably infinite and introduces the *infinite hidden Markov models.*

HMMs are widely used in a variety of fields for modeling dependence in data and for analyzing (observed) phenomena depending on an underlying and not observable system. An HMM generalizes the classical mixture model and the component populations, from one observation to the next, are selected according to an unobserved Markov chain. Moreover, they are used to study situations where the hidden Markov chain has a physical meaning (it is not just a tool for introducing dependence in data).

Besides economics (Hamilton 1989, 1990; Albert and Chib, 1993), HMMs have been applied to some specific areas such as signal processing (Juang and Rabiner, 1991), biology (Leroux and Puterman, 1992), genetics (Churchill, 1989; Liu *et al.*, 1999; Guha *et al.*, 2008).

### 1.3.1 The likelihood function

Let $\boldsymbol{\vartheta}$ be a vector containing the model parameters, arising from the transition mechanism (*i.e.* the transition matrix $A$ and the initial state probability distribution $\pi$) and from the emission mechanism (*i.e.* the emission matrix $B$ or the emission parameters $\boldsymbol{\xi}$).

Consider a sequence of length $T$ and let $(\boldsymbol{y}, \boldsymbol{X})$ be the complete-data, $(Y_1 = y_1, \ldots, Y_T = y_T, X_0 = x_0, X_1 = x_1, \ldots, X_T = x_T)$; then the complete-data likelihood function $p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta})$ is given by

$$p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})p(\boldsymbol{X}|\boldsymbol{\vartheta}). \tag{1.1}$$

The density $p(\boldsymbol{X}|\boldsymbol{\vartheta})$ depends on $A$ only and so

$$\begin{aligned} p(\boldsymbol{X}|A) &= \prod_{t=1}^{T} p(X_t|X_{t-1}, A)p(X_0|A) = \pi_{x_0} \prod_{t=1}^{T} a_{x_{t-1},x_t} \\ &= \pi_{x_0} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{i,j}^{n_{ij}} \end{aligned} \tag{1.2}$$

where $K$ is the number of possible states ($i.e.$ $K = |\mathsf{X}|$) and $n_{ij} = \#\{1 \leq t \leq T : X_{t-1} = i, X_t = j\}$, $\forall i, j \in \mathsf{X}$.

The probability distribution $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})$, in 1.1, depends on the hypothesized emission structure, coming from the assumed finite or parametric HMM; moreover, because conditionally on $\boldsymbol{X}$ the random variables $Y_1, \ldots, Y_T$ are independent, $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{X}, \boldsymbol{\vartheta})$.

Therefore equation (1.1) is

$$p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}) = \pi_{x_0} \prod_{k=1}^{K} \left( \prod_{t:X_t=k} b_k(y_t) \right) \prod_{i=1}^{K} \prod_{j=1}^{K} a_{i,j}^{n_{ij}}$$

in the finite HMM and

$$p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}) = \pi_{x_0} \prod_{k=1}^{K} \left( \prod_{t:X_t=k} p(y_t|\xi_k) \right) \prod_{i=1}^{K} \prod_{j=1}^{K} a_{i,j}^{n_{ij}} \tag{1.3}$$

in the parametric case. For example, when a Poisson HMM is considered, equation (1.3) becomes

$$p(\boldsymbol{y}, \boldsymbol{X}|\lambda_1, \ldots, \lambda_K, A, \pi) \propto \pi_{x_0} \prod_{k=1}^{K} \lambda_k^{n_k \overline{y}_k} e^{-n_k \lambda_k} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{i,j}^{n_{ij}},$$

where $n_k = \#\{1 \leq t \leq T : X_t = k\}$ and $\overline{y}_k$ is the mean of all observations, when $X_t = k$.

Note that we are assuming that the parametric distribution depends on a single parameter ($\xi_k$ is not a vector in (1.3)), but it could be a two or more parameters family.

Summing the complete-data likelihood function over all possible hidden state sequences we obtain the likelihood function

$$p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \sum_{\boldsymbol{X} \in \mathsf{X}^{T+1}} p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}), \tag{1.4}$$

where $\mathsf{X}^{T+1}$ is the space of all possible realizations of $\boldsymbol{X}$. The sum in (1.4) is over $K^{T+1}$ elements and becomes infeasible for practical evaluation of the likelihood function, even for small values of $K$ and $T$.

A solution to the problem of computing the likelihood function (1.4) is provided by the forward-backward recursions of Baum *et al.* (1970) (for a clear presentation of the procedure see the important tutorial of Rabiner, 1989).

## 1.4 Bayesian inference

When an HMM is considered, problems of interest are inference on the model parameters $\boldsymbol{\vartheta}$ and on the unobserved chain $\{X_t\}$. For many years HMMs have been implemented using recursive algorithms developed for parameter estimation (Baum and Petrie, 1966; Baum *et al.*, 1970) and for restoring the hidden Markov chain (Viterbi, 1967). More recently, these models have been studied from a Bayesian point of view (among others Robert *et al.* 1993, 2000; Chib, 1996).

In a Bayesian approach, model parameters are random quantities, on which a prior has to be assigned; a standard prior assumption is that the emission parameters ($B$ or $\boldsymbol{\xi}$) are a priori independent of the transition matrix $A$: $p(\boldsymbol{\vartheta}) = p(A)p(B)$ or $p(\boldsymbol{\vartheta}) = p(A)p(\xi_1, \ldots, \xi_K)$. Inference on the model parameters and on the hidden chain is based on the posterior distribution

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{\vartheta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})p(\boldsymbol{X}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}) \\ &\propto p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}). \end{aligned} \tag{1.5}$$

Sampling from the posterior (1.5) is commonly carried out by the Markov Chain Monte Carlo (MCMC) sampling scheme. MCMC sampler is a powerful and widely used method for iteratively sampling from posterior distributions. The original MCMC algorithm was introduced by Metropolis *et al.* (1953) for the purpose of optimization on a discrete state space. The Metropolis algorithm was later generalized by Hastings (1970) and Peskun (1973, 1981) to statistical simulation. Despite several other papers that highlighted its usefulness in specific settings (see, for example, Geman and Geman, 1984; Tanner and Wong, 1987; Besag, 1989), the starting point for an intensive use of MCMC methods by the statistical community can be traced to the presentation of the Gibbs sampler by Gelfand and Smith (1990). Several other algorithmic approaches are available, such as, slice (Neal, 2003), and adaptive rejection sampling (Gilks and Wild, 1992).

HMM's missing-data structure naturally admits posterior samplers that alternate between simulating $\boldsymbol{X}$ given $\boldsymbol{\vartheta}$ and $\boldsymbol{y}$, and simulating $\boldsymbol{\vartheta}$ given the complete-data $\boldsymbol{X}$ and $\boldsymbol{y}$; then we can consider the following (general) Gibbs sampling algorithm.

**Algorithm 1.1** Gibbs sampling

Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \dots, L_0, \dots, L$.

1. Sample $\boldsymbol{\vartheta}$ from the complete-data posterior $p(\boldsymbol{\vartheta}|\boldsymbol{X}^{(l-1)}, \boldsymbol{y})$ and store the values.
2. Conditional on knowing the model parameters $\boldsymbol{\vartheta}^{(l)}$, sample a path $\boldsymbol{X}$ of the hidden Markov chain from the conditional posterior $p(\boldsymbol{X}|\boldsymbol{\vartheta}^{(l)}, \boldsymbol{y})$ and store all generated states.
3. Increase $l$ and return to step 1.

$L_0$ is the number of burn-in samples to be discarded from the estimate. In the next Sections we will explain in detail each step in Algorithm 1.1; as we will explain in Sections 1.4.1 and 1.4.2, given the independence assumption between transition and emission parameters, step 1. can be divided in two sub-points.

### 1.4.1 Sampling the transition matrix

We fix the initial state at 1, *i.e.* $X_0 = 1$ (for a discussion on the choices of the initial state distribution $\pi$ and related prior specification see Cappé *et al.*, 2005, Subsection 13.1.2 or Frühwirth-Schnatter, 2006, Subsection 10.3.4).

Let the rows of the transition matrix be independent a priori, each following a conjugate Dirichlet prior:

$$p(A) = \prod_{i=1}^{K} \mathrm{Dir}(\boldsymbol{a}_i|\alpha_{i1}, \dots, \alpha_{iK}),$$

where $\boldsymbol{a}_i = (a_{i,1}, \dots, a_{i,K})$ is the $i$th row of the transition matrix $A$. Then the rows are independent a posteriori, and, for a given trajectory $\boldsymbol{X}$ of the hidden Markov chain, are drawn from the Dirichlet distribution

$$\text{Dir}(\boldsymbol{a}_i|\alpha_{i1} + n_{i1}, \dots, \alpha_{ij} + n_{ij}, \dots \alpha_{iK} + n_{iK}), \tag{1.6}$$

where $n_{ij} = \#\{0 \leq t \leq T-1 : X_t = i, X_{t+1} = j\}$, $i, j \in \{1, \dots, K\}$.

### 1.4.2 Sampling the emission parameters

Sampling the emission parameters depends on if a finite or a parametric HMM is hypothesized and, in the latter case, on the chosen parametric family.

#### Finite hidden Markov model

Recall that in a finite HMM the state space of the Markov chain, $\mathsf{X}$, and the set in which the observable process takes its values, $\mathsf{Y}$, are finite sets; let $\mathsf{Y} = \{0, 1, \dots, q\}$. The rows of the emission matrix are assumed independent a priori, each following a conjugate Dirichlet prior:

$$p(B) = \prod_{i=1}^{K} \text{Dir}(\boldsymbol{b}_i|\beta_{i0}, \dots, \beta_{iq}),$$

where $\boldsymbol{b}_i = (b_i(0), \dots, b_i(q))$ is the $i$th row of the emission matrix $B$. Therefore, as for the transition matrix, the rows are independently drawn from the following Dirichlet distribution

$$\text{Dir}(\boldsymbol{b}_i|\beta_{i0} + e_{i0}, \dots, \beta_{iy} + e_{iy}, \dots, \beta_{iq} + e_{iq}), \tag{1.7}$$

where $e_{iy} = \#\{1 \leq t \leq T : X_t = i, Y_t = y\}$, with $i \in \{1, \dots, K\}$ and $y \in \mathsf{Y}$.

#### Parametric hidden Markov model

Consider for example a Poisson HMM, with conditional probability, $Y_t|X_t = x_t \sim \text{Pois}(y_t|\lambda_{x_t})$, assume prior independence of the means and a conjugate Gamma prior, $\lambda_k \sim \Gamma(\lambda_k|a_0, b_0)$, for $\lambda_k$. Then means are a posteriori independent and $\lambda_k$ is drawn from the posterior distribution

$$p(\lambda_k|\boldsymbol{X}, \boldsymbol{y}) \sim \Gamma(\lambda_k|a_0 + n_k\overline{y}_k, b_0 + n_k) \tag{1.8}$$

where $n_k = \#\{1 \leq t \leq T : X_t = k\}$ and $\overline{y}_k$ is the mean of the observations when $X_t = k$.

Of course different parametric models could be considered; in Chapter 2 we will consider also the negative binomial and the compound Poisson HMMs.

### 1.4.3 Sampling paths of the hidden Markov chain

Now we consider how sample a path of the hidden Markov chain $\boldsymbol{X}$ from the conditional distribution $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$.

Early papers (Robert *et al.*, 1993) use a sampling scheme, called *single updating of the hidden chain*, that samples the state $X_t$ conditional on all other states of the hidden chain. A more efficient way to sample $\boldsymbol{X}$ is the *global updating of the hidden chain* (see Cappé *et al.*, 2005), where the trajectory of the hidden chain is updated as a whole from its conditional distribution given the data and the model parameters $\boldsymbol{\vartheta}$.

#### Global updating of the hidden chain

Global updating of the hidden chain is based on writing the joint posterior distribution $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$ as

$$p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta}) = \prod_{t=1}^{T} p(X_t|X_{t-1}, \boldsymbol{y}, \boldsymbol{\vartheta}). \tag{1.9}$$

Now, let $\boldsymbol{y}_{h:k} = (Y_h = y_h, \ldots, Y_k = y_k)$ with $h \leq k$; so

$$\begin{aligned}
&p(X_t = j|X_{t-1} = i, \boldsymbol{y}, \boldsymbol{\vartheta}) = \\
&\quad = \frac{p(y_t|X_{t-1} = i, X_t = j, \boldsymbol{y}_{t+1:T}, \boldsymbol{\vartheta})p(X_t = j|X_{t-1} = i, \boldsymbol{y}_{t+1:T}, \boldsymbol{\vartheta})}{p(y_t|X_{t-1} = i, \boldsymbol{y}_{t+1:T}, \boldsymbol{\vartheta})} \\
&\quad = \frac{p(y_t|X_{t-1} = i, X_t = j, \boldsymbol{\vartheta})p(\boldsymbol{y}_{t+1:T}|X_{t-1} = i, X_t = j, \boldsymbol{\vartheta})}{p(\boldsymbol{y}_{t+1:T}|X_{t-1} = i, X_t = j, \boldsymbol{\vartheta})} \\
&\quad \cdot \frac{p(X_t = j|X_{t-1} = i, \boldsymbol{\vartheta})}{p(y_t|X_{t-1} = i, \boldsymbol{y}_{t+1:T}, \boldsymbol{\vartheta})} \\
&\quad \propto p(y_t|X_t = j, \boldsymbol{\vartheta})p(X_t = j|X_{t-1} = i, \boldsymbol{\vartheta})p(\boldsymbol{y}_{t+1:T}|X_t = j, \boldsymbol{\vartheta}) \\
&\quad \propto p(y_t|X_t = j, \boldsymbol{\vartheta})a_{i,j}p(\boldsymbol{y}_{t+1:T}|X_t = j, \boldsymbol{\vartheta})
\end{aligned}$$

where $p(\boldsymbol{y}_{t+1:T}|X_t = j, \boldsymbol{\vartheta})$ is the so called *backward variable* and it is the probability of the partial observation sequence from $t+1$ to T, given state $X_t = j$ and the model parameters $\boldsymbol{\vartheta}$.

Let $p(\boldsymbol{y}_{t+1:T}|X_t = j, \boldsymbol{\vartheta}) = \beta_t(j)$ and solve inductively as follows:

a) Initialize with

$$\beta_T(j) = 1 \quad 1 \leq j \leq K$$

b) and for $t = T-1, T-2, \dots, 1$, $1 \leq i \leq K$

$$\beta_t(i) = \sum_{j=1}^{K} a_{i,j} p(y_{t+1}|X_{t+1} = j)\beta_{t+1}(j). \tag{1.10}$$

When $p(y_{t+1}|X_{t+1} = j)$ is a density distribution, it is not necessarily bounded by 1. Then the backward variables may converge, at a geometric rate, to either zero or infinity. For this reason the introduction of a *scaling* factor is needed; then we scale $\beta_t(i)$ by multiplying each variable by a scale coefficient $\frac{1}{\sum_j^K \beta_t(j)}$, that depends only on $t$; each scale factor effectively restores the magnitude of the $\beta_t(i)$ terms to 1.

The conditional distribution (1.9) becomes

$$p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta}) \propto \prod_{t=1}^{T} p(y_t|X_t = x_t, \boldsymbol{\vartheta}) a_{x_{t-1},x_t} \beta_t(x_t) \tag{1.11}$$

and $X_t$, for $1 \leq t \leq T$, can be sampled from

$$Pr(X_t = j|X_{t-1} = x_{t-1}, \boldsymbol{y}, \boldsymbol{\vartheta}) = \frac{p(y_t|X_t = j, \boldsymbol{\vartheta}) a_{x_{t-1},j} \beta_t(j)}{\sum_{i=1}^{K} p(y_t|X_t = i, \boldsymbol{\vartheta}) a_{x_{t-1},i} \beta_t(i)} \tag{1.12}$$

where we recall that $X_0 = 1$.

Obviously, the sampling probability (1.12) specializes for the finite and parametric HMM, by substituting the corresponding emission probabilities $p(y_t|X_t = i, \boldsymbol{\vartheta})$.

### Single updating of the hidden chain

The local (or single) updating of the hidden chain is an alternative method of sampling a path of the chain from the conditional distribution $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$; it consists in sampling the state $X_t$ conditional on all other states from the conditional posterior distribution $Pr(X_t = j|\boldsymbol{X}_{-t}, \boldsymbol{y}, \boldsymbol{\vartheta})$, where $\boldsymbol{X}_{-t}$ denotes the whole path of $\boldsymbol{X}$ without the element $X_t$.

Then, the chain is sampled from

$$Pr(X_t = x_t|\boldsymbol{X}_{-t}, \boldsymbol{y}, \boldsymbol{\vartheta}) \propto a_{x_{t-1},x_t} a_{x_t,x_{t+1}} p(y_t|X_t = x_t) \tag{1.13}$$

for $1 \leq t \leq T - 1$, with $X_0 = 1$ and from

$$Pr(X_T = x_T|\boldsymbol{X}_{-T}, \boldsymbol{y}, \boldsymbol{\vartheta}) \propto a_{x_{T-1},x_T} p(y_T|X_T = x_T) \tag{1.14}$$

for $t = T$.

A computational advantage of the local updating sampler over the global updating one is that the time-consuming computing of the backward variables is avoided; a theoretical disadvantage of local updating, however, is that the auto-covariance function of any complete-data sufficient statistics drawn is equal to the auto-covariance function of the same statistics under global updating plus a penalty term (Scott, 2002). Hence the local updating sampler should mix and explore the posterior surface much more slowly than when global updating is used.

### 1.4.4 Label-switching

As arises from the previous discussion, sampling model parameters $\boldsymbol{\vartheta}$ from its complete data posterior should be trivial once $\boldsymbol{X}$ is drawn, but the draw is complicated by an identifiability issue known as *label-switching*. The HMM likelihood, in fact, is invariant under permutations of the state labels and, if either the prior distributions are exchangeable (*i.e.* they are invariant under permutations of the components), the posterior will also be exchangeable. We underline that the priors hypothesized in Sections 1.4.1 and 1.4.2 are exchangeable.

Consider a Poisson HMM with two possible states; swap value of $\lambda_1$ and $\lambda_2$, relabel all points currently in state 1 as state 2, and vice versa. The complete-data likelihood assumes the same value with new and old labels and so (if the priors are exchangeable) their marginal posterior densities are identical. Label switching is not a problem related to the sampling strategy, but it is an intrinsic property of the model and its prior.

Lack of identifiability also creates a difficulty with the maximum a posteriori estimator, in fact the exchangeability property implies that there are a multiple of $K!$ modes of the posterior surface.

Because the data contains no information about the order of the state labels, labels may only be identified in the posterior distribution by putting constraints on the prior; it is common to assume constraints ordering parameters. For example, in the Poisson HMM with 4 states, we could require the means to appear in ascending order, that is $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$.

Choosing parameters constraints can be informative, because we construct a new prior that is zero in regions where the constrains does not hold; moreover, as pointed out in Celeux *et al.* (2000), ordering different sets of the model parameters produces different posterior means. Scientific insight about the chain may suggest an order for the parameters. Consider the Business Cycle example presented in Section 1.1, where an HMM with a 2 state Markov chain is assumed, state 1 means that the economy is in an expanding status, while state 0 means that the economy is in a contracting status; assume that the gross national product given that the chain is in state $i$ is distributed according to a Gaussian distribution with mean $\mu_i$ and variance 1; then $\mu_0 > \mu_1$ is nonsensical. From a practical point of view, in a MCMC simulation, ordering can be imposed at each step of the sampler.

Lack of identifiability can also be avoided by using a loss function that is invariant under permutation of the labels; for instance, in case of mixture, Celeux *et al.* (2000) employed a loss function based on the Kullback-Liebler divergence.

A totally different estimate method is the *Variational method*, also called *Ensamble learning* (Beal, 2003) that approximates the posterior distribution with a simpler and tractable (lower or upper) bound and then optimizes this bound.

### 1.4.5 The MCMC algorithm and the Bayesian estimation

Taking into account previous considerations, we can state, in Algorithm 1.2, the MCMC sampling scheme for drawing from posterior density $p(\boldsymbol{X}, \boldsymbol{\vartheta}|\boldsymbol{y})$, in the case of finite HMM:

**Algorithm 1.2** MCMC algorithm

```
Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \ldots, L_0, \ldots, L$.

  1. Sample each row of $A$ from the complete-data posterior distribution $p(A|\boldsymbol{X}^{(l-1)})$ in equation (1.6), and store the values.

  2. Sample the emission parameter from the complete-data posterior in equation (1.7), $p(B|\boldsymbol{y}, \boldsymbol{X}^{(l-1)})$ and store the values.

  3. Conditional of knowing the model parameters $\boldsymbol{\vartheta}^{(l)}$

       a) Compute the means of the observable values in each state, $\boldsymbol{\mu}^{(l)} = (\mu_1^{(l)}, \mu_2^{(l)}, \ldots, \mu_K^{(l)})$, with $\mu_i^{(l)} = \sum_{j=0}^{q} j \cdot b_i^{(l)}(j)$, and
            – if $\mu_1^{(l)} < \mu_2^{(l)} < \ldots < \mu_K^{(l)}$ go to step b)
            – else order the mean vector $\rho(\boldsymbol{\mu})$ and $\rho(A)$, $\rho(B)$

       b) Compute the scaled backward variables, like in equation (1.10)

       c) Sample a path $\boldsymbol{X}$ of the hidden Markov chain from the conditional posterior $p(\boldsymbol{X}|\boldsymbol{\vartheta}^{(l)}, \boldsymbol{y})$ in equation (1.12), trough the global updating sampler and store all states.

  4. Increase $l$ and return to step 1.
```

Obviously the algorithm specializes for the parametric HMM; in particular, in the Poisson HMM, the ordering step 3.a) is directly achieved by the mean of the Poisson distributions.

Posterior draws produced by the MCMC sampler, provided that a sufficiently large number $L_0$ of draws are discarded, are used for statistical inference (for a discussion on different methods to estimated the hidden chain see for example Scott, 2002).

Considering a quadratic loss function (*i.e.* $\text{loss}(\boldsymbol{\vartheta}, \hat{\boldsymbol{\vartheta}}) = \|\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}\|^2$), the model parameters' estimate is the posterior expectation, that is approximated by averaging over the draws from the posterior distribution:

$$\hat{\boldsymbol{\vartheta}} = \frac{1}{L - L_0} \sum_{l=L_0}^{L} \boldsymbol{\vartheta}^{(l)}.$$

Point estimations of the hidden Markov chain $\boldsymbol{X}$ may be obtained by considering a 0/1 loss function, minimized by the mode of the marginal posterior probability, also called the *maximum a posteriori* (MAP) estimator. Therefore the estimated Markov chain is:

$$\boldsymbol{X}_t^{\text{MAP}} = \underset{l=L_0:L}{\text{mode}} \boldsymbol{X}_t^{(l)},$$

for $t = 1, \dots, T$

## 1.5 Choosing the number of states of the hidden Markov chain

Before considering the choice of the number of states of the Markov chain, let us briefly present the dataset provided by the AEEG.

First analyzes implemented by the Italian Authority were referred to a dataset relative to more than 300 territorial districts, covering the entire national territory. During the consultation process for the third regulatory period, 2008-2011, additional elements emerged; in particular a more appropriate definition of the spatial units was introduced. The proposal was to explore a geographical area larger than the district.

The Italian distribution sector includes one large utility, that serves more than 80% of the consumers and a number of local companies. Telecontrol centers are closely related to the technical structure of the network; therefore they appeared as the preferable choice. However, it was observed that they might be modified over time by the company to include different groups of consumers. For this reason, the company for local utilities and the province for the big one were preferred as the new spatial units for the analysis.

The propose to identify exceptional events by an HMM was developed during this consultation process. Then we dispose of two datasets: one containing the hourly number of interruptions for 34 telecontrol centers in years 2004-2005 and the other one relative to 113 province and company combinations for the three year time span 2004-2006. We underly that utilities are the main object of our analysis while telecontrol centers or province and company combinations are the considered spatial units; in other words the analyzed object is the same while the point of view (or better the statistical units) changes.

The main common characteristic between sequences in the two datasets is that the great majority of the observations (more than 75% for the telecontrol centers and about 90% for the province/company) is equal to 0; a quite large number of faults is equal to 1 and so on in a decreasing order.

Consider Figure 1.2 showing observations for a center labeled as dg4, for the period from 13th to 26th August 2004; it represents a typical trend noticeable in the datasets. Very often the observations are equal to 0, there are several hours with a quite small number of interruptions (the lower peaks); moreover we can distinguish peaks with different height and width or, in other words, there are periods with interruptions of different order of magnitude and involving various hours.

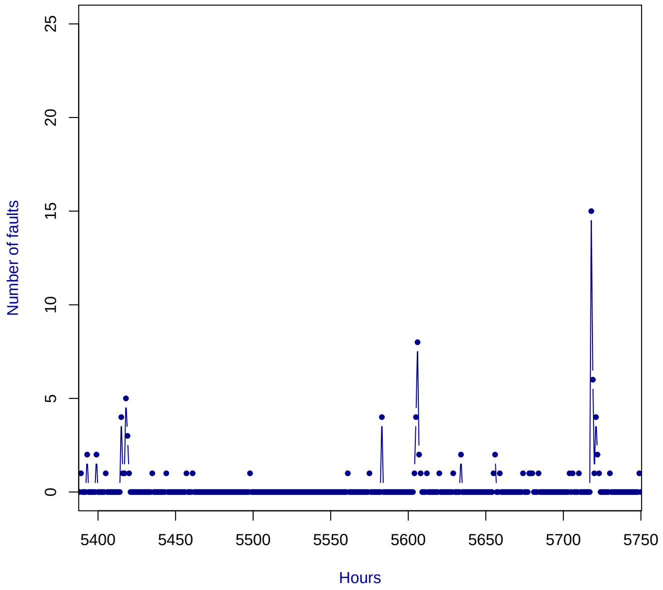


**Figure 1.2:** Time series' typical trend. Observations from 13th to 26th August 2004 for the center dg4.

Coming back to the choice of the number of possible state, in order to reach the set goal we will consider an HMM with a four state Markov chain. States have a physical meaning: state 1 indicates that the system is in a normal operating status while state 4 indicates an exceptional operating status; states 2 and 3 are transitional and refer to an increasing degree of perturbation of the system operating status, as yet non exceptional.

The choice to consider four possible states is the results of a "Goldilocks selection": we first tried to consider two states (a normal and an exceptional state) and three states; however results suggest that the problem is more complicate and needs a more complex structure. In fact with, for example, two states we are not able to discriminate between interruptions holding one hour (and then not exceptional) and faults during in time (and the caused by an exceptional event).

## 1.6 Studying the “Exceptional excursions”: Phase-type distributions

As we said in Section 1.1 the Authority has become more interested in controlling the efficiency and effectiveness of utility restoration schemes. In an HMM, the transition matrix $A$ reassumes information regarding the transition dynamic of the underlying process and in particular dynamic related to the exceptional state. Then, studying the transition matrix $A$, we can achieve some information related to the exceptional periods.

First of all, it is known that the time spent by the Markov chain $\{X_t\}$ in a state $i$, say $T(i)$, has a geometric distribution with parameter $1 - a_{i,i}$, where $a_{i,i} = P(X_{t+1} = i|X_t = i)$; therefore the expected number of instant the chain passes in state $i$ is $E(T(i)) = \frac{1}{1-a_{i,i}}$.

Moreover, when an exceptional event occurs (*i.e.* the underlying chain is in state 4), it could be interesting to analyze the requested time for re-establishing the normal situation (*i.e.* the chain comes back to state 1); we define now an *exceptional excursion.*

**Definition 1.1.** *Let an exceptional excursion be a sequence beginning in the first state after a state 1 (it could be state 2, 3 or 4), ending in state 1 and containing at least one state 4.*

In other words an exceptional excursion represents the time passed by the chain outside the normal state 1, when an exceptional event occurs.

In order to analyze the length of an exceptional excursion, we consider the *Phase-type distribution*, that is the distribution of the number of steps from a Markov chain starting until absorption into absorbing state (for a complete and clear explanation see Neuts, 1994).

**Definition 1.2.** *Consider a discrete-time Markov chain with $m+1$ states, where $m \geq 1$. The states $1, \ldots, m$ are transient and state $m+1$ is an absorbing state. The process has an initial probability of starting in any of the $m+1$ phases given by the probability vector $(\boldsymbol{\alpha}, \alpha_{m+1})$.*

*This process can be written in the form of a transition probability matrix*

$$P = \begin{pmatrix} T & T^o \\ 0 & 1 \end{pmatrix}$$

*where $T$ is an $m \times m$ substochastic matrix, $T^o + T\boldsymbol{e} = 1$ and $\boldsymbol{e}$ is the column vector with all its components equal to one.*

*The distribution of the number of steps $S$ until the process reaches the absorbing state is said to be discretely Phase-type distributed and it is represented by the pair $(\boldsymbol{\alpha}, T)$.*

*Moreover, the probability density $\{p_s\}$ of the Phase-type distribution is given by*

$$\begin{aligned} p_0 &= \alpha_{m+1} \\ p_s &= \boldsymbol{\alpha} T^{s-1} T^0, \quad \textit{for} \quad s \geq 1. \end{aligned}$$

We can now state the following result.

**Proposition 1.1.** *The length of an exceptional excursion is discretely Phase-type distributed.*

*Proof.* Let the normal state 1 be the absorbing state and, without loss of generality, fix at 1 the starting time of the exceptional excursion; consider the two stopping times $E = \min\{t > 1 : X_t = 1\}$, that is the instant the chain reaches the absorbing state 1 and $H = \min\{t \geq 1 : X_t = 4\}$, that is the first time the chain enters in state 4. In order to prove the statement we need to verify that

$$P(X_1 = x_1, \ldots, X_E = 1 | X_0 = 1, X_1 \neq 1, X_h = 4 \text{ for at least one } 0 < h < E) \tag{1.15}$$

has a Markov structure.

In the following we will write "$X_h = 4$, $0 < h < E$" to indicate "$X_h = 4$ for at least one $0 < h < E$". We have

$$P(X_1 = x_1, \ldots, X_E = 1 | X_0 = 1, X_1 \neq 1, X_h = 4, 0 < h < E) =$$

$$\begin{aligned}
&= \frac{P(X_1 = x_1, \ldots, X_E = 1, X_h = 4, 0 < h < E | X_0 = 1, X_1 \neq 1)}{P(X_h = 4, 0 < h < E | X_0 = 1, X_1 \neq 1)} \\
&\propto P(X_1 = x_1 | X_0 = 1, X_1 \neq 1) \\
&\cdot P(\boldsymbol{X}_{2:E} = \boldsymbol{x}_{2:E}, X_h = 4, 0 < h < E | X_1 = x_1, X_0 = 1, X_1 \neq 1) \\
&\propto \frac{a_{1,x_1}}{\sum_{i \in \mathsf{X} \setminus \{1\}} a_{1,i}} \mathbf{1}_{\mathsf{X} \setminus \{1\}}(x_1) \\
&\cdot P(\boldsymbol{X}_{2:E} = \boldsymbol{x}_{2:E}, X_h = 4, 0 < h < E | X_1 = x_1, X_0 = 1, X_1 \neq 1) \quad (1.16)
\end{aligned}$$

where $\mathbf{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$, $\mathsf{X}$ is the state space of the Markov chain, and, as usual, $\boldsymbol{X}_{t:s} = (X_t, X_{t+1}, \ldots, X_s)$, with $t < s$.

Consider the second term in equation (1.16),

$$P(\boldsymbol{X}_{2:E} = \boldsymbol{x}_{2:E}, X_h = 4, 0 < h < E | X_1 = x_1, X_0 = 1, X_1 \neq 1) =$$

$$\begin{aligned}
&= P(\boldsymbol{X}_{2:H} = \boldsymbol{x}_{2:H}, \boldsymbol{X}_{H+1:E} = \boldsymbol{x}_{H+1:E}, H < E | X_1 = x_1, X_1 \neq 1) \\
&= P(\boldsymbol{X}_{2:H} = \boldsymbol{x}_{2:H}, H < E | X_1 = x_1, X_1 \neq 1) \\
&\cdot P(\boldsymbol{X}_{H+1:E} = \boldsymbol{x}_{H+1:E}, H < E | X_1 \neq 1, \boldsymbol{X}_{1:H} = \boldsymbol{x}_{1:H}) \\
&= \text{by the strong Markov property} \\
&= P(\boldsymbol{X}_{2:H} = \boldsymbol{x}_{2:H}, H < E | X_1 = x_1, X_1 \neq 1) \prod_{i=H+1}^{E} a_{i-1,i}. \quad (1.17)
\end{aligned}$$

Consider the first term in equation (1.17); being $H < E$ and $H = \min\{t \geq 1 : X_t = 4\}$, we know that from time 1 to $H-1$, in the exceptional excursion there are no 1 and no 4. Therefore equation (1.15) becomes

$$\begin{aligned}
&P(X_1 = x_1, \ldots, X_E = 1 | X_0 = 1, X_1 \neq 1, X_h = 4 \text{ for at least one } 0 < h < E) \\
&\quad = \frac{a_{1,x_1}}{\sum_{i \in \mathsf{X} \setminus \{1\}} a_{1,i}} \mathbf{1}_{\mathsf{X} \setminus \{1\}}(x_1) \prod_{l=2}^{H} \frac{a_{x_{l-1},x_l}}{\sum_{j \in \mathsf{X} \setminus \{1\}} a_{x_{l-1},j}} \mathbf{1}_{\mathsf{X} \setminus \{1\}}(x_l) \prod_{i=H+1}^{E} a_{i-1,i}
\end{aligned}$$

and this complete the proof. □

We can now obtain the pair $(\alpha^*, T^*)$, representing the Phase-type distribution. Consider a Markov chain with absorbing state 1 and transient states $\{2^*, 3^*, 2, 3, 4\}$, where $2^*$ and $3^*$ are states 2 and 3 conditioning on the fact that in following there is at least a 4; indicate the corresponding state space with $\mathsf{X}^*$. Then the initial probability vector is

$$\boldsymbol{\alpha}^* = \left( \frac{a_{1,2}}{a_{1,2}+a_{1,3}+a_{1,4}}, \frac{a_{1,3}}{a_{1,2}+a_{1,3}+a_{1,4}}, 0, 0, \frac{a_{1,4}}{a_{1,2}+a_{1,3}+a_{1,4}} \right)$$

and $\alpha_1 = 0$. The first element in $\boldsymbol{\alpha}^*$ is, for example, the probability to have $X_1 = 2^*$, given that $X_0 = 1$ and $X_1 \neq 1$; null values are relative to states 2 and 3, in fact, given the definition of an exceptional excursion, it is not possible to have $X_1 = 2$ or $X_1 = 3$ (we can have $X_1 = 2^*, X_1 = 3^*$ or $X_1 = 4$).

Using similar considerations we obtain the following transition probability matrix

$$P^* = \left( \begin{array}{ccccc|c} \frac{a_{2,2}}{a_{2,2}+a_{2,3}+a_{2,4}} & \frac{a_{2,3}}{a_{2,2}+a_{2,3}+a_{2,4}} & 0 & 0 & \frac{a_{2,4}}{a_{2,2}+a_{2,3}+a_{2,4}} & 0 \\ \frac{a_{3,2}}{a_{3,2}+a_{3,3}+a_{3,4}} & \frac{a_{3,3}}{a_{3,2}+a_{3,3}+a_{3,4}} & 0 & 0 & \frac{a_{3,4}}{a_{3,2}+a_{3,3}+a_{3,4}} & 0 \\ 0 & 0 & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,1} \\ 0 & 0 & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,1} \\ 0 & 0 & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,1} \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right)$$

## 1.7 Outline of the work

After opting to analyze the utilities' performance by an HMM with a four state Markov chain, we need to decide on considering a parametric or a finite HMM.

In Chapter 2 we consider different parametric models; output does not completely accomplish our goal and the analysis of the results suggests us to consider a finite HMM. Application of this model in Chapter 3 (Accoto *et al.*, 2008) yields interesting results, also compared with what obtained by the AEEG.

Dataset employed in the implementation of these models (parametric and finite) is relative to the telecontrol centers; as we said in Section 1.5, a new definition of the spatial units was introduced and the dataset relative to the province and company combinations was available. In Chapter 4, after analyzing all province and company combinations for year 2004, we consider a Cluster analysis, in order to investigate eventual similarities in the underlying systems.

In Chapter 5 we define and apply, after presenting a method to make inference, a model-based clustering method, alternative to the distance-based one proposed in Chapter 4; hence we will introduce the *Hidden mixture Markov Model*.

This work could be virtually divided into two parts: one containing what so far presented and another small one representing the starting point of a research topic propose. In fact, in Chapter 6 we will present the Reinforced Urn Processes (Muliere *et al.*, 2000) and how they could be applied in the prior specification when an HMM, from a Bayesian point of view, is considered.

# Chapter 2

# Parametric hidden Markov models for identifying exceptional events

## 2.1 Introduction

The interpretation of the faults in the electricity distribution as a signal of an underlying process leads us to consider an HMM, for identifying interruptions due to an exceptional event.

HMMs represent a class of different possible model. The choice of the number of possible states of the hidden Markov chain (see Section 1.5) restricts this class; however, especially regarding implications on the emission mechanism, we could consider two different types of HMM: the parametric HMM and the finite HMM. The main difference between these two models is that in the parametric HMM the observed value (given the state of the underlying process) is distributed according to a parametric distribution, while in the finite HMM a matrix (called emission matrix) contains the conditional probabilities to observe the different values, given the state of the hidden chain. Of course the choice between the parametric and the finite HMM has implications in terms of the number of emission parameters to be estimated. If we consider a Poisson distribution, we have to estimate $K$ Poisson means, whilst if we consider a finite HMM, with $\mathsf{Y} = \{0, \ldots, q\}$, we need to estimate $K \times (q + 1)$ emission probabilities.

