 Vol. 23: 419–441, 1996

Fitting Phase-type Distributions via the EM Algorithm

SØREN ASMUSSEN
Aalborg University

OLLE NERMAN and MARITA OLSSON
University of Göteborg

ABSTRACT. Estimation from sample data and density approximation with phase-type distributions are considered. Maximum likelihood estimation via the EM algorithm is discussed and performed for some data sets. An extended EM algorithm is used to minimize the information divergence (maximize the relative entropy) in the density approximation case. Fits to Weibull, log normal, and Erlang distributions are used as illustrations of the latter.

Key words: Coxian distribution, density estimation, EM algorithm, hidden Markov chain, I-divergence, phase-type distributions, relative entropy

1. Introduction

Phase-type distributions are defined as distributions of absorption times Y in Markov processes with $p < \infty$ transient states (the phases) and one absorbing state 0. The simplest examples are mixtures and convolutions of exponential distributions (in particular Erlang distributions, defined as gamma distributions with integer parameter). More generally, the class comprises all series/parallel arrangements of exponential distributions, possibly with feedback.

There are several motivations for using phase-type distributions as statistical models. The most established ones come from their role as the computational vehicle of much of applied probability: very often, problems which have an explicit solution assuming exponential distributions are algorithmically tractable when one replaces the exponential distribution with a phase-type distribution. For typical examples, see e.g. Neuts (1981), Sengupta (1989), Asmussen (1992) in queueing, Asmussen & Rolski (1991) in insurance risk theory, Kao (1988), Lipsky (1992), Asmussen & Bladt (1996) in renewal theory, and Bobbio *et al.* (1980), Jonsson *et al.* (1994) in reliability. Assume e.g. that for design purposes the engineer needs the mean waiting time in a data queue. He would then fit a phase-type distribution to the observed service times and compute the exact mean waiting time in the queue with this fitted service time distribution. In such situations, the phase-type model is crucial because otherwise the mean waiting time is not available in closed form. Furthermore, one may argue that there is no essential loss in generality in the phase-type setup: the class of phase-type distributions (with p taking any finite value) is dense and hence any distribution on $[0, \infty)$ can, at least in principle, be approximated arbitrarily close by a phase-type distribution. Some main studies of the estimation problem originating from this framework are Bux & Herzog (1977), Johnsson & Taaffe (1989, 1990a, b), Bobbio & Cumani (1990) and Bobbio & Telek (1994).

The relevance of phase-type distributions can also be argued in more traditional statistical settings. Due to the denseness, one can view phase-type modelling as a semi-parametric density estimation procedure with a built-in smoothing (the degree of smoothness being determined by the value of p). In such applications, the phases have no physical interpretation and the

phase-type modelling is purely descriptive. However, in other areas such as demography (see Hoem, 1969), drug kinetics, epidemiology, etc, the probabilistic interpretation fits in nicely with standard Markovian modelling. Explicit use of phase-type ideas to generalize exponential times to more arbitrary residence times can be found in Faddy (1990, 1993, 1994), where maximum likelihood estimation for some substructures of phase-type distributions, permitting a restricted form of feedback, is also investigated. Another recent contribution to this area is found in Aalen (1993, 1995), where acyclic phase-type models in survival analysis are discussed and several examples of phase-type modelling in survival analysis are given.

The literature on estimation of (and approximation by) general phase-type distributions is meagre and not always satisfying from a statistical point of view. Up to now estimation in connection with subclasses have mainly been considered. In Bux & Herzog (1977) minimization of the maximal absolute value of the difference between the empirical distribution and Coxian distributions, in a fixed finite set of points, is used as a fitting criteria. (For a definition of Coxian distributions see section 2). Mixtures of Erlang distributions are fitted by a variety of methods including moment matching and non-linear programming which is used in Johnson & Taaffe (1989, 1990a, b), and Johnson (1990). Such restrictions on the class of phase-type distributions are, however, not natural in all applications; a particular drawback is that assuming a special structure like a mixture of Erlang distributions, leads to large values of p and thereby a high complexity in the applied probability algorithms. Numerical maximum likelihood methods for Coxian distributions, using non-linear constrained optimization, have been implemented recently in Bobbio & Cumani (1990), and Bobbio & Telek (1994); this approach appears in many ways to be the most satisfying developed so far, the main restriction being that only Coxian distributions are allowed.

In this paper (which is based on ideas first sketched in Asmussen & Nerman (1991)), we present a general statistical approach to estimation theory for phase-type distributions. The idea is quite straightforward: the class of phase-type distributions may for a fixed p be viewed as a multi-parameter exponential family, provided the whole of the underlying absorbing Markov process is observed. Since the data in practice consist of i.i.d. replications of the absorption times $Y_1, \ldots, Y_n$ of Y, we are in the setting of incomplete observations and may try to implement the EM algorithm.

The idea to use the EM algorithm in connection with finite state space Markov chains is certainly not new. In fact, one of the roots of the algorithm, Baum *et al.* (1970), is from Markov chain theory. Another variant of the EM algorithm which is of particular relevance for us, was developed in Sundberg (1974, 1976), in connection with partial observations of samples from the exponential family. (Inspiration to the development also came from missing observation problems, Orchard & Woodbury (1972), and estimation of mixture distributions, Redner & Walker (1984).) A classical reference on the EM algorithm in general is Dempster *et al.* (1977) and the discussion contributions there. Convergence criteria and problems with convergence to saddle points and local maxima are discussed in Wu (1983), where some mistakes in Dempster *et al.* (1977) are also pointed out.