We start considering the parametric models. The nature of the data under analysis (*i.e.* the number of faults) naturally leads to consider a Poisson HMM (Section 2.2). An accurate analysis of the results obtained by the application of this model suggests us to consider two different emission distributions: the negative binomial and the compound Poisson distributions (Sections 2.3 and 2.4).

## 2.2 Poisson hidden Markov model

### 2.2.1 Model and results

Consider a Poisson HMM:

$$\begin{aligned} Y_t|X_t = i &\sim \text{Pois}(y_t|\lambda_i), \\ \{X_t\}|A &\sim \text{Markov chain}(A); \end{aligned}$$

where Markov chain$(A)$ is for "Markov chain with transition matrix $A$", $A = \{a_{i,j}\}$ with $a_{i,j} = P(X_{k+1} = j|X_k = i)$, $i, j \in \mathsf{X} = \{1, \ldots, K\}$, and $K$ is the number of possible states.

In Chapter 1 we presented inference for the Poisson HMM as an example of a parametric HMM. Fix $X_0 = 1$, let the rows of the transition matrix be independent a priori, each following a conjugate Dirichlet prior, $p(A) = \prod_{j=1}^{K} \text{Dir}(\boldsymbol{a}_j|\alpha_{j1}, \ldots, \alpha_{jK})$ and assume prior independence of the means and a conjugate Gamma prior, $\lambda_k \sim \text{Gamma}(\lambda_k|a_0, b_0)$. Then, given a data sequence $\boldsymbol{y} = (Y_1 = y_1, \ldots, Y_T = y_T)$, the Gibbs sampling in Algorithm 1.1 specializes in the following Algorithm 2.1 (see Section 1.4 for details).

**Algorithm 2.1** Gibbs sampling for a Poisson HMM

```
Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \ldots, L_0, \ldots, L$.

  1. Sample each row of $A$ from the Dirichlet posterior distribution
  2. Sample the Poisson mean from the Gamma posterior
  3. Conditional of knowing the model parameters
       a) – if $\lambda_1^{(l)} < \lambda_2^{(l)} < \ldots < \lambda_K^{(l)}$ go to step b)
          – else order the mean vector $\rho(\boldsymbol{\lambda})$ and $\rho(A)$
       b) Compute the scaled backward variables
       c) Sample a path $\boldsymbol{X}$ of the hidden Markov chain trough the global updating sampler
  4. Increase $l$ and return to step 1.
```

Each combination of telecontrol center and year is studied as an independent global system; the hourly number of faults generated by the system in a year is modeled by means of an HMM with a four state Markov chain; in the following we will focus our attention on a telecontrol center codified with dr3, for year 2004. Moreover, in the prior specification, we consider equal to 1 parameters for the Dirichlet and the Gamma distributions.

We now briefly consider convergence of the MCMC algorithm; before starting we have to observe that unfortunately diagnostics only say if stationarity has not been achieved. Diagnostics considered in the following are implemented in the R package (R Development Core Team, 2005) coda (Plummer *et al.*, 2008).
Consider Figure 2.1 showing the first 5 000 generated values of probability of staying in each state and the generated Poisson means. By this plot it seems that the chain reached the stationarity, in fact there is not a trend.

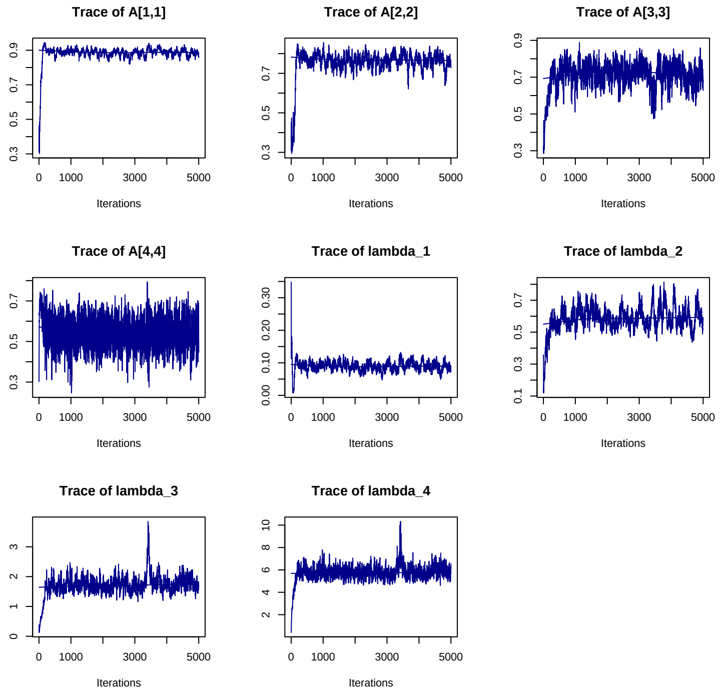


**Figure 2.1:** First 5 000 generated values from the Gibbs sampler, for center dr3, year 2004. In the plots' title $A[i, i]$ refers to the probability of staying in state $i$, $a_{i,i}$ and lambda_i is for $\lambda_i$.

Apart from these qualitative diagnostics, different quantitative diagnostics have been proposed in literature; for example the Geweke's diagnostic and the Heidelberger-Welch diagnostic. Geweke (1992) proposed a convergence diagnostic for Markov chains based on a test for equality of the means of the first and last part of a Markov chain. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. The Heidelberger-Welch (1983) convergence test uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution. The test is successively applied, firstly to the whole chain, then after discarding the first 10%, 20%, . . . of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded. The latter outcome constitutes 'failure' of the stationarity test and indicates that a longer MCMC run is needed. In our case, also these two diagnostics confirm that the chain is stationary.

Finally we use the Raftery and Lewis's diagnostic (1992a, 1992b, 1995) to obtain indications relative to the number of iterations to consider in the algorithm ($L$) and how many generated values we need to discard from the estimation ($L_0$). This diagnostic, in fact estimates how long the chain needs to run in order to estimate quantiles ($q$), within a specified accurancy ($r$) with some specified probability ($s$). The minimum length is the required sample size for a chain with no correlation between consecutive samples (of course MCMC draws are not independent). Positive autocorrelation will increase the required sample size above this minimum value. The number of burn-in iterations to be discarded at the beginning of the chain is also calculated.

Table 2.1 contains the results of the Raftery and Lewis's test when respectively, the fist, the second and the third quartile are considered, with $r = 0.05$ and $s = 0.95$. The maximum values for the burn-in and the total number of iterations are respective 117 and 12 460; however the "dependence factor", that is an estimate of the extent to which autocorrelation inflates the required sample size, is quite large (in general values greater than 5 are considered considerable); for this reason we will consider $L = 30\,000$ and $L_0 = 2\,000$.

| | Burn-in | Total | Lower bound | Dependence Factor |
|---|---|---|---|---|
| A[1,1] | 108 | 9708 | 289 | 33.60 |
| A[2,2] | 78 | 6408 | 289 | 22.20 |
| A[3,3] | 88 | 6732 | 289 | 23.30 |
| A[4,4] | 10 | 884 | 289 | 3.06 |
| lambda_1 | 72 | 6872 | 289 | 23.80 |
| lambda_2 | 104 | 9880 | 289 | 34.20 |
| lambda_3 | 40 | 3575 | 289 | 12.40 |
| lambda_4 | 30 | 2700 | 289 | 9.34 |
| A[1,1] | 100 | 12460 | 385 | 32.40 |
| A[2,2] | 80 | 9792 | 385 | 25.40 |
| A[3,3] | 54 | 6912 | 385 | 18.00 |
| A[4,4] | 8 | 1176 | 385 | 3.05 |
| lambda_1 | 78 | 10296 | 385 | 26.70 |
| lambda_2 | 117 | 13932 | 385 | 36.20 |
| lambda_3 | 40 | 5090 | 385 | 13.20 |
| lambda_4 | 36 | 4548 | 385 | 11.80 |
| A[1,1] | 80 | 6672 | 289 | 23.10 |
| A[2,2] | 63 | 5789 | 289 | 20.00 |
| A[3,3] | 35 | 3300 | 289 | 11.40 |
| A[4,4] | 10 | 894 | 289 | 3.09 |
| lambda_1 | 70 | 6048 | 289 | 20.90 |
| lambda_2 | 110 | 9526 | 289 | 33.00 |
| lambda_3 | 56 | 5026 | 289 | 17.40 |
| lambda_4 | 36 | 3090 | 289 | 10.70 |

**Table 2.1:** Results relative to the Raftery and Lewis's diagnostic for $q = 0.25$ (top), $q = 0.5$ (middle) and $q = 0.75$ (bottom).

We now consider results obtained by the MCMC generated values (see Section 1.4.5 for details on the adopted Bayesian estimation).
The estimated transition matrix is

$$\hat{A} = \begin{pmatrix} 0.8888 & 0.1038 & 0.0054 & 0.002 \\ 0.208 & 0.7668 & 0.0191 & 0.006 \\ 0.0392 & 0.188 & 0.7163 & 0.0565 \\ 0.1528 & 0.0577 & 0.251 & 0.5385 \end{pmatrix}$$

and the estimated mean vector is

$$\hat{\boldsymbol{\lambda}} = (0.091, 0.596, 1.754, 5.817).$$

In the following, with an abuse of terminology, we will say "observation $y_t$ *classified* in state $i$" to indicate that at time $t$ the underlying estimated Markov chain is in state $i$ and it emits the observation $y_t$.
To better evaluate the model consider results concerning the estimated hidden chain, summarized in Table 2.2, where each value $n_{xy}$ is the number of observations equal to $y$, classified in state $x$, *i.e.* $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 14 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5798 | 395 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 844 | 978 | 318 | 54 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 31 | 74 | 88 | 59 | 25 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 5 | 14 | 20 | 14 | 12 | 9 | 6 | 1 | 2 | 1 | 1 | 1 | 1 |

**Table 2.2:** Telecontrol center dr3, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

Also considering what obtained by the AEEG method, it seems that too many observations (88) are considered as due to an exceptional event, *i.e.* they are classified in state 4. Moreover consider the following subsequences, where a quite large number of interruptions, followed and preceded by no interruptions are classified in the exceptional state; these situations violate the natural idea, also emerged by the AEEG method (see the end of Section 1.2), that an exceptional event generate a large number of interruptions protracting in time.

| $\boldsymbol{X}^{\mathrm{MAP}}$ | 1 | 4 | 1 |
|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 5 | 0 |

(2.1)

| $\boldsymbol{X}^{\mathrm{MAP}}$ | 1 | 4 | 1 |
|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 6 | 0 |

| $\boldsymbol{X}^{\mathrm{MAP}}$ | 1 | 4 | 1 |
|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 8 | 0 |

We now analyze results in order to try to propose an improvement of the model. An HMM is a generalization of a Mixture model, where the components are not selected independently, but according to an underlying Markov chain. In Figure 2.2 the Poisson distributions (mixture components) with mean $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$ are plotted; it highlights that it is highly probable that the exceptional state (state 4) emits observations larger than 4 or 5 (*i.e.* the tail of the $\mathrm{Pois}(\cdot|\lambda_3)$ is neglectable at those values).

Let us explain this consideration and try to interpret results contained in those subsequences previously considered. In the single updating scheme, presented in Section 1.4.3, the chain is sampled from:

$$P(X_t = x_t|\boldsymbol{X}_{-t}, \boldsymbol{y}, \boldsymbol{\vartheta}) \propto a_{x_{t-1},x_t} a_{x_t,x_{t+1}} \mathrm{Pois}(y_t|\lambda_{x_t}) \tag{2.2}$$

where $\boldsymbol{X}_{-t} = (X_1 = x_1, \dots, X_{t-1} = x_{t-1}, X_{t+1} = x_{t+1}, \dots, X_T = x_T)$. Then, consider for example subsequence 2.1; given that the chain is in state 1 and the corresponding observation is equal to 0, in the next time even if the probability to go in state 4 ($a_{1,4}$) is small, the emission probability, corresponding to a number of interruption greater than 4 or 5, is considerably larger in state 4 than in other states ($\mathrm{Pois}(4|\lambda_4) > \mathrm{Pois}(4|\lambda_i)$, $i = 1, 2, 3$); then the sampled state of the chain will be the exceptional state 4.

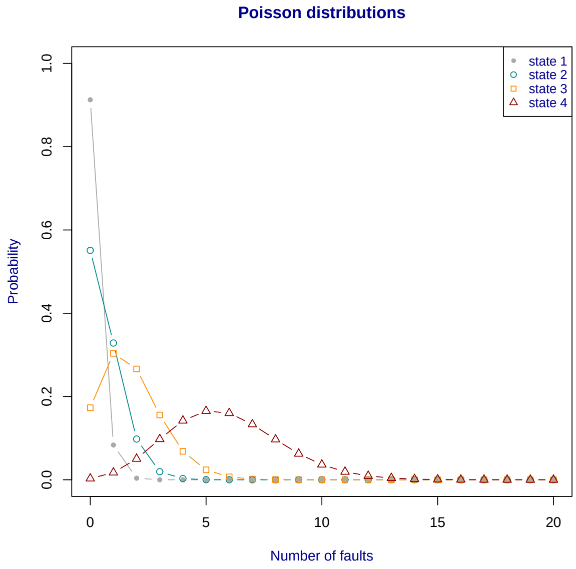


**Figure 2.2:** Center dr3, year 2004: Poisson probability mass functions with means $\lambda_1 = 0.091$, $\lambda_2 = 0.596$, $\lambda_3 = 1.754$ and $\lambda_4 = 5.817$.

Those considerations induce us to consider other distributions more "heavy tailed": the *negative binomial* and the *compound Poisson distributions*.

Consider a random variable (r.v.) $Z \sim \text{NegBin}(r, p)$, where $0 < p < 1$ and $r > 0$; then

$$P(Z = z) = \frac{\Gamma(z + r)}{\Gamma(r) z!} p^r (1 - p)^z, \qquad z = 0, 1, 2, \ldots.$$

Furthermore $E(Z) = r\frac{1-p}{p}$ and $V(Z) = r\frac{1-p}{p^2}$.

The compound Poisson distribution arises in a model formed by supposing that objects (for example earthquakes or faults in the electricity distribution) occur in cluster, the number of clusters having a Poisson distribution, while the number of objects per cluster varies according to a distribution $Q$ (see Johnson *et al.* (1992), Chapter 9). More formally, let $N \sim \text{Pois}(\lambda)$ and consider $W_1, W_2, \ldots$ independent and identically distributed (i.i.d.) r.v. with common distribution $Q$ independent of $N$; then $Y = \sum_{i=1}^{N} W_i$ has a compound Poisson distribution $\text{CP}(\lambda, Q)$. The parameter $\lambda$ is called the *rate* of $\text{CP}(\lambda, Q)$ and $Q$ is the *base distribution*; moreover, if $E(W_i) = \mu$ and $E(W_i^2) = m_2$, then $E(Y) = \lambda\mu$ and $V(Y) = \lambda m_2$.

Let $W_i \sim \mathrm{geo}(1-p)$, $0 < p \leq 1$, with probability mass function $P(W = w) = p^{w-1}(1-p)$, $w = 1, 2, \ldots$; then the $\mathrm{CP}(\lambda, \mathrm{geo}(1-p))$ is also called the *Pólya-Aeppli distribution* with $E(Y) = \frac{\lambda}{1-p}$ and $V(Y) = \frac{\lambda p(1+p)}{(1-p)^2}$. Furthermore, for $y = 1, 2, \ldots$ and with $q = 1 - p$

$$\begin{aligned} P(Y = y) &= \sum_{n=1}^{\infty} P(Y = y|N = n)P(N = n) \\ &= \sum_{n=1}^{y} P\left(\sum_{i=1}^{n} W_i = y|N = n\right) P(N = n) \\ &= e^{-\lambda} \sum_{n=1}^{y} \frac{\lambda^n}{n!} \frac{\Gamma(y)}{(y-n)!\Gamma(n)} q^n (1-q)^{y-n} \end{aligned} \tag{2.3}$$

and

$$P(Y = 0) = P(N = 0) = e^{-\lambda}.$$

The last equality in (2.3) is due to the fact that $P(N = n)$ is $\mathrm{Pois}(n|\lambda)$ by assumption and, denoted by $W^0$ and $W^1$ r.v. geometrically distributed with probability of success $q$ and discrete support respectively starting in 0 and 1, $P\left(\sum_{i=1}^{n} W_i^1 = y|N = n\right) = P\left(\sum_{i=1}^{n} W_i^0 + n = y|N = n\right)$ and, finally, $\sum_{i=1}^{n} W_i^0 \sim \mathrm{NegBin}(n, q)$.

By considering $\frac{\Gamma(0)}{\Gamma(0)} = 1$ we have, for $y = 0, 1, 2, \ldots$

$$P(Y = y) = e^{-\lambda} \sum_{n=0}^{y} \frac{\lambda^n}{n!} \frac{\Gamma(y)}{(y-n)!\Gamma(n)} q^n (1-q)^{y-n}. \tag{2.4}$$

In fact when $y = 0$ the sum is equal to 1 and when $y \neq 0$ the first term of the sum (*i.e.* with $n = 0$) is equal to 0 $\left(\frac{1}{y\Gamma(0)} = \frac{1}{\infty} = 0\right)$.

In the following we will indicate with $\mathrm{CP}(\lambda, q)$ the Pólya-Aeppli distribution $\mathrm{CP}(\lambda, \mathrm{geo}(1-p))$.

The Figure 2.3 shows, for each state, the three probability mass functions with the same mean; to be precise each plot contains the $\mathrm{Pois}(\lambda_i)$, the $\mathrm{NegBin}(\lambda_i, 0.5)$ and the $\mathrm{CP}(\lambda_i, 0.5)$ distributions, $i = 1, \ldots, 4$ and $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$. The great differences are related to states 3 and 4: both distributions (the negative binomial and the compound Poisson) are more skewed to the right, but the negative binomial concentrates more mass on smaller values while the compound Poisson is shifted on larger values.

Figure 2.4 represents a different way (similar to that in Figure 2.2) to look at the model, in terms of mixture components, when the negative binomial and the compound Poisson distributions are considered.

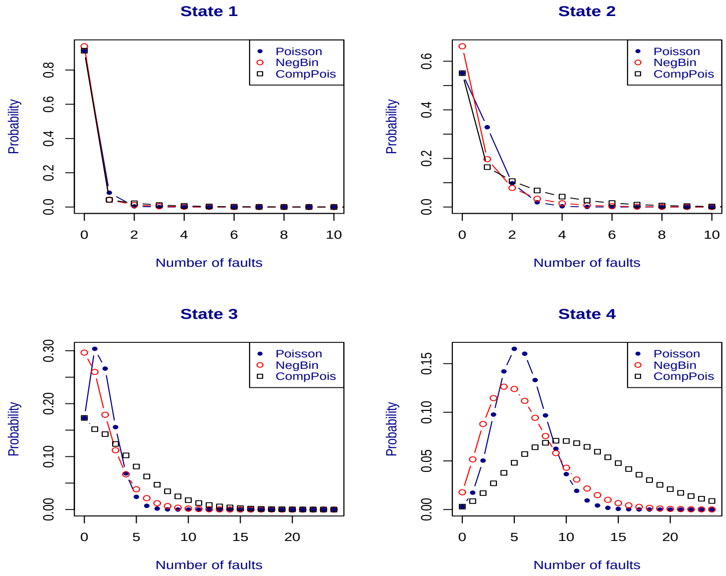


**Figure 2.3:** Poisson, negative binomial and compound Poisson probability mass functions. Top-left: with mean equal to 0.091. Top-right: with mean equal to 0.596. Bottom-left: with mean equal to 1.754. Bottom-right: with mean equal to 5.817.

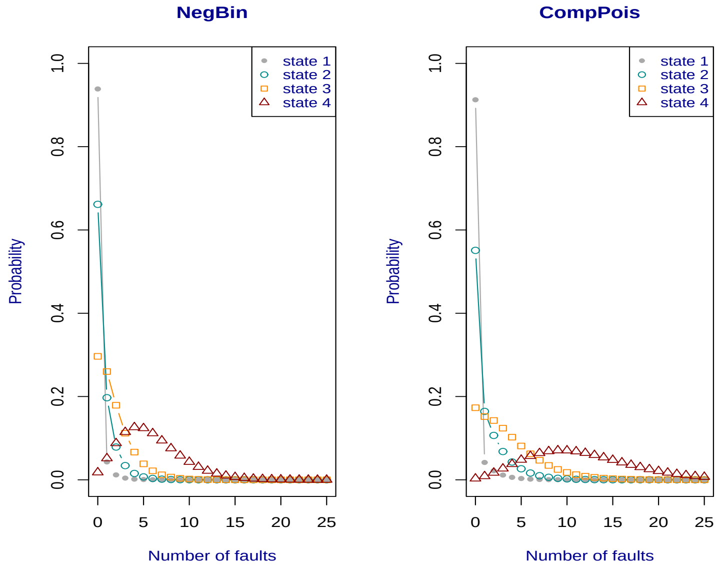


**Figure 2.4:** Negative binomial (Left) and compound Poisson (Right) probability mass functions with means $\lambda_1 = 0.091$, $\lambda_2 = 0.596$, $\lambda_3 = 1.754$ and $\lambda_4 = 5.817$.

### 2.2.2 Zero-inflated Poisson hidden Markov model

As we said in Section 1.5, for every utility more than 90% of observations is equal to 0. Then before considering the introduced models (the negative binomial and the compound Poisson HMMs) we take into account this feature by considering the zero-inflated Poisson HMM.

In general, consider a r.v. $W$ distributed according to a zero-inflated distribution, then

$$P(W = w) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & w = 0; \\ (1-\pi)\text{Pois}(w|\lambda), & w > 0 \end{cases}$$

or

$$P(W = w) = \pi\text{Pois}(w|0) + (1-\pi)\text{Pois}(w|\lambda),$$

since $\text{Pois}(w|0) = 0$ for all $w > 0$ and $\text{Pois}(0|0) = 1$.

Consider a zero-inflated Poisson HMM and let the state 1 be the “zero-inflated state”, that is $P(Y_t = 0|X_t = 1) = 1$.

Inference on the transition matrix and on the Poisson means $\lambda_i$, $i = 2, \ldots, K$, is the same as described in Algorithm 2.1, while sampling from the posterior

$p(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\vartheta})$ changes. Consider the sampling probability for the single updating scheme recalled in equation (2.2). In the zero-inflated Poisson HMM we need to specialize (2.2) and to characterize it for the state 1. In fact

$$P(X_t = 1|\boldsymbol{X}_{-t}, y_t \neq 0, \boldsymbol{\vartheta}) = 0,$$

$$\begin{aligned} P(X_t = 1|\boldsymbol{X}_{-t}, y_t = 0, \boldsymbol{\vartheta}) &\propto a_{x_{t-1},1} a_{1,x_{t+1}} \text{Pois}(0|0) \\ &\propto a_{x_{t-1},1} a_{1,x_{t+1}} \end{aligned}$$

and

$$\begin{aligned} P(X_t = j|\boldsymbol{X}_{-t}, y_t = 0, \boldsymbol{\vartheta}) &\propto a_{x_{t-1},j} a_{j,x_{t+1}} \text{Pois}(0|\lambda_{x_t}) \\ &\propto a_{x_{t-1},j} a_{j,x_{t+1}} e^{-\lambda_{x_t}} \end{aligned}$$

for $j = 2, \ldots, K$. Of course these considerations hold also for the global updating scheme, and in particular for the sampling probabilities computation and in the backward variables determination.

We applied the zero-inflated Poisson HMM, with both $K = 4$ and $K = 5$. Results related to the estimated hidden chain are summarized in Tables 2.3 and 2.4. When $K = 4$ results for the exceptional state are quite similar to what obtained by the Poisson HMM (see Table 2.2); when $K = 5$ only 38 observations are classified in the 0-inflated state 1. Then we might conclude that, even if we take into account the feature that the great majority of the observations is equal to 0, changes in terms of the exceptional events do not represent an improvement of the Poisson HMM.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 14 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4046 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2529 | 1273 | 253 | 33 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 98 | 174 | 150 | 74 | 29 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 3 | 11 | 22 | 25 | 15 | 12 | 9 | 6 | 1 | 2 | 1 | 1 | 1 | 1 |

**Table 2.3:** Telecontrol center dr3, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\text{MAP}} = x, Y_t = y\}$, $K = 4$.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 14 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 5275 | 414 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1323 | 942 | 312 | 53 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 37 | 91 | 92 | 59 | 27 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 | 6 | 14 | 21 | 15 | 12 | 9 | 6 | 1 | 2 | 1 | 1 | 1 | 1 |

**Table 2.4:** Telecontrol center dr3, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$, $K = 5$.

## 2.3 Negative binomial hidden Markov model

### 2.3.1 Model and inference

Consider a negative binomial HMM:

$$\begin{array}{rcl} Y_t | X_t = x_t & \sim & \mathrm{NegBin}(r_{x_t}, p_{x_t}), \\ \{X_t\} | A & \sim & \text{Markov chain}(A); \end{array}$$

then

$$P(Y_t = y_t | X_t = i, r_i, p_i) \quad = \quad \frac{\Gamma(y_t + r_i)}{\Gamma(r_i) y_t!} p_i^{r_i} (1 - p_i)^{y_t}.$$

Consider a sequence of length $T$ and let $(\boldsymbol{y}, \boldsymbol{X})$ be the complete-data, $(Y_1 = y_1, \ldots, Y_T = y_T, X_0 = x_0, X_1 = x_1, \ldots, X_T = x_T)$, so the complete-data likelihood function $p(\boldsymbol{y}, \boldsymbol{X} | \boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta} = (r_1, \ldots, r_K, p_1, \ldots, p_K, A)$, is given by

$$p(\boldsymbol{y}, \boldsymbol{X} | \boldsymbol{\vartheta}) = p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\vartheta}) p(\boldsymbol{X} | \boldsymbol{\vartheta}).$$

The density $p(\boldsymbol{X} | \boldsymbol{\vartheta})$ is given in equation (1.2):

$$p(\boldsymbol{X} | A) \quad = \quad \pi_{x_0} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{i,j}^{n_{ij}},$$

while

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\vartheta}) &= \prod_{t=1}^{T} p(y_t|\boldsymbol{X},\boldsymbol{r},\boldsymbol{p}) = \prod_{t=1}^{T} p(y_t|X_t=i,r_i,p_i) \\ &= \prod_{t=1}^{T} \frac{\Gamma(y_t+r_i)}{\Gamma(r_i)y_t!} p_i^{r_i}(1-p_i)^{y_t} \\ &= \prod_{i=1}^{K} \prod_{\{t:X_t=i\}} \left(\frac{\Gamma(y_t+r_i)}{\Gamma(r_i)y_t!}\right) p_i^{n_i r_i}(1-p_i)^{S_i}, \end{aligned}$$

where $n_i = \#\{1 \leq t \leq T : X_t = i\}$ and $S_i$ is the sum of observations when $X_t = i$.

In order to implement the Gibbs sampling in Algorithm 1.1 we need to compute the complete-data posterior distribution of the model parameters

$$p(\boldsymbol{\vartheta}|\boldsymbol{X},\boldsymbol{y}) \propto p(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}).$$

As pointed out by Cappé (2002), negative binomial parameters $r$ and $p$ are not independent and this has implications on the specification of the prior $p(\boldsymbol{\vartheta})$. However, by simulations, he concludes that "it is not unrealistic to assume that the mean $\mu = r\frac{1-p}{p}$ and the dispersion $p$ are independent". Then we can assume that a priori

$$p(\mu_i, p_i) = \text{Gamma}(\mu_i|a_\mu, b_\mu)\text{Beta}(p_i|a_p, b_p),$$

which gives, after the transformation $r_i = \mu_i \frac{p_i}{1-p_i}$,

$$p(r_1, p_1, \ldots, r_K, p_K) = \prod_{i=1}^{K} \text{Gamma}\left(r_i|a_\mu, b_\mu \frac{1-p_i}{p_i}\right) \text{Beta}(p_i|a_p, b_p).$$

Finally, prior on the transition matrix $A$ is

$$p(A) = \prod_{j=1}^{K} \text{Dir}(\boldsymbol{a}_j|\alpha_{j1}, \ldots, \alpha_{jK}).$$

Following inference presented in Cappé (2002), we have that parameters $p_1, \ldots, p_K$ are conditionally independent with fully conditional distribution given by

$$\begin{aligned} p(p_i|\boldsymbol{X},\boldsymbol{y},r_i) &\propto \text{Beta}\left(p_i|r_i n_i + a_p, S_i + b_p\right) \\ &\cdot \text{Gamma}\left(r_i|a_\mu, b_\mu \frac{1-p_i}{p_i}\right) \end{aligned} \tag{2.5}$$

where $n_i = \#\{1 \leq t \leq T : X_t = i\}$ and $S_i$ is the sum of observations when $X_t = i$.

The first term in (2.5) corresponds to the product of the likelihood by the marginal prior on $p_i$, whereas the second term is the prior on $r_i$ given $p_i$. In practical situation variations of the second term are rather small; then an efficient simulation procedure consists of using a Metropolis-Hastings step, where the proposed update $\bar{p}_i$ is distributed according to a Beta $(\bar{p}_i | r_i n_i + a_p, S_i + b_p)$ distribution and accepted with probability $\min(1, A_p)$, where

$$A_p = \left( \frac{\bar{p}_i}{1-\bar{p}_i} \frac{1-p_i}{p_i} \right)^{-a_\mu} e^{-r_i \left( b_\mu \frac{1-\bar{p}_i}{\bar{p}_i} - b_\mu \frac{1-p_i}{p_i} \right)}.$$

Before considering the full conditional distribution for $r_i$ we need to introduce a computational remark; the log-likelihood of i.i.d. negative binomial observations, $(z_1, \ldots, z_T)$ can be computed in two different ways:

$$\log(z_1, \ldots, z_T | p, r) = Tr\log(p) + S\log(1-p) - T\log(\Gamma(r)) + \sum_{t=1}^{T} \log(\Gamma(z_t + r))$$

where $S = \sum_{t=1}^{T} z_t$, or

$$\log(z_1, \ldots, z_T | p, r) = Tr\log(p) + S\log(1-p) + \sum_{m=1}^{M} C_m \log(r + m - 1)$$

where $M = \max(z_1, \ldots, z_T)$, $C_m = \#\{1 \leq t \leq T : z_t \geq m\}$ are the rank statistics and with the convention that the sum is null if $M = 0$.

We can now state the full conditional distribution for $r_i$:

$$\begin{aligned} p(r_i | \boldsymbol{X}, \boldsymbol{y}, p_i) \quad \propto \quad & r_i^{a_\mu - 1} \left\{ \prod_{m=1}^{M_i} (r_i + m - 1)^{C_{m,i}} \right\} \\ \cdot \quad & \exp\left\{ - \left[ b_\mu \frac{1-p_i}{p_i} - n_i \log\frac{1}{p_i} \right] r_i \right\}, \end{aligned} \tag{2.6}$$

where $M_i$ denotes the maximum value of observations when $X_t = i$ and $C_{m,i}$ are the corresponding rank statistics. The full conditional in (2.6) is closely fitted by a Gamma distribution; then we will use a Metropolis-Hastings algorithm with a Gamma proposal tuned to match the mode and the log-curvature of the full conditional. Differentiating we obtain

$$\begin{aligned}\frac{\partial \log p(r_i|\boldsymbol{X},\boldsymbol{y},p_i)}{\partial r_i} &= -\left(b_\mu \frac{1-p_i}{p_i} - n_i \log\frac{1}{p_i}\right) \\ &+ \frac{a_\mu - 1}{r_i} + \sum_{m=1}^{M_i} \frac{C_{m,i}}{r_i + m - 1}\end{aligned} \tag{2.7}$$

$$\frac{\partial^2 \log p(r_i|\boldsymbol{X},\boldsymbol{y},p_i)}{\partial^2 r_i} = -\left(\frac{a_\mu - 1}{r_i^2} + \sum_{m=1}^{M_i} \frac{C_{m,i}}{(r_i + m - 1)^2}\right). \tag{2.8}$$

Find the mode $\kappa$ of the full conditional distribution (2.6) and compute the log-curvature at the mode, $\iota$,

$$\iota = -\frac{\partial^2 \log p(r_i|\boldsymbol{X},\boldsymbol{y},p_i)}{\partial^2 r_i}$$

according to (2.8). Use a Gamma distribution with mode and log-spread matched to $\kappa$ and $\iota$ with parameters $a_r^* = 1 + \kappa^2\iota$ and $b_r^* = \kappa\iota$; sample a Gamma$(a_r^*, b_r^*)$ distributed proposal $\bar{r}_i$, which is accepted with probability $\min(1, A_r)$ where

$$\begin{aligned}A_r &= \frac{\bar{r}_i}{r_i}^{a_\mu - a_r^*} \left\{\prod_{m=1}^{M_i}\left(\frac{\bar{r}_i + m - 1}{r_i + m - 1}\right)^{C_{m,i}}\right\} \\ &\cdot \exp\left\{-\left[b_\mu \frac{1-p_i}{p_i} - n_i \log\frac{1}{p_i} - b_r^*\right](\bar{r}_i - r_i)\right\}.\end{aligned}$$

Then, inference is based on the following algorithm

**Algorithm 2.2** MCMC for a negative binomial HMM

```
Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \ldots, L_0, \ldots, L$.
  1. Sample each row of $A$ from the Dirichlet posterior distribution
  2. Sample the emission parameters trough the Metropolis-Hastings algorithm
  3. Conditional of knowing the model parameters
     a) – if $\mu_1^{(l)} = r_1^{(l)} \frac{1-p_1^{(l)}}{p_1^{(l)}} < \ldots < \mu_K^{(l)} = r_K^{(l)} \frac{1-p_K^{(l)}}{p_K^{(l)}}$ go to step b)
        – else order the mean vector $\rho(\boldsymbol{\mu})$ and $\rho(A), \rho(\boldsymbol{r}), \rho(\boldsymbol{p})$
     b) Compute the scaled backward variables
     c) Sample a path $\boldsymbol{X}$ of the hidden Markov chain trough the global updating sampler
  4. Increase $l$ and return to step 1.
```

### 2.3.2 Results

In order to implement the MCMC algorithm we need to specify the parameters for the prior distributions, the number of iterations to consider and the number of generated values to be discarded. Set parameters of the Dirichlet and Beta distributions equal to 1; regarding the Gamma prior on the mean, we set $a_\mu = 0.1$ and $b_\mu = \frac{a_\mu}{\mu_y}$, where $\mu_y$ is the mean of the observed values.

As pointed out by Cappé (2002), MCMC sampler has a very slow convergence. This feature is also verified in this study, in fact all stationarity diagnostics (qualitative and quantitative), introduced in Section 2.2 point out that the chain has not achieved the stationarity even after 50 000 iterations. Moreover, the Raftery-Lewis's diagnostic estimates a very large dependence factor. This could be due to the dependence of the parameters.

To obviate this convergence problem we hypothesized a common $p$ for each possible state. In this case the mixing problem seems to be kept down (even if not completely solved) and the dependence factors are smaller. Anyway during the sampling process when $p$ is almost equal to 1 (and this happens quite often), values of $r$ in states 3 and 4 become very large; in fact $\mu = r\frac{1-p}{p}$ and if $p$ is very close to 1, in order to increase the mean value, $r$ becomes very large. When $p$ is a little bit smaller than 1, values of $r$ precipitate to smaller values. As a result of these two situations the algorithm slowly converges and the estimated chain (with $L_0 = 30\,000$ and $L = 55\,000$) is very similar to what obtained by the Poisson HMM; this could be due to the fact that

$$\lim_{r\to\infty} \mathrm{NegBin}\left(r, \frac{r}{\lambda + r}\right) = \mathrm{Pois}(\lambda).$$

## 2.4 Compound Poisson hidden Markov model

### 2.4.1 Model and inference

Consider a compound Poisson HMM (more precisely we should say a Pólya-Aeppli HMM):

$$\begin{aligned} Y_t|X_t = x_t &\sim \mathrm{CP}(\lambda_{x_t}, q_{x_t}), \\ \{X_t\}|A &\sim \text{Markov chain}(A). \end{aligned}$$

Then

$$P(Y_t = y_t|X_t = i, \lambda_i, q_i) = e^{-\lambda_i}\sum_{n=0}^{y_t}\frac{\lambda_i^n}{n!}\frac{\Gamma(y_t)}{(y_t-n)!\Gamma(n)}q_i^n(1-q_i)^{y_t-n},$$

with $\frac{\Gamma(0)}{\Gamma(0)} = 1$.

In the complete-data likelihood function is $p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})p(\boldsymbol{X}|\boldsymbol{\vartheta})$, where

$p(\boldsymbol{X}|\boldsymbol{\vartheta})$ is, as before,

$$p(\boldsymbol{X}|A) = \pi_{x_0}\prod_{i=1}^{K}\prod_{j=1}^{K}a_{i,j}^{n_{ij}},$$

while

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta}) = \prod_{i=1}^{K}\prod_{\{t:X_t=i\}} e^{-\lambda_i}\sum_{n=0}^{y_t}\frac{\lambda_i^n}{n!}\frac{\Gamma(y_t)}{(y_t-n)!\Gamma(n)}q_i^n(1-q_i)^{y_t-n}.$$

The complete-data posterior distribution of the model parameters is

$$p(\boldsymbol{\vartheta}|\boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}) \tag{2.9}$$

where, assuming independence between parameters, the prior distribution is $p(\boldsymbol{\vartheta}) = p(A)p(\boldsymbol{\lambda})p(\boldsymbol{q})$; note that this independence hypothesis is not a strong assumption, in fact given the interpretation of the compound Poisson distribution (in page 33), it is possible to assume that the number of clusters is independent of the number of objects per cluster.
Prior on the transition matrix is

$$p(A) = \prod_{j=1}^{K} \mathrm{Dir}(\boldsymbol{a}_j|\alpha_{j1}, \ldots, \alpha_{jK});$$

assume prior independence of the Poisson means and a conjugate Gamma prior on each $\lambda_i$,

$$p(\boldsymbol{\lambda}) = \prod_{i=1}^{K} p(\lambda_i) = \prod_{i=1}^{K} \mathrm{Gamma}(\lambda_i|a_0, b_0);$$

finally,

$$p(\boldsymbol{q}) = \prod_{i=1}^{K} p(q_i) = \prod_{i=1}^{K} \mathrm{Beta}(q_i|c_0, d_0).$$

Fix $X_0 = 1$ and consider equation (2.9),

$$\begin{aligned}
p(\boldsymbol{\vartheta}|\boldsymbol{X},\boldsymbol{y}) &\propto \prod_{i=1}^{K}\left(\prod_{j=1}^{K} a_{i,j}^{n_{ij}} \prod_{\{t:X_t=i\}} e^{-\lambda_i}\sum_{n=0}^{y_t}\frac{\lambda_i^n}{n!}\frac{\Gamma(y_t)}{(y_t-n)!\Gamma(n)}q_i^n(1-q_i)^{y_t-n}\right.\\
&\cdot \left. a_{i,1}^{\alpha_{i1}-1}\cdots a_{i,K}^{\alpha_{iK}-1}\lambda_i^{a_0-1}e^{-b_0\lambda_i}q_i^{c_0-1}(1-q_i)^{d_0-1}\right)\\
&\propto \prod_{i=1}^{K}\mathrm{Dir}(\boldsymbol{a}_i|\alpha_{i1}+n_{i1},\ldots,\alpha_{iK}+n_{iK})\lambda_i^{a_0-1}e^{-b_0\lambda_i}q_i^{c_0-1}(1-q_i)^{d_0-1}\\
&\cdot \prod_{\{t:X_t=i\}} e^{-\lambda_i}(1-q_i)^{y_t}\sum_{n=0}^{y_t}\left(\frac{\lambda_i q_i}{1-q_i}\right)^n\frac{\Gamma(y_t)}{n!(y_t-n)!\Gamma(n)}\\
&= \prod_{i=1}^{K}\mathrm{Dir}(\boldsymbol{a}_i|\alpha_{i1}+n_{i1},\ldots,\alpha_{iK}+n_{iK})\lambda_i^{a_0-1}e^{-(n_i+b_0)\lambda_i}\\
&\cdot q_i^{c_0-1}(1-q_i)^{S_i+d_0-1}\prod_{\{t:X_t=i\}}\sum_{n=0}^{y_t}\left(\frac{\lambda_i q_i}{1-q_i}\right)^n\frac{\Gamma(y_t)}{n!(y_t-n)!\Gamma(n)}\\
&= \prod_{i=1}^{K}\mathrm{Dir}(\boldsymbol{a}_i|\alpha_{i1}+n_{i1},\ldots,\alpha_{iK}+n_{iK})e^{-(n_i+b_0)\lambda_i}\lambda_i^{a_0-1}\\
&\cdot q_i^{c_0-1}(1-q_i)^{S_i+d_0-1}\prod_{\tilde{y}\in\mathsf{Y}^i}\left[\sum_{n=0}^{\tilde{y}}\left(\frac{\lambda_i q_i}{1-q_i}\right)^n\frac{\Gamma(\tilde{y})}{n!(\tilde{y}-n)!\Gamma(n)}\right]^{n_{\tilde{y},i}}
\end{aligned}$$

where, for $i,j\in\{1,2,\ldots,K\}$,

$$\begin{aligned}
n_{ij} &= \#\{1\le t\le T-1 : X_t=i, X_{t+1}=j\}\\
n_i &= \#\{1\le t\le T : X_t=i\}\\
S_i &= \sum_{\{t:X_t=i\}} y_t\\
\mathsf{Y}^i &= \{\text{set of different observed values when } X_t=i\}\\
n_{\tilde{y},i} &= \#\{1\le t\le T : Y_t=\tilde{y}, \tilde{y}\in\mathsf{Y}^i, i=1,\ldots,K\}
\end{aligned}$$

The rows of the transition matrix are independent a posteriori, and, for a given trajectory $\boldsymbol{X}$ of the hidden Markov chain, are drawn from the posterior Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{a}_i|\alpha_{i1}+n_{i1},\ldots,\alpha_{iK}+n_{iK})$, $i=1,\ldots,K$. Moreover, for all $i$,

$$\begin{aligned} p(\lambda_i|\boldsymbol{X},\boldsymbol{y},q_i) &\propto e^{-(b_0+n_i)\lambda_i}\lambda_i^{a_0-1} \\ &\cdot \prod_{\tilde{y}\in\mathsf{Y}^i}\left[\sum_{n=0}^{\tilde{y}}\lambda_i^n\frac{\Gamma(\tilde{y})q_i^n}{n!(\tilde{y}-n)!\Gamma(n)(1-q_i)^n}\right]^{n_{\tilde{y},i}} \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} p(q_i|\boldsymbol{X},\boldsymbol{y},\lambda_i) &\propto q_i^{c_0-1}(1-q_i)^{S_i+d_0-1} \\ &\cdot \prod_{\tilde{y}\in\mathsf{Y}^i}\left[\sum_{n=0}^{\tilde{y}}\frac{q_i^n}{(1-q_i)^n}\frac{\lambda_i^n\Gamma(\tilde{y})}{n!(\tilde{y}-n)!\Gamma(n)}\right]^{n_{\tilde{y},i}} \end{aligned} \tag{2.11}$$

Drawing some plots it appears that the full conditionals (2.10) and (2.11) are closely fitted by, respectively, a Gamma and a Beta proposal. Thus we use a Metropolis-Hasting algorithm with Gamma and Beta proposals tuned to match the mode and the log-curvature of the full conditionals.

Let

$$\sum_{n=0}^{\tilde{y}}\lambda_i^n\frac{\Gamma(\tilde{y})q_i^n}{n!(\tilde{y}-n)!\Gamma(n)(1-q_i)^n} = f(\lambda_i,\tilde{y});$$

the logarithm of the full conditional $p(\lambda_i|\boldsymbol{X},\boldsymbol{y},q_i)$ is

$$\begin{aligned} \log p(\lambda_i|\boldsymbol{X},\boldsymbol{y},q_i) &\propto -(b_0+n_i)\lambda_i+(a_0-1)\log\lambda_i \\ &+ \sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i}\log f(\lambda_i,\tilde{y}). \end{aligned} \tag{2.12}$$

Differentiating (2.12) yields

$$\frac{\partial\log p(\lambda_i|\boldsymbol{X},\boldsymbol{y},q_i)}{\partial\lambda_i} = -(b_0+n_i)+\frac{a_0-1}{\lambda_i}+\sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i}\frac{f'(\lambda_i,\tilde{y})}{f(\lambda_i,\tilde{y})}$$

and

$$\frac{\partial^2\log p(\lambda_i|\boldsymbol{X},\boldsymbol{y},q_i)}{\partial^2\lambda_i} = -\frac{a_0-1}{\lambda_i^2}+\sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i}\frac{f''(\lambda_i,\tilde{y})f(\lambda_i,\tilde{y})-(f'(\lambda_i,\tilde{y}))^2}{f(\lambda_i,\tilde{y})^2} \tag{2.13}$$

where

$$\begin{aligned} f'(\lambda_i, \tilde{y}) &= \sum_{n=0}^{\tilde{y}} \frac{n\lambda_i^{n-1}\Gamma(\tilde{y})q_i^n}{n!(\tilde{y}-n)!\Gamma(n)(1-q_i)^n} \\ f''(\lambda_i, \tilde{y}) &= \sum_{n=0}^{\tilde{y}} \frac{n(n-1)\lambda_i^{n-2}\Gamma(\tilde{y})q_i^n}{n!(\tilde{y}-n)!\Gamma(n)(1-q_i)^n}. \end{aligned}$$

As for the negative binomial case, find the mode $\mu$ of the full conditional distribution (2.10) and compute the log-curvature at the mode, $\upsilon$,

$$\upsilon = -\frac{\partial^2 \log p(\lambda_i|\boldsymbol{X}, \boldsymbol{y}, q_i)}{\partial^2 \lambda_i}$$

according to (2.13). Use a Gamma distribution with mode and log-spread matched to $\mu$ and $\upsilon$ with parameters $\gamma = 1 + \mu^2\upsilon$ and $\delta = \mu\upsilon$; sample a Gamma$(\gamma, \delta)$ distributed proposal $\bar{\lambda}_i$, which is accepted with probability $\min(1, A_\lambda)$ where

$$A_\lambda = e^{-(n_i + b_0 - \delta)(\bar{\lambda}_i - \lambda_i)} \left(\frac{\bar{\lambda}_i}{\lambda_i}\right)^{a_0-\gamma} \prod_{\tilde{y}\in\mathsf{Y}^i} \left[\frac{f(\bar{\lambda}_i, \tilde{y})}{f(\lambda_i, \tilde{y})}\right]^{n_{\tilde{y},i}}.$$

Similar procedure for $\boldsymbol{q}$. Let

$$\sum_{n=0}^{\tilde{y}} \frac{q_i^n}{(1-q_i)^n} \frac{\lambda_i^n \Gamma(\tilde{y})}{n!(\tilde{y}-n)!\Gamma(n)} = g(q_i, \tilde{y})$$

and consider the logarithm of the full conditional (2.11)

$$\begin{aligned} \log p(q_i|\boldsymbol{X}, \boldsymbol{y}, \lambda_i) \propto\ & (S_i + d_0 - 1)\log(1-q_i) + (c_0 - 1)\log q_i \\ & + \sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i} \log g(q_i, \tilde{y}). \end{aligned} \tag{2.14}$$

First and second derivatives of (2.14) are

$$\frac{\partial \log p(q_i|\boldsymbol{X}, \boldsymbol{y}, \lambda_i)}{\partial q_i} \propto -\frac{S_i + d_0 - 1}{1-q_i} + \frac{c_0 - 1}{q_i} + \sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i} \frac{g'(q_i, \tilde{y})}{g(q_i, \tilde{y})} \tag{2.15}$$

and

$$\begin{aligned} \frac{\partial^2 \log p(q_i|\boldsymbol{X}, \boldsymbol{y}, \lambda_i)}{\partial^2 q_i} \propto\ & -\frac{S_i + d_0 - 1}{(1-q_i)^2} - \frac{c_0 - 1}{q_i^2} \\ & + \sum_{\tilde{y}\in\mathsf{Y}^i} n_{\tilde{y},i} \frac{g''(q_i, \tilde{y})g(q_i, \tilde{y}) - (g'(q_i, \tilde{y}))^2}{g(q_i, \tilde{y})^2} \end{aligned} \tag{2.16}$$

where

$$\begin{aligned} g'(q_i, \tilde{y}) &= \sum_{n=0}^{\tilde{y}} \frac{n q_i^{n-1} \lambda_i^n \Gamma(\tilde{y})}{(1-q_i)^{n+1} n!(\tilde{y}-n)!\Gamma(n)} \\ g''(q_i, \tilde{y}) &= \sum_{n=0}^{\tilde{y}} \frac{n q_i^{n-2} (n + 2q_i - 1) \lambda_i^n \Gamma(\tilde{y})}{(1-q_i)^{n+2} n!(\tilde{y}-n)!\Gamma(n)}. \end{aligned}$$

Find the mode $\rho$ of the full conditional distribution (2.11) and compute the log-curvature at the mode, $\eta$,

$$\eta = -\frac{\partial^2 \log p(q_i|\boldsymbol{X}, \boldsymbol{y}, \lambda_i)}{\partial^2 q_i}$$

according to (2.16). Use a Beta distribution with mode and log-spread matched to $\rho$ and $\eta$ with parameters $\varphi = \eta\rho^2(1-\rho) + 1$ and $\omega = \eta\rho(1-\rho)^2 + 1$; sample a Beta$(\varphi, \omega)$ distributed proposal $\bar{q}_i$, which is accepted with probability $\min(1, A_q)$, where

$$A_q = \left(\frac{\bar{q}_i}{q_i}\right)^{c_0 - \varphi} \left(\frac{1-\bar{q}_i}{1-q_i}\right)^{S_i + d_0 - \omega} \prod_{\tilde{y} \in \mathsf{Y}^i} \left[\frac{g(\bar{q}_i, \tilde{y})}{g(q_i, \tilde{y})}\right]^{n_{\tilde{y},i}}.$$

In each step of the MCMC sampler, given the model parameters, a path of the hidden chain is sampled from $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$, through the global updating scheme (see Section 1.4).

Then, inference for a compound Poisson HMM is based on the following algorithm

**Algorithm 2.3** MCMC for a compound Poisson HMM

```
Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \ldots, L_0, \ldots, L$.

1. Sample each row of $A$ from the Dirichlet posterior distribution
2. Sample the emission parameters trough the Metropolis-Hastings algorithm
3. Conditional of knowing the model parameters
    a) – if $\mu_1^{(l)} = \frac{\lambda_1^{(l)}}{q_1^{(l)}} < \ldots < \mu_K^{(l)} = \frac{\lambda_K^{(l)}}{q_K^{(l)}}$ go to step b)
       – else order the mean vector $\rho(\boldsymbol{\mu})$ and $\rho(A), \rho(\boldsymbol{\lambda}), \rho(\boldsymbol{q})$
    b) Compute the scaled backward variables
    c) Sample a path $\boldsymbol{X}$ of the hidden Markov chain trough the global updating sampler
4. Increase $l$ and return to step 1.
```

### 2.4.2 Results

For the implementation of the MCMC we set parameters of the Dirichlet distributions equal to 1, parameters for the Beta and Gamma prior, respectively on each probability $q_i$ and on each Poisson mean $\lambda_i$, $i = 1, \ldots, 4$, equal to 1.1.

The algorithm seems to reach the stationarity after 20 000 iterations. This is confirmed by the Geweke's and the Heidelberger-Welch diagnostics.

Considering 50 000 iterations of the MCMC algorithm, $L = 50\,000$, and discarding the first 20 000 generated values, $L_0 = 20\,000$, we obtain the following results:

$$\hat{A} = \begin{pmatrix} 0.9151 & 0.0761 & 0.0069 & 0.0019 \\ 0.1558 & 0.8208 & 0.0179 & 0.0055 \\ 0.0408 & 0.1777 & 0.7394 & 0.0421 \\ 0.0854 & 0.0571 & 0.1889 & 0.6686 \end{pmatrix},$$

$$\hat{\boldsymbol{q}} = (0.9416, 0.9425, 0.8858, 0.6604)$$

and

$$\hat{\boldsymbol{\lambda}} = (0.107, 0.52, 1.43, 3.263).$$

Before considering results related to the estimated hidden Markov chain, let us analyze implications of estimated model parameters. In states 1 and 2, because probabilities are almost 1, emission distribution are approximately Poisson.
In fact, consider the compound Poisson distribution

$$P(Y = y) \quad = \quad e^{-\lambda} \sum_{n=0}^{y} \frac{\lambda^n}{n!} \frac{\Gamma(y)}{(y-n)!\Gamma(n)} q^n (1-q)^{y-n}. \tag{2.17}$$

We know that when $n = 0$ the sum is equal to 1; if $q \to 1$ other addends tend to 0, unless the last one, when $n = y$, (in fact $(1-q)^0 = 1$); then

$$P(Y = y) \overset{q \to 1}{\approx} \frac{e^{-\lambda}\lambda^y}{y!}.$$

Also probability for the state 3, $q_3$, is quite close to 1; this consideration does not hold for state 4. Given the estimated emission parameters, the mean of the (mixture) distribution in each state is equal to

$$\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{q}}} = (0.1136, 0.5517, 1.6140, 4.9410)$$

while we recall that estimated means in the Poisson HMM, say $\hat{\boldsymbol{\lambda}}_{\text{POIS}}$, were

$$\hat{\boldsymbol{\lambda}}_{\text{POIS}} = (0.091, 0.596, 1.754, 5.817).$$

Means of the distributions in states 1, 2 and 3 are quite similar (in the compound Poisson and the Poisson case), while they differ in state 4; in particular the mean of the distribution in the exceptional state for the compound Poisson HMM is smaller than in the Poisson HMM.

All those considerations explain results related to the estimated hidden Markov chain, summarized in Table 2.5, where more observations are classified in the exceptional state than in the Poisson HMM (see Table 2.2).

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 14 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5645 | 463 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 992 | 899 | 289 | 54 | 18 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 36 | 84 | 80 | 52 | 21 | 14 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 2 | 10 | 12 | 17 | 14 | 12 | 12 | 9 | 6 | 1 | 2 | 1 | 1 | 1 | 1 |

**Table 2.5:** Telecontrol center dr3, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

## 2.5 Conclusions

In this Chapter we investigated the possibility to consider a parametric HMM in the analysis of electrical faults.

Starting with the Poisson HMM and analyzing the results we found that the instance that too many observations were classified in the exceptional state could be due to the fact that the Poisson distribution in state 3 concentrates small probability mass on large values.

Then we introduced the negative binomial and the compound Poisson HMMs. In the negative binomial HMM, results obtained by adapting inference presented in Cappé (2002) showed a very slow convergence of the sampling algorithm. Hypothesizing a common probability $p$ we improved the mixing problem, but results were not better than the Poisson HMM.

We developed inference for the compound Poisson HMM, but results did not satisfy our expectations. Anyway the compound Poisson model has an interpretation more suitable for the analyzed problem and even if the algorithm needs quite a lot of iterations to reach the stationarity, it does not seem to suffer the mixing problem, encountered in the negative binomial model.

In the next Chapter we will investigate the applicability of the finite HMM.

# Chapter 3

# A finite hidden Markov model for the analysis of electricity supply

## 3.1 Introduction and data inspection

We now apply a finite HMM in order to identify the exceptional events. We briefly recall notation and assumptions for the finite HMM: for each telecontrol center the observed number of electrical service faults $\{Y_t\}_{t>0}$ depends on a four state hidden Markov chain $\{X_t\}_{t\geq 0}$. Fix $X_0 = 1$, then the model can be characterized by the transition matrix $A = \{a_{i,j}\}$, with $a_{i,j} = P(X_{k+1} = j | X_k = i)$, $i, j \in \mathsf{X}$, where $\mathsf{X} = \{1, \ldots, 4\}$ is the state space of the Markov chain, and the emission matrix $B = \{b_i(y)\}$, with the conditional probabilities $b_i(y) = P(Y_k = y | X_k = i)$, $i \in \mathsf{X}$, $y \in \mathsf{Y}$, where $\mathsf{Y} = \{0, \ldots, q\}$ is the set of the observable values.

As before we will study each combination of telecontrol center and year separately from the others; first of all we analyze two (randomly chosen) telecontrol centers, denominated dg4 and dr3, for years 2004 and 2005. As 2004 was a leap year, we have 8 784 and 8 760 observations respectively for year 2004 and 2005.

Consider Table 3.1, where the telecontrol centers' principal characteristics are summarized. The great majority of the observations is equal to 0, but there are differences between the two centers: for center dg4, both years, about 90% of the observations are equal to 0, while for center dr3 the percentage is about 75%. Moreover, in center dr3 a quite large number of observations is equal to 1, 2 and 3 and the mean of the observed values is higher than the mean for center dg4.

Finally, we underline that, taking into account these features, for center dg4 year 2005 seems to be a little bit better than year 2004; the opposite for center dr3.

No information regarding the telecontrol centers, such as location, size or number of served consumers, are available. However, the final goal of the method is to identify exceptional events only on the basis of the observed performance, without taking into account external information (see Section 1.2). Moreover, some information are "indirectly" taken into account by the model, during the estimation process. In fact, because each telecontrol center is analyzed separately from the others, the models evaluates as exceptional, interruptions of different order of magnitude, regarding telecontrol center with different behavior; this fact could be a way to incorporate information about the size of the center and/or to reward "good" centers. Consider for example a center with at most 5 hourly interruptions. If the maximum is observed during an instability condition, it could be evaluated as due to an exceptional event by the model; the two situations are possible: telecontrol center is small or it serves a small number of consumers, and then 5 interruptions represent an emergence or the telecontrol center has an efficient global system and it is rewarded, by considering exceptional an observation that, compared with what observed in other centers, could be considered small.

In the following we will consider observations greater than 9 just as "many interruptions"; because the finite HMM could be considered as a nonparametric model, the introduction of this threshold does not affect the estimating method and it allows us to compare not only different telecontrol centers, but also the same center in different years. Finally, we underline that this hypothesis involves a quite small number of observations (see the last column in Table 3.1).

| | | Mean | $y = 0$ | $y \in \{1, 2, 3\}$ | $y > 9$ |
|---|---|---|---|---|---|
| dg4 | 2004 | 0.16 | 89% | 10% | 0.08% |
| | 2005 | 0.12 | 90% | 10% | 0.01% |
| dr3 | 2004 | 0.39 | 76% | 22% | 0.08% |
| | 2005 | 0.42 | 75% | 24% | 0.2% |

**Table 3.1:** Summarizing table with telecontrol centers' principal characteristics.

## 3.2 Model specification and results

In this Section we present and discuss results concerning the estimated model parameters and the hidden chain, obtained respectively by the posterior mean and the *maximum a posteriori* (MAP) estimation (see Section 1.4.5). Moreover, in order to obtain the estimated Phase-type distribution (see Section 1.6), we compute, for any MCMC iteration, $l = 1, \ldots, L$, $\{p_s^{(l)}\}$, $s \geq 1$; then the estimated probability density is $\{\hat{p}_s\}$, where $\forall s \geq 1$

$$\hat{p}_s = \frac{1}{L - L_0} \sum_{l=L_0}^{L} p_s^{(l)},$$

where, as usual, $L_0$ is the number of generated values to be discarded from the estimation.

We consider as priors on each row of the transition matrix and the emission matrix a Dirichlet distribution with all parameters equal to 1; we recall that considering observations greater than 9 as "many interruptions" the emission matrix is $B = \{b_i(y)\}$, $y \in \{0, 1, \ldots, 10^+\}$, where $10^+$ is for values from 10 to the maximum observed value.

Considering convergence diagnostics presented in Section 2.2.1, it seems that the chain reached the stationarity. All estimates are obtained considering $L = 30\,000$ generated values from the Gibbs sampler (see Section 1.4.5 for details), after the initial $L_0 = 2\,000$ initial draws have been removed.

As for the parametric models, we will say "observation $y_t$ *classified* in state $i$" to indicate that at time $t$ the underlying estimated Markov chain is in state $i$, $X_t^{\text{MAP}} = i$, and it emits the observation $y_t$.

Let us now present the results; we will first consider the telecontrol center dg4, for years 2004 and 2005 and then the center dr3. For each combination of center and year we will compare results with what obtained with the AEEG method, presented in Section 1.2. Moreover, at the end of the discussion relative to each center, a comparison between its performance in the two years will be provided; at the end of Section 3.2.2 we will propose a way for comparing the telecontrol centers.

Results related to telecontrol center dg4, year 2004 are presented and discussed in detail, while for year 2005 and for telecontrol center dr3 (years 2004, 2005) comments are introduced when necessary to underline differences and similarities eventually emerged.

## 3.2.1 Telecontrol center dg4

### Year 2004

Model parameters for a finite HMM are the transition and emission matrices. Figures 3.1 and 3.2 show the estimated transition and (transpose) emission matrices, when the center dg4, year 2004, is analyzed; the value in parenthesis is the estimate's standard deviation. We recall that the emission matrix has eleven columns, because we are considering observations greater than 9 in the same way.

$$\hat{A} = \begin{pmatrix} \underset{(0.0048)}{0.983} & \underset{(0.0048)}{0.0156} & \underset{(0.0007)}{0.0009} & \underset{(0.0004)}{0.0005} \\ \underset{(0.0419)}{0.2246} & \underset{(0.0457)}{0.7481} & \underset{(0.0135)}{0.0183} & \underset{(0.0083)}{0.009} \\ \underset{(0.0799)}{0.0867} & \underset{(0.1607)}{0.3392} & \underset{(0.1819)}{0.309} & \underset{(0.194)}{0.2651} \\ \underset{(0.0594)}{0.062} & \underset{(0.1343)}{0.1845} & \underset{(0.2057)}{0.303} & \underset{(0.2165)}{0.4505} \end{pmatrix}$$

**Figure 3.1:** Estimated transition matrix for center dg4, year 2004; the value in parenthesis is the estimate's standard deviation.

$$\hat{B}^T = \begin{pmatrix}
\underset{(0.0057)}{0.9283} & \underset{(0.0654)}{0.4534} & \underset{(0.0544)}{0.0545} & \underset{(0.0397)}{0.0403} \\
\underset{(0.0051)}{0.0631} & \underset{(0.0405)}{0.397} & \underset{(0.0705)}{0.0812} & \underset{(0.0536)}{0.0591} \\
\underset{(0.0012)}{0.0064} & \underset{(0.0248)}{0.0974} & \underset{(0.0954)}{0.1591} & \underset{(0.0685)}{0.0859} \\
\underset{(0.0004)}{0.001} & \underset{(0.0127)}{0.0251} & \underset{(0.0958)}{0.1815} & \underset{(0.0818)}{0.1323} \\
\underset{(0.0002)}{0.0004} & \underset{(0.005)}{0.0052} & \underset{(0.08)}{0.1339} & \underset{(0.0765)}{0.1067} \\
\underset{(0.0001)}{0.0002} & \underset{(0.0048)}{0.0055} & \underset{(0.0679)}{0.1044} & \underset{(0.0672)}{0.1023} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0037)}{0.0046} & \underset{(0.0516)}{0.0659} & \underset{(0.0548)}{0.0746} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0025)}{0.0029} & \underset{(0.032)}{0.0335} & \underset{(0.038)}{0.0455} \\
\underset{(0.0001)}{0.0002} & \underset{(0.0041)}{0.005} & \underset{(0.0585)}{0.0696} & \underset{(0.0738)}{0.1264} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0018)}{0.0018} & \underset{(0.0443)}{0.0433} & \underset{(0.0647)}{0.1021} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0022)}{0.0021} & \underset{(0.0583)}{0.0731} & \underset{(0.0723)}{0.1248}
\end{pmatrix}$$

**Figure 3.2:** Estimated emission matrix for center dg4, year 2004; the value in parenthesis is the estimate's standard deviation.

Consider Figure 3.1; the probability of staying in states 1, $a_{1,1}$, is almost equal to 1: when the global system is in state 1, it means that it is in the normal working status and this is the most common situation. Also the probability of staying in the exceptional state, $a_{4,4}$ is quite high: this means that, when the system enters in an exceptional operating status, it is highly probable that it remains in it for some instants, before visiting other states.

Regarding the emission mechanism, by Figure 3.2 it emerges that: states 1 and 2 mainly "emit" observations equal to 0 and 1, when the chain is in state 3 it is highly probable to have from 3 to 5 faults and when the underling system is in state 4, it is likely to observe higher values.

As we recalled in Section 2.2.1 an HMM is a generalization of a mixture model. Each row of the estimated emission matrix (each column in Figure 3.2) is a probability distribution and represents the mixture component. Each plot in Figure 3.3 compares each probability mass functions with a Poisson distribution with the same mean (*i.e.* each mean is equal to $\sum_{j=0}^{10} j \cdot b_i(j)$, with $i = 1, 2, 3, 4$);

a similar comparison, but from a different point of view, is provided in Figure 3.4. Distributions in state 1 and 2 are similar, while distributions in states 3 and 4 are quite different; in particular, for the finite HMM, distributions in state 3 and 4 concentrate more mass on larger values, than the Poisson distributions. The irregular form of the estimated mixture components in states 3 and 4 underlines that the finite HMM is more flexible than the parametric one and suggests the key to understand the reason that the parametric models did not provide satisfactory results.

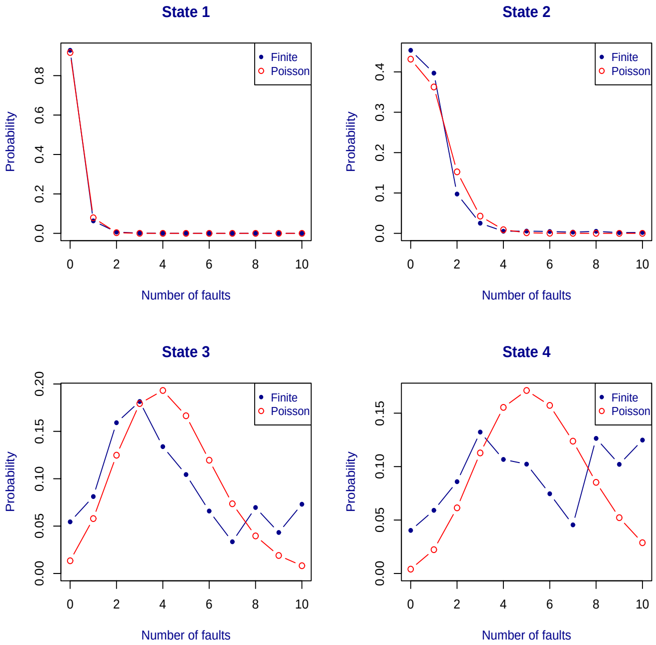


**Figure 3.3:** Finite HMM mixture component and Poisson distribution with the same mean for state 1 (top-left), state 2 (top-right), state 3 (bottom-left) and state 4 (bottom-right).

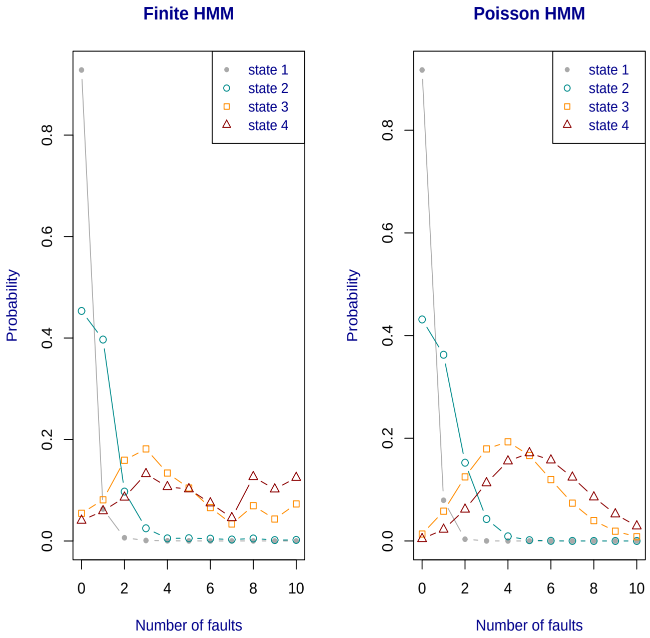


**Figure 3.4:** Finite HMM mixture components and Poisson distributions, with the same means.

We now analyze in detail the estimated underlying Markov chain; to better evaluate the model we consider the values really observed (in fact, we recall that in the MCMC implementation we relabelled observations greater than 9 with 10).

Results concerning the hidden chain are represented in the Figure 3.5, where observations have different colors based on the state of the Markov chain, and in Table 3.2, where each value $n_{xy}$ is the number of observations equal to $y$, classified in state $x$, *i.e.* $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

Figure 3.5 could be misleading because the time dependence is hidden in the sense that, even if the number of faults were observed in different hours, they seem to be on the same line and then relative to the same time. Anyway, by this plot it is possible to see when exceptional events occurred; for example for center dg4 the exceptional periods are related to the second half of the year, unless a peak in the first three months.

By Table 3.2 we would like to underline that observations classified in state 4 are not that one bigger than a threshold (as in the AEEG method), but the scenario is more uneven.

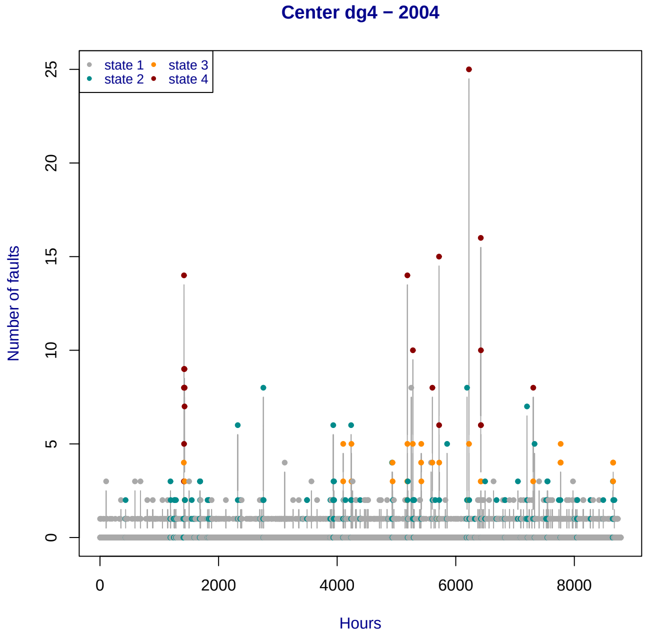


**Figure 3.5:** Telecontrol center dg4, year 2004: observations with different colors based on the state of the estimated underling Markov chain.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 14 | 15 | 16 | 25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7739 | 563 | 58 | 9 | 2 | 0 | 0 | 0 | $\underline{1}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 94 | 187 | 55 | 12 | 1 | 2 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 3 | 10 | 10 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 2 | 0 | 1 | 3 | 1 | 6 | 4 | 2 | 2 | 1 | 1 | 1 |

**Table 3.2:** Telecontrol center dg4, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\text{MAP}} = x, Y_t = y\}$. There is an observation (underlined in the table) high, but classified in state 1: this is due to the fact that the previous and following observations are equal to 0.

Consider Table 3.2; the great majority of observations (about 95%) are classified in state 1, principally equal to 0, intermediate values are assigned to state 3, 24 observations are declared as due to an exceptional event. There is a quite high number of interruptions (8, observation relative to 6th August, 5 PM) classified in state 1; it is due to the fact that the previous and the following observations are equal to 0 and that the probability of staying in state 1 is much higher than the emission probability, corresponding to 8, in other states (3 and 4). Note that the fact that the model does not consider as exceptional that observation is coherent with the idea, incorporated in the AEEG method, that an exceptional event causes several interruptions protracting in time.

Let us now focus our attention on the exceptional observations that are the core of the analysis. The majority of the exceptional hours (13) are related to February, while the others are referred to the second part of the year (in particular to the period August - October).

The exceptional hours belong to eight exceptional excursions (introduced in Section 1.6), represented in Figure 3.6. In each plot the title indicates the month and the abscissa's labels the day and the hour when the exceptional excursion starts and finishes; for example first plot (top-left) shows an exceptional excursion occurred in February, started the 28th at noon and ended the 29th at 5 PM.

The exceptional excursions describe the behavior of the system near an exceptional event. By Figure 3.6 we can see that the chain gradually reaches and leaves the exceptional state 4; therefore it seems that before and after an exceptional event an instability condition, causing an intermediate number of faults, occurs. Note that this feature is also taken into account by the AEEG method (see Section 1.2) that considers exceptional also the 3 hours before the beginning and after the end of an exceptional interval.

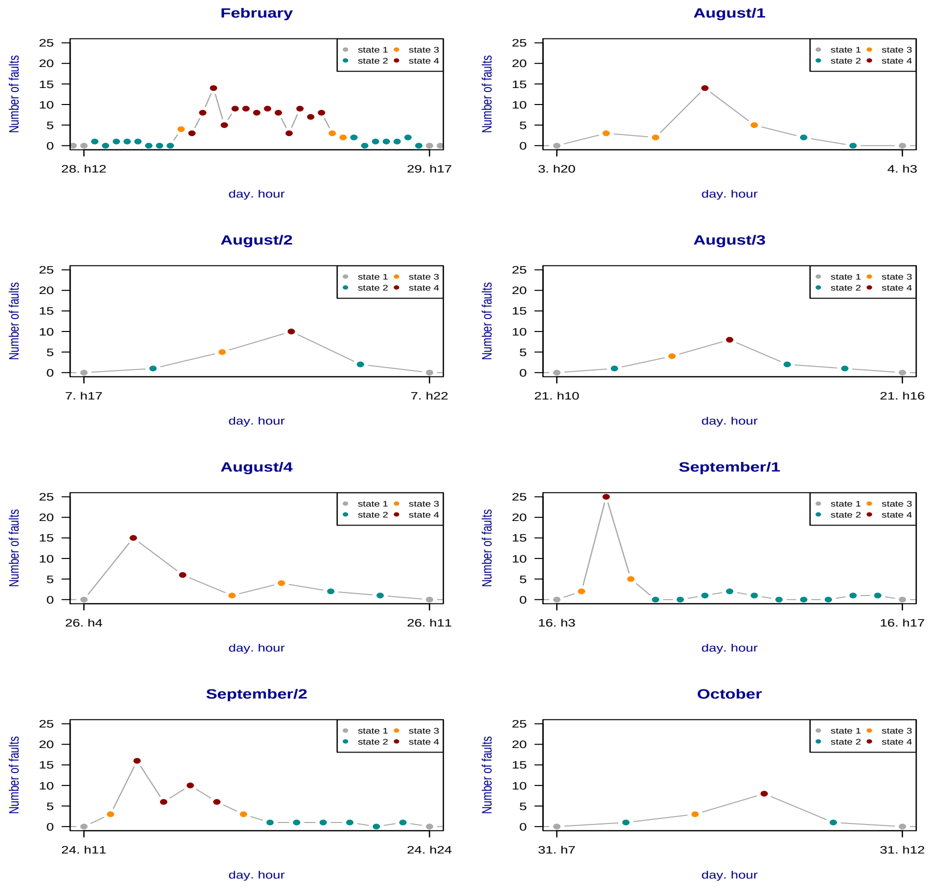


**Figure 3.6:** Telecontrol center dg4, year 2004: exceptional excursions.

We now compare results obtained by the finite HMM and by the AEEG methodology (see Section 1.2).