Our paper is organized as follows: in section 2, we give a short introduction to phase-type distributions following such standard sources as Neuts (1981). In section 3, we describe the EM algorithm in detail, with some of the calculations for the E-step being deferred to the Appendix. For the approximation of a theoretical density by a phase-type density we also consider the infinite sample analogue of maximum likelihood estimation: minimization of the information divergence (relative entropy or Kullback–Leibler information). Computationally, this turns out to be almost equivalent to the EM algorithm for a sample. Section 4 contains descriptions of the most important substructures of phase-type distributions. In section 5, a number of examples performed with the EMPHT-program (an

implementation of the proposed algorithm which is described in detail in Häggström *et al.* (1992)) are shown. In the Appendix we derive some key formulas used in the E-step of the EM algorithm, and we also discuss the theoretical basis of the information divergence variant of the EM algorithm in general.

We conclude this section by pointing out some further related work and references: Lang & Arthur (1994) present a careful experimental evaluation of various methods and packages for fitting phase-type distributions, including the approach of the present paper as implemented in Häggström *et al.* (1992) (their main criticism concerns speed but as remarked later, we have not yet implemented the various possibilities for speeding up the algorithm). The algorithm presented in this paper has been extended to handle right-censored observations and interval-censored observations, which is presented in a companion paper (Olsson, 1996). A numerical algorithm for maximum likelihood estimation of so-called generalized mixed exponential distributions (permitting negative mixing weights) is treated in Harris & Sykes (1987). This class is, just like phase-type distributions, a subclass of the so-called generalized Coxian distributions: all distributions on $[0, \infty)$ which have rational Laplace transforms (Cox, 1953). Some of its distributions are not representable as phase-type distributions and, vice versa, some phase-type distributions with cyclic Markov representation are not general mixed exponential distributions. See also Ruhe (1980) for another contribution to this problem area, and for the use of the EM algorithm in mixing models in general see Redner & Walker (1984) and its references. In the series of papers collected in Ryden (1993), estimation theory for Markov modulated point processes is considered, a problem which in applied probability can be seen as the natural next step after phase-type fitting of one-dimensional distributions. One of the papers, in fact, also implements the EM algorithm. The situation in Ryden (1993) falls within the framework of hidden Markov models which has been studied in general in Leroux (1992). It should be noted, however, that in our observation scheme there are no dependencies of the type occurring in hidden Markov models and which are the main problem there.

2. Phase-type distributions

Consider a Markov process J_u on a finite state space $\{0, 1, \ldots, p\}$, where 0 is absorbing and the other states are transient. The absorbing state makes it possible to block partition the infinitesimal generator $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{pmatrix} 0 & 0, \ldots, 0 \\ \mathbf{t} & \mathbf{T} \end{pmatrix},$$

where t_i (the ith element of $\mathbf{t}$, the *exit vector*) is the conditional intensity of absorption in 0 when J_u is in state i. The $(p \times p)$-dimensional matrix $\mathbf{T}$ (called the phase-type generator) is always non-singular and thus invertible. Further, it is clear that $\mathbf{t} = -\mathbf{Te}$, where $\mathbf{e} = (1, \ldots, 1)'$, since each row in $\mathbf{Q}$ sums to zero.

A random variable Y, distributed as the absorption time $\inf\{v > 0: J_v = 0\}$ corresponding to an initial distribution $\boldsymbol{\pi}$ (defined as a row vector), is said to be *phase-type distributed* $(\boldsymbol{\pi}, \mathbf{T})$. The statistical parameters are thus $\boldsymbol{\pi}$ and $\mathbf{T}$. We treat p as fixed and do not discuss the choice of p in this paper (although one could use, for example, Akaike's information criterion AIC, to compare different choices of p).

The transition matrices

$$\mathbf{P}_v = \exp(\mathbf{Q}v\} = \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n v^n}{n!}$$

of the Markov process can also be block partitioned,

$$\mathbf{P}_v = \begin{pmatrix} 1 & 0, \ldots, 0 \\ \mathbf{e} - \exp\{\mathbf{T}v\}\mathbf{e} & \exp\{\mathbf{T}v\} \end{pmatrix},$$

which immediately gives us an expression for the distribution function $F(y)$

$$F(y) = 1 - \boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{e}.$$

Some further basic analytical characteristics of a phase-type distribution are:

(i) the density $f(y) = \boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{t}$
(ii) the failure rate $r(y) = \boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{t} / \boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{e}$
(iii) the Laplace transform $\int_0^\infty \exp\{-sy\}F(dy) = \boldsymbol{\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$
(iv) the nth moment $m_n = \int_0^\infty y^n F(dy) = (-1)^n n! \boldsymbol{\pi}\mathbf{T}^{-n}\mathbf{e}$.

We call the phase-type distribution *Coxian* if

$$\pi_1 = 1, \quad -t_{i,i} = t_{i,i+1} + t_i \text{ for } i = 1, \ldots, p-1, \quad \text{and} \quad -t_{p,p} = t_p.$$

In the rest of this section we discuss the parameterization problem, which is important from a statistical point of view and certainly is a non-trivial problem. The parameterization with $\boldsymbol{\pi}$ and $\mathbf{T}$ is by no means unique. First, there is a trivial aliasing due to the arbitrary ordering of the states; simultaneous permutations of rows and columns in $\mathbf{T}$ and of the elements in $\boldsymbol{\pi}$ (with the same permutation) result in the same phase-type distribution. Second, there is a more subtle unidentifiability problem present; if e.g. $\mathbf{t} = -\mathbf{T}\mathbf{e} = \lambda\mathbf{e}$ for $\lambda > 0$, then the corresponding phase-type distribution is exponentially distributed with parameter λ, irrespective of the choice of $\boldsymbol{\pi}$.