Results for the finite HMM are obtained considering the hourly number of faults and without considering information relative to other years. Then in order to compare results we compute the *exceptionality threshold* $q_\alpha$ using the 6 hours time-interval data (as in the AEEG method), but without considering the regression function; moreover we declare (label) the hours belonging to the obtained exceptional periods as “AEEG exceptional”.

Exceptional excursions are related to the realization of the Markov chain (they are not referred to the time), while by the AEEG method we obtain the exceptional hours. Then in order to compare results, given an exceptional excursion we label as “HMM Exceptional” - HE - the hours corresponding to that excursion.

By Figure 3.7 we can see that there is a concordance between the HMM and the AEEG exceptional hours (respectively the first and the second row). Given the large number of observations (8 784) and the differences, also conceptual, in the applied methods it is almost impossible that the exceptional periods perfectly overlap.

By Figure 3.7 we can deduce that some EPs have no intersection with any HE. Subsequences corresponding to those EPs are presented in the following, where the estimated hidden chain and the observed values are considered:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}^{\text{MAP}}$ | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| $\boldsymbol{y}$ | 0 | 1 | 2 | 8 | 0 | 1 | 1 | 1 | 2 | 0 | 0 | 0 |

in April

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}^{\text{MAP}}$ | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| $\boldsymbol{y}$ | 0 | 1 | 1 | 2 | 8 | 1 | 1 | 0 | 0 | 1 |

in September

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}^{\text{MAP}}$ | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 |
| $\boldsymbol{y}$ | 0 | 0 | 0 | 2 | 0 | 4 | 4 | 5 | 4 | 0 | 0 | 0 | 0 |

in November and

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}^{\text{MAP}}$ | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| $\boldsymbol{y}$ | 3 | 1 | 1 | 3 | 4 | 4 | 2 | 1 | 0 | 1 | 0 | 1 |

in December.
The Markov chain “recognizes” an instability condition (in fact the chain is in states 2 or 3), but it does not consider the event exceptional.

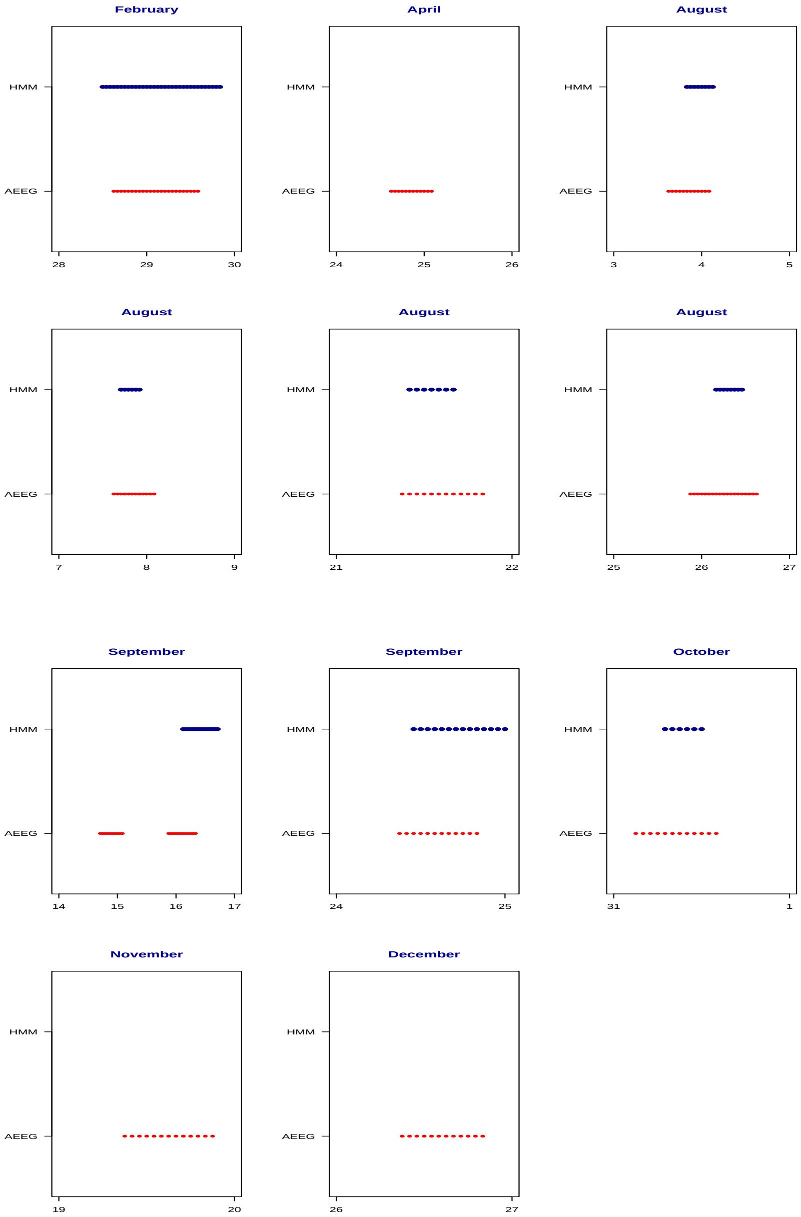


**Figure 3.7:** Telecontrol center dg4, year 2004: hours in the exceptional excursions and the AEEG exceptional hours. The plot's title indicates the month and the abscissa's label the considered days.

In conclusion, the HMM-based method supports the methodology adopted by the AEEG. In particular it identifies as exceptional, the events that cause a large number of interruptions protracting in time; moreover it gathers the feature that an exceptional situation (for example a particularly serious meteorological disturbance) is preceded and followed by an instable situation. Moreover, by the HMM approach we can also deduce considerations related to the behavior of the system when an exceptional event occurs; this dealing in fact is described by the transition matrix of the underlying Markov chain. As we said in Section 1.6 the number of hours that the chain passes in state 4 has a geometric distribution with parameter $1 - a_{4,4}$; for telecontol center dg4, year 2004, the expected value is about 3. Moreover, considering the distribution of the length of the exceptional excursions we analyze the time needed to the system to reestablish the normal operating status. Figure 3.8 shows the estimated Phase-type distribution (presented and discussed in Section 1.6); in the legend the fundamental proprieties of the distribution. Comparing the estimated Phase-type distribution and the identified exceptional excursions, we can see that the mean of the Phase-type distribution is much larger, while the mode is more of comparable size.

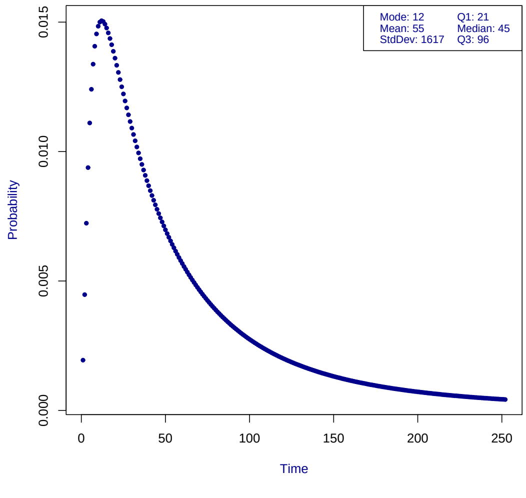


**Figure 3.8:** Telecontrol center dg4, year 2004: estimated Phase-type distribution; in the legend the fundamental properties.

### Year 2005

Consider now the results for telecontrol center dg4, year 2005. Considerations similar to those described for the year 2004 are valid for the estimated transition and emission matrices (displayed in Figures 3.9 and 3.10).

Results concerning the estimated hidden chain are represented in Figure 3.11, where, as before, observations have different colors based on the state of the hidden Markov chain, and in Table 3.3, where each value $n_{xy}$ is the number of observations equal to $y$, classified in state $x$, *i.e.* $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

Just two observations are classified as due to an exceptional event and they are in an exceptional excursion represented in Figure 3.12.

$$\hat{A} = \begin{pmatrix} \underset{(0.0045)}{0.9881} & \underset{(0.0045)}{0.0113} & \underset{(0.0003)}{0.0003} & \underset{(0.0002)}{0.0003} \\ \underset{(0.0528)}{0.2763} & \underset{(0.0554)}{0.6994} & \underset{(0.0134)}{0.0139} & \underset{(0.01)}{0.0104} \\ \underset{(0.1334)}{0.1506} & \underset{(0.1933)}{0.3349} & \underset{(0.1847)}{0.2737} & \underset{(0.1836)}{0.2407} \\ \underset{(0.1354)}{0.1512} & \underset{(0.1893)}{0.2941} & \underset{(0.2001)}{0.2897} & \underset{(0.1801)}{0.2651} \end{pmatrix}$$

**Figure 3.9:** Estimated transition matrix for center dg4, year 2005; the value in parenthesis is the estimate's standard deviation.

$$\hat{B}^T = \begin{pmatrix}
\underset{(0.0053)}{0.9219} & \underset{(0.0753)}{0.3777} & \underset{(0.0934)}{0.1044} & \underset{(0.0784)}{0.0861} \\
\underset{(0.0045)}{0.0706} & \underset{(0.046)}{0.401} & \underset{(0.0983)}{0.1149} & \underset{(0.0845)}{0.0951} \\
\underset{(0.0013)}{0.0062} & \underset{(0.0375)}{0.1648} & \underset{(0.0926)}{0.1115} & \underset{(0.0869)}{0.1011} \\
\underset{(0.0003)}{0.0005} & \underset{(0.0154)}{0.0239} & \underset{(0.1076)}{0.1418} & \underset{(0.1027)}{0.1297} \\
\underset{(0.0001)}{0.0002} & \underset{(0.0074)}{0.0077} & \underset{(0.1038)}{0.1522} & \underset{(0.109)}{0.1564} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0073)}{0.0086} & \underset{(0.0702)}{0.0769} & \underset{(0.0743)}{0.0822} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0038)}{0.0035} & \underset{(0.0613)}{0.0681} & \underset{(0.0666)}{0.0767} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0033)}{0.0032} & \underset{(0.0489)}{0.0481} & \underset{(0.0543)}{0.0546} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0036)}{0.0033} & \underset{(0.0615)}{0.0672} & \underset{(0.0693)}{0.0809} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0033)}{0.0031} & \underset{(0.0494)}{0.0485} & \underset{(0.0542)}{0.0544} \\
\underset{(0.0001)}{0.0001} & \underset{(0.0036)}{0.0032} & \underset{(0.0604)}{0.0664} & \underset{(0.0694)}{0.0828}
\end{pmatrix}$$

**Figure 3.10:** Estimated emission matrix for center dg4, year 2005; the value in parenthesis is the estimate's standard deviation.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 7850 | 646 | 55 | 4 | 0 | 0 | 0 | 0 | 0 |
| 2 | 42 | 88 | 53 | 8 | 2 | 3 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 2 | 4 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Table 3.3:** Telecontrol center dg4, year 2005: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\text{MAP}} = x, Y_t = y\}$.

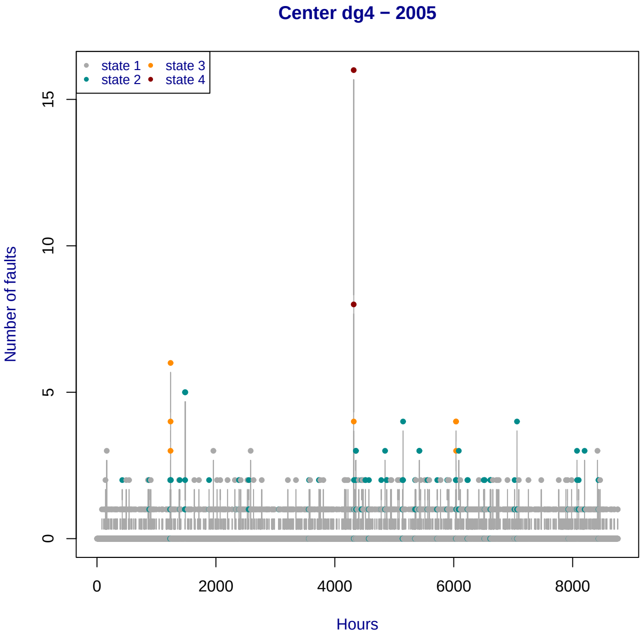


**Figure 3.11:** Telecontrol center dg4, year 2005: observations with different colors based on the state of the underlying Markov chain.

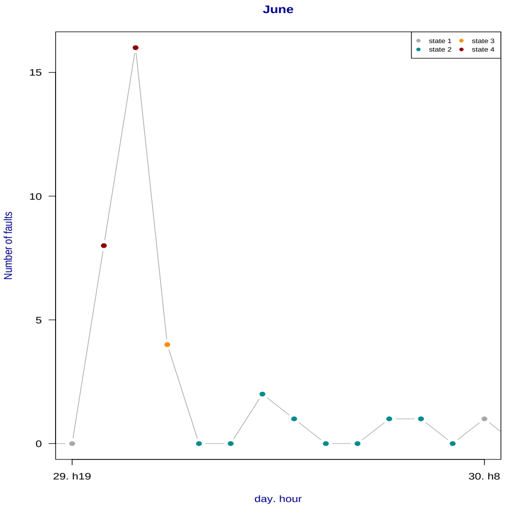


**Figure 3.12:** Telecontrol center dg4, year 2005: exceptional excursion; title indicates the month and the abscissa's labels the day and the hour when the exceptional excursion starts and finishes.

By the AEEG methodology more than one period are considered exceptional (see Figure 3.13); observations corresponding to those periods are presented in the following subsequences:

| $\boldsymbol{X}^{\text{MAP}}$ | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 1 | 3 | 4 | 6 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 |

| $\boldsymbol{X}^{\text{MAP}}$ | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 0 | 0 | 0 | 1 | 1 | 5 | 1 | 2 | 5 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| $\boldsymbol{X}^{\text{MAP}}$ | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{y}$ | 0 | 2 | 0 | 2 | 4 | 4 | 3 | 0 | 1 | 2 | 0 | 0 | 0 |

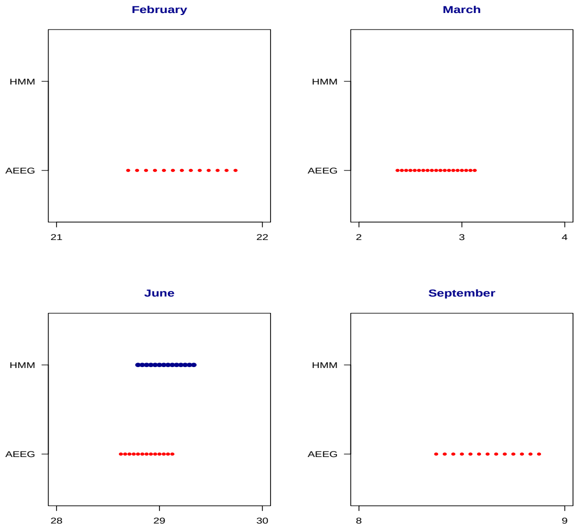


**Figure 3.13:** Telecontrol center dg4, year 2005: hours in the exceptional excursion and the AEEG exceptional hours. The plot's title indicates the month and the abscissa's label the considered days.

Considering the estimated transition matrix we can deduce that the expected number of hours that the underlying Markov chain passes in the exceptional state is about 2; the estimated Phase-type distribution, describing the length of the exceptional excursions, for center dg4, year 2005, is shown in Figure 3.14; as before in the legend the fundamental proprieties of the distribution.

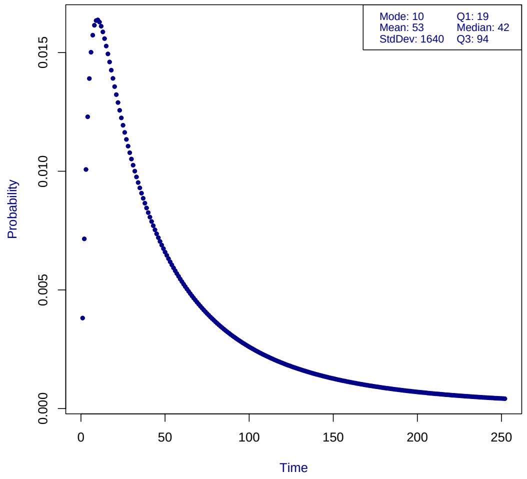


**Figure 3.14:** Telecontrol center dg4, year 2005: estimated Phase-type distribution; in the legend the fundamental properties.

Comparing the estimated transition matrices for center dg4, years 2004 and 2005 (see Figures 3.1 and 3.9), we could understand if the transition dynamic of the underlying system is changing; however even if we could check that transition probabilities are "more or less similar" it is more difficult to understand if situation is improving or not. The Phase-type distribution, obtained by the transition probabilities, provides a more visible comparison and it has a technical interpretation (in fact we recall that it is the distribution of the time needed to the system to reestablish the normal operating status). Consider Figure 3.15, where the estimated Phase-type distributions, for both years, are plotted: they are quite similar, but the distribution relative to year 2005 concentrates more probability on smaller values showing a (little bit) higher probability of having shorter exceptional excursions. Note that the difference between the two years is not so significant as we could expect just considering the number of observations classified in state 4 or the exceptional excursions. In fact, we recall that 24 observations were considered as due to an exceptional event in 2004, while in 2005 just 2; moreover, those exceptional hours are in 8 exceptional excursions in 2004 and in just 1 in 2005. This is due to the fact that given a transition matrix generated paths could be very different and then even if the estimated Markov chains are quite different it does not necessarily imply that the relative transition matrices are different.

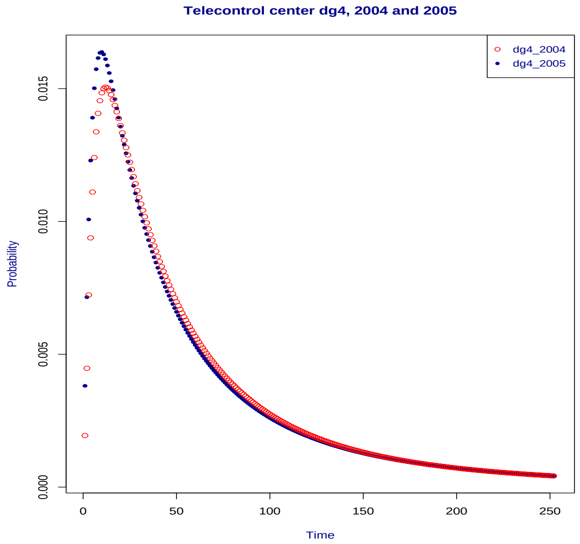


**Figure 3.15:** Telecontrol center dg4: estimated Phase-type distributions for years 2004 and 2005.

## 3.2.2 Telecontrol center dr3

Consider now the results concerning telecontrol center dr3.

### Year 2004

As we have already underlined, observed values for this center are different from that for center dg4 and this obviously affects the estimate of the model parameters and the hidden chain.

In particular, as it can be seen in Figure 3.16, the instance that there is a larger number of observations equal to 1, 2 and 3 influences the probabilities of staying in state 2 and 3, that unlike center dg4, are quite high. This remark is better explained considering the estimated emission matrix as well (Figure 3.17) and noticing that when the chain is in states 2 and 3 it is highly probable to have 1, 2, or 3 faults and because this often happens in the data, the resulting probabilities of staying in states 2 and 3 become high.

$$\hat{A} = \begin{pmatrix} \underset{(0.0102)}{0.9552} & \underset{(0.0102)}{0.0429} & \underset{(0.0011)}{0.0014} & \underset{(0.0005)}{0.0005} \\ \underset{(0.025)}{0.1739} & \underset{(0.0265)}{0.8049} & \underset{(0.0083)}{0.0159} & \underset{(0.0051)}{0.0053} \\ \underset{(0.0464)}{0.0463} & \underset{(0.117)}{0.2614} & \underset{(0.1862)}{0.5129} & \underset{(0.1518)}{0.1794} \\ \underset{(0.0487)}{0.0482} & \underset{(0.1055)}{0.1271} & \underset{(0.1582)}{0.2965} & \underset{(0.1698)}{0.5282} \end{pmatrix}$$

**Figure 3.16:** Estimated transiton matrix for center dr3, year 2004; the value in parenthesis is the estimate's standard deviation.

$$\hat{B}^T = \begin{pmatrix} \underset{(0.0119)}{0.8584} & \underset{(0.0446)}{0.4295} & \underset{(0.0446)}{0.0543} & \underset{(0.0448)}{0.0404} \\ \underset{(0.0091)}{0.1182} & \underset{(0.0222)}{0.3583} & \underset{(0.067)}{0.1013} & \underset{(0.0603)}{0.0504} \\ \underset{(0.0034)}{0.0181} & \underset{(0.0192)}{0.1444} & \underset{(0.0834)}{0.2209} & \underset{(0.1057)}{0.0798} \\ \underset{(0.0011)}{0.002} & \underset{(0.0107)}{0.044} & \underset{(0.0807)}{0.1838} & \underset{(0.0834)}{0.1263} \\ \underset{(0.0006)}{0.0012} & \underset{(0.0053)}{0.0121} & \underset{(0.0731)}{0.1618} & \underset{(0.0848)}{0.131} \\ \underset{(0.0005)}{0.0009} & \underset{(0.0034)}{0.0061} & \underset{(0.0536)}{0.0899} & \underset{(0.0756)}{0.1314} \\ \underset{(0.0002)}{0.0003} & \underset{(0.0016)}{0.0022} & \underset{(0.0404)}{0.0699} & \underset{(0.0502)}{0.0602} \\ \underset{(0.0002)}{0.0003} & \underset{(0.0008)}{0.0009} & \underset{(0.0351)}{0.0382} & \underset{(0.0666)}{0.1218} \\ \underset{(0.0002)}{0.0004} & \underset{(0.0007)}{0.0008} & \underset{(0.0241)}{0.0278} & \underset{(0.0449)}{0.0687} \\ \underset{(0.0001)}{0.0001} & \underset{(0.0008)}{0.0008} & \underset{(0.0246)}{0.0309} & \underset{(0.041)}{0.0545} \\ \underset{(0.0001)}{0.0001} & \underset{(0.0008)}{0.0009} & \underset{(0.0249)}{0.0212} & \underset{(0.0621)}{0.0953} \end{pmatrix}$$

**Figure 3.17:** Estimated emission matrix for center dr3, year 2004; the value in parenthesis is the estimate's standard deviation.

Consider now the estimated hidden chain, and the summarizing Figure 3.18 and Table 3.4, that respectively represent the observations with different colors based on the underlying chain, and the number of observations equal to $y$ classified in state $x$.

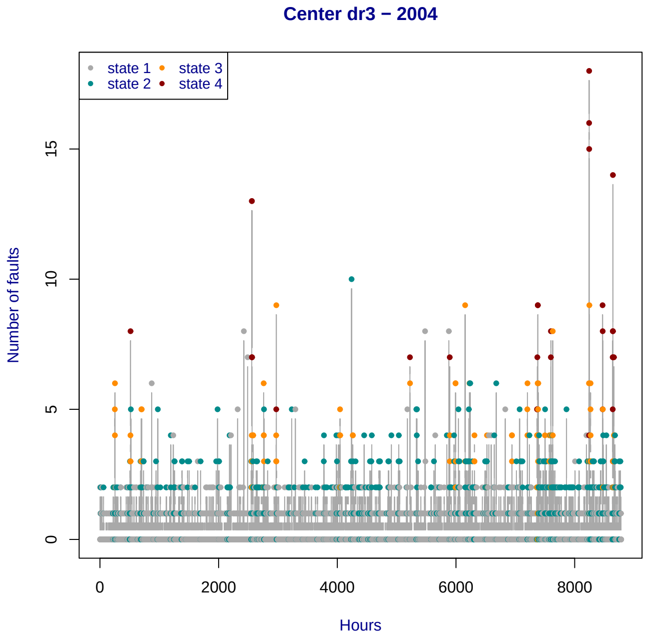


**Figure 3.18:** Telecontrol center dr3, year 2004: observations with different colors based on the state of the underlying Markov chain.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 14 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6312 | 870 | 112 | 4 | 6 | 4 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 360 | 575 | 271 | 88 | 20 | 13 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 3 | 23 | 25 | 29 | 9 | 11 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 1 | 8 | 0 | 11 | 5 | 3 | 0 | 2 | 1 | 1 | 1 | 1 |

**Table 3.4:** Telecontrol center dr3, year 2004: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\text{MAP}} = x, Y_t = y\}$.

Figure 3.19 shows the exceptional excursions containing the 35 hours classified as due to an exceptional event; as it can be also seen in Figure 3.4 exceptionality are distributed in the whole year.

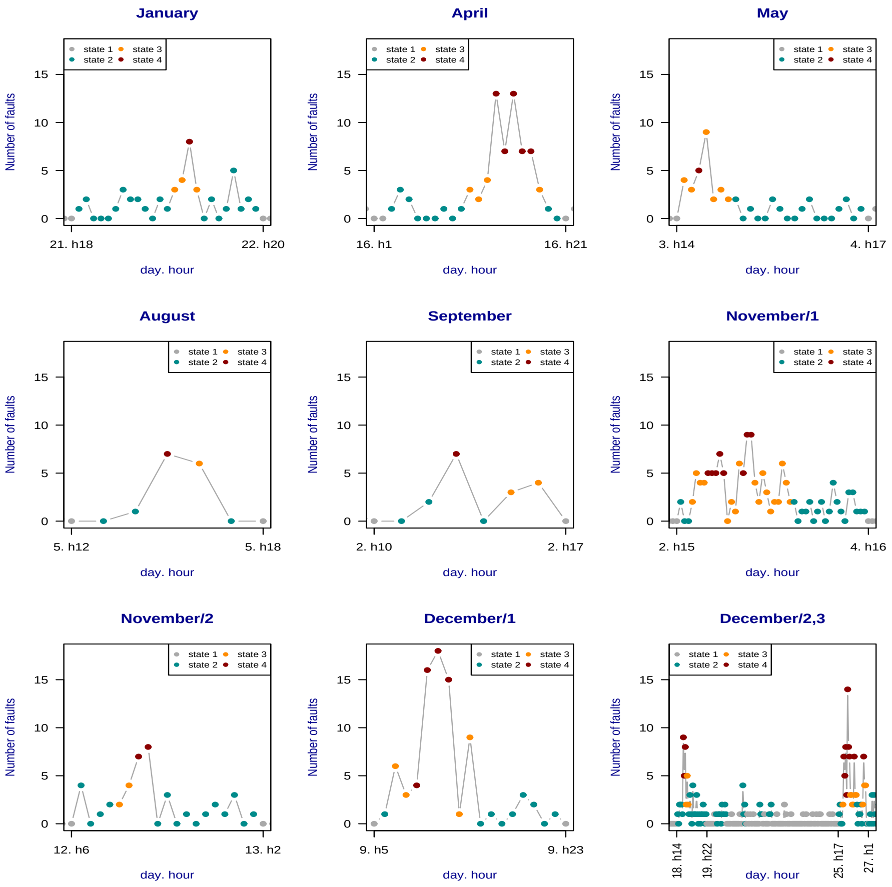


**Figure 3.19:** Telecontrol center dr3, year 2004: exceptional excursions; in each plot title indicates the month and the abscissa's labels the day and the hour when the exceptional excursion starts and finishes. Note that the last two excursions, related to the second part of December, are plotted together in the last picture (bottom-right).

Comparing results with what obtained by the AEEG methodology we can see that, on the contrary of what we had for the telecontrol center dg4 (both years), by the HMM more periods are considered exceptional than the AEEG method does (see Figure 3.20). Of course we do not consider observations and the estimated state in those periods, because they are the corresponding exceptional excursions (those relative to January, August, September and November) plotted in Figure 3.19.

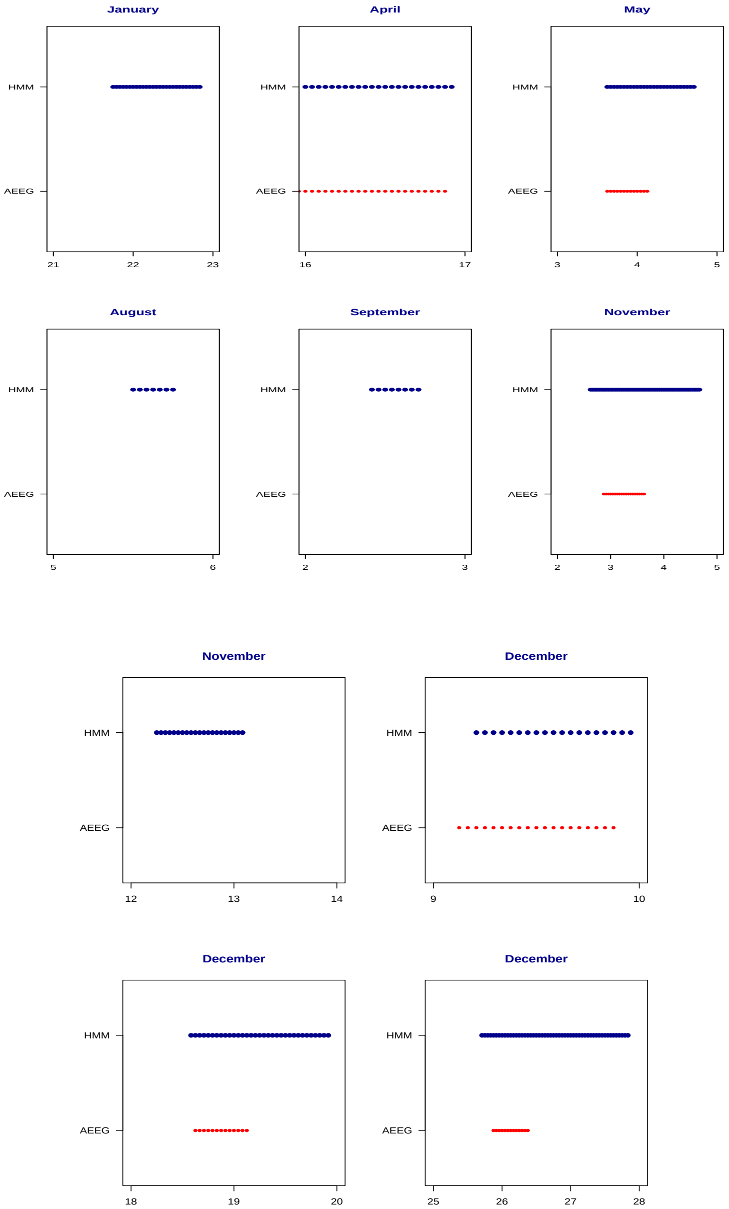


**Figure 3.20:** Telecontrol center dr3, year 2004: hours in the exceptional excursions and AEEG exceptional hours. The plot's title indicates the month and the abscissa's label the considered days.

The expected number of hours that the underlying Markov chain passes in the exceptional state is about 2. Figure 3.21 represents the estimated Phase-type distribution, describing the distribution of the length of the exceptional excursions, resulting from the $L - L_0$ generated values by the MCMC; in the legend the fundamental proprieties of the distribution.

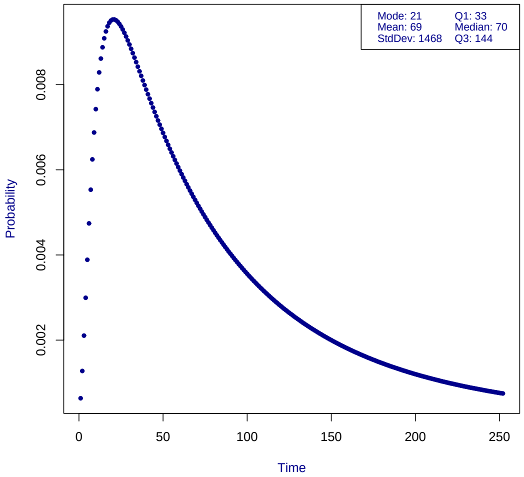


**Figure 3.21:** Telecontrol center dr3, year 2004: estimated Phase-type distribution; in the legend the fundamental properties.

### Year 2005

Considerations on the estimated model parameters (see Figures 3.23 and 3.22) discussed for year 2004, remain valid for the year 2005.

From Figure 3.24 it emerges that the exceptional hours are concentrated in the last part of the year; in fact, considering the exceptional excursions (represented in Figure 3.25), it can be seen that the period containing exceptional excursions is from July to December.

$$\hat{B}^T = \begin{pmatrix} \underset{(0.0094)}{0.8298} & \underset{(0.0337)}{0.393} & \underset{(0.0578)}{0.0598} & \underset{(0.0287)}{0.0292} \\ \underset{(0.0069)}{0.1406} & \underset{(0.0197)}{0.3323} & \underset{(0.0674)}{0.0726} & \underset{(0.0319)}{0.033} \\ \underset{(0.0032)}{0.0221} & \underset{(0.0177)}{0.1712} & \underset{(0.094)}{0.1239} & \underset{(0.0425)}{0.0544} \\ \underset{(0.0012)}{0.003} & \underset{(0.0105)}{0.0667} & \underset{(0.0741)}{0.098} & \underset{(0.0481)}{0.074} \\ \underset{(0.0007)}{0.0017} & \underset{(0.0063)}{0.0193} & \underset{(0.0949)}{0.1756} & \underset{(0.0546)}{0.0692} \\ \underset{(0.0004)}{0.0007} & \underset{(0.0033)}{0.0065} & \underset{(0.0801)}{0.1405} & \underset{(0.0522)}{0.0694} \\ \underset{(0.0003)}{0.0004} & \underset{(0.003)}{0.005} & \underset{(0.0848)}{0.1394} & \underset{(0.0594)}{0.0997} \\ \underset{(0.0002)}{0.0003} & \underset{(0.0013)}{0.0016} & \underset{(0.0561)}{0.0608} & \underset{(0.0654)}{0.177} \\ \underset{(0.0003)}{0.0007} & \underset{(0.001)}{0.0011} & \underset{(0.0404)}{0.0423} & \underset{(0.0392)}{0.0624} \\ \underset{(0.0002)}{0.0003} & \underset{(0.0012)}{0.0015} & \underset{(0.0361)}{0.0349} & \underset{(0.0403)}{0.0711} \\ \underset{(0.0002)}{0.0004} & \underset{(0.0014)}{0.0018} & \underset{(0.0539)}{0.0522} & \underset{(0.078)}{0.2606} \end{pmatrix}$$

**Figure 3.22:** Estimated emission matrix for center dr3, year 2005; the value in parenthesis is the estimate's standard deviation.

$$\hat{A} = \begin{pmatrix} \underset{(0.0073)}{0.9606} & \underset{(0.0073)}{0.0382} & \underset{(0.0009)}{0.0009} & \underset{(0.0002)}{0.0003} \\ \underset{(0.0226)}{0.1801} & \underset{(0.0248)}{0.7997} & \underset{(0.0103)}{0.0153} & \underset{(0.0039)}{0.0049} \\ \underset{(0.1020)}{0.1064} & \underset{(0.1774)}{0.4657} & \underset{(0.1432)}{0.232} & \underset{(0.1379)}{0.1959} \\ \underset{(0.0435)}{0.0436} & \underset{(0.1007)}{0.1377} & \underset{(0.1322)}{0.2151} & \underset{(0.1299)}{0.6036} \end{pmatrix}$$

**Figure 3.23:** Estimated transition matrix for center dr3, year 2005; the value in parenthesis is the estimate's standard deviation.

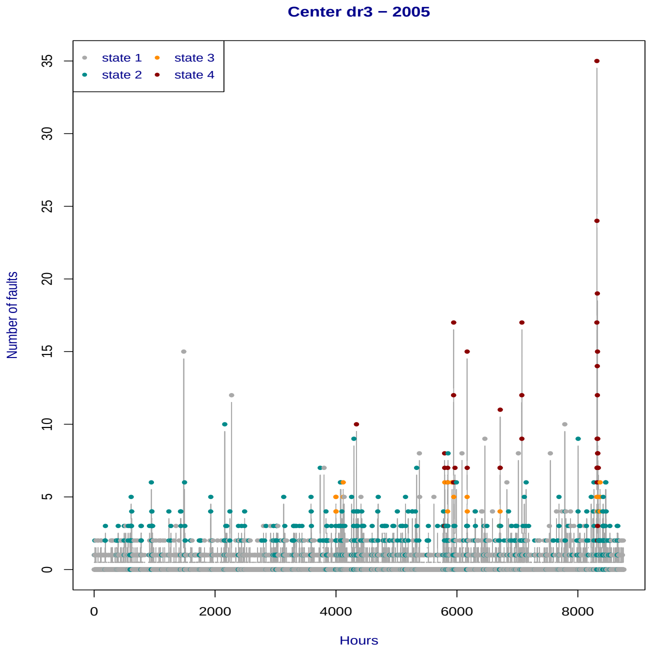


**Figure 3.24:** Telecontrol center dr3, year 2005: observations with different colors based on the state of the underlying Markov chain.

| $n_{xy}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 15 | 17 | 19 | 24 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6276 | 1108 | 156 | 12 | 10 | 4 | 1 | 1 | 4 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 273 | 410 | 268 | 113 | 34 | 11 | 9 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 8 | 8 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 2 | 0 | 0 | 4 | 10 | 3 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 1 | 1 | 1 |

**Table 3.5:** Telecontrol center dr3, year 2005: summarizing table containing $n_{xy} = \#\{1 \leq t \leq T : X_t^{\mathrm{MAP}} = x, Y_t = y\}$.

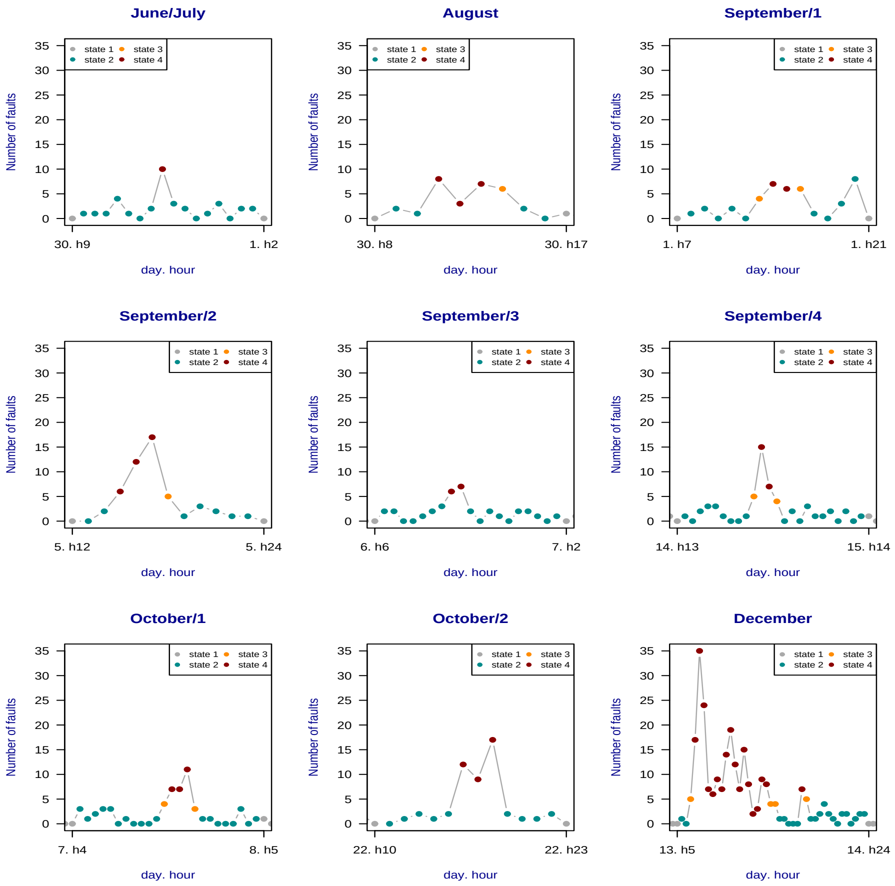


**Figure 3.25:** Telecontrol center dr3, year 2005: exceptional excursions; in each plot title indicates the month and the abscissa's labels the day and the hour when the exceptional excursion starts and finishes.

As for 2004, by the HMM more periods are considered exceptional than the AEEG method does (see Figure 3.26).

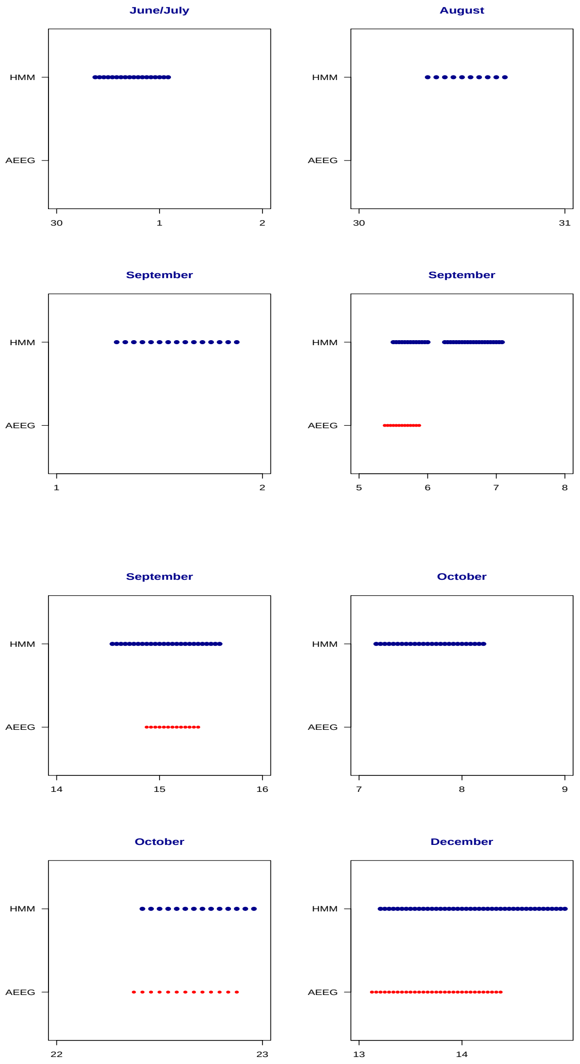


**Figure 3.26:** Telecontrol center dr3, year 2005: hours in the exceptional excursions and AEEG exceptional hours. The plot's title indicates the month and the abscissa's label the considered days.

The expected number of hours that the chain passes in the exceptional state 4 is about 3; moreover the estimated Phase-type distribution is represented in Figure 3.27).

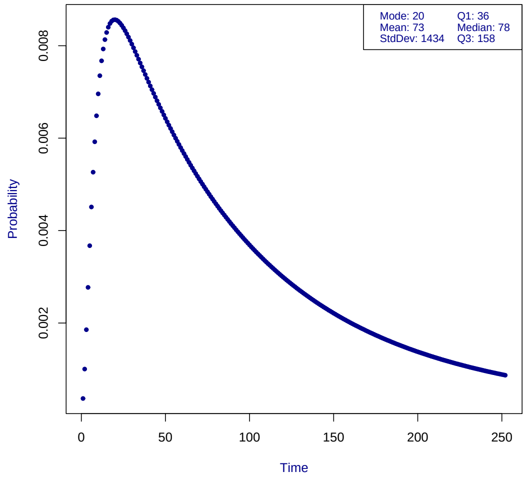


**Figure 3.27:** Telecontrol center dr3, year 2005: estimated Phase-type distribution; in the legend the fundamental properties.

As for telecontrol center dg4, in Figure 3.28 we plot Phase-type distributions for years 2004 and 2005: the form of the distribution is quite similar in both years, but in year 2004 more probability is concentrated on small values, indicating that, in 2004 it is more probable to have a faster recovering of the system.

With similar motivations and purposes we can compare Phase-type distributions for telecontrol center dg4 and dr3 both years; by Figure 3.29 it can be seen that the distribution for center dr3 concentrates more probability on higher values. The fact that center dr3 needs more time to reestablish the normal situation could be due to the geographical position, exposed to particularly persistent phenomena, or to a weakness of the center, unable to tackle an exceptional situation.

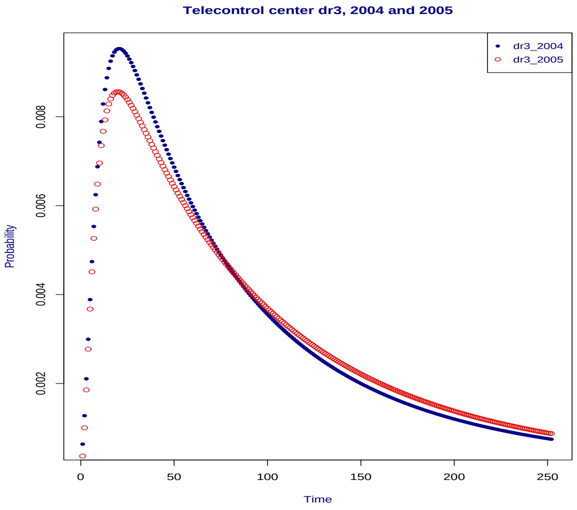


**Figure 3.28:** Telecontrol center dr3: estimated Phase-type distributions in years 2004 and 2005.

As we said in Section 1.1, Regulators have become more interested in controlling and evaluating the effectiveness and efficiency of utility restoration schemes. Then by the comparison by means of Phase-type distributions we provide a method for controlling the center's restoration scheme between years (see Figures 3.15 and 3.28) but also a way for evaluating the restoration scheme between telecontrol centers (see Figure 3.29).

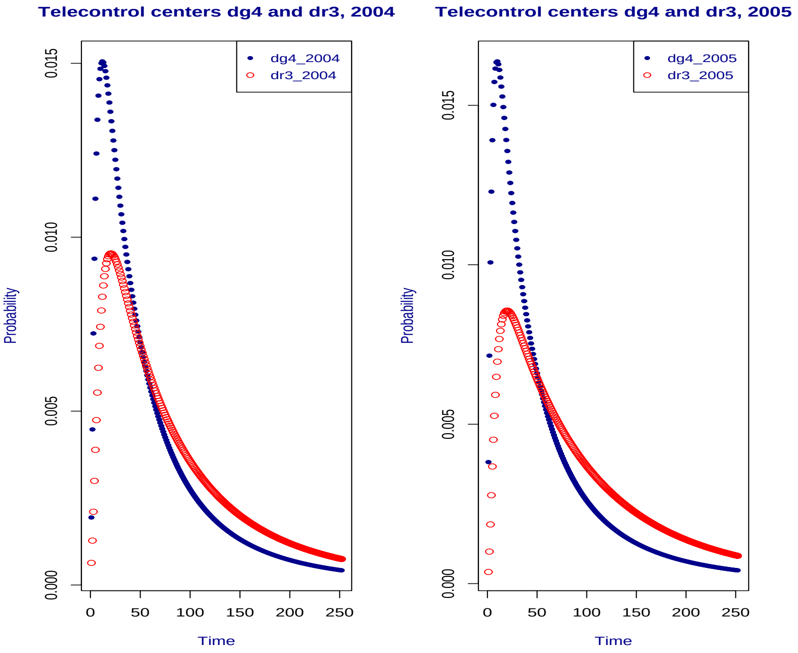


**Figure 3.29:** Telecontrol centers dg4 and dr3: estimated Phase-type distributions in years 2004 (Left) and 2005 (Right).

# Chapter 4

# Studying electricity distribution utilities and clustering them via hidden Markov models

## 4.1 Introduction

As we stated in Section 1.7, next analyzes will be performed considering, as spatial units, province and company combinations instead of telecontrol centers, so far studied. Analysis of results for a finite HMM applied to the telecontrol centers (see Chapter 3) showed that the estimated hidden chain is able to identify exceptional events. This encouraged us to identify exceptional hours in province and company combinations by a finite HMM.

After analyzing each combination separately from the others we would like to understand if provinces or companies are in some sense similar. Exceptional events are the central point of our study and, given the interpretation of the model, the hidden process is the mechanism that manages the occurrence of the exceptional operating status experienced by the system. For this reason we will investigate, by means of a Cluster analysis, if provinces or companies are similar with respect to the underlying process.

The use of HMMs for clustering sequences appears to have first been mentioned in Juang and Rabiner (1985) and subsequently used in the context of discovering subfamilies of protein sequences in Krogh *et al.* (1994). Methods for clustering can be coarsely categorized into two classes: distance and model-based approaches. A method for calculating distance for clustering using HMMs was proposed in Falkhausen *et al.* (1995), distance-based clustering methods using HMMs are investigated, for example, in Bicego *et al.* (2003); model-based clustering methods with HMMs are, among the others, in Smyth (1997) and Li and Biswas (2000); finally in Panuccio *et al.* (2002) a model-based approach for calculating distance measures is considered.

In Section 4.2, after considering the model specification, we will present and discuss results related to the analysis of all the province and company combinations, for year 2004; in Section 4.3 we will present the proposed clustering method and finally in Section 4.4 what obtained by the clustering analysis.

## 4.2 Model specification and general results

We briefly recall notation and assumptions for the finite HMM considered in the study presented in Chapter 3: for each province and company combination the observed number of electrical service faults $\{Y_t\}_{t>0}$ depends on a four state hidden Markov chain $\{X_t\}_{t\geq 0}$. If we assume that $X_0 = 1$, the model can be characterized by the transition matrix $A = \{a_{i,j}\}$, with $a_{i,j} = P(X_{k+1} = j | X_k = i)$, $i, j \in \mathsf{X}$, where $\mathsf{X}$ is the state space of the Markov chain and the emission matrix $B = \{b_i(y)\}$, with the conditional probabilities $b_i(y) = P(Y_k = y | X_k = i)$, $i \in \mathsf{X}$, $y \in \mathsf{Y}$, where $\mathsf{Y}$ is the set of the observable values.

Observations greater than 9 are considered as "many interruptions", that is $\mathsf{Y} = \{0, 1, \ldots, 10^+\}$; as we already underlined in Section 3.1 this assumption regards a small number of observations and permits the comparison not only between provinces or companies, but also between the same province or company in different years. Finally states have a physical meaning, in particular state 1 indicates the normal operating status, state 4 the exceptional one, while states 2 and 3 refer to an increasing degree of perturbation of the system operating status.

The data set includes the hourly number of faults relative to the three year time span 2004 - 2006 for a total of 113 province and company combinations. As for the telecontrol centers, each combination of province and company is studied separately from the others.

The 34 telecontrol centers (studied in Chapter 3) cover the whole national territory. Given a year, the total number of observed faults is (more or less) the same either if we consider, as spatial units, the 34 telecontrol centers or the 113 province and company combinations. This implies that a telecontrol center could serve more than one province; note that this consideration does not hold for companies, because they have one telecontrol center each and their territory is always confined within a single province. Then we can expect that data for the provinces are more sparse than data previously considered for the centers. We incorporate this consideration considering a more vague prior on each row of the emission matrix. More precisely, for centers we considered independent Dirichlet distributions on each row, with all parameters equal to 1 (see Section 3.2), while for analyzing provinces data, parameters of the Dirichlet distribution are set equal to 4/11. The expected value of each emission probability is, as with parameters previously hypothesized, equal to 1/11, but the variance is larger (approximatively 0.016 versus 0.007); in fact we recall that if $Z = (Z_1, \ldots, Z_g) \sim \text{Dir}(c_1, \ldots, c_g)$ and $c_0 = \sum_{j=1}^{g} c_j$ then $E(Z_i) = \frac{c_i}{c_0}$ and $V(Z_i) = \frac{c_i(c_0-c_i)}{c_0^2(c_0+1)}$.

Finally, the prior on the transition matrix remains the same as for the telecontrol center's analysis, *i.e.* independent Dirichlet distribution on each row, with all parameters equal to 1.

Before considering the clustering analysis, let us concentrate our attention on the exceptional events, estimated by the finite HMM, in all the 113 province and company combinations.

First of all, considering together the 113 combinations for the whole year 2004 (*i.e.* a total of $113 \times 8\,784 = 992\,592$ observations), we have that the great majority of observations - more than 97% - are classified in the normal state (*i.e* the estimated underlying Markov chain is in state 1), the 2% and the 0.1% of the interruptions are classified in the transitional states (respectively 2 and 3), while just about the 0.09% of the faults - in particular 881 observations - is classified as due to an exceptional event (*i.e.* the estimated hidden chain is in state 4). This is not an astonishing result, given that for each province or company, about 90% of observations is equal to 0.

In order to understand if at the same hour the exceptionality involves more than one province, consider the number of provinces and companies that contemporaneously are in the exceptional operating status; in other words for each of the 8 784 hours in 2004 in which at least one province or company was in the exceptional state, we count the number of provinces or companies that experienced an exceptional operating status. By Table 4.1 we can see that at most 7 provinces or companies are contemporaneously in the exceptional status, whilst often (in 312 hours) the exceptionality concerns one province or company.

| # of provinces | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Hours | 312 | 108 | 39 | 26 | 14 | 8 | 2 |

**Table 4.1:** Number of province and company combinations contemporaneously in the exceptional state: in 312 hours 1 province/company was in the exceptional operating status, in 108 hours two provinces/companies were contemporaneously in the exceptional state and so on.

Note that in Table 4.1, the point of view is different: before we considered the data points (992 592 observations) while now we are considering the time points (8 784 hours); then we have 881 exceptional observations and 509 exceptional hours.

Let us concentrate our attention on province and company combinations that were the only one in the exceptional operating status in those 312 “single exceptional hours”. In Figure 4.1 we can recognize the Italy; each point represents a province’s capital or the reference province’s capital for the company; red circles are centered on the province/company and the diameter is proportional to the number of hours the province/company was the only one in the exceptional state.

Without considering any technical or morphological information, we could expect that an exceptional event (think to a meteorological phenomenon) involves more than one province/company; consequently we could conclude that provinces or companies that more often are the only one in the exceptional operating status are particularly sensible to changes in underlying conditions. Finally, note that provinces or companies with “single exceptional hours” are mainly placed in the South part of Italy.

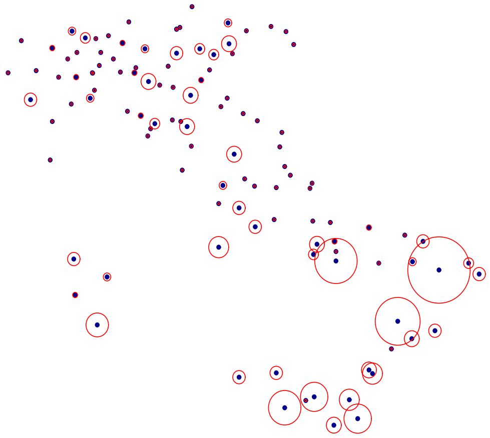


**Figure 4.1:** Distribution of the number of times in which each province or company was the only one in the exceptional operating status: each point represents the province/company and the red circle centered on it has a diameter proportional to the number of times the province/company was the only one in the exceptional state. Companies are plotted a little bit displaced with respect to their province reference, in order to have a more readable plot.

The situation just described changes if we consider the exceptional excursions instead of the exceptional events. Given that an exceptional event occurs, we enlarge the interval in order to also consider the instability condition preceding and following an exceptional event. Therefore it becomes more likely that different provinces/companies are contemporaneously in an exceptional excursion. In fact, by table 4.2 we can see that different provinces/companies experienced an exceptional instability situation in the same period; in particular in 1 hour (January 30th, 5 AM), 15 provinces/companies were managing an exceptional situation. The corresponding of Figure 4.1, when the exceptional excursions are considered, is shown in Figure 4.2. The plot shows a more uniform situation, with fewer provinces/companies that were the only one in an exceptional excursion; however the feature that provinces or companies with "single exceptional excursion" are mainly placed in the South part of Italy still holds.

| # of prov. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hours | 796 | 291 | 170 | 111 | 85 | 81 | 57 | 24 | 11 | 5 | 10 | 8 | 7 | 2 | 1 |

**Table 4.2:** Number of province and company combinations contemporaneously in an exceptional excursion: in 796 hours, 1 province/company was in an exceptional instability situation and so on.

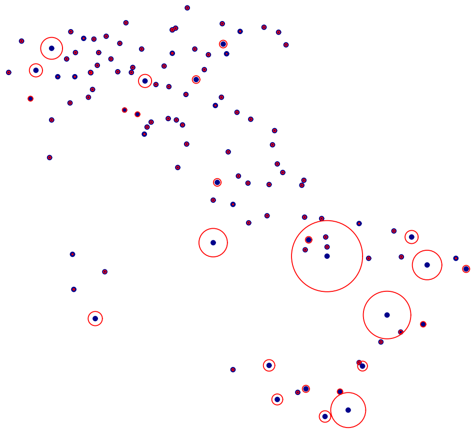
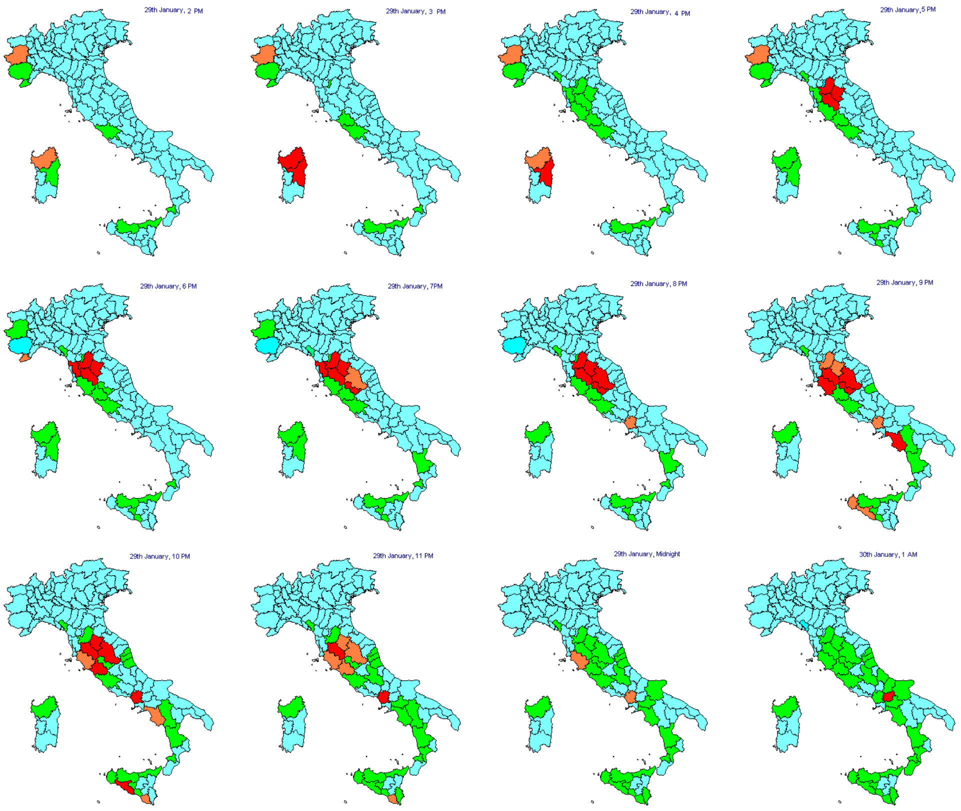


**Figure 4.2:** Distribution of the number of times in which each province or company was the only one in an exceptional excursion: each point represents the province/company and the red circle centered on it has a diameter proportional to the number of times the province/company was the only one in the exceptional state. Companies are plotted a little bit displaced with respect to their province reference, in order to have a more readable plot.

Information related to the spatial position of provinces and companies permit us to make a spatio-temporal analysis of the exceptionality. In Figure 4.3 each plot shows the situation at a certain time, the administrative area covered by a province has a different color on the basis of the state of the estimated underlying Markov chain (at that certain time); in other words each plot represents a photo on the operating situation. In each plot if the system is in state 1 the area is colored in pale blue, if it is in state 2 in green, in state 3 in orange and in state 4 in red. The Figure shows the temporal evolution of the situation since January 29th 2 PM to January 30th 1 PM; concentrating our attention on the red areas that indicate an exceptional operating status, we can notice that the instability started from the North/West and with the passage of time involved the Center and finally the South-East part of Italy. Therefore it seems that with the model (even if it was estimated separately for each province or company) we are gathering an unsettled situation that plausibly was a meteorological phenomenon. We recall that the identification of the exceptional event by a *Force Majeure* attribution was experienced in the first regulatory period 2000 - 2003, but the application of this criterion resulted in some practical cases quite difficult and ambiguous (see Section 1.2).


29th January, 2 PM
29th January, 3 PM
29th January, 4 PM
29th January,5 PM
29th January, 6 PM
29th January, 7PM
29th January, 8 PM
29th January, 9 PM
29th January, 10 PM
29th January, 11 PM
29th January, Midnight
30th January, 1 AM

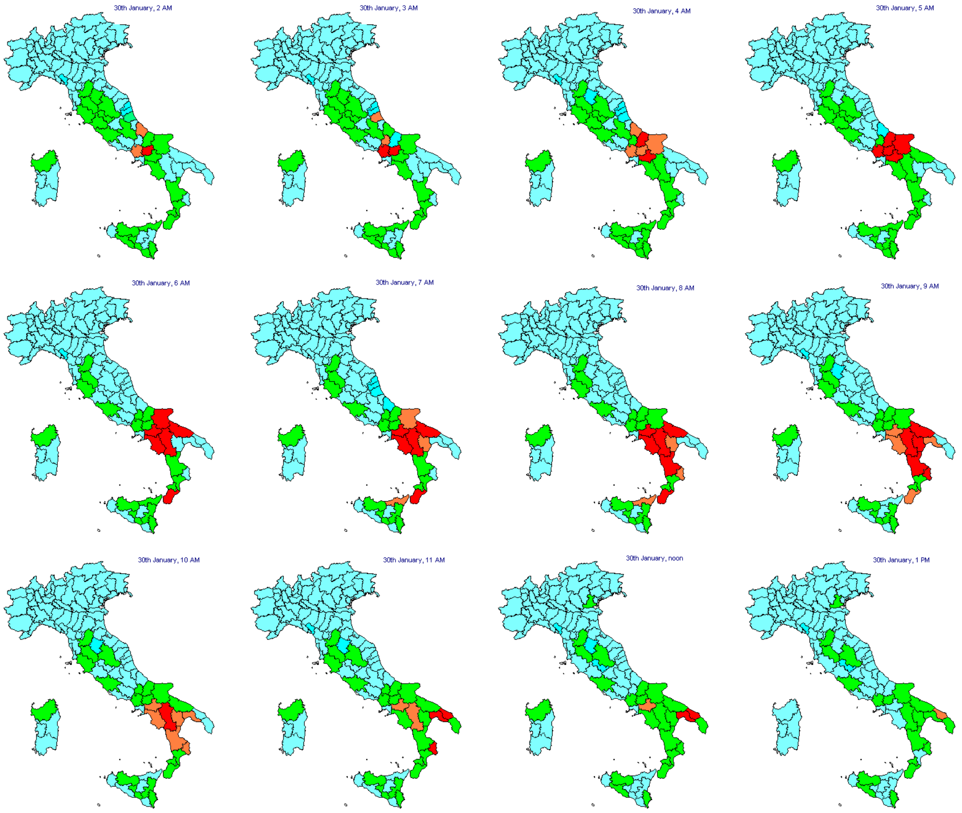


**Figure 4.3:** Spatio-temporal evolution of the exceptional events: the administrative area covered by each province has a different color on the basis of the state of the estimated underlying Markov chain; if the system is in state 1 the area is colored in pale blue, if it is in state 2 in green, in state 3 in orange and in state 4 in red. In particular plots show the situation from January 29th 2 PM to January 30th 1PM; time evolves from top to bottom and from left to right.

Consider Table 4.3 containing the number of observations classified as due to an exceptional event by the model for each Region and each month.
First of all we can notice that in the South part of Italy about half of the total number of exceptional events (433 on 881) occurred, while respectively the 32% and the 19% of the exceptional observations are relative to the North and Center part of Italy. In particular in Sicilia the largest number of exceptional events occurred, followed by Puglia, Piemonte and Toscana.

Regarding the temporal aspect we can notice that in the Autumn/Winter period a large number of exceptional events occurred; in particular the largest number of exceptional observations is related to February, followed by November, September and January. Moreover exceptional events in January and February are mainly concentrated respectively in the South/Center and the North part of Italy; the same structure is not so evident in the last part of the year, where exceptional situations are distributed on all the national territory.

| | | *Jan* | *Feb* | *Mar* | *Apr* | *May* | *June* | *July* | *Aug* | *Sept* | *Oct* | *Nov* | *Dec* | *Total* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **North** | *Piemonte* | 0 | 63 | 1 | 0 | 0 | 1 | 9 | 8 | 0 | 0 | 0 | 2 | **84** |
| | *Liguria* | 0 | 6 | 1 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 13 |
| | *Lombardia* | 0 | 17 | 15 | 0 | 0 | 0 | 6 | 3 | 3 | 0 | 10 | 0 | 54 |
| | *Trentino A.A.* | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 3 |
| | *Veneto* | 0 | 37 | 0 | 0 | 0 | 2 | 3 | 8 | 15 | 2 | 3 | 4 | 74 |
| | *Friuli V.G.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 5 |
| | *Emilia Romagna* | 0 | 29 | 1 | 2 | 1 | 1 | 3 | 1 | 4 | 0 | 6 | 1 | 49 |
| **Center** | *Toscana* | 23 | 2 | 7 | 0 | 0 | 0 | 0 | 5 | 15 | 5 | 25 | 0 | **82** |
| | *Umbria* | 6 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 20 | 0 | 3 | 0 | 34 |
| | *Marche* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 6 |
| | *Lazio* | 1 | 0 | 0 | 0 | 2 | 1 | 9 | 2 | 7 | 7 | 1 | 9 | 39 |
| | *Abruzzo* | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 5 |
| **South** | *Molise* | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 6 |
| | *Campania* | 20 | 1 | 1 | 0 | 4 | 0 | 2 | 4 | 7 | 5 | 21 | 6 | 71 |
| | *Basilicata* | 16 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 19 |
| | *Puglia* | 14 | 6 | 4 | 0 | 2 | 9 | 25 | 5 | 4 | 5 | 11 | 8 | **93** |
| | *Calabria* | 17 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 2 | 4 | 15 | 15 | 61 |
| | *Sicilia* | 5 | 0 | 4 | 20 | 4 | 16 | 0 | 1 | 23 | 5 | 26 | 39 | **143** |
| | *Sardegna* | 5 | 0 | 0 | 2 | 1 | 0 | 0 | 9 | 7 | 2 | 3 | 11 | 40 |
| | *Total* | **109** | **162** | 35 | 25 | 22 | 31 | 58 | 55 | **116** | 36 | **131** | 101 | |

**Table 4.3:** Spatio-temporal distribution of the exceptional hours: for each Region in the North, Center and South part of Italy we consider the number of hours in the exceptional operating status occurred in each month. In boldface the four largest total number of exceptional events by Region and by year.

## 4.3 The proposed clustering method

In this Section we present the clustering method introduced, in order to attend our goal. All clustering algorithms begin by measuring the *(dis)similarity* between the objects to be clustered.

As a by product of the application of an HMM we have, for each province and company, estimated values for the transition matrix $A$, the emission matrix $B$, the hidden Markov chain and the Phase-type distribution (see Section 1.6). In order to reach the set goal we will consider estimated parameters related to the hidden Markov chain. The Phase-type distribution has a "technical interpretation" (in fact it is the distribution of the time needed for each system to reestablish the normal situation) and it also permits a more visible comparison between provinces/companies; nevertheless it is obtained by a transformation of the transition probabilities and then it "contains less information" than the transition matrix.

Therefore the transition matrix seems to be more adequate for our purpose. Of course, given a transition matrix, generated paths can be different; in other words, even if two transition matrices are quite similar the generated chains might be different. Then a measure of the distance between probability distributions and a dissimilarity measure between the estimated Markov chains, able to underline if there are periods with a similar behavior, need to be introduced.

### Dissimilarity measures

In literature different probability metrics are been proposed; for a complete a clear presentation see Gibbs and Su (2002). Because each transition matrix is a collection of $K$ probability distributions (where we recall $K = 4$ is the number of possible states) and rows with the same index are probability distributions conditional on the same event (that is "the chain is in state $i$"), we consider as a measure of dissimilarity the average of the symmetrized Kullback-Liebler distance between corresponding rows (Ramoni *et al.*, 2002). Let $a^q_{i,j}$ and $a^r_{i,j}$ be the transition probabilities from $i$ to $j$ in two transition matrices $A^q$ and $A^r$ (corresponding to province and companies combinations labeled with $q$ and $r$).

The Kullback-Liebler divergence between rows $i$, $\boldsymbol{a}_i^q$ and $\boldsymbol{a}_i^r$, of these matrices is

$$d_p(\boldsymbol{a}_i^q, \boldsymbol{a}_i^r) = \sum_{j=1}^{K} a_{i,j}^q \log \frac{a_{i,j}^q}{a_{i,j}^r}. \tag{4.1}$$

The distance in equation (4.1) is not symmetric because $d_p(\boldsymbol{a}_i^q, \boldsymbol{a}_i^r) \neq d_p(\boldsymbol{a}_i^r, \boldsymbol{a}_i^q)$; the symmetric version of it is defined as $D_p(\boldsymbol{a}_i^q, \boldsymbol{a}_i^r) = [d_p(\boldsymbol{a}_i^q, \boldsymbol{a}_i^r) + d_p(\boldsymbol{a}_i^r, \boldsymbol{a}_i^q)]/2$. Then the average distance between provinces or companies labeled with $q$ and $r$, with respect to the transition matrices $A^q$ and $A^r$ is

$$D_p(q, r) = \frac{1}{K} \sum_{i=1}^{K} D_p(\boldsymbol{a}_i^q, \boldsymbol{a}_i^r). \tag{4.2}$$

Note that the distance becomes 0 if and only if $A^q = A^r$ and it is otherwise positive.

Consider now dissimilarity measure between the estimated Markov chains. Given the physical interpretation of the states of the hidden Markov chain, as increasing degree of system perturbation, we could consider a path of the chain as a sequence of ordinal values.

First of all note that it could happen that an exceptional event starts in two provinces in close, but different, hours. In the previous analysis by studying the estimated Markov chain we verified that the system, before/after dealing with an exceptional event, experiences situations with an increasing/decreasing degree of perturbation. We analyzed this feature introducing the concept of exceptional excursion (see Section 1.6).

Consider for example the following subsequences of the estimated Markov chain relative to provinces Agrigento (AG) and Caltanissetta (CL), starting and ending at the same time (September 6th, 1 PM and 7th, 3 PM in year 2004):

$$\begin{aligned} \boldsymbol{X}^{\mathrm{AG}} &= (4,3,3,3,4,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1) \\ \boldsymbol{X}^{\mathrm{CL}} &= (1,2,4,4,3,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1); \end{aligned}$$

in particular note that $\boldsymbol{X}^{\mathrm{AG}}$ is an exceptional excursion (*i.e.* it contains all the states visited by the chain after leaving and before reentering the normal state 1). It seems that provinces experienced the same instability condition, even if in

Agrigento it started a little bit before and it ended later than in Caltanissetta. Then we could consider similar $\boldsymbol{X}^{\mathrm{AG}}$ and $\boldsymbol{X}^{\mathrm{CL}}$ when the exceptional excursions occurred at the same time.
Let us indicate with $\mathsf{E}_j$ the set containing the exceptional excursions $EE_j$ that occurred in the estimated Markov chain for province/company $j$. To every exceptional excursion $EE_j$ an interval $h_j$ indicating when $EE_j$ started and ended is associated; we indicate with $\mathsf{H}_j$ the set containing the time intervals $h_j$.
Then for each $j = 1, \ldots, 113$, given a path of length $T$ of the chain, $\boldsymbol{X}^j = (x_1^j, \ldots, x_T^j)$, at each time $t = 1, \ldots, T$ set

$$Z_t^j = \begin{cases} 4 & \text{if } t \in h_j \\ x_t^j & \text{otherwise;} \end{cases}$$

in other words we relabel states in an exceptional excursion as exceptional. Note that this relabelling process is conceptually similar to the AEEG procedure that considers exceptional the three hours intervals preceding and following an exceptional interval - EI (see Section 1.2).

Given the relabelled paths $\boldsymbol{Z}^q$ and $\boldsymbol{Z}^r$, relative to provinces or companies $q$ and $r$, consider the Spearman rank correlation coefficient $\rho(\boldsymbol{Z}^q, \boldsymbol{Z}^r)$ (see Lehmann, 2006). The Spearman coefficient is a nonparametric measure of the strength of the associations between two variables, when data in the form of rank orders are available; like the Pearson correlation coefficient, the Spearman coefficient lies between -1 and +1, but it does not search for a linear relation between the variables.

Correlation coefficients, say $c(i, j)$, can be converted to dissimilarities, say $d(i, j)$, by setting $d^1(i, j) = (1 - c(i, j))/2$ or $d^2(i, j) = 1 - |c(i, j)|$; Lance and Williams (1979) compared these formulas and concluded that $d^1(i, j)$ is the best. Then given the Spearman correlation coefficient, we set the dissimilarity between $q$ and $r$ with respect to the estimated chain as

$$D_o(q, r) = \frac{1 - \rho(\boldsymbol{Z}^q, \boldsymbol{Z}^r)}{2}. \tag{4.3}$$

The two dissimilarity measures introduced in (4.2) and (4.3), $D_p(q, r)$ and $D_o(q, r)$, quantify distance between two different features of the provinces/companies.

More specifically, $D_p(q, r)$ measures the dissimilarity with respect to the possible transition dynamic of the system between the possible states, while $D_o(q, r)$ takes into account the trajectories actually generated by the transition matrices. In other words $D_p(q, r)$ measures the "potential" dissimilarity between provinces/companies, while $D_o(q, r)$ quantifies the "actual, observed" dissimilarities. The adjective "observed", referred to the hidden chain, creates an oxymoron, but it is just to underline the difference between what we can potentially obtain and what we effectively have from the estimates (see comparison between 2004 and 2005 for telecontrol center denominated dg4 at the end of Section 3.2.1).