A remarkable result from Cumani (1982) and Dehon & Latouche (1982) (see also Ó Cinnéide, 1987), is that any phase-type distribution having an acyclic Markov chain representation can be uniquely represented by a Coxian distribution with stochastically increasing states, i.e. $-t_{1,1} \geqslant -t_{2,2} \geqslant \cdots \geqslant -t_{p,p}$. Such a process starts in state 1, and can only jump from state i to $i+1$ or to the absorption state 0. Thus, the true parameter dimension for acyclic phase type distributions of order p is $2p-1$ (note that start in 0 is not allowed).

In fact, also the full class of phase-type distributions of order p has a parametrization in $2p-1$ dimensions: it follows from the Cayley–Hamilton theorem that there is at least one sequence $\lambda_0, \lambda_1 \ldots, \lambda_{p-1}$ such that

$$\mathbf{T}^{-p}\mathbf{e} = \sum_{i=0}^{p-1} \lambda_i \mathbf{T}^{-i}\mathbf{e}.$$

If we fix such a sequence, then these coefficients together with the first $p-1$ moments determine all moments recursively. This is seen by multiplying the relation above by $\boldsymbol{\pi}\mathbf{T}^{-n}$ from the left:

$$\frac{(-1)^{n+p} m_{n+p}}{(n+p)!} = \sum_{i=0}^{p-1} \lambda_i \frac{(-1)^{n+i} m_{n+i}}{(n+i)!} \quad n = 0, 1, 2, \ldots,$$

where $m_0 = 1$. Now, since the Laplace transform near zero is determined by all the moments, it follows that $\lambda_0, \ldots, \lambda_{p-1}, m_1, \ldots, m_{p-1}$ determine the distribution.

The method we will use to estimate (or approximate by) phase-type distributions depends on the parameters of the Markov process. However, since we are primarily interested in estimable quantities as the distribution, density or failure rate function, the drawback of using over-parameterization is not that great.

3. The EM algorithm for phase-type distributions

The EM (expectation-maximization) algorithm is an iterative method for maximum likelihood estimation (Dempster *et al.*, 1977; Wu, 1983). Its area of applications concerns incomplete data, i.e. data which can be thought of as partial observations of a larger experiment, where a more specified course of events can be observed than in the experiment actually performed.

Suppose that $Y = u(X)$, with density g_γ, is observed (for a many to one mapping u) and think of X, with density f_γ, as the result of the larger unobserved experiment. Then step $n+1$ in the EM algorithm consists of finding a value γ_{n+1} which maximizes

$$\gamma \to \mathbb{E}_{\gamma_n}[\log f_\gamma(X) \mid u(X) = y],$$

where y is the observed data and γ_n the current estimate after n steps of the algorithm; the evaluation of the conditional expectation is the *E-step*, and the maximization is the *M-step*.

Denote by k_γ the conditional density of X given $u(X)$. Using the logarithm of the relation

$$f_\gamma(x) = g_\gamma(u(x))k_\gamma(x \mid u(x))$$

and Jensen's inequality, it is straightforward to see that the likelihood increases in each step:

$$g_{\gamma_{n+1}}(y) \geq g_{\gamma_n}(y).$$

Thus, if γ_n converges we can hope for a convergence to the maximum likelihood estimate $\hat{\gamma}$. However, convergence might be hard to prove, and worse, convergence may take place to local maxima or even saddle points (see Wu, 1983).

In our case when $\mathbf{X}$ belongs to a multi-dimensional exponential family with density

$$f_\gamma(\mathbf{x}) = \exp\{\boldsymbol{\theta}(\gamma)'\mathbf{S}(\mathbf{x}) + d(\boldsymbol{\theta}(\gamma))\},$$

(cf. (1) in the next section), the E-step consists of calculating

$$\mathbb{E}_{\gamma_n}[\mathbf{S}(\mathbf{X}) \mid u(\mathbf{X}) = \mathbf{y}].$$

In the M-step, the likelihood f is maximized by using this expectation as the observed value of $\mathbf{S}(\mathbf{X})$.

3.1. *Construction of the complete sample*

The connection between Markov processes and phase-type distributions makes it natural to consider the incomplete data approach to find a way of calculating maximum likelihood estimates. An observation y of the time to absorption can be regarded as an incomplete observation of the Markov process J_u. It is incomplete in the sense that it only tells us when the process hits 0, and does not provide any information about how it got there, where it started, which of the states it visited and for how long.

Observing the whole Markov process is equivalent to observing the embedded Markov chain $I_0, I_1, \ldots, I_{M-1}$ ($I_M = 0$) and the sojourn times $S_0, S_1, \ldots, S_{M-1}$ ($S_M = \infty$), where M is the number of jumps until J_t hits the absorbing state 0.

Thus, given an observation y of the phase-type distribution, a complete observation of the process J_u on the interval $(0, y]$ can be represented by $x = (i_0, \ldots, i_{m-1}, s_0, \ldots, s_{m-1})$, where the sojourn times must satisfy $y = s_0 + \cdots + s_{m-1}$.