Then two clustering methods will be implemented for grouping provinces/companies "potentially" and "actually" similar.

### Clustering algorithm

Given a generic dissimilarity matrix $\Delta = \{\delta(i, j)\}$, with $i, j = 1, \ldots, 113$, a classification method needs to be chosen in order to classify provinces/companies in the same cluster. Two kinds of clustering algorithms, namely partitioning and hierarchical methods, are usually considered in the classification literature (see for example Mardia *et al.*, 1979).

Very briefly *hierarchical algorithms* find successive clusters using previously established clusters, while *partitioning methods* try to find a partition in $k$ (fixed) groups maximizing a measure of adequacy of the partition. Then hierarchical methods do not need the specification of the number of groups, but they can never repair what was done in previous steps; moreover a partitioning method tries to select the best clustering with $k$ groups, which is not the goal of a hierarchical method.

In order to reach our goal, we will use a two steps procedure: we will first choose the number of groups by an *agglomerative hierarchical method*, with the distance between clusters calculated by the *complete linkage method*, and then we will cluster provinces/companies by the *Partitioning Around Medoids - PAM* algorithm (also called *k-medoids method*, Kaufman and Rousseeuw, 1990).

Briefly this means:

**Step 1**

- ▷ start considering $n = 113$ clusters each containing just one province/company (or company)
- ▷ at the first step fuse the two nearest (with the smallest dissimilarity) provinces/companies obtaining $n - 1$ clusters
- ▷ at the second step fuse in the same cluster the two nearest of the $n - 1$ clusters to form $n - 2$ clusters
- ▷ continue in this manner until at the $(n - 1)$th step the two clusters left are fused into a single cluster of $n$ provinces/companies
- ▷ at each step recalculate the dissimilarity matrix by the complete linkage method: consider two clusters $A$ and $B$ containing respectively $n_A$ and $n_B$ provinces/companies; then the distance between $A$ and $B$ is defined by

$$\delta(A, B) = \max_{i \in A,\, h \in B} \delta(i, h).$$

This aggregation process may be represented by a two-dimensional diagram, called *dendrogram*, which illustrates the fusion made at each stage of the analysis. The horizontal axis displays the labels of the points (provinces/companies), whereas the vertical axis gives the distance between the clusters. “Cutting the tree” at a level we choose the number of clusters $k$ to be considered in Step 2.

**Step 2**

- ▷ select $k$ objects to be the initial cluster medoids
- ▷ assign each remaining objects to the nearest representative object
- ▷ recalculate the position of the $k$ medoids by minimizing the average dissimilarity of the medoid to all the other objects of the same cluster
- ▷ continue until the medoids become fixed.

Note that the *k-medoids method* is similar to the *k-means method*. The main difference is that, rather than using the mean of each cluster as the centroid of the cluster as in the $k$-means, it finds an observation to be the centroid; moreover the $k$-medoids is more robust with respect to outliers and it also deals with dissimilarity coefficients.

Output concerning each cluster obtained by the PAM method can be graphically represented by the Silhouettes introduced by Rouseeuw (1987); briefly Silhouettes provides a measure of how well a data point was classified when it was assigned to a cluster by according to both the tightness of the clusters and the separation between them. Consider an object $i$ and denote by $A$ the cluster to which it has been assigned; calculate

$$a(i) = \text{average dissimilarity of } i \text{ to all other objects of } A;$$

consider any cluster $C$ different from $A$ and define

$$\delta(i, C) = \text{average dissimilarity of } i \text{ to all other objects of } C.$$

After computing $\delta(i, C)$ for all clusters $C \neq A$, let

$$b(i) = \min_{C \neq A} d(i, C).$$

The cluster $B$ such that $\delta(i, B) = b(i)$ is called the *neighbor* of object $i$ and it is like the second-best choice for object $i$. The silhouette $s(i)$ is obtained as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4.4}$$

and then

$$-1 \leq s(i) \leq 1.$$

Moreover by the definition of $s(i)$ we can deduce that observations with a large $s(i)$ (almost 1) are very well clustered, a small $s(i)$ (around 0) means that the observation lies between two clusters, and observations with a negative $s(i)$ are probably placed in the wrong cluster.

*k-medoids methods* are implemented in the R package cluster (Maechler *et al.*, 2005), that also provides the silhouette plot and other diagnostic tools.

## 4.4 The clustering results

We will first present what obtained by clustering province and company combinations by means of the estimated transition matrices and then clustering results relative to the estimated hidden Markov chain.

### Clustering by means of the transitional dynamic

The clustering process obtained by the application of the Agglomerative Hierarchical algorithm, with complete linkage, is plotted in Figure 4.9; this dendrogram suggests us to consider $k = 3$ clusters in the implementation of the PAM method.

Figure 4.4 shows the silhouette plot, where, for each cluster $C_j$, $j = 1, \ldots, k$, the silhouette $s(i)$, $i \in C_j$ is plotted by a bar in decreasing order; then, because values of $s(i)$ are positive and quite close to 1, we can deduce that provinces are well classified.

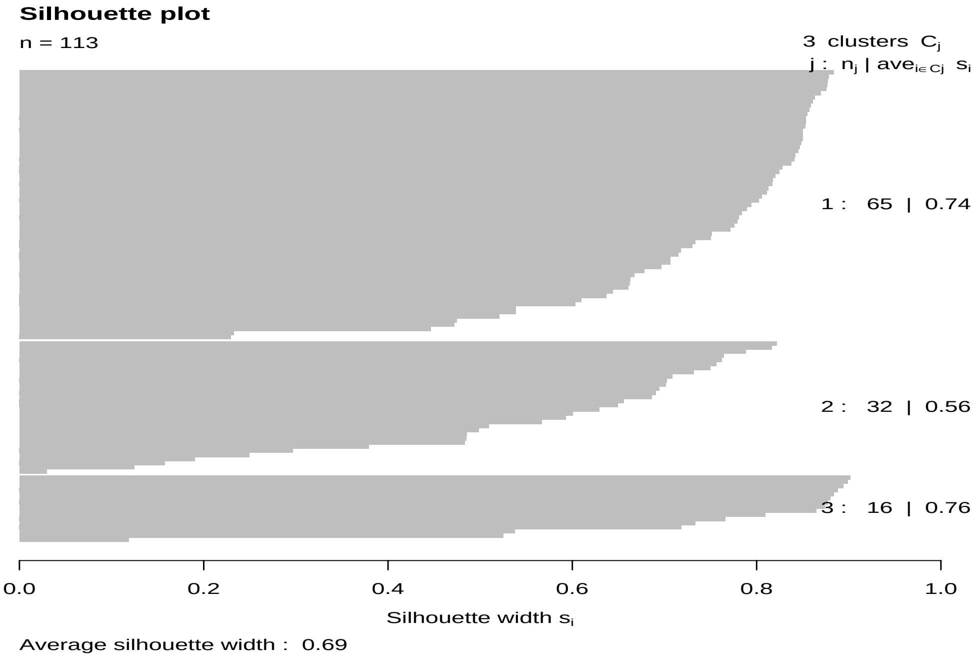


**Figure 4.4:** Silhouette plot for the clustering by the transition matrices. For each cluster the silhouette is plotted by a bar in decreasing order.

The obtained clusters are described in Figure 4.5, where the administrative area covered by each province is colored with a different color on the basis of the cluster to which it has been assigned. There are provinces served by two distribution utilities (province and company); when there is a disagreement in terms of assigned cluster the corresponding area is striped and colored with both colors, otherwise the number “2” indicates that the two utilities have been classified in the same cluster.

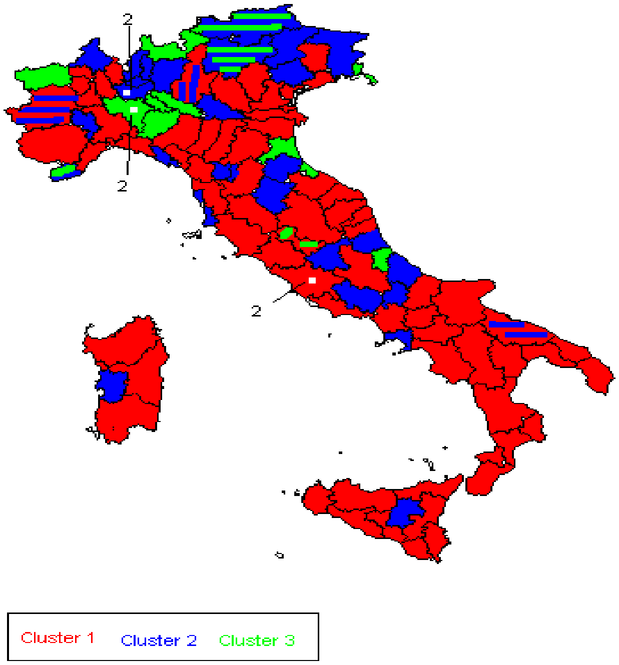


**Figure 4.5:** Clustering results by transitional dynamic. The administrative area covered by each province is colored with a different color on the basis of the cluster to which it has been assigned; striped areas are referred to disagreements in terms of assigned clusters, while the number “2” indicates an agreement. White points just indicate the indicators' starting point.

In order to understand which provinces or companies are classified in each cluster we consider the Phase-type distribution, that, we recall, represents the number of hours needed to the system to reestablish the normal operating situation. Figure 4.6 shows Phase type distributions for medoid of each cluster; considering the graph where the distributions are plotted together (bottom-right) we can see that for provinces in cluster 1 (red points) the Phase-type distribution concentrates more mass on larger values (that is they need more time to reestablish the normal situation), followed by provinces in cluster 2 (blue points) and in cluster 3 (green points).

Therefore a general classification could be:

| | | |
|---|---|---|
| provinces or companies in cluster 1 | → | "exceptional persistent" |
| provinces or companies in cluster 2 | → | "exceptional transitional" |
| provinces or companies in cluster 3 | → | "fast recovering" |

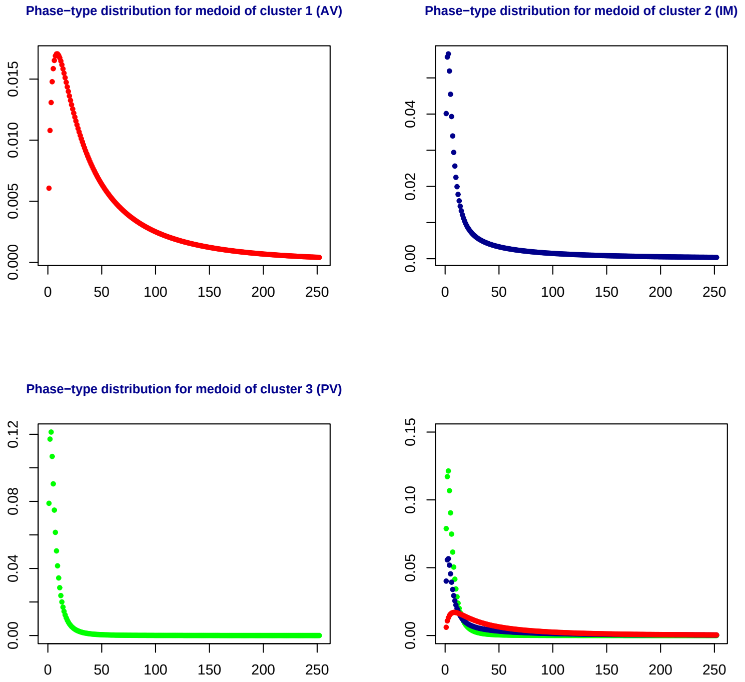


**Figure 4.6:** Phase-type distributions for medoids of each cluster; in the plot on bottom-right the three distributions are plotted together.

### Clustering by means of the underlying process

We now consider the clustering by means of the estimated hidden Markov chains. In order to calculate the Spearman correlation coefficient we will not consider province and company combinations whose estimated Markov chain was always in the normal state 1.

Dendrogram in Figure 4.10 obtained by the application of the Hierarchical Agglomerative method does not clearly identify a number of clusters; anyway it seems to suggest to consider $k = 3$ for the application of the PAM method.

Silhouette plot, shown in Figure 4.7, indicates that clusters are not strongly defined and that there are provinces or companies in each cluster that have been classified in the wrong cluster by the PAM algorithm. Provinces or companies in each cluster are represented in Figure 4.8, plot on the left. As we said the algorithm provides the neighbor cluster; then if we move provinces with a negative silhouette to the neighbor we obtain a more uniform situation, shown in Figure 4.8, plot on the right. We underline that moving objects with negative silhouette value in the neighbor cluster is not a general rule (in fact, the neighbor cluster is the second-best choice); however we have information coming from the problem under analysis and this helps us for having more understandable results.

Regarding the final interpretation, we recall that the dissimilarity measure used in this clustering analysis, $D_o(q, r)$, defined in equation (4.3), quantifies the dissimilarity between provinces related to operating status actually experienced by the provinces/companies. Then by Figure 4.8, plot on the right, we could conclude that utilities are affected by some “spatial dependence”; in fact Cluster 1 mainly contains provinces/companies in the South, Cluster 2 provinces/companies in the North and Cluster 3 provinces/companies in the Center part of Italy. Considering this clustering evidence along with what caught by analyzing Figure 4.3, it seems that exceptional events spread geographically. However, given information at disposal, we could just presume that exceptionality are caused by some external factor, such for example a very bad weather. Nothing can be said regarding eventual technical aspect, such as interconnection between the electricity transmission network, that could cause this geographical structure.

Of course, results obtained by clustering provinces and companies by means of the transition matrices and by means of the underlying estimated chains are quite different (compare Figure 4.5 and Figure 4.8). In fact, as we said in Section

4.3, given a transition matrix, generated paths can be very different. Therefore, in the first case, if two provinces or companies are in the same cluster, this means that the systems are similar with respect to the potential transition dynamic between different states; in other words it means that they have more or less the same probability to pass, for example, from the normal operating state 1 to state 2. While if two provinces or companies are in the same cluster, when we are considering the underlying process, this means that they actually experienced more or less the same operating situation.

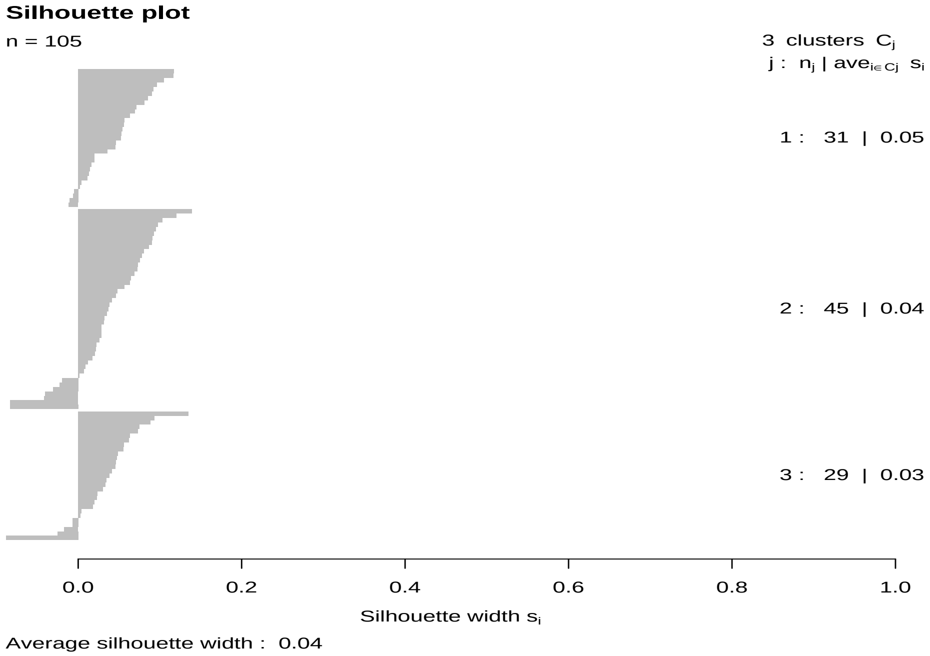


**Figure 4.7:** Silhouette plot for the clustering by the estimated hidden Markov chains. For each cluster the silhouette is plotted by a bar in decreasing order. Negative values indicate a misclassification.

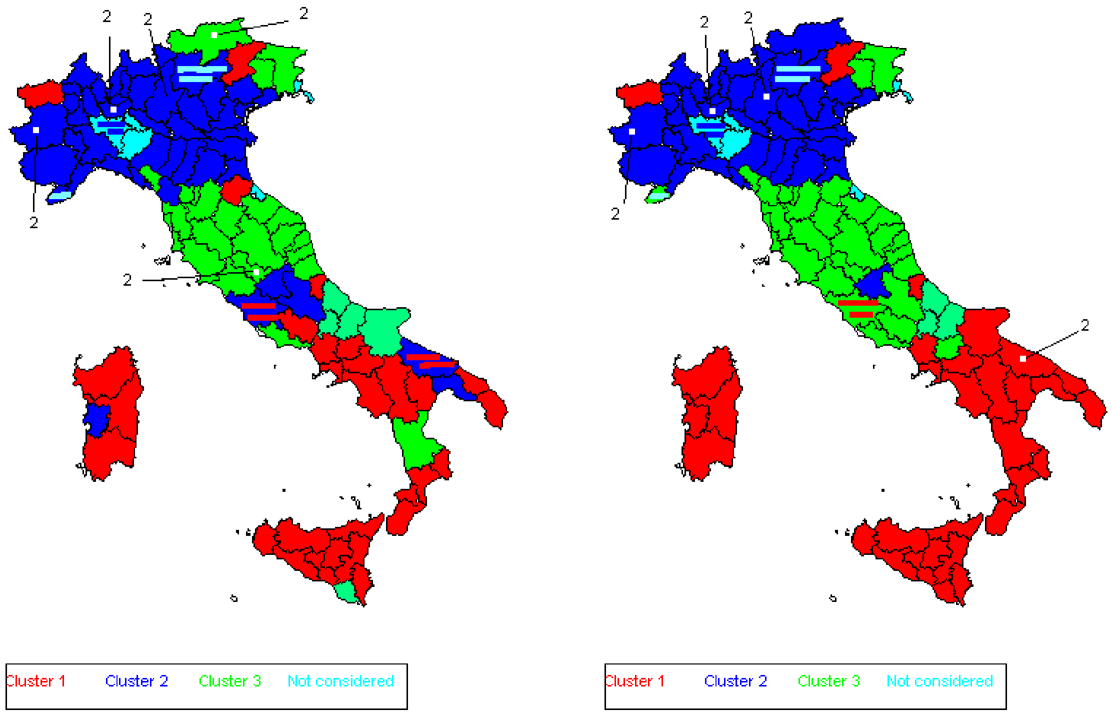


**Figure 4.8:** Clustering results by estimated Markov chains. The administrative area covered by each province is colored with a different color on the basis of the cluster to which it has been assigned; striped areas are referred to disagreements in terms of assigned clusters, while the number "2" indicates an agreement. Provinces not considered in the cluster analysis are colored in pale blue. White little squares just indicate the indicators' starting point. Left: clusters before the reallocation of misclassified provinces or companies to the neighbor cluster. Right: clusters after the reallocation of the misclassified provinces or companies.

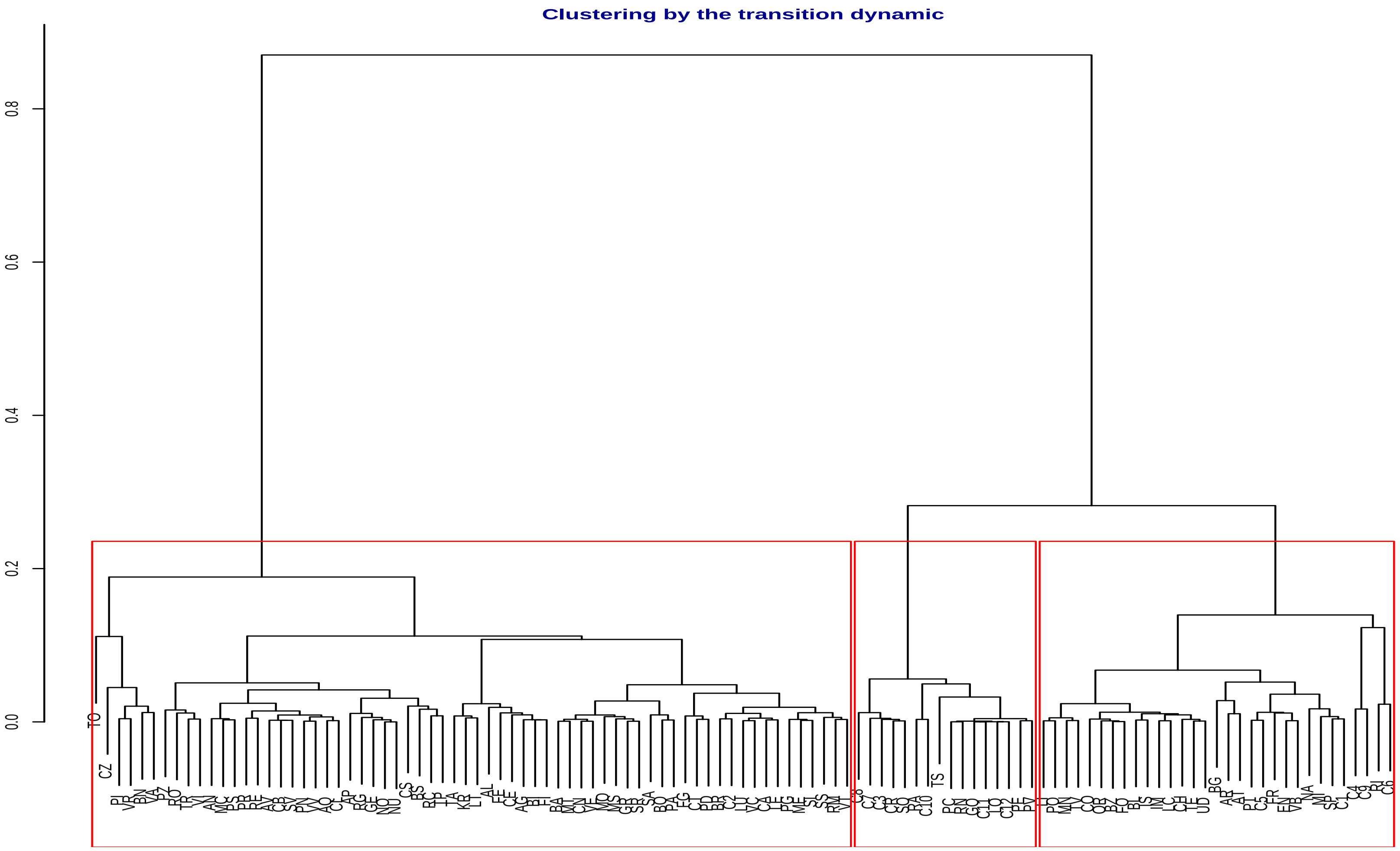


**Figure 4.9:** Dendrogram for the hierarchical clustering method, with complete linkage, when the estimated transition matrices are considered. In the following companies are codified with "C" followed by a number.

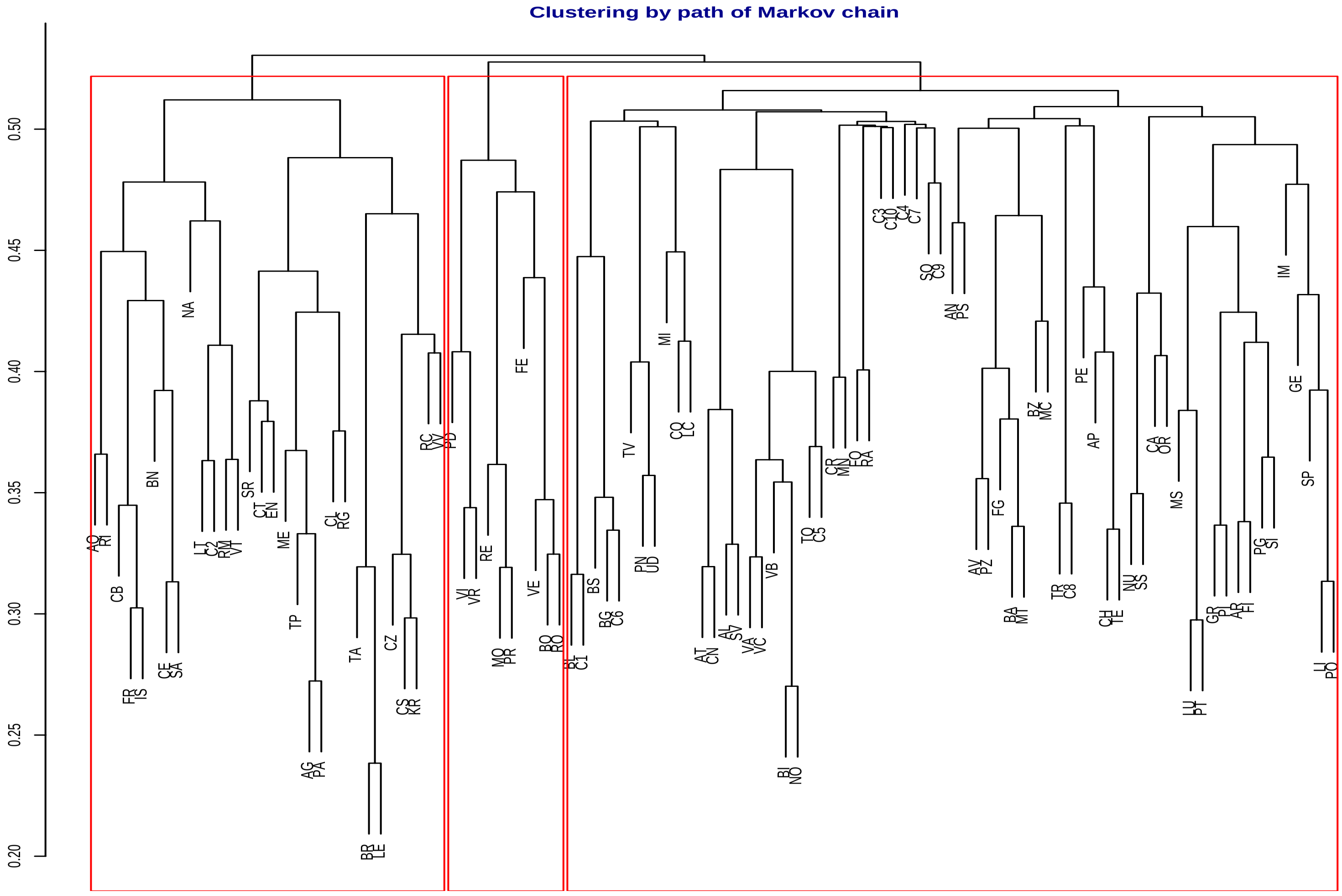


**Figure 4.10:** Dendrogram for the hierarchical clustering method, with complete linkage, when the estimated hidden Markov chains are considered. In the following companies are codified with "C" followed by a number.

# Chapter 5

# A model-based clustering method: the hidden mixture Markov model

## 5.1 Introduction

In the previous Chapter 4 we analyzed each province and company combination separately from the others and then we introduced a method for clustering them; given our purpose and the physical meaning of the adopted model we proposed a method based on estimated values of the transition matrices and the underlying Markov chain.

Another way for identifying groups, alternative to the data-driven Cluster analysis previously introduced, is to consider Mixture models; this model-based approach assumes that data are drawn from a mixture of underlying probability distributions and observations drawn from the same probability distribution belong to the same cluster. However a difficulty in mixture analysis is choosing the number of mixture components. The Dirichlet process mixture model (Antoniak, 1974; Escobar and West, 1995; Neal, 2000; Ischwaran and James, 2001), using the intrinsic clustering property of the Dirichlet process allows for this possibility: an infinite mixture model (*i.e.* a mixture model with an infinite number of possible components) is considered and the Dirichlet process will reveal the proper number of groups existing in data.

The Dirichlet process has been applied as prior in Bayesian HMMs in different ways and with different purposes. In the *infinite hidden Markov model* (Beal *et al.*, 2002), also called *hidden Markov Dirichlet process* (Xing and Sohn, 2004) or *hierarchical Dirichlet process hidden Markov model* (Teh *et al.*, 2006), an observed sequence is studied by an HMM, the state space of the hidden Markov chain is assumed to be countably infinite and each row of the infinite dimensional transition and emission matrices is modeled by a Dirichlet process. While in the *hidden Markov mixture models* (Yuting *et al.*, 2007) $N$ sequences of observations are considered, each data sequence drawn from a mixture of HMMs; the assumed Dirichlet process as common prior on the parameters of the individual HMMs will reveal the number of HMMs that explains the complexity in data.

In Section 5.2 we will overview fundamentals of the Dirichlet process and the Dirichlet process mixture model useful for our discussion; then in Section 5.3 we will briefly present how these processes were been used in an HMM framework. Finally in Section 5.4 we will introduce a *hidden mixture Markov model* for clustering provinces, by assuming that the underlying process is drawn from a mixture of Markov chains, with exchangeable transition matrices modeled by a Dirichlet process prior.

## 5.2 Dirichlet process and Dirichlet process mixture models

The Dirichlet process is a random measure on measures, *i.e.* each draw from a Dirichlet process is itself a measure. Distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters, thus the classification as a nonparametric model. The nonparametric nature in the Dirichlet process mixture model translates to mixture models with a countably infinite number of possible components.

### Dirichlet process

The Dirichlet process was first formalized by Ferguson (1973) as a flexible prior for Bayesian nonparametrics.

A random distribution $G$ is distributed according to a Dirichlet process if its marginal distributions are Dirichlet distributed. Specifically, let $G_0$ be a distribution over $\Theta$ and $\gamma$ be a positive real number. Then for any finite measurable partition $C_1, \ldots, C_r$ of $\Theta$, $G$ is Dirichlet process distributed with base distribution $G_0$ and scaling parameter $\gamma$, written $\mathrm{DP}(\gamma, G_0)$, if

$$(G(C_1), \ldots, G(C_r)) \sim \mathrm{Dir}(\gamma G_0(C_1), \ldots, \gamma G_0(C_r)). \tag{5.1}$$

The Dirichlet process provides a *conjugate* family of priors over distributions that is closed under posterior updates given observations: let $\theta_1, ..., \theta_n$ be a sequence of independent draws from $G$, then the posterior distribution of $G$ given values of $\theta_1, ..., \theta_n$ is a Dirichlet process with updated scaling parameter $\gamma + n$ and base distribution $\frac{\gamma G_0 + \sum_{i=1}^n \delta(\theta_i)}{\gamma + n}$, where $\delta(\theta_i)$ denotes a point mass located at $\theta_i$.

Then the posterior Dirichlet process can be rewritten as

$$G|\theta_1, ..., \theta_n \sim \mathrm{DP}\left(\gamma + n, \frac{\gamma}{\gamma + n} G_0 + \frac{n}{\gamma + n} \frac{\sum_{i=1}^n \delta(\theta_i)}{n}\right); \tag{5.2}$$

the posterior base distribution is a weighted average between the prior base distribution $G_0$ and the empirical distribution $\frac{\sum_{i=1}^n \delta(\theta_i)}{n}$. The weight associated with the prior base distribution is proportional to $\gamma$, while the empirical distribution has weight proportional to the number of observations $n$. Then if the Dirichlet process is used as nonparametric prior over distributions in a Bayesian nonparametric model, $\gamma$ represents the strength associated with the prior; for this reason $\gamma$ is also called the strength parameter.

This interpretation of the scaling parameter is more evident in the formulation in Blackwell and MacQueen (1973), where a Pólya urn scheme is introduced in order to generate an exchangeable sequence of random variables, whose de Finetti measure is a Dirichlet process. Consider an urn containing balls with colors given by the values in the (infinitely countable) space $\Theta$; the number of balls of colors $\theta \in \Theta$ initially contained in the urn is equal to $\gamma G_0(\theta)$. At each stage $n \geq 1$ a ball is sampled from the urn and replaced in it along with another ball of the same color. Then, if we denote with $P(\theta_i = j)$ the probability of drawing a ball of color $j \in \Theta$ at the step $i$, we obtain

$$P(\theta_1 = j) = \frac{\gamma G_0(j)}{\sum_{\theta \in \Theta} \gamma G_0(\theta)} = \frac{\gamma G_0(j)}{\gamma} = G_0(j)$$

$$P(\theta_2 = j|\theta_1) = \frac{\gamma G_0(j) + \delta(\theta_1 = j)}{\gamma + 1}$$

and so on till we get

$$P(\theta_{n+1} = j|\theta_1, \dots, \theta_n) = \frac{\gamma G_0(j) + \sum_{i=1}^{n} \delta(\theta_i = j)}{\gamma + n}.$$

With $G$ marginalized out we obtain the predictive distribution

$$\theta_{n+1}|\theta_1, \dots, \theta_n \sim \frac{\gamma}{\gamma + n} G_0 + \frac{1}{\gamma + n} \sum_{i=1}^{n} \delta(\theta_i). \tag{5.3}$$

The predictive distribution (5.3) makes more evident the *clustering property* of Dirichlet process: since the values of draws can be repeated, let $\theta_1^*, \dots, \theta_m^*$ be the different values among $\theta_1, \dots, \theta_n$, and $n_h$ be the number of repeats of $\theta_h^*$; then the predictive distribution can be equivalently written as

$$\theta_{n+1}|\theta_1, \dots, \theta_n \sim \frac{\gamma}{\gamma + n} G_0 + \frac{1}{\gamma + n} \sum_{h=1}^{m} n_h \delta(\theta_h^*). \tag{5.4}$$

Then $\theta_{n+1}$ is either equal to a previously seen $\theta_h^*$ with probability proportional to $n_h$ or it is a value independently drawn from $G_0$; thereof the name of $\gamma$ as the innovation parameter.

Sethuraman (1994) provides a constructive definition of $G$ in terms of a stick-breaking construction; this construction is given as follows:

$$v_h|\gamma \sim \text{Beta}(1, \gamma), \qquad p_h = v_h \prod_{k=1}^{h-1} (1 - v_k), \qquad \theta_h|G_0 \sim G_0 \tag{5.5}$$

then

$$G = \sum_{h=1}^{\infty} p_h \delta(\theta_h) \sim \text{DP}(\gamma, G_0), \tag{5.6}$$

with the convention that $\prod_{i=1}^{0} = 1$. The mixing weights $p_h$ for $\theta_h$ are given by successively breaking a unit length stick into an infinite number of pieces: starting with a stick of length 1, break it at $v_1$, assign $p_1$ to be the length of the broken stick and recursively break the other portion to obtain $p_2, p_3$ and so on.

By (5.6) we can notice that support of $G$ consists of an infinite set of atoms located at $\theta_h$, drawn independently from $G_0$; then measures drawn from a Dirichlet process are *discrete* (with probability one).

The construction in (5.6) can be truncated at $S$ by setting $v_S = 1$; Ishwaran and James (2001) gave conditions for choosing $S$ in order to obtain a good approximation of measure in (5.6).

### Dirichlet process mixture models

Roughly speaking a Dirichlet process mixture model arises when a Dirichlet process is introduced as prior on parameters of an infinite mixture model. Consider a set of observations $z_1, \dots, z_n$ and assume that they are exchangeable, or equivalently, that they are independently and identically drawn from some unknown distribution. The distribution from which the $z_i$ are drawn is a mixture of distributions of the form $F(\theta_i)$, parameterized by $\theta_i$, which are drawn independently and identically from a Dirichlet process $G$, with parameters $\gamma$ and $G_0$:

$$\begin{aligned} z_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G|\gamma, G_0 &\sim \mathrm{DP}(\gamma, G_0). \end{aligned} \tag{5.7}$$

Because $G$ is discrete, multiple $\theta_i$'s can take the same value simultaneously, and the model (5.7) can be seen as a mixture model, where $z_i$'s with the same value of $\theta_i$ belong to the same cluster.

In the following we present the representations in terms of Pólya urn mechanism and of stick breaking constructions, which are the core of two possible methods for sampling from the posterior distribution of the parameters, presented in the next Section 5.2.1.

If we integrate over $G$ the model in (5.7) we obtain a representation in terms of successive conditional distributions, arising from the Pólya urn scheme:

$$\theta_i|\theta_1, \dots, \theta_{i-1} \sim \frac{\gamma}{i-1+\gamma} G_0 + \frac{1}{i-1+\gamma} \sum_{h=1}^{i-1} \delta(\theta_h). \tag{5.8}$$

Consider the stick breaking construction in (5.5) and (5.6), let $c_i$ be a cluster assignment variable (label), which takes value $h$ with probability $p_h$. Then (5.7) can be equivalently expressed, in the usual representation of mixture models, as

$$z_i|c_i, \boldsymbol{\theta} \sim F(\theta_{c_i}) \tag{5.9}$$

where

$$\theta_h | G_0 \sim G_0,$$

$$v_h | \gamma \sim \text{Beta}(1, \gamma), \qquad p_h = v_h \prod_{k=1}^{h-1} (1 - v_k),$$

$$c_i | \boldsymbol{p} \sim \text{Mult}(\boldsymbol{p}), \qquad G = \sum_{h=1}^{\infty} p_h \delta(\theta_h);$$

Mult($\boldsymbol{p}$) is the multinomial distribution with parameter vector $\boldsymbol{p}$. From the perspective of infinite mixture models, $\boldsymbol{p} = \{p_i\}_{i=1,\dots,\infty}$ comprise the infinite mixing proportions and $\boldsymbol{\theta} = \{\theta_i\}_{i=1,\dots,\infty}$ are the infinite number of mixture components.

### 5.2.1 Gibbs sampling for Dirichlet process mixture models

In the Dirichlet process mixture model the posterior distribution of the parameters of the component distributions is intractable to compute. However Markov chain Monte Carlo (MCMC) methods have been developed for sampling from these posteriors.

Making a general distinction we could say that the two different ways of sampling presented in literature can be classified as: one based on the Pólya urn representation in (5.8) (Escobar, 1994; MacEachern, 1994; Escobar and West, 1995) and the other one based of the stick breaking construction (Ishwaran and James, 2001) in (5.9). Each of those methods allows for the case with non-conjugate priors (West *et al.*, 1994; MacEachern and Müller, 1998; Walker and Damie, 1998; Neal, 2000).

Following the notation in Ishwaran and James (2001) we will refer to the first method as *Pólya urn Gibbs sampler* and to the second one as *blocked Gibbs sampler*; both methods will be presented only in the conjugate case.

#### Pólya urn Gibbs sampler

The simplest method for sampling from the posterior distribution of $\theta_1, \dots, \theta_n$ is Gibbs sampling (Neal, 2000). This method repeatedly draw values for each $\theta_i$ from its conditional distribution given both data and the $\theta_j$, with $j \neq i$, written

$\boldsymbol{\theta}^{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. Given the exchangeability we can write the prior in (5.8) as

$$\theta_i|\boldsymbol{\theta}^{-i} \quad \sim \quad \frac{\gamma}{n-1+\gamma}G_0 + \frac{1}{n-1+\gamma}\sum_{\substack{j=1\\j\neq i}}^{n}\delta(\theta_j); \tag{5.10}$$

then combining (5.10) with the likelihood, written $F(z_i, \theta_i)$, we obtain

$$\theta_i|\boldsymbol{\theta}^{-i}, z_i \sim b\gamma G_0(\theta_i)F(z_i, \theta_i) + b\sum_{\substack{j=1\\j\neq i}}^{n} F(z_i, \theta_j)\delta(\theta_j). \tag{5.11}$$

The quantity $b$ in (5.11) is the normalizing constant given by

$$b = \left(\gamma q_0 + \sum_{\substack{j=1\\j\neq i}}^{n} F(z_i, \theta_j)\right)^{-1}$$

where

$$q_0 = \int_\theta F(z_i, \theta)dG_0(\theta).$$

Equation (5.11) can be written in terms of the posterior $H(\theta_i|z_i) = \frac{G_0(\theta_i)F(z_i,\theta_i)}{\int_\theta F(z_i,\theta)dG_0(\theta)}$ as

$$\theta_i|\boldsymbol{\theta}^{-i}, z_i \sim b\gamma q_0 H(\theta_i|z_i) + b\sum_{\substack{j=1\\j\neq i}}^{n} F(z_i, \theta_j)\delta(\theta_j). \tag{5.12}$$

For this Gibbs sampling method to be feasible, computing the integral in $q_0$ and sampling from $H$ must be feasible operations. This is generally be so when $G_0$ is the conjugate prior for the likelihood.

This method is used by Escobar (1994) and by Escobar and West (1995), but MacEachern (1994) suggested a more efficient algorithm that works in terms of the number of distinct values $m$, the labels $\boldsymbol{c} = (c_1, ..., c_n)$ and the set of distinct parameter values $\boldsymbol{\theta}^* = (\theta^*_1, \ldots, \theta^*_m)$. First of all, equation (5.12) can also be written in terms of $\boldsymbol{\theta}^*$ as

$$\theta_i|\boldsymbol{\theta}^{-i}, z_i \sim b\gamma q_0 H(\theta_i|z_i) + b\sum_{\substack{j=1\\j\neq i}}^{m} n_j^{-i}\delta(\theta^*_j)F(z_i, \theta^*_j), \tag{5.13}$$

where $n_j^{-i}$ is the number of parameters equal to $\theta_j^*$ in $(\theta_1^*, \dots, \theta_{i-1}^*, \theta_{i+1}^*, \dots, \theta_m^*)$. Then Gibbs sampling for $c_i$ is based on

$$P(c_i = c|\boldsymbol{c}^{-i}, z_i, \boldsymbol{\theta}^*) = bn_c^{-i} F(z_i, \theta_c^*) \tag{5.14}$$

if $c = c_j$ for some $j \neq i$, otherwise

$$P(c_i \neq c_j, \forall\, j \neq i|\boldsymbol{c}^{-i}, z_i, \boldsymbol{\theta}^*) = b\gamma \int_{\theta^*} F(z_i, \theta^*) dG_0(\theta^*). \tag{5.15}$$

When a value for $c_i$ different from any other $c_j$ is sampled, a value for $\theta_{c_i}^*$ is chosen from $H(\theta^*|z_i)$, the posterior distribution of $\theta^*$ based on the prior $G_0$ and the single observation $z_i$.

Then the Gibbs sampling method can be summarized as follows: consider $\boldsymbol{c} = (c_1, \dots, c_n)$ and $\boldsymbol{\theta}^* = \{\theta_c^* : c \in \boldsymbol{c}\}$, then

**Algorithm 5.1** Pólya urn Gibbs sampler

Iterate $L$ times between the following steps:

1. for $i = 1, \dots, n$ if the present value of $c_i$ is associated with no other observation ($n_{c_i}^{-i} = 0$), remove $\theta_{c_i}^*$ and draw a new value for $c_i$ from equations (5.14) and (5.15); if the new $c_i$ is not associated with any other observation, draw a value for $\theta_{c_i}^*$ from the posterior $H$ and add $c_i$ to indicators $\boldsymbol{c}$;
2. for all $c \in \boldsymbol{c}$ draw a new value for the parameter from the posterior $\theta_c^*|z_i$ such that $c_i = c$.

### Blocked Gibbs sampler

In the previous Pólya urn Gibbs sampler the $\boldsymbol{c}$ variables are updated one at a time, which can slow down the algorithm; starting by this consideration Ishwaran and James (2001) proposed the blocked Gibbs sampler.

Consider the stick-breaking construction in (5.5) and (5.6) truncated in $S$, where the Beta variables $\boldsymbol{v} = (v_1, \dots, v_S)$ determine the mixing proportions, $p_1, \dots, p_S$, by the expression $p_h = v_h \prod_{k=1}^{h-1}(1 - v_k)$ and the parameters $\boldsymbol{\theta} =$

$(\theta_1, ..., \theta_S)$ are associated with $S$ mixture components. The "truncated version" of the Dirichlet process mixture model in (5.9) can be written as:

$$\begin{aligned} z_i|\boldsymbol{c}, \boldsymbol{\theta} &\sim F(\theta_{c_i}) \\ \theta_h|G_0 &\sim G_0 \\ v_h|\gamma &\sim \text{Beta}(1, \gamma) \\ c_i|\boldsymbol{p} &\sim \text{Mult}(\boldsymbol{p}) \end{aligned} \tag{5.16}$$

for $i = 1, \dots, n$ and $h = 1, \dots, S$. Note that introducing a truncation level we fix the maximum number of possible mixture components.

So far we have considered distinct values of the parameters by defining $\boldsymbol{\theta}^*$; the concept remains similar if applied to the labels: let $\{c_1^*, \dots, c_m^*\}$ denote the set of current $m$ unique values of $\boldsymbol{c}$. Then the blocked Gibbs sampler can be summarized as follows:

**Algorithm 5.2** Blocked Gibbs sampler

Iterate $L$ times between the following steps:

1. Simulate $\theta_k \sim G_0$ for each $k \in \boldsymbol{c} \setminus \{c_1^*, \dots, c_m^*\}$; conditioned on $\boldsymbol{c}$ and $\boldsymbol{z}$, the mixture components $\theta_k$, for $k = 1, \dots, m$, are sampled from the posterior distribution $p(\theta_k|\boldsymbol{c}, \boldsymbol{z})$.

2. Conditioned on $\boldsymbol{v}$, $\boldsymbol{\theta}$ and data $\boldsymbol{z}$ the labels $c_i$, for $i = 1, \dots, n$, are sampled independently from

   $$c_i|\boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \sim \prod_{k=1}^{S} p_{k,i}\delta(k)$$

   where

   $$p_{1,i}, \dots, p_{S,i} \sim (p_1 F(z_i, \theta_1), \dots, p_S F(z_i, \theta_S))$$

   and $p_j = v_j \prod_{h=1}^{j-1}(1 - v_h)$, for $j = 1, \dots, S$.

3. Conditioned on $\boldsymbol{c}$ and $\boldsymbol{z}$, the $v_j$ variables, for $j = 1, \dots, S - 1$, (while $v_S = 1$) are independently sampled from Beta$(1 + n_j, \gamma + \sum_{l=j+1}^{S} n_l)$, where $n_j$ is the number of $c_i$ equal to $j$.

## 5.3 Dirichlet process and Hidden Markov Models

In the discussion regarding the Dirichlet process essentially two important properties emerged: the nonparametric nature and the clustering property. Making a very general classification we could say that the infinite hidden Markov model profits by the first property while the hidden Markov mixture model by the second.

We recall the notation relative to the HMM: let $\{Y_t\}_{t>0}$ be the observable process taking values in $\mathsf{Y}$, with $|\mathsf{Y}| = q$ and $\{X_t\}_{t\geq 0}$ be the hidden Markov chain with state space $\mathsf{X} = \{1, \dots, K\}$. As in the previous presentation we fix $X_0 = 1$; then the HMM can be characterized by the transition matrix $A = \{a_{i,j}\}$, where $a_{i,j} = P(X_{k+1} = j | X_k = i)$, $i, j \in \mathsf{X}$ and the emission matrix $B = \{b_i(y)\}$, with the conditional probabilities $b_i(y) = P(Y_k = y | X_k = i)$, $i \in \mathsf{X},\ y \in \mathsf{Y}$.

We now briefly present those models, then we introduce the hidden mixture Markov model and describe a method for making inference.

### Infinite hidden Markov models

An HMM represents a dynamic variant of the finite mixture model, in which there is one mixture component corresponding to each value of the hidden process: the state at time $t$, $x_t$, indexes a specific row of the transition matrix, with the probabilities in this row serving as the mixing proportions for the choice of the state at time $t+1$, $x_{t+1}$; in a similar way given $x_{t+1}$, by the corresponding row of the emission matrix, the observation $y_{t+1}$ is drawn.

In the infinite hidden Markov model a countable infinite state space of the Markov chain is considered; then this model could be seen as a dynamic variant of an infinite mixture model. As for the infinite mixture model, the infinite hidden Markov model was introduced for avoiding the choice of the number of possible states of the hidden chain. So far we did not discuss the topic of the unknown number of states in an HMM, because we considered cases in which it is gathered from the physical interpretation of the underlying process; in particular we fixed the number of states by attending the final goal of identifying exceptional events. However methods for dealing with this problem by a Bayesian point of view were proposed in literature and they are essentially based on the reversible jump method (see, Robert *et al.*, 2000; Cappé *et al.*, 2005 - Chapter 13) or on the Bayes factor method (Kass and Raftery, 1995); moreover a different approach was proposed in Chopin (2001) and Chopin and Pelgrin (2004).

For dealing with this nonparametric variant of the HMM each row of the transition and emission matrices are modeled by a Dirichlet process. By the Blackwell and MacQueen construction (in Section 5.2) we realized that new draws are more likely to be one of the most popular state. Consider now the same construction with the infinite state space case; that is consider an urn for each possible state (because we have a Dirichlet process for each row of the transition matrix) with balls from an infinite collection of colors (corresponding to the states). When a ball is sampled, given that we can choose between a huge number of balls, is highly likely that a ball with a color never seen before could be sampled. Then with the "simple" Dirichlet process prior, with probability 1, a new state is visited from the chain.

Beal *et al.* (2002) deals with this feature by considering a two-level Dirichlet process hierarchy (hierarchical Dirichlet process): at the first level the probability of transitioning from state $i$ to state $j$ is proportional to the number of times the same transition is observed at other times, while with probability proportional to $\gamma_0$ an "oracle" process is invoked. At this second level, the probability of transitioning to state $j$ is proportional to the number of times state $j$ has been chosen by the oracle (regardless of the previous state), while the probability of transitioning to a novel state is proportional to $\gamma$. The same for the emission process.

### Hidden Markov mixture models

In an hidden Markov mixture model the mixture structure is introduced in an higher level, in fact it assumes that the HMM generating data is itself chosen between a mixture of models. More precisely, the hidden Markov mixture model with $m$ components, as defined in Yuting *et al.* (2007), may be written as

$$p(\boldsymbol{z}|\pi_1, \dots, \pi_S, \Theta_1, \dots, \Theta_S) = \sum_{k=1}^{S} \pi_k p(\boldsymbol{z}|\Theta_k), \tag{5.17}$$

where $\boldsymbol{z} = \{z_t\}_{t=1,\dots,T}$ is a sequence of observations, $p(\boldsymbol{z}|\Theta_k)$ represents the $k$th HMM component with associated parameters $\Theta_k$ and $\pi_k$ represents the mixing weight for the $k$th HMM, with $\sum_{k=1}^{S} \pi_k = 1$.

Let $\mathbf{Z} = \{\boldsymbol{z}_d\}_{d=1,\dots,D}$ be the set of $D$ sequences of data. Each $\boldsymbol{z}_d$ is assumed to be drawn from an HMM with parameters $\Theta_d = (A_d, B_d)$; *i.e.* $\boldsymbol{y}_d \sim \mathcal{H}(\Theta_d)$, where $\mathcal{H}(\Theta_d)$ indicates the HMM. Assume that the set of associated parameters $\{\Theta_d\}_{d=1,\dots,D}$ are independently and identically drawn from a shared prior $G$, *i.e.* they are exchangeable. The distribution $G$ is assumed to be drawn by a Dirichlet process; then, given the clustering property of the Dirichlet process, the parameters $\{\Theta_d\}_{d=1,\dots,D}$ will be clustered and each such cluster corresponds to an HMM mixture component in (5.17). Finally inference is made by a variational Bayes approach (Blei and Jordan, 2004).

## 5.4 The hidden mixture Markov model

Coming back to our purpose of clustering provinces/companies, we said that we are searching for clusters with provinces/companies similar with respect to the transition dynamic.

Let $\mathbf{Y} = \{\boldsymbol{y}_d\}_{d=1,\dots,D}$ be the set containing sequences (of length $T$) relative to the $D = 113$ province and company combinations. Each sequence of data $\boldsymbol{y}_d = (Y_{d,1} = y_{d,1}, \dots, Y_{d,T} = y_{d,T})$ is assumed to be drawn from an HMM with parameters $(A_d, B_d)$, *i.e.* $\boldsymbol{y}_d \sim \mathcal{H}(A_d, B_d)$. Indicate with $\boldsymbol{X}_d = (X_{d,0} = x_{d,0}, \dots, X_{d,T} = x_{d,T})$ the underlying Markov sequence associated with province $d$, where, as before, $X_{d,0} = 1$ for all $d = 1, \dots, D$; in the following this assumption will be taken for granted. Assume that the set of associated transition matrices $\{A_d\}_{d=1,\dots,D}$ are independently and identically drawn from a shared prior $G$, *i.e.* they are exchangeable. Note that we are not assuming that provinces (in the administrative sense) are exchangeable; we are assuming that the transition dynamics between different operating status of the underlying systems are exchangeable.

Assume that the distribution $G$ is drawn from a Dirichlet process with concentration parameter $\gamma$ and base distribution $G_0$; each emission matrix $B_d$ is assumed to be a priori distributed as a product of independent Dirichlet distributions (the same hypothesis introduced for the analysis in Chapter 4).

Then the hidden mixture Markov model with Dirichlet (process and distributions) priors can be written as:

$$\begin{aligned} \boldsymbol{y}_d | A_d, B_d &\sim \mathcal{H}(A_d, B_d) \\ A_d | G &\sim G \\ G &\sim \mathrm{DP}(\gamma, G_0) \\ B_d | \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K &\sim \prod_{i=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_i) \end{aligned} \tag{5.18}$$

where $K$ is the number of possible states of the underlying Markov chain and $\boldsymbol{\beta}_i$ is the vector of length $q+1$ (where $q+1 = |\mathsf{Y}|$, with $\mathsf{Y} = \{0, 1, \ldots, q\}$ is the set of different observable values) containing parameters of the Dirichlet distribution, corresponding to the $i$th row of the emission matrix; we underly that we are considering the same Dirichlet parameters for each province/company.

Consider the "truncated version" of the Sethuraman construction in (5.5) and (5.6); introduce labels $\boldsymbol{c} = \{c_d\}_{d=1,\ldots,D}$, where $c_d = h$ indicates that $A_d$ takes the value $A_h$, with $h = 1, \ldots, S$. Then the model in (5.18) can also be written as

$$\begin{aligned} \boldsymbol{y}_d | \boldsymbol{c}, \{A_h\}_{h=1,\ldots,S}, B_d &\sim \mathcal{H}(A_{c_d}, B_d) \\ c_d | \boldsymbol{p} &\sim \mathrm{Mult}(\boldsymbol{p}) \\ v_h | \gamma &\sim \mathrm{Beta}(1, \gamma) \\ A_h | G_0 &\sim G_0 \\ B_d | \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K &\sim \prod_{i=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_i) \end{aligned} \tag{5.19}$$

where $\boldsymbol{p} = (p_1, \ldots, p_S)$ is determined by $(v_1, \ldots, v_S)$.

For computation convenience we use conjugate priors and assume that

$$G_0 = \prod_{i=1}^{K} \mathrm{Dir}(\boldsymbol{\alpha}_i)$$

where $\boldsymbol{\alpha}_i$ is the vector of length $K$ with parameters of the Dirichlet distribution corresponding to the $i$th row of the transition matrix.

### 5.4.1 Inference by the blocked Gibbs sampler

Consider now inference for the hidden mixture Markov model. First of all we recall the Gibbs sampler for the standard (*i.e.* without the mixture level) finite HMM, when a sequence of data $\boldsymbol{y}$ is analyzed (see Section 1.4, Algorithm 1.2):

**Algorithm 5.3** Gibbs sampler for the finite hidden Markov model

```
Iterate $L$ times between the following steps:
    1. Parameter simulation conditional on the state sequence $\boldsymbol{X}$
        1.a) Sample $A$ from the complete-data posterior distribution $p(A|\boldsymbol{X})$.
        1.b) Sample the emission matrix $B$ from the complete-data posterior $p(B|\boldsymbol{y}, \boldsymbol{X})$.
    2. Conditional of knowing the model parameters $A, B$ sample a path $\boldsymbol{X}$ of the hidden Markov chain from the conditional posterior $p(\boldsymbol{X}|A, B, \boldsymbol{y})$.
```

In a hidden mixture Markov model $D$ sequences of data are considered (and not just one); for each province/company $d = 1, \ldots, D$ the estimating algorithm is similar to that one just presented, but (of course) sampling the transition matrix (step 1.a)) changes, in order to consider the mixture structure.

The model in (5.19) can be represented in the following way, where the nature of the HMM is underlined. For $d = 1, \ldots, D$ consider

- the hidden process:

$$\begin{aligned} \boldsymbol{X}_d|\boldsymbol{c}, \{A_h\}_{h=1,\ldots,S} &\sim \text{Markov chain } (A_{c_d}) \\ A_h|G_0 &\sim G_0 \\ v_h|\gamma &\sim \text{Beta}(1, \gamma) \\ c_d|\boldsymbol{p} &\sim \text{Mult}(\boldsymbol{p}) \end{aligned} \tag{5.20}$$

  where $h = 1, \ldots, S$;

- the observable process:

$$\begin{aligned} \boldsymbol{y}_d|\boldsymbol{X}_d, B_d &\sim B_d \qquad (5.21)\\ B_d|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K &\sim \prod_{i=1}^{K} \text{Dir}(\boldsymbol{\beta}_i). \end{aligned}$$

The hidden process in (5.20) has a structure similar to the model (5.16) and then inference can be made by the blocked Gibbs sampler.

As before let $\{c_1^*, \ldots, c_m^*\}$ denote the set of current $m$ unique values of $\boldsymbol{c}$. Given that $G_0$ is a conjugate prior for the transition matrix, for each $h = 1, \ldots, m$ the posterior distribution is

$$p\left(A_h|\{\boldsymbol{X}_d \text{ s.t. } c_d = h\}\right) = \prod_{i=1}^{K} \text{Dir}\left(\boldsymbol{\alpha}_i + \sum_{d:c_d=h} \boldsymbol{n}_{d,i}\right), \qquad (5.22)$$

where $\boldsymbol{n}_{d,i} = \{n_{d,ij}\}_{j\in\mathsf{X}}$, with $n_{d,ij} = \#\{0 \leq t \leq T-1 : X_{d,t} = i, X_{d,t+1} = j\}$, $i, j \in \{1, \ldots, K\}$.

The posterior distribution for the emission matrix associated with a province/company $d$ is

$$p(B_d|\boldsymbol{X}_d, \boldsymbol{y}_d) = \prod_{i=1}^{K} \text{Dir}(\boldsymbol{\beta}_i + \boldsymbol{e}_{d,i}), \qquad (5.23)$$

where $\boldsymbol{e}_{d,i} = \{e_{d,iy}\}_{y\in\mathsf{Y}}$, with $e_{d,iy} = \#\{1 \leq t \leq T : X_{d,t} = i, Y_{d,t} = y\}$, for $i \in \{1, \ldots, K\}$ and $y \in \mathsf{Y}$.

The blocked Gibbs sampler (see Algorithm 5.2) involves a quantity, $F(z_i, \theta_h)$, that is the likelihood of the observation $z_i$ when the mixture component is $\theta_h$. In this context it becomes $F(\boldsymbol{X}_d, A_h)$, that is the likelihood of a sequence of states $\boldsymbol{X}_d = \{X_{d,t}\}_{t=0,\ldots,T}$, associated with the province/company $d$, when the transition matrix matrix is $A_h$. Let us indicate $A_h = \{a_{j,k}^h\}_{j,k\in\mathsf{X}}$; then, for $h = 1, \ldots, S$

$$F(\boldsymbol{X}_d, A_h) = \prod_{j=1}^{K}\prod_{k=1}^{K} (a_{j,k}^h)^{n_{d,jk}}, \qquad (5.24)$$

where $n_{d,jk} = \#\{0 \leq t \leq T-1 : X_{d,t} = j, X_{d,t+1} = k\}$, $d = 1, \ldots, D$, $j, k \in \mathsf{X}$.

So now we can state the MCMC algorithm for estimating the hidden mixture Markov model.

**Algorithm 5.4** Blocked Gibbs sampler for the hidden mixture Markov model

```
Iterate $L$ times between the following steps:
  1. Transition matrix simulation conditional on the state sequence $\boldsymbol{X}_d$:
     1.a) Simulate each row $i$, $i = 1, \dots, K$ of $A_h$ from the prior Dir$(\boldsymbol{\alpha}_i)$ for each $h \in \boldsymbol{c} \setminus \{c_1^*, \dots, c_m^*\}$; for $h = 1, \dots, m$ draw each row $i$ of $A_h$ from the posterior Dirichlet distribution with parameters $\boldsymbol{\alpha}_i + \sum_{d:c_d=h} \boldsymbol{n}_{d,i}$ (see equation (5.22))
     1.b) Conditioned on $\boldsymbol{v} = (v_1, \dots, v_S)$, $A_1, \dots, A_S$ and $\boldsymbol{X}_d$ the label $c_d$ is sampled from
          $P(c_d = h|\boldsymbol{v}, A_1, \dots, A_S, \boldsymbol{X}_d) \propto p_h F(\boldsymbol{X}_d, A_h)$
          where $p_h = v_h \prod_{j=1}^{h-1}(1 - v_j)$, for $h = 1, \dots, S$ and $F(\boldsymbol{X}_d, A_h)$ is given in equation (5.24).
     1.c) Conditioned on $\boldsymbol{c}$ and $\boldsymbol{X}_d$, the $v_j$ variables, for $j = 1, \dots, S-1$, (while $v_S = 1$) are independently sampled from Beta$(1 + n_j, \gamma + \sum_{l=j+1}^{S} n_l)$, where $n_j$ is the number of $c_i$ equal to $j$.
  2. Sample each row $i$ of the the emission matrix $B_d$ from the posterior Dirichlet distribution with parameters $\boldsymbol{\beta}_i + \boldsymbol{e}_{d,i}$ (see equation (5.23)).
  3. Conditional of knowing the model parameters $A_{c_d}, B_d$ sample a path $\boldsymbol{X}_d$ of the hidden Markov chain from the conditional posterior $p(\boldsymbol{X}_d|A_{c_d}, B_d, \boldsymbol{y}_d)$.
  4. Repeat steps from 1. to 3. for each $d = 1, \dots, D$.
```

Note that because $S$ is the maximum number of possible mixture components, the relation $m \leq S \leq D$ holds.

### 5.4.2 Model specification and results

In this Section we will present results obtained by the application of the hidden mixture Markov model to the available datasets. Before that, we need to introduce some comments, especially related to the model specification.

The implementation of Algorithm 5.4, for the analysis of the hourly number of interruptions for the 113 province and company combinations, is very time consuming. For this reason we reduce the dimensionality of the problem by considering the 6 hours time-interval data (used in the AEEG method).

This (computationally motivated) assumption induces us to rethink the specification of the finite HMM and in particular the choice of the number of possible states of the hidden Markov chain. The physical interpretation of the underlying process, as the system operating status, and the final goal to compare groups resulting from this model-based method with the Cluster analysis presented in Chapter 4, lead up to still consider $K = 4$. Of course estimated model parameters remain not comparable and also the Phase-type distribution needs to be interpreted and treated in a different way. In fact, we could expect that the exceptional excursions will be shorter than that one so far obtained. We underline that the application of the finite HMM, with 4 states, to each province and company combination (*i.e.* each province/company is studied separately from the others), using the 6 hours time interval data gives interesting results in terms of observations deemed as exceptional.

As previously pointed out $S$ is also the number of possible clusters; for this reason and given the obtained clustering results (Section 4.4) we set $S = 10$.

As in Chapter 4, we consider, for each province and company combination, observations greater than 9 in the same way; note that also considering this dataset, this assumption involves a small number of observations (almost the 0.2% of the total). Moreover, priors on the model parameters are the same so far considered (see Section 4.2).

For each $d = 1, \ldots, D$ path of the hidden Markov chain $\boldsymbol{X}_d$ are sampled by the single updating scheme, in order to avoid the time consuming calculation of the backward variables, required in the global updating scheme, so far considered (see Section 1.4.3 for details).

We set Dirichlet distribution parameters and the innovation parameter $\gamma$ equal to 1.

Results suggest that there exists a big cluster containing all the province and company combinations. How to interpret this result? First of all the dataset used in the application of the hidden mixture Markov model is different from the one used in Chapter 4 (6 hours time-interval data and hourly observations). Anyway if we run the MCMC sampler with the hourly number of interruptions, after few iterations (more or less 10) all the province/company are classified in the first cluster (*i.e.* $c_d = 1$, for all $d = 1, \ldots, 113$); for the other cluster the corresponding transition matrix is generated from the prior, without any information from the data, and essentially from a product of uniform distribution (because the Dirichlet parameters are equal to 1). Then it becomes improbable that in the following iterations the classification will change; of course this is a general (naive) consideration because the chain generated by the algorithm is not stationary and everything could happen during the sampling process.

Moreover the estimated transition matrices seen in Chapter 3 (Figures 3.1, 3.9, 3.16. 3.23) have the same structure, that is the probabilities of staying in state 1 and 4 are quite large, probabilities to go from state 1 to 4 and viceversa are small and so on. Then it could be that the model recognizes this general structure and it does not catch other differences.

# Chapter 6

# Urn processes and prior specification in Bayesian hidden Markov models

## 6.1 Introduction

As we anticipated in Section 1.7, in this Chapter we will consider Reinforced Urn Processes, introduced by Muliere *et al.* (2000), in the prior specification when an HMM is analyzed by a Bayesian point of view. A perspective on the prior specification aims to characterize the prior through assumptions on the process $\{X_t\}$; this is what in the de Finetti school is often referred as the "predictive approach". In HMMs this means characterizing the prior on the transition probability of the Markov chain $\{X_t\}$ from assumptions on the predictive structure of the process.

Reinforced urn processes (RUPs) adopt this predictive perspective. A RUP $\{X_t\}_{t \geq 0}$ is a random walk on a state space of Pólya's urn and defines a partially exchangeable sequence (in the sense of Diaconis and Freedman, 1980); when it is recurrent, $\{X_t\}$ is, conditionally on $\mathbf{M}$, a Markov chain with transition matrix $\mathbf{M}$ and $\mathbf{M}$ has a probability law induced by the RUP, namely, the rows of $\mathbf{M}$ are independent Dirichlet processes.

RUPs have been applied in a variety of interesting fields such as survival analysis (Muliere *et al.*, 2000), credit default probability estimation (Amerio *et al.*, 2004) and clinical trials (Mezzetti *et al.*, 2007); to our knowledge they have not been considered for HMMs. Our aim is to consider RUPs and their extensions to construct priors for Bayesian HMMs.

In Section 6.2, we review basic notions of RUPs and their properties. In Section 6.3 and 6.4 we begin to explore how RUP can be of interest for respectively parametric and finite HMMs.

## 6.2 Reinforced Urn Processes

In this Section, we introduce Reinforced Urn Processes - RUPs - as presented in Muliere *et al.* (2000). These processes represent a generalization of the result of Blackwell and MacQueen (1973), who consider a Pólya's urn with a continuum of colors generating an infinite exchangeable sequence of real random variables, whose de Finetti measure is a Dirichlet process. RUP is a class of partially exchangeable processes on a state space of Pólya's urns. By specifying appropriate elements some processes used in Bayesian nonparametrics, like Pólya trees and beta-Stacy processes, can be recovered from RUPs.

**Definition 6.1.** *Consider:*

1. *A countable state space* $\mathsf{X}$
2. *A finite set of colors* $\mathcal{C} = \{c_1, \dots, c_k\}$, *with cardinality* $|\mathcal{C}| = k \geq 1$
3. *An initial urn composition function* $U : \mathsf{X} \to \mathbb{R}_+^k$ *such that for all* $x \in \mathsf{X}$, *there is an urn with composition*

$$U(x) = (n_x(c_1), \dots, n_x(c_k))$$

*where* $n_x(c)$ *represents the number of balls of color* $c$ *contained in the urn, and*

$$\sum_{i=1}^{k} n_x(c_i) > 0$$

4. *A law of motion* $q : \mathsf{X} \times \mathcal{C} \to \mathsf{X}$ *such that,* $\forall x, z \in \mathsf{X}$ *there is at most one color* $c(x, z) \in \mathcal{C}$ *such that* $q(x, c(x, z)) = z$

*The stochastic process* $\{X_t\}$ *on* $\mathsf{X}$ *is constructed as follows: fix an initial state* $x_0 \in \mathsf{X}$*; for all* $t \geq 1$*, if* $X_{t-1} = x \in \mathsf{X}$*, a ball is picked at random from the urn associated with* $x$ *and returned to it along with another of the same color. If* $c \in \mathcal{C}$ *is the color of the sampled ball, set* $X_t = q(x, c)$*.*
*The process is called* $\{X_t\} \in RUP(\mathsf{X}, \mathcal{C}, U, q)$ *with initial state* $x_0$*.*

In other words, with every state $x \in \mathsf{X}$ is associated an urn and we construct a reinforced random walk with the law of motion $q$. For example, we sample a ball from the urn associated with $x_0$, *i.e.* $X_0 = x_0$; if it is of color $c_0$, we return the ball in the urn with another of the same color (reinforcement), set $X_1 = x_1 = q(x_0, c_0)$ and move to $x_1$; next we pick at random a ball from the urn associated with $x_1$; say it is of color $c_1$ we return it in the urn along with another of the same color, set $X_2 = x_2 = q(X_1, c_1)$ and move to $x_2$ and so on.

The resulting process, $\{X_t\} \in \mathrm{RUP}(\mathsf{X}, \mathcal{C}, U, q)$, satisfies some properties.

$\{X_t\}$ is partially exchangeable in the sense of Diaconis and Freedman (1980). Following these authors, two finite sequences $\sigma$ and $\rho$ of elements of $\mathsf{X}$ are *equivalent* if they begin with the same element and, for every $x, z \in \mathsf{X}$ the number of transitions from $x$ to $z$ is the same in both sequences. A process defined on $\mathsf{X}$ is partially exchangeable (in the sense of Diaconis and Freedman, 1980) if for all $n \geq 0$ and all equivalent sequences, $\sigma = (s_0, \ldots, s_t)$ and $\rho = (r_0, \ldots, r_t)$, of elements of $\mathsf{X}$, $P(X_0 = s_0, \ldots, X_t = s_t) = P(X_0 = r_0, \ldots, X_t = r_t)$. As proved by Muliere *et al.* (2000) (Theorem 2.3) given a finite sequence of elements of $\mathsf{X}$, the finite-dimensional law of the RUP $\{X_t\}$ depends only on the number of transitions from a state to another one in $\mathsf{X}$ and therefore it is partially exchangeable.

By Theorem 7 of Diaconis and Freedman (1980) if a partially exchangeable process is also recurrent then it is a mixture of Markov chains.

Therefore a recurrent process $\{X_t\} \in \mathrm{RUP}(\mathsf{X}, \mathcal{C}, U, q)$ is a mixture of Markov chains. More precisely, for all $x \in \mathsf{X}$, set $R_x = \{z \in \mathsf{X} : n_x(c(x,z)) > 0\}$ the set of states $z \in \mathsf{X}$ that the process reaches from $x$; define $R^{(0)} = \{x_0\}$ and $R^{(n)} = \bigcup_{i \in R^{(n-1)}} R_i$ the set of states that the RUP, starting in $x_0$, can reach with positive probability in $n$ steps and finally $R = \bigcup_{n=0}^{\infty} R^{(n)}$. Note that by an appropriate choice of the law of motion $q$ and the urn composition function $U$, we can have $R = \mathsf{X}$ or $R \subset \mathsf{X}$.
Since with probability one the states visited by the process $\{X_t\}$ are elements of

$R$, there is a probability distribution $\mu$ on the set $\mathcal{M}$ of stochastic matrices on $R \times R$, such that, for all $t \geq 1$ and all finite sequences $(x_0, \ldots, x_t)$ of elements in $R$,

$$P(X_0 = x_0, \ldots, X_t = x_t) = \int_{\mathcal{M}} \prod_{j=0}^{t-1} \mathfrak{m}(x_j, x_{j+1}) \mu(d\mathfrak{m}).$$

Let $\mathbf{M}$ be a random element of $\mathcal{M}$ with probability distribution $\mathfrak{m}$; for all $x \in R$, let $\mathbf{M}(x)$ the $x$th row of $\mathbf{M}$ and $\alpha(x)$ be the measure on $R$ which assigns mass $n_x(c)$ to $q(x, c)$ for each $c \in \mathcal{C}$ such that $n_x(c) > 0$ and mass 0 to other elements of $R$.

**Theorem 6.1** (Muliere *et al.* (2000))**.** *If $\{X_t\}$ is recurrent, the rows of $\mathbf{M}$ are mutually independent random probability distributions on $R$ and, for all $x \in R$, the law of $\mathbf{M}(x)$ is that of a Dirichlet process with parameter $\alpha(x)$.*

Moreover if $\{X_t\} \in \mathrm{RUP}(\mathsf{X}, \mathcal{C}, U, q)$ is recurrent, the sequence $\{B_t\}$ of $x_0$-blocks (*i.e.* finite sequences which begin by $x_0$ and contain no further $x_0$, as defined in Diaconis and Freedman, 1980) is exchangeable. This implies that if $\psi$ is a measurable function, the sequence $\{\psi(B_t)\}$ is also exchangeable; $\psi(B_t)$ can be, for example, the length of $B_t$ or the last coordinate.

The de Finetti measure of the exchangeable sequence $\{\psi(B_t)\}$ is characterized by the properties of the recurrent urn process which generated it. Important examples are, for instance, beta-Stacy priors (Walker and Muliere, 1997) and Pólya tree priors (Mauldin *et al.*, 1992).

## 6.3 Reinforced Urn Processes for parametric hidden Markov models

In this Section we consider RUPs in order to construct priors when a parametric HMM is studied from a Bayesian point of view. More specifically we want to take a predictive approach and to characterize the distribution of the rows of the transition matrix $A$ by assumptions on the underlying Markov chain $\{X_t\}$.

In the following we will indicate with Markov chain $(\pi, A)$ a Markov chain with initial state distribution $\pi$ and transition matrix $A$.

Consider a parametric HMM

$$\begin{aligned} Y_t|X_t = i &\sim f(y|\xi_i) \\ \{X_t\}|\pi, A &\text{ is } \text{Markov chain } (\pi, A) \text{ on a countable state space } \mathsf{X} \end{aligned} \tag{6.1}$$

Unknown model parameters are $\boldsymbol{\vartheta} = (\pi, A, \boldsymbol{\xi})$ and a prior on them needs to be specified; let $\xi_i \overset{\text{i.i.d}}{\sim} G_0$ and define a RUP in order to construct the prior on $\pi$ and $A$.

Consider the state space $\mathsf{X}$, a finite set of colors $\mathcal{C}$ and a Pólya's urn $\mathcal{U}_i$ for each possible state $i \in \mathsf{X}$, containing $n_i(c) \geq 0$ balls of color $c \in \mathcal{C}$. Set $X_0 = x_0$ and construct the process $\{X_t\}$ by a law of motion $q$.