Let

$$p_{jk} = \mathbb{P}(I_{n+1} = k \mid I_n = j) = \begin{cases} t_{jk}/\lambda_j & j, k = 1, \ldots, p \\ t_j/\lambda_j & j = 1, \ldots, p, \quad \text{and } k = 0, \end{cases}$$

where λ_i is the intensity of the holding time in state i, that is $\lambda_i = -t_{ii}$, and t_i represents the ith element of the exit vector $\mathbf{t}$. Then the density of the observation x is

$$f(x; \boldsymbol{\pi}, \mathbf{T}) = \pi_{i_0}\lambda_{i_0} \exp\{-\lambda_{i_0}s_0\}p_{i_0i_1} \ldots \lambda_{i_{m-1}} \exp\{-\lambda_{i_{m-1}}s_{m-1}\}p_{i_{m-1}0}$$
$$= \pi_{i_0} \exp\{-\lambda_{i_0}s_0\}t_{i_0i_1} \ldots \exp\{-\lambda_{i_{m-1}}s_{m-1}\}t_{i_{m-1}}.$$

Now, suppose that we have n independent replications of the process, $J_u^{[1]}, \ldots, J_u^{[n]}$, and let $I_0^{[v]}, \ldots, I_{M^{[v]}-1}^{[v]}$ denote the embedded Markov chain and $S_0^{[v]}, \ldots, S_{M^{[v]}-1}^{[v]}$ the holding times for the vth process. Hence, a sample of size n is represented by

$$x = (i_0^{[1]}, \ldots, i_{m^{[1]}-1}^{[1]}, s_0^{[1]}, \ldots, s_{m^{[1]}-1}^{[1]}, \ldots, i_0^{[n]}, \ldots, i_{m^{[n]}-1}^{[n]}, s_0^{[n]}, \ldots, s_{m^{[n]}-1}^{[n]}).$$

This is the complete data-set which we will use in the EM algorithm to try to find the maximum likelihood estimate of $(\boldsymbol{\pi}, \mathbf{T})$ from the observed sample

$$\mathbf{y} = (y_1, \ldots, y_n) = (s_0^{[1]} + \cdots + s_{m^{[1]}-1}^{[1]}, \ldots, s_0^{[n]} + \cdots + s_{m^{[n]}-1}^{[n]}).$$

The density of the complete sample $\mathbf{x}$ can be written in the form

$$f(\mathbf{x}; \boldsymbol{\pi}, \mathbf{T}) = \prod_{i=1}^{p} \pi_i^{B_i} \prod_{i=1}^{p} \exp\{t_{ii}Z_i\} \prod_{i=1}^{p} \prod_{\substack{j=0 \\ j \neq i}}^{p} t_{ij}^{N_{ij}} \tag{1}$$

where

$$B_i = \sum_{v=1}^{p} \mathbf{1}_{\{I_0^{[v]} = i\}} = \text{the number of Markov processes starting in state } i,\ i = 1, \ldots, p.$$

$$Z_i = \sum_{v=1}^{n} \prod_{k=0}^{m^{(v)}-1} \mathbf{1}_{\{I_k^{[v]} = i\}} S_k^{[v]} = \text{the total time spent in state } i,\ i = 1, \ldots, p.$$

$$N_{ij} = \sum_{v=1}^{n} \sum_{k=0}^{m^{[v]}-1} \mathbf{1}_{\{I_k^{[v]} = i,\ I_{k+1}^{[v]} = j\}}$$
$$= \text{the total number of jumps from state } i \text{ to state } j, \text{ for } i \neq j,\ i = 1, \ldots, p,$$
$$\text{and } j = 0, 1, \ldots, p.$$

The density $f(\mathbf{x}; \boldsymbol{\pi}, \mathbf{T})$ is a member of a curved multi-parameter exponential family with sufficient statistic

$$\mathbf{S} = ((B_i)_{i=1,\ldots,p}, (Z_i)_{i=1,\ldots,p}, (N_{ij})_{i=1,\ldots,p, j=0,\ldots,p, i\neq j}).$$

It follows either by general theory for exponential families or by explicit calculations (using $-(t_i + \Sigma_{j\neq i} t_{ij}) = t_{ii}$) that the maximum likelihood estimates, based on the fictitious sample $\mathbf{x}$, are

$$\hat{\pi}_i = \frac{B_i}{n}, \quad \hat{t}_{ij} = \frac{N_{ij}}{Z_i}, \quad \hat{t}_i = \frac{N_{i0}}{Z_i}, \quad \hat{t}_{ii} = -\left(\hat{t}_i + \sum_{\substack{j=1 \\ j\neq i}}^{p} \hat{t}_{ij}\right), \quad i, j = 1, \ldots, p. \tag{2}$$

See Albert (1961) or Basawa & Rao (1980) for a detailed account of how to derive the maximum likelihood estimate of the intensity matrix of a finite state Markov process.

3.2. The E- and M-steps

The first step of each iteration, the E-step, consists of calculating the conditional expectation of the sufficient statistic $\mathbf{S}$, given the observed sample $\mathbf{y}$ and the current estimates of $(\boldsymbol{\pi}, \mathbf{T})$, say $(\boldsymbol{\pi}, \mathbf{T})^{(k)}$. Then in the M-step the likelihood (1) is maximized, using the conditional

expectation of $\mathbf{S}$ as its observed value. Hence, we get the new estimates of $(\boldsymbol{\pi}, \mathbf{T})$ simply by replacing the statistics in (2) with their conditional expectations evaluated in the E-step.

Note that the single statistics in $\mathbf{S}$ are all sums over the sample, which means that conditioning on the sample $\mathbf{y}$ reduces to conditioning on one observation in each summand. Letting $B_i^{[v]}$, $Z_i^{[v]}$, and $N_{ij}^{[v]}$ be the contributions to $\mathbf{S}$ from the vth observed process, then the $k+1$st iteration of the algorithm becomes

E-step: Calculate

$$B_i^{(k+1)} = \sum_{v=1}^{n} \mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k)}}[B_i^{[v]} \mid y_v] \quad \text{for } i = 1, \ldots, p$$

$$Z_i^{(k+1)} = \sum_{v=1}^{n} \mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k)}}[Z_i^{[v]} \mid y_v] \quad \text{for } i = 1, \ldots, p$$

$$N_{ij}^{(k+1)} = \sum_{v=1}^{n} \mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k)}}[N_{ij}^{[v]} \mid y_v] \quad \text{for } j \neq i,\ i = 1, \ldots, p, \text{ and } j = 0, 1, \ldots, p.$$

M-step: The new estimates are given by

$$\pi_i^{(k+1)} = \frac{B_i^{(k+1)}}{n}, \quad t_{ij}^{(k+1)} = \frac{N_{ij}^{(k+1)}}{Z_i^{(k+1)}}, \quad t_i^{(k+1)} = \frac{N_{i0}^{(k+1)}}{Z_i^{(k+1)}}, \quad t_{ii}^{(k+1)} = -\left(t_i^{(k+1)} + \sum_{\substack{j=1 \\ j \neq i}}^{p} t_{ij}^{(k+1)}\right).$$

The difficult part of the EM algorithm is in our case the E-step, which is computationally heavy. In the appendix we show that

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})}[B_i^{[v]} \mid Y = y_v] &= \frac{\pi_i b_i(y_v \mid \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_v \mid \mathbf{T})} \\
\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})}[Z_i^{[v]} \mid Y = y_v] &= \frac{c_i(y_v; i \mid \boldsymbol{\pi}, \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_v \mid \mathbf{T})} \\
\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})}[N_{ij}^{[v]} \mid Y = y_v] &= \frac{t_{ij} c_j(y_v; i \mid \boldsymbol{\pi}, \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_v \mid \mathbf{T})} \\
\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})}[N_{i0}^{[v]} \mid Y = y_v] &= \frac{t_i a_i(y_v \mid \boldsymbol{\pi}, \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y_v \mid \mathbf{T})}
\end{aligned} \tag{3}$$

where $\mathbf{e}_i$ is the ith unit vector and $\mathbf{a}(y \mid \boldsymbol{\pi}, \mathbf{T})$, $\mathbf{b}(y \mid \mathbf{T})$, $\mathbf{c}(y; 1 \mid \boldsymbol{\pi}, \mathbf{T}), \ldots, \mathbf{c}(y; p \mid \boldsymbol{\pi}, \mathbf{T})$ are p-dimensional vector functions defined by

$$\mathbf{a}(y \mid \boldsymbol{\pi}, \mathbf{T}) = \boldsymbol{\pi} \exp\{\mathbf{T}y\}$$

$$\mathbf{b}(y \mid \mathbf{T}) = \exp\{\mathbf{T}y\}\mathbf{t}$$

$$\mathbf{c}(y; i \mid \boldsymbol{\pi}, \mathbf{T}) = \int_0^y \boldsymbol{\pi} \exp\{\mathbf{T}u\}\mathbf{e}_i \exp\{\mathbf{T}(y-u)\}\mathbf{t}\, du \quad i = 1, \ldots, p.$$

For fixed $\boldsymbol{\pi}$ and $\mathbf{T}$, these functions satisfy a $p(p+2)$-dimensional linear system of homogeneous differential equations. Let $a_i(y \mid \boldsymbol{\pi}, \mathbf{T})$ be the ith element of the vector function $\mathbf{a}(y \mid \boldsymbol{\pi}, \mathbf{T})$, $b_i(y \mid \boldsymbol{\pi}, \mathbf{T})$ the ith element of the vector function $\mathbf{b}(y \mid \boldsymbol{\pi}, \mathbf{T})$ and so on, then the system can be written as

$$\mathbf{a}'(y \mid \boldsymbol{\pi}, \mathbf{T}) = \mathbf{a}(y \mid \boldsymbol{\pi}, \mathbf{T})\mathbf{T}$$

$$\mathbf{b}'(y \mid \mathbf{T}) = \mathbf{T}\mathbf{b}(y \mid \mathbf{T})$$

$$\mathbf{c}'(y; i \mid \boldsymbol{\pi}, \mathbf{T}) = \mathbf{T}\mathbf{c}(y; i \mid \boldsymbol{\pi}, \mathbf{T}) + a_i(y \mid \boldsymbol{\pi}, \mathbf{T})\mathbf{t} \quad i = 1, \ldots, p.$$

Combining these equations with the initial conditions $\mathbf{a}(0 \mid \boldsymbol{\pi}, \mathbf{T}) = \boldsymbol{\pi}$, $\mathbf{b}(0 \mid \mathbf{T}) = \mathbf{t}$, and $\mathbf{c}(0; i \mid \boldsymbol{\pi}, \mathbf{T}) = \mathbf{0}$ for $i = 1, \ldots, p$, we can solve the system numerically with high precision, using some standard method, see e.g. Moler & Van Loan (1978). In the EMPHT-program, the Runge–Kutta method of fourth order is implemented for this purpose.