As in Section 6.2, let $R^{(n)}$ be the set of states that the process reaches with positive probability in $n$ steps and $R = \bigcup_{n=0}^{\infty} R^{(n)}$. So we can state the following result

**Proposition 6.1.** *The HMM's underlying process $\{X_t\} \in RUP(\mathsf{X}, \mathcal{C}, U, q)$ with initial state $x_0$ and then the process is partially exchangeable. When it is recurrent the rows of $A$ are mutually independent random distributions on $R$ and, for each $x \in R$, the law of the $x$th row of $A$, $A(x)$, is that of a Dirichlet process, whose parameter $\alpha(x)$ is the measure which assigns mass $n_x(c)$ to $q(x, c)$ $\forall c \in \mathcal{C}$ such that $n_x(c) > 0$ and mass 0 otherwise.*

Note that, setting $X_0 = x_0$, we have that the HMM's initial state distribution $\pi = \delta(x_0)$, where $\delta(x)$ denotes the distribution concentrated at point $x$.

In the previous presentation we have not specified if $\mathsf{X}$ is finite or not; however, as we already said, the state space of the hidden Markov chain is often assumed to be a *finite* set, say $\mathsf{X} = \{1, \ldots, K\}$.

In this case we can be more precise in the specification of the RUP's elements, in particular of the set of colors $\mathcal{C}$ and the law of motion $q$.

Consider a Pólya's urn $\mathcal{U}_i$ for each possible state $i \in \mathsf{X}$, containing $n_i(x) \geq 0$ balls labeled with each $x \in \mathsf{X}$, *i.e.* $\mathcal{C} = \mathsf{X}$. For all $n \geq 1$, if $X_{n-1} = i \in \mathsf{X}$, a ball is picked at random from the Pólya's urn associated with $i$; if $j \in \mathcal{C}$ is the label of the sampled ball, we set $X_t = q(i, j) = j$. In other words, the law of motion $q$ simply says that the next state is the label of the picked ball.

The generated process is partially exchangeable and then we can consider the finite version of Proposition 6.1:

**Proposition 6.2.** *If $\{X_t\}$ is recurrent, the rows of $A$ are independent and, since the state space $\mathsf{X}$ is finite, each row $A(x)$ is a Dirichlet distribution with $\alpha(x) = (\alpha_1(x), \dots, \alpha_K(x)) = (n_x(1), \dots, n_x(K))$.*

Consider now a more general HMM formulation (than the one in (6.1))

$$\begin{array}{rcl} Y_t|X_t = \xi_i & \sim & f(y|\xi_i) \qquad\qquad\qquad\qquad\qquad\qquad (6.2) \\ \{X_t\}|\pi, A, \xi_1, \dots, \xi_K & \text{is} & \text{Markov chain } (\pi, A) \text{ on } \mathsf{X} = \{\xi_1, \dots, \xi_K\} \text{ (random)} \end{array}$$

In other words the underlying Markov chain is (directly) defined on the parameter space and this permits us to avoid the label-switching problem (see Section 1.4.4). Of course if we consider labels

$$S_t = j \Longleftrightarrow X_t = \xi_j$$

we obtain the usual formulation (6.1), with $Y_t|S_t = j \sim f(y|\xi_j)$.

As before, given the state space $\mathsf{X} = \{\xi_1, \dots, \xi_K\}$ the RUP's construction for the prior on the transition matrix remains the same as described above and the law of each row $A(\xi_i)$ is a Dirichlet distribution, with parameter $\alpha(\xi_i)$, where the $j$th component, say $\alpha_j(\xi_i)$, is equal to $n_{\xi_i}(\xi_j) \geq 0$, $\xi_j \in \mathcal{C}$ (in fact we have $\mathcal{C} = \mathsf{X}$). Consider a finite sequence of length $T$, $\boldsymbol{X} = (X_0 = x_0, \dots, X_T = x_T)$. Then

$$P(X_{T+1} = \xi_i|\boldsymbol{X}) = \frac{n_{x_T}(\xi_i) + t(x_T, \xi_i)}{\sum_{j=1}^K n_{x_T}(\xi_j) + t(x_T)},$$

where, for $x, z \in \mathsf{X}$, $t(x, z)$ counts the number of transitions from $x$ to $z$ and $t(x)$ is the number of transition from state $x$ in the sequence $\boldsymbol{X}$, *i.e.* $t(x) = \sum_{z \in \mathsf{X}} t(x, z)$.

Now let $(\boldsymbol{y}, \boldsymbol{X})$ be the complete data, $(Y_0 = y_0, \dots, Y_T = y_T, X_0 = x_0, \dots, X_T = x_T)$; inference on the model parameters and on the hidden chain is based on the posterior distribution

$$p(\boldsymbol{X}, \boldsymbol{\vartheta}|\boldsymbol{y}) \;\; \propto \;\; p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}), \qquad (6.3)$$

with

$$p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\vartheta}) = \delta_{x_0} \prod_{k=1}^K \left( \prod_{t:X_t = \xi_t} f(y_t|\xi_t) \right) \prod_{i=1}^K \prod_{j=1}^K a_{\xi_i, \xi_j}^{t(\xi_i, \xi_j)},$$

and, assuming independence between the transition matrix and the emission parameters,

$$p(\boldsymbol{\vartheta}) = \prod_{i=1}^{K} G_0(\xi_i)\,\mathrm{Dir}(\alpha_1(\xi_i), \ldots, \alpha_j(\xi_i), \ldots, \alpha_K(\xi_i)).$$

Sampling from the posterior (6.3) is commonly carried out by the Markov Chain Monte Carlo (MCMC) sampling scheme; in particular a Gibbs sampling algorithm, similar to that one presented in Section 1.4, can be considered:

**Algorithm 6.1** Gibbs sampling algorithm

```
Start with some state sequence $\boldsymbol{X}^{(0)}$ and repeat the following steps for $l = 1, \ldots, L_0, \ldots, L$.

  1. Parameter simulation conditional on the state sequence $\boldsymbol{X}^{(l-1)}$
     1.a) Sample $A$ from the complete-data posterior distribution $p(A|\boldsymbol{X}^{(l-1)})$ and store the values.
     1.b) Sample the emission parameter from the complete-data posterior $p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{X}^{(l-1)})$ and store the values.
  2. Conditional of knowing the model parameters $\boldsymbol{\vartheta}^{(l)}$, sample a path $\boldsymbol{X}$ of the hidden Markov chain from the conditional posterior $p(\boldsymbol{X}|\boldsymbol{\vartheta}^{(l)}, \boldsymbol{y})$ and store all generated states.
  3. Increase $l$ and return to step 1.
```

$L_0$ is the number of burn-in samples to be discarded from the estimate.

The full conditional distribution of $A$ is

$$p(A|\boldsymbol{X}^{(m-1)}) = \prod_{i=1}^{K} \mathrm{Dir}(\alpha_1(\xi_i) + t(\xi_i, \xi_1), \ldots, \alpha_K(\xi_i) + t(\xi_i, \xi_K)).$$

The full conditional distribution $p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{X}^{(m-1)})$ depends on the prior $G_0$ on $\xi_i$; if for example we consider a Poisson HMM (*i.e.* $Y_t|X_t = \xi_i \sim \mathrm{Poi}(\xi_i)$) and $G_0$ is a conjugate Gamma prior, $\Gamma(a_0, b_0)$, then

$$p(\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{K} \Gamma(a_0 + t(\xi_i)\,\overline{y}_i, b_0 + t(\xi_i))$$

where $t(\xi_i) = \#\{1 \leq t \leq T : X_t = \xi_i\}$ and $\overline{y}_i$ is the mean of the observations when $X_t = \xi_i$.

Finally a path of the hidden chain is sampled from $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$, through the *global updating* scheme (see Cappé *et al.*, 2005).

Let $\boldsymbol{y}_{t+1:T} = (y_{t+1}, \dots, y_T)$ and consider the *backward variable*, $\beta_t(\xi_j) = p(\boldsymbol{y}_{t+1:T}|X_t = \xi_j, \boldsymbol{\vartheta})$, inductively computed as follows:

a) Initialize with

$$\beta_T(\xi_j) = 1, \quad 1 \leq j \leq p,$$

b) and for $t = T-1, T-2, \dots, 1$, $1 \leq i \leq p$,

$$\beta_t(\xi_i) = \sum_{j=1}^{K} a_{i,j} p(y_{t+1}|X_{t+1} = \xi_j)\beta_{t+1}(\xi_j).$$

Then the conditional distribution $p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta})$ is

$$p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\vartheta}) \propto \prod_{t=1}^{T} p(y_t|X_t = x_t) a_{x_{t-1},x_t} \beta_t(x_t)$$

and $X_t$ is sampled from

$$Pr(X_t = \xi_j|X_{t-1} = x_{t-1}, \boldsymbol{y}, \boldsymbol{\vartheta}) = \frac{p(y_t|X_t = \xi_j) a_{x_{t-1},\xi_j} \beta_t(\xi_j)}{\sum_{i=1}^{K} p(y_t|X_t = \xi_i) a_{x_{t-1},\xi_i} \beta_t(\xi_i)}$$

for $2 \leq t \leq T$,

$$Pr(X_1 = \xi_j|X_0 = x_0, \boldsymbol{y}, \boldsymbol{\vartheta}) = \frac{p(y_1|X_1 = \xi_j) a_{x_0,\xi_j} \beta_1(\xi_j)}{\sum_{i=1}^{K} p(y_1|X_1 = \xi_i) a_{x_0,\xi_i} \beta_1(\xi_i)}$$

for $t = 1$.

Note that no ordering step (often considered in order to avoid the label switching problem, see Sections 1.4.4 and 1.4.5) was introduced in the Algorithm 6.1. The label switching problem is a feature that HMM shares with Mixture models; in this last framework it is studied in Teicher (1963), Yakowitz and Spragins (1968), Chandra (1977), Redner and Walker (1984), and Crawford (1994); the temporal spectrum of the existing works shows that it is a challenging problem. In the HMM framework the label switching is studied in Chopin (2007).

## 6.4 Reinforced Urn Processes for finite hidden Markov models

Let us consider RUPs in order to construct a prior for the unknown parameters in a finite HMM:

$$\begin{aligned} P(Y_t = y | X_t = i) &= b_i(y) \quad y \in \mathsf{Y}, i \in \mathsf{X} \\ \{X_t\}|\pi, A \quad &\text{is} \quad \text{Markov chain } (\pi, A) \text{ on } \mathsf{X} \end{aligned} \tag{6.4}$$

The HMM's structure can be traced out by a RUP's construction; however in this case two processes are considered and then a (slightly) different RUP needs to be introduced.

**Definition 6.2.** *Consider:*

1. *Two finite state spaces* $\mathsf{X} = \{1, \ldots, K\}$ *and* $\mathsf{Y} = \{1, \ldots, q\}$
2. *Two sets of labels* $\mathcal{L}_1 = \mathsf{X}$ *and* $\mathcal{L}_2 = \mathsf{Y}$
3. *Two initial urn composition functions* $U_1 : \mathsf{X} \to \mathbb{R}_+^K$ *and* $U_2 : \mathsf{X} \to \mathbb{R}_+^q$ *such that for all* $x \in \mathsf{X}$*, there are two urns with composition*

$$U_1(x) = (n_x(1), \ldots, n_x(p)) = \boldsymbol{n}_x$$

$$U_2(x) = (m_x(1), \ldots, m_x(q)) = \boldsymbol{m}_x$$

*where* $n_x(i)$ *represents the number of balls labeled with* $i \in \mathcal{L}_1 = \mathsf{X}$ *contained in one urn, and* $m_x(y)$ *represents the number of balls labeled with* $y \in \mathcal{L}_2 = \mathsf{Y}$ *in the other one;*

$$\sum_{i=1}^{K} n_x(i) > 0, \qquad \sum_{i=1}^{q} m_x(i) > 0$$

4. *A law of motion* $\tau : \mathsf{X} \times \mathcal{L}_1 \to \mathsf{X}$ *and a law of emission* $\epsilon : \mathsf{X} \times \mathcal{L}_2 \to \mathsf{Y}$ *such that,* $\forall x, z \in \mathsf{X}$*,* $\tau(x, z) = z$ *and,* $\forall x \in \mathsf{X}, y \in \mathsf{Y}$*,* $\epsilon(x, y) = y$*.*

*The bi-dimensional stochastic process* $\{Z_t\} = \{X_t, Y_t\}$ *is constructed as follows: fix an initial state* $x_0 \in \mathsf{X}$*; for all* $t \geq 1$*, if* $X_{t-1} = x \in \mathsf{X}$*, a ball is picked at*

*random from each of the two urns associated with $x$ and returned to them along with another with the same label. If $i \in \mathcal{L}_1$ and $y \in \mathcal{L}_2$ are the labels of the sampled balls, set $X_t = \tau(x, i) = i$ and $Y_t = \epsilon(x, y) = y$.*
*The process is called $\{Z_t\} \in$ bi-RUP$(\mathsf{X}, \mathsf{Y}, \mathcal{L}_1, \mathcal{L}_2, U_1, U_2, \tau, \epsilon)$ with initial state $x_0$.*

In other words, two urns are associated with every state $x \in \mathsf{X}$, a "transition urn" $\mathcal{U}_x$ and an "emission urn" $\mathcal{U}_y(x)$; at each time $t \geq 0$ the ball drawn from the urn $\mathcal{U}_x$ gives the transition to the next state at time $t+1$, through the law of motion $\tau$, while the ball drawn from the urn $\mathcal{U}_y(x)$ gives the emitted value at the same time $t$, through the law of emission $\epsilon$. The described sampling scheme is graphically represented in Figure 6.1.

$$\begin{tikzcd}
\{Y_t\} & y_0 & y_1 & y_2 & \cdots \\
 & \mathcal{U}_y(x_0) \arrow[u, "\epsilon"] & \mathcal{U}_y(x_1) \arrow[u, "\epsilon"] & \mathcal{U}_y(x_2) \arrow[u, "\epsilon"] & \\
 & \mathcal{U}_{x_0} \arrow[dr, "\tau"] & \mathcal{U}_{x_1} \arrow[dr, "\tau"] & \mathcal{U}_{x_2} \arrow[dr, "\tau"] & \\
\{X_t\} & x_0 & x_1 & x_2 & x_3 \\
t & 0 & 1 & 2 & 3 \quad \ldots
\end{tikzcd}$$

**Figure 6.1:** Bi-RUP's sampling scheme: fix $x_0 \in \mathsf{X}$ with which two urns are associated, $\mathcal{U}_{x_0}$ and $\mathcal{U}_y(x_0)$; sample a ball from $\mathcal{U}_y(x_0)$ and if the label of the sampled ball is $y_0$ set $Y_0 = y_0 = \epsilon(x_0, y_0)$; sample a ball from $\mathcal{U}_{x_0}$ and if the label of the sampled ball is $x_1$ move to $X_1 = x_1 = \tau(x_0, x_1)$. At time $t = 1$ sample a ball from $\mathcal{U}_y(x_1)$ and set $Y_1 = y_1 = \epsilon(x_1, y_1)$; sample a ball from $\mathcal{U}_{x_1}$ and move to $X_2 = x_2 = \tau(x_1, x_2)$ and so on.

For any $T \geq 0$ and any finite sequence $\boldsymbol{Z} = (Z_0 = z_0, \ldots, Z_T = z_T) = (\boldsymbol{X}, \boldsymbol{y})$

$$P(\boldsymbol{Z}) = P(\boldsymbol{y}|\boldsymbol{X})P(\boldsymbol{X}). \tag{6.5}$$

Let $t^n(i,j)$ and $e^n(i,y)$ respectively be the number of transitions from $i \in \mathsf{X}$ to $j \in \mathsf{X}$ and the number of emitted values equal to $y \in \mathsf{Y}$ from state $i \in \mathsf{X}$, in a sequence of length $n$; then

$$P(\boldsymbol{X}) = \prod_{i=0}^{T-1} \frac{n_{x_i}(x_{i+1}) + t^{i-1}(x_i, x_{i+1})}{\sum_{x \in \mathsf{X}} (n_{x_i}(x) + t^{i-1}(x_i, x))} \tag{6.6}$$

$$P(\boldsymbol{y}|\boldsymbol{X}) = \prod_{i=0}^{T} \frac{m_{x_i}(y_i) + e^{i-1}(x_i, y_i)}{\sum_{y \in \mathsf{Y}}(m_{x_i}(y) + e^{i-1}(x_i, y))} \tag{6.7}$$

with $t^{-1}(\cdot, \cdot) = e^{-1}(\cdot, \cdot) = 0$.

We can now state the following result.

**Theorem 6.2.** *The process $\{X_t\}$ is partially exchangeable.*

*Proof.* In order to show that the generated process $\{X_t\}$ is partially exchangeable we need to prove that, given two equivalent sequences, $\sigma = (s_0, \ldots, s_T)$ and $\rho = (r_0, \ldots, r_T)$, of elements of $\mathsf{X}$,

$$P(X_0 = s_0, \ldots, X_T = s_T) = P(X_0 = r_0, \ldots, X_T = r_T). \tag{6.8}$$

By the definition in Diaconis and Freedman (1980), two sequences are equivalent if they begin with the same element and, for every $i, j \in \mathsf{X}$, the number of transition from state $i$ to $j$ is the same in both sequence. Then because $P(\boldsymbol{X})$ only depends on the number of transition (see equation (6.6)) condition (6.8) follows. □

**Theorem 6.3.** *The bidimensional process $\{Z_t\} = \{X_t, Y_t\}$ generated by the bi-RUP($\mathsf{X}, \mathsf{Y}, \mathcal{L}_1, \mathcal{L}_2, U_1, U_2, \tau, \epsilon$) is a finite HMM. Moreover:*

- *if the process $\{X_t\}$ is recurrent then it is a mixture of Markov chains and each xth row of the transition matrix is independently distributed as a Dirichlet distribution with parameters $\boldsymbol{n}_x$;*

- *the conditional probabilities $P(Y_t|X_t = x)$, $x \in \mathsf{X}$ are Dirichlet distributions with parameters $\boldsymbol{m}_x$, $x \in \mathsf{X}$*

*Proof.* In order to prove that the generated process is an HMM we need to show that the process $\{X_t\}$ is a Markov chain and that conditional on $\{X_t\}$, $\{Y_t\}$ is a sequence of independent random variables such that the conditional distribution of $Y_t$ only depends on $X_t$.

In Theorem 6.2 we have shown that the process $\{X_t\}$ is partially exchangeable; then, by Diaconis and Freedman (1980), if it is recurrent it is a mixture of Markov chains. While the dependence structure of the process $\{Y_t\}$ (*i.e.* dependence of $\{X_t\}$ and conditional independence of $\{Y_t\}$) comes from the construction of the process.

Moreover, given that in the construction of the bi-RUP the emission process does not affect the transition matrix, the process $\{X_t\} \in$ RUP. Then by Proposition 6.2 (if it is recurrent) the rows of the transition matrix $A$ are independent and each row $A(x)$ is a Dirichlet distribution with parameters $\boldsymbol{n}_x$.

Let us now consider the last part of the Theorem. Instead of sampling a ball, at each time, from both the transition urn $\mathcal{U}_x$ and the emission urn $\mathcal{U}_y(x)$ we can first sample the whole state sequence from 1 to $T$ and subsequently for $t = 1, \dots, T$ draw from the emission urn associated with $x_t$. Moreover, given the state sequence $\boldsymbol{X}$, for each state $i \in \mathsf{X}$ we can compute $t(i)$, that is the number of transitions in state $i$; the quantity $t(i)$ is also the number of times we draw from the emission urn associate with $i$. Then given $i$ the $t(i)$ draws from $\mathcal{U}_y(i)$ are independent of draws from others $\mathcal{U}_y(j)$, with $j \neq i$ and moreover they form a Pólya sequence and therefore they are exchangeable. Finally given results in Blakwell and MacQueen (1973) the conditional probability $P(Y_t|X_t = i)$ is a Dirichlet distribution with parameters $\boldsymbol{m}_i$. □

We can also consider the predictive distribution

$$\begin{aligned} P(Z_{T+1} = z_{T+1}|\boldsymbol{Z}) &= P(Y_{T+1} = y_{T+1}, X_{T+1} = x_{T+1}|\boldsymbol{y}, \boldsymbol{X}) \\ &= P(Y_{T+1} = y_{T+1}|\boldsymbol{y}, \boldsymbol{X}_{0:T+1} = \boldsymbol{x}_{0:T+1}) \\ &\quad P(X_{T+1} = x_{T+1}|\boldsymbol{y}, \boldsymbol{X}), \end{aligned}$$

with

$$P(X_{T+1} = x_{T+1}|\boldsymbol{y}, \boldsymbol{X}) = \frac{n_{x_T}(x_{T+1}) + t(x_T, x_{T+1})}{\sum_{x \in \mathsf{X}}(n_{x_T}(x) + t(x_T, x))}$$

and

$$P(Y_{T+1} = y_{T+1}|\boldsymbol{y}, \boldsymbol{X}_{0:T+1}) = \frac{m_{x_{T+1}}(y_{T+1}) + e(x_{T+1}, y_{T+1})}{\sum_{y \in \mathsf{Y}}(m_{x_{T+1}}(y) + e(x_{T+1}, y))}, \tag{6.9}$$

where $\boldsymbol{X}_{0:T+1} = (X_0 = x_0, \dots, X_{T+1} = x_{T+1})$.

## 6.5 Conclusions

In this Chapter we adopted a completely different approach in the prior specification of the HMM parameters; in fact by fixing the predictive structure of the process, we characterized the prior on parameters arising from the parametric and the finite HMMs.

Results stated in Proposition 6.2 and in Theorem 6.3 contains priors usually used in the Bayesian HMM framework (and that we also used in the first five Chapters of this work). However the starting motive for thinking in RUP's terms was to highlight assumptions on the processes implied by the "usual priors" (in terms for example of the recurrence of the process).

Moreover by RUP's theory we know that the so called $x_0$-blocks are exchangeable and if the RUP's parameters are set in a suitable way we have the de Finetti measure; then the idea, actually not investigated in the present work, was to point out also this implication and to study the distribution of these $x_0$-blocks.

# Chapter 7

# Conclusions and further research topics

We have presented different hidden Markov models for identifying exceptional events in electricity distribution: the Poisson HMM, the zero-inflated Poisson HMM, the Negative Binomial HMM, the compound Poisson HMM and the finite HMM. The application of a finite HMM to two telecontrol centers showed that the model fits well the data and the estimated hidden chain is able to identify exceptional events as those where a large number of faults protracting in time occurs. Comparison with the AEEG method is provided and, even if there is not a perfect agreement between periods declared exceptional by the two methods, we can conclude that results obtained by the HMM are satisfactory.

Inspection of the posterior distribution of the transition matrix showed that before and after entering the exceptional operating status, the system is in a transitional operating status; we called *exceptional excursion* the sequence of states visited by the system before reentering the normal state. The time length of exceptional excursions has a discrete Phase-type distribution. In our context this distribution represents the distribution of the time needed to the system to reestablish the normal operating status; then we employed the estimated Phase-type distributions to achieve information related to the efficiency and effectiveness of utility restoration schemes and to make a comparison between the behavior of the utility in different years and between utilities.

We analyzed all province and company combinations, for year 2004; we employed the obtained results to understand if there are similarities between the underlying processes, that manage the occurrence of the exceptional events. In particular, we used the estimated transition matrices and the estimated paths of Markov chain. Two dissimilarity measures have been proposed (based on the Kullback-Leibler distance and the Spearman correlation coefficient) and used in a two steps clustering algorithm.

Clustering results by means of the transition matrices showed that three groups could be identified and the analysis of the Phase-type distributions helped us to characterize province/company in each cluster as: "exceptional persistent", "exceptional transitional"and "fast recovering". Clusters obtained by means of the estimated paths of the hidden chain mainly contain province/company in the North, Center and South part of Italy, suggesting a spatial dependence between provinces.

Moreover, we introduced the hidden mixture Markov model, a model-based approach for clustering province/company, by means of the transitional dynamic. Results suggest that there is a single big cluster.

In order to improve results in the parametric models we could consider combinations of the proposed models; for example we could assume that faults in state 1 and 2 are distributed according to Poisson distributions and observations in states 3 and 4 are emitted by negative binomial or compound Poisson distributions. However we need to be careful when we consider the ordering step in the sampling algorithm, introduced for avoiding the label switching problem; in fact we could obtain nonsensical results and ordered parameters vector could not respect the model assumption. Roughly speaking, in the case with the first two states emitting observations according to a Poisson distribution, after the ordering step we could have that the Poisson means are referred to states 3 and/or 4.

Let us briefly recall interpretation of the compound Poisson distribution for the electricity distribution problem: it arises in a model formed by supposing that the faults occur in cluster, the number of clusters having a Poisson distribution, while the number of interruptions per cluster varies according to a geometric distribution. In other words the Poisson distribution says for how long the instability situation occurs (in fact it says how many geometric values we have to add)

while the number of interruptions (for each hour) is distributed according to a geometric distribution. Instability period does not necessarily mean exceptional event; in the four state Markov chain the instability period refers to states 2, 3 and 4.
But what about observations equal to zero? They are not due to an instability period. Then we could assume that observations in a normal situation (state 1) are distributed according to a Poisson distribution, while given that an instability period occurred, its duration is managed by a Poisson distribution, the number of faults is managed by the geometric distribution, and parameters vary on the base that a minimum, medium or exceptional instability period occurred.

In the performed analyzes we considered data relative to year 2004 for the province and company combinations; however we have at disposal also data for 2005 and 2006. Then we should trace out analysis performed for 2004 also for the other two available years, especially to understand if there are changes in terms of the generated clusters.
In the clustering method we employed a modification of the Spearman correlation coefficient to evaluate distance between the estimated paths of the Markov chain; however another possible way for computing distances between two estimated paths, that we could investigate, is the sequence alignment. In bioinformatics, a sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences; a general global alignment technique is called the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and is based on dynamic programming.
Results from the Cluster analysis by means of the estimated path of the Markov chains suggested that provinces/companies are affected by spatial dependence; then we should try to introduce a model with spatial dependence.

To better understand results from the hidden mixture Markov model we could analyze each province and company combinations separately from the others using the 6 hours time interval data, perform the Cluster analysis by means of the transition matrix and see if there are changes in the conclusions. Moreover we know that the parameter $\gamma$ in the Dirichlet process is also called the innovation parameter, because it manages the generation of new values; then we could consider larger values of this parameter (in the considered application it was set to 1) using also the information deriving from the performed Cluster analysis, that revealed the presence of three clusters.

Another further aspect that could be very interesting to investigate is the predictive problem.

Finally, we considered RUPs for the prior specification in Bayesian HMMs when a predictive perspective is adopted; in particular we applied those processes in order to avoid the label switching problem in the parametric HMM and we proved that the process generated by the introduced bi-RUP is a finite HMM.

Of course the RUP's general theory remains the same also if we consider an infinite state space of the Markov chain; however in this case we need to specify in a suitable way RUP's parameters in order to have the recurrence of the generated process.
An Hoppe's urn (Hoppe, 1984 and 1987) is a Pólya-like urn containing one black ball with positive mass and various numbers of other balls having assorted colors (non-black) each of mass one. At each instant a ball is drawn at random and if the selected ball is black it is returned together with one additional ball of a color not already in the urn. In the HMM framework, the assumption of considering an infinite state space of the Markov chain is introduced in order to avoid the choice of the number of possible states; then the generalization of a RUP with an Hoppe's urn instead of a Pólya's urn could reveal the number of states supported by the data.
We know that in an HMM, observations depend on the underlying process. Then we could consider a RUP where for each $x \in \mathsf{X}$ is associated an urn, but the state space is $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$; in other words we could introduce a RUP were the number of urns is lesser than the number of possible states and investigate properties of the generated process.
In RUP's discussion we mentioned that the sequence of $x_0$-blocks is exchangeable and changing RUP's parameters important processes as beta-Stacy and Pólya trees priors are recovered. How translates this feature in the HMM framework?

Could we obtain the same results in terms of exceptional excursions and Phase-type distributions by the RUP setting?

# References

Accoto, N., Rydén, T. and Secchi, P. (2008). A Bayesian hidden Markov model for identifying exceptional events in electricity distribution. In: *Atti della XLIV Riunione Scentifica SIS*, Società Italiana di Statistica 2008. Arcavacata di Rende (CS). 25-27 Giugno 2008. CLEUP (ITALY).

AEEG, Autorità per l'energia elettrica e il gas, *Regulatory Order n. 333/07*, Available (in Italian) on the web site www.autorita.energia.it.

Albert, J.H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11, 1-15.

Amerio, E., Muliere, P. and Secchi, P. (2004). Reinforced urn processes for modelling credit default distribution, *Internat. J. of Theoret. and Appl. Finance*, 7

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with application to Bayesian nonparametric problems. *Annals of Statistics*, 2, 1152-1174.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics*, 37(6), 1554-1563.

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

Beal, M.J., Ghahramani, Z. and Rasmussen, C. E. (2002). The Infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, 14, 577-584.

Beal, M.J., (2003). *Variational Algorithms for Approximate Bayesian inference*, PHD Thesis, Gatsby Computational Neuroscience Unit, University College London
URL: `http://www.cse.buffalo.edu/faculty/mbeal/papers.html`

Besag, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics*, 16, 395-407.

Bicego, M., Murino, V. and Figueiredo, M.A.T. (2003). *Similarity-Based Clustering of Sequences Using Hidden Markov Models*, Lecture Notes in Computer Science, Springer, Berlin.

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1, 353-355.

Blei, D. and Jordan, M. (2004). Variational methods for the Dirichlet process. ACM International Conference Proceeding Series; Vol. 69, Proceedings of the twenty-first international conference on Machine learning.

Cappé, O., (2002). A Bayesian approach for simultaneous segmentation and classification of count data. *IEEE transactions on signal processing*, 50, 2, 173-449.

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

CEER (2005). *Third benchmarking report on quality of electricity supply*, Council of European Energy Regulators, Brussels.

Celeux, G., Hurn, M. and Robert, C.P. (2000). Computational and Inferential Difficulties with Mixture prior distributions. *Journal of the American Statistical Association*, 95, 957-970.

Chandra, S. (1977). On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 4, 105-112.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75, 79-97.

Chopin, N. (2001). Sequential inference and state number determination for discrete state-space models through particle filtering. CREST Working Paper 2001-34, INSEE, Paris.

Chopin, N. and Pelgrin, F. (2004) Bayesian inference and state number determination for hidden Markov models: an application to the information content of the yield curve about inflation. *Journal of Econometrics*, 123, 327-344.

Chopin, N. (2007). Inference and model choice for sequentially ordered hidden Markov models. *Journal of the Royal Statistical Society. Series B, statistical methodology*, 69, 2, 269-284.

Christie, R.D. (2003). Statistical classification of major event days in distribution system reliability. *IEEE Transactions on Power Delivery*, 18, 4, 1336-1341.

Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematica Biology*, 51, 79-97.

Crawford, S. L. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89, 259-267.

Diaconis, P. and Freedman, D. (1980). De Finetti's Theorem for Markov chains. *Annals of Probability*, 8, 115-130.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 275-285.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 99, 205-215.

Falkhausen, M., Reininger, H. and Wolf, D. (1995). Calculation of distance measures between hidden Markov models. *Proc. Eurospeech*, 1487-1490.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer Series in Statistics.

Fumagalli, E., Lo Schiavo, L., Salvati, S. and Secchi, P. (2006). Statistical identification of major event days: an application to continuity of supply regulation in Italy. *IEEE Transactions on Power Delivery*, 21, 2, 761-767.

Fumagalli, E., Lo Schiavo, L. and Delestre, F. (2007). *Service quality regulation in electricity distribution and retail*, Springer-Verlag.

Fumagalli, E., Lo Schiavo, L., Paganoni, A. and Secchi, P. (2008). Statistical analyses of exceptional events: the Italian Regulatory experiences. *IEEE Transactions on Power Delivery*, forthcoming.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geweke, J., (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics*, 4, Oxford University Press.

Gibbs, A.L. and Su, F.E. (2002). On choosing and bounding probability metrics, *International Statistical Review*, 70, 3, 419-435.

Gilks, W.R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41(2), 337-348.

Guha, S., Li., Yi and Neuberg, D. (2008). Bayesian Hidden Markov Modeling of Array CGH Data. *Journal of the American Statistical Association*, 103, 482, 485-497.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357-384.

Hamilton, J.D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45, 39-70.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57, 97-109.

Heidelberger, P. and Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. *Opns Res.*, 31, 1109-44.

Hoppe, F. M. (1984). Pólya-like urns and Ewens' sampling formula. *Journal of Mathematical Biology*, 20, 91-94.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 20, 91-94.

Ischwaran, J. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-174.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distribution*, Wiley, New York.

Juang, B. H. and Rabiner, L. R. (1985). A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64, 2, 391-408.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Models for Speech Recognition. *Technometrics* 33, 251-272.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data: an introduction to Cluster Analysis.* Wiley series in probability and mathematical statistics. John Wiley and Sons Inc.

Krogh, A., Brown, M., Mian, I.S., Sjlander, K. and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235, 1501-1531.

Lance, G.N. and Williams, W.T. (1979). INVER: A program for the computation of distance-measures between attributes of mixed types. *Australian Computer Journal*, 11, 27-28.

Lehmann, E.L. (2006). *Nonparametrics: statistical methods based on ranks*, Originally published by Prentice-Hall, 1st ed. 1975. Revised edition 2006, Springer-Verlag.

Leroux, B.G. and Puterman, M.L. (1992). Maximum-Penalized-Likelihood estimation for independent and Markov dependent Mixture models. *Biometrics*, 48, 545-558.

Li, C. and Biswas, G. (2000). A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models. *International Conference on Machine Learning*, Stanford, California, 543-550.

Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1999). Markovian structures in biological sequence alignments. *Journal of the American Statistical Association*, 94, 1-15.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23, 727-741.

MacEachern, S. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics*, 7, 223-238.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2005). Cluster Analysis Basics and Extensions; unpublished

Mardia, K. V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, London.

Mauldin, R.D., Sudderth, W.D. and Williams, S.C. (1992). Pólya trees and random distributions. *Annals of Statistics*, 20, 1203-1221.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1087-1092..

Mezzetti, M., Muliere, P. and Bulla, P. (2007). An application of reinforced urn processes to determining maximum tolerated dose. *Statistics and Probability Letters*, 77 (7), 740-747.

Muliere, P., Secchi, P. and Walker, S. (2000). Urn schemes and reinforced random walks. *Stochastic Processes and their Applications*, 88, 58-79.

Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, Vol. 9, No. 2, 249-265.

Neal, R.M. (2003). Slice Sampling. *Annals of Statistics*, 31, 705-767.

Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 (3), 443-53.

Neuts, M.F. (1994). *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*, Corrected reprint of the 1981 original, Dover Publications, Inc., New York.

Panuccio, A., Bicego, M. and Murino, V. (2002). A hidden markov model-based approach to sequential data clustering. In the book *Structural, Syntactic and Statistical Pattern Recognition*, Springer.

Peskun, P.H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60, 607-612.

Peskun, P.H. (1981) Guidelines for chosing the transition matrix in Monte Carlo methods using Markov chains. *Journal of Computational Physics*, 40, 327-344.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2008). coda: Output analysis and diagnostics for MCMC. R package version 0.13-3.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabiner, L.R., (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77, 257-285.

Raftery, A.E. and Lewis, S.M. (1992a). How many iterations in the Gibbs Sampler?. *Bayesian Statistics*, 4, Oxford University Press.

Raftery, A.E. and Lewis, S.M. (1992b). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7, 493-497.

Raftery, A.E. and Lewis, S.M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Practical Markov Chain Monte Carlo*, Chapman and Hall.

Ramoni, M., Sebastiani, P. and Cohen, P. (2002). Bayesian clustering by dynamics. *Machine Learning*, 47, 91-121.

Redner, R. A. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26, 195-239.

Robert, C.P., Celeux, G. and Diebold, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics and Probability Letters*, 16, 77-83.

Robert, C.P., Rydén, T. and Titterington, M. (2000). Bayesian inference in hidden Markov models through reversible jump Markov chain Monte Carlo. *Journal of the Royal Statistal Society Ser. B*, 62, 57-75.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Sethuraman, J. (1994). A constructive definition of the Dirichlet prior. *Statistica Sinica*, 2, 639-650.

Scott, S. L. (2002). Bayesian methods for Hidden Markov Models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97, 337- 351

Smyth, P. (1997). Clustering sequences with hidden Markov models. *Advances in Neural Information Processing Systems*, MIT Press, 648-654.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-550.

Teh, Y., Jordan, M.I., Beal, M. and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, 1566-1581.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34, 1265-1269.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260-269.

Walker, S. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes, in *Practical Nonparametric and Semiparametric Bayesian*, eds. D. Dey, P. Müller and D. Sinha, 243-254, Springer-Verlag, New York.

Warren, C.A., Bouford, J.D., Christie, R.D., Kowalewski, D., McDaniel, J., Robinson, R., Schepers, D.J., Viglietta, J. and Williams, C. (2003). Classification of major event days. In proceeding of *Power Engineering Society General Meeting*, IEEE, 1, 466-471.

West, M., Müller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models with applications in regression and density estimation, in *Aspects of Uncertainty*, eds. P. R. Freeman and A. F. Smith, 363– 386, John Wiley, New York.

Xing, E. P. and Sohn, K. (2007). Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2(2).

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39, 209-214.

Yuting, Qi, Paisley, J.W. and Carin, L. (2007). Dirichlet Process HMM Mixture Models with Application to Music Analysis. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2, 465-468.