An interesting property of the algorithm presented above is that the mean of the fitted phase-type distribution is the same as the mean of the sample (or the theoretical mean if the fit is to another distribution). This is not a general feature of the EM algorithm, but is true in our case because the observations are linear functions of the canonical sufficient statistics of the underlying exponential family, $y_v = \Sigma_{i=1}^{p} Z_i^{[v]}$. In every iteration of the E- and M-step, the new estimates of $(\boldsymbol{\pi}, \mathbf{T})$ are the solution of

$$\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k)}}\left[\sum_{v=1}^{n}\sum_{i=1}^{p} Z_i^{[v]} \mid \boldsymbol{y}\right] = \mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k+1)}}\left[\sum_{v=1}^{n}\sum_{i=1}^{p} Z_i^{[v]}\right].$$

The left side above equals $\Sigma_{v=1}^{n} y_v$, and the right side equals $n\mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k+1)}}[Y]$, and the result follows.

3.3. *Fitting continuous distributions*

The EM algorithm can, after minor adjustments, be used to fit a phase-type distribution to a theoretically given distribution. We let $f_{(\boldsymbol{\pi}, \mathbf{T})}$ be a density of a phase-type distribution and h the density of the given distribution. By fitting $f_{(\boldsymbol{\pi}, \mathbf{T})}$ to h we mean minimizing the information divergence (relative entropy or Kullback–Leibler information), which is equivalent to maximizing $\int \log(f_{(\boldsymbol{\pi}, \mathbf{T})}(y))h(y)\,dy$. This is a natural analogue to maximizing the log-likelihood function when we fit $f_{(\boldsymbol{\pi}, \mathbf{T})}$ to a sample $\mathbf{y}$, interpreting $1/n\ \Sigma_{v=1}^{n}$ as an integral w.r.t. the empirical distribution. Thus, we can also generate the EM algorithm. Details and further theoretical motivations are given in the Appendix for a general class of densities g_γ.

When we fit $f_{(\boldsymbol{\pi}, \mathbf{T})}$ to a density h, the E-step consists of calculating

$$\bar{B}_i^{(k+1)} = \int_0^\infty \mathbb{E}_{(\boldsymbol{\pi}, \mathbf{T})^{(k)}}[B_i \mid y]h(y)\,\mathrm{d}y \quad \text{for } i = 1, \ldots, p$$

and corresponding formulas for $\bar{Z}_i^{(k+1)}$ and $\bar{N}_{ij}^{(k+1)}$. The new estimates are calculated in the M-step:

$$\pi_i^{(k+1)} = \bar{B}_i^{(k+1)}, \quad t_{ij}^{(k+1)} = \frac{\bar{N}_{ij}^{(k+1)}}{\bar{Z}_i^{(k+1)}}, \quad t_i^{(k+1)} = \frac{\bar{N}_{i0}^{(k+1)}}{\bar{Z}_i^{(k+1)}}, \quad t_{ii}^{(t+1)} = -\left(t_i^{(k+1)} + \sum_{\substack{j=1 \\ j \neq i}}^{p} t_{ij}^{(k+1)}\right).$$

In the EMPHT-program the integrals in the E-step are approximated by a weighted sum over a finite grid: $\int_0^\infty \ldots h(y)\,dy = \Sigma_{v=1}^{n} \ldots w_v y_v$. Hence, the difference between fitting f to a sample and to a distribution is computationally very small. In fact, by assigning weights $w_v = 1$ to each observation, fitting to a sample becomes a special case of fitting to a distribution.

4. Special structures

One of the advantages of using the EM algorithm to estimate $(\boldsymbol{\pi}, \mathbf{T})$ is that it preserves the structures of zeros in $\boldsymbol{\pi}$ and $\mathbf{T}$. That is, once an element has been estimated to be zero, it will remain zero thereafter. This is easily seen in the formulas (3) of the conditional expectations. Probabilistically it means that the conditional probability of an impossible event remains equal to zero.

Hence, if one wants a phase-type fit within a special class having some elements fixed to zero, one needs only to start the EM algorithm with $(\boldsymbol{\pi}, \mathbf{T})^{(0)}$ in that class. The most common special classes, or substructures, are

(i) hyper exponential, i.e. a (finite) mixture of exponentials: the Markov process may start in any state and is absorbed without visiting any other state, i.e. $\boldsymbol{\pi}$ is arbitrary, while T is diagonal;
(ii) sum of exponentials (general Erlang): starts in state one, jumps only from state i to $i+1$, and is absorbed from state p;
(iii) Coxian: same as sum of exponentials except that absorption is allowed from any state.

In the EMPHT-program the user can choose among five different pre-specified structures (the three described above included). Also, it is possible to specify any other structure by assigning the initial values of $\boldsymbol{\pi}$ and $\mathbf{T}$, instead of using a random initialization.

The reasons for paying particular attention to such special structures is in part historical. One may note, though, that in applied probability algorithms for queueing or renewal theory, the complexity is determined by the number p of phases alone, and there is no simplification by assuming, say, a Coxian structure. In most of our experimental work, we found a Coxian distribution to provide almost as good a fit as a general phase-type distribution with the same p; for one exception, see the Erlang distribution with feedback in section 5.3. One advantage of special structure is that the fitting algorithm is faster for a given p, and for a given amount of allocated computer time one can thereby work with a larger number of phases and possibly obtain a better fit (for an example, see the geyser data in section 5.5).

As was pointed out in section 1, Cumani (1982) and Dehon & Latouche (1982) have shown that all phase-type distributions corresponding to acyclic distributions (that is a distribution whose generator is upper triangular), coincide with the Coxian distributions. Still, there may be reasons to consider general acyclic distributions when fitting a data-set using the EM algorithm, since the complete data models will not be the same in such a distribution as in a Coxian. Therefore the EM-steps will not develop in the same way, and the algorithm may end up in different distributions (if there exist local maxima or saddle points of the likelihood) depending on from which structure it was started.

Another possibility not exploited in this paper, is to have restrictions on the relation between elements within $\boldsymbol{\pi}$ and $\mathbf{T}$. For example, one can assume that all t_{ii} are equal in the sum of the exponential structure to derive an Erlang distribution. However, such restrictions require a modification of the M-step, which is not yet implemented in our computer programs.

5. Examples

To get some idea of how the algorithm works in our case, we have performed a sequence of illustrative examples, in which we try to illustrate graphically both the dynamics of the algorithm and how the resulting approximation works out. We start the series of examples with a sequence of fits to theoretical densities. A somewhat haphazard collection of phase-type orders have been tried on three different theoretical distributions: Weibull, log normal and an Erlang distribution with feedback. The performance of the algorithm is illustrated with plots of the densities and failure rates of the theoretical distributions together with their EM-approximations after various numbers of iterations. Some of the theoretical densities are chosen from a set of standards worked out by the participants in a workshop on phase-type fitting at Aalborg University in 1991 and used also in Bobbio & Telek (1994).

Table 1. *Time (in seconds) needed to perform 10 iterations*

Structure	p	Time	Structure	p	Time
General PH	2	0.3	Coxian	5	4
	5	8		10	40
	10	170		30	5100

The time it takes to perform one iteration of the EM algorithm when using the EMPHT-program depends on several factors; the main ones being the value of p, which structure is fitted, the size of the sample, and the step-length used in the Runge–Kutta function (which can be chosen by the user). In Table 1 we give some examples of the CPU-time (40 MHz Viking SPARC-processor) in seconds, needed to perform 10 iterations of the EM algorithm for fitting phase-type distributions of some different orders and structures. All these distributions have been fitted to the same sample of 100 observations, and the step-length in the Runge–Kutta function (in the EMPHT-program) has been set to its default value throughout.

The number of iterations needed to fit a phase-type distribution reasonably well to a sample or to another distribution, depend mostly on how many parameters there are to be estimated; the larger the order of the distribution is, the more iterations are needed, and a Coxian structure needs fewer iterations than a general phase-type structure of the same order. In the examples to follow, we have performed 1000–10 000 iterations to find reasonable fits. We have not used any strict criteria for deciding how many iterations should be done in the different examples, but stopped the fitting when the changes in the successive parameter estimates have become negligible.

5.1. Fits to a Weibull distribution

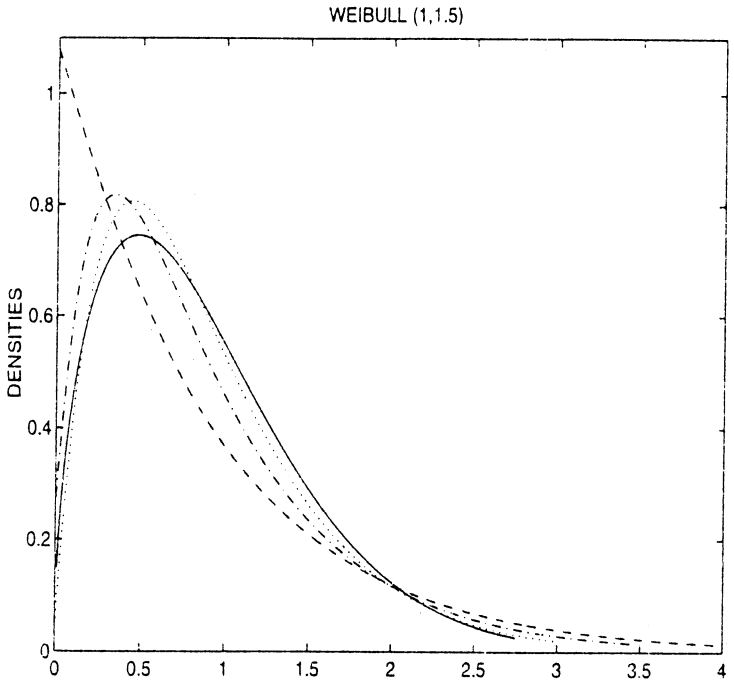
The chosen Weibull density has scale parameter equal to 1 and index equal to 1.5, i.e.

$$f(x) = 1.5x^{0.5} \exp\{-x^{1.5}\}.$$

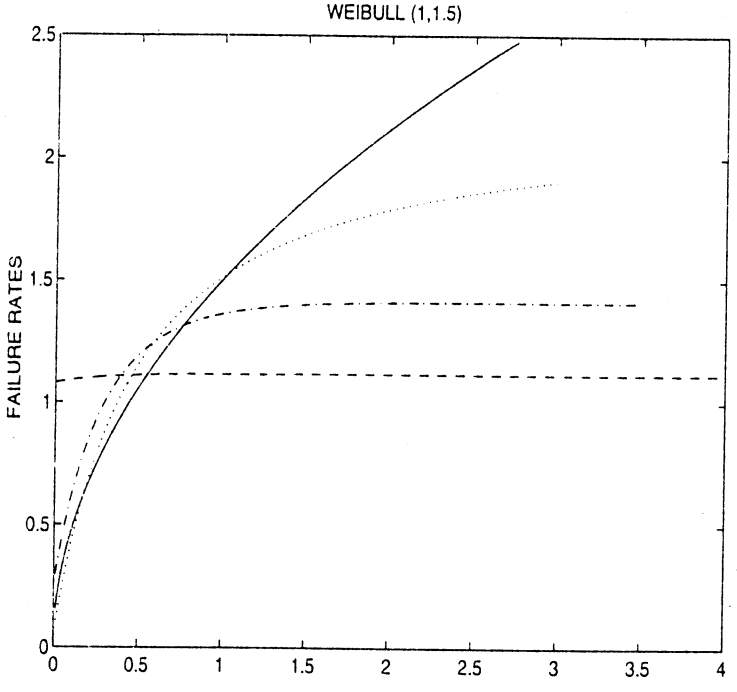
The first pair of figures show the successive fits of a phase-type distribution of order 2. In Fig. 2 we show the fits of phase-type distribution of order 6. We also illustrate (in Fig. 3) how the likelihood of the successive estimates grows toward the maximum likelihood value. The "likelihood" should be interpreted as an approximation of the integral of the logarithm of the fitted phase-type density times the theoretical density in question, (the last part of the information divergence, see section 6.2). By "maximum likelihood value" we mean this likelihood based on the last iteration of the EM algorithm.

5.2. Fits to two log normal distributions

We have used two different log normal distributions; the first with parameters $\mu = -0.32$, $\sigma^2 = 0.8$, the second with $\mu = -1.62$, $\sigma^2 = 1.8$. Both distributions have mean equal to 1, but the second has a standard deviation which is about 5 times the standard deviation of the first distribution. The second log normal distribution was very well approximated by a phase-type distribution of order 2, while a phase-type distribution of order 4 was required to get a reasonable fit of the first log normal distribution (see Fig. 4).



(a)



(b)

Fig. 1. Approximations of Weibull (1, 1.5). The density (a) and the failure rate (b) of a phase-type fit of order $p = 2$ after 1 ---, 25 - · -, and 1000 · · · iterations.

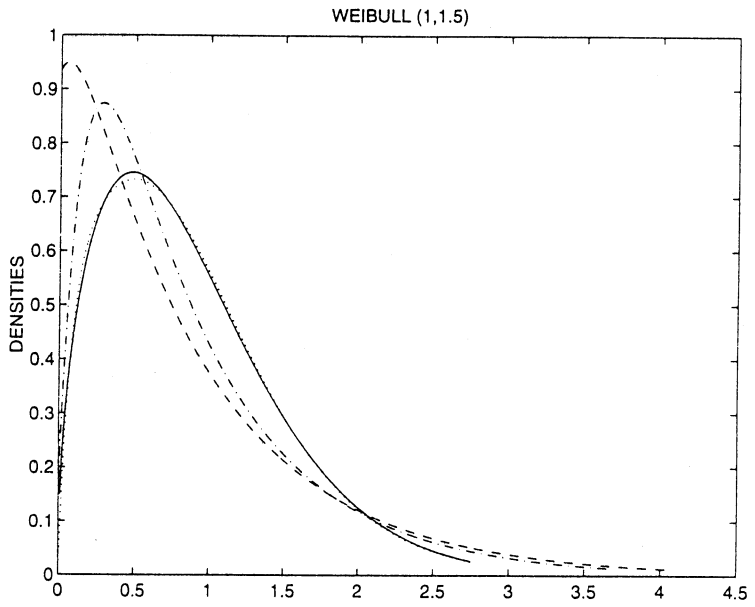


Fig. 2. Approximation of Weibull (1, 1.5). A phase-type fit of order $p = 6$ after 1 ---, 25 -·-, and 10 000 ··· iterations.

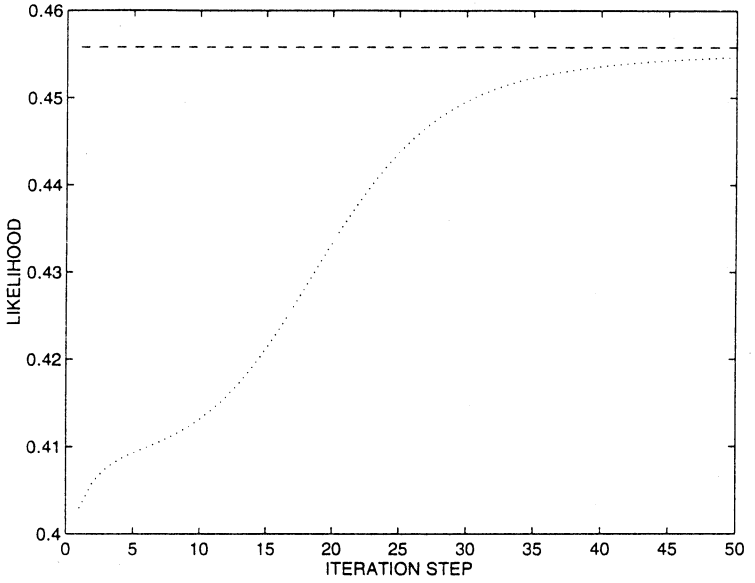


Fig. 3. Approximation of Weibull (1, 1.5). The dotted line is the value of the likelihood function of the phase-type fit of order 2, during the first 50 iterations. The dashed line shows the value of the likelihood function after 1000 iterations.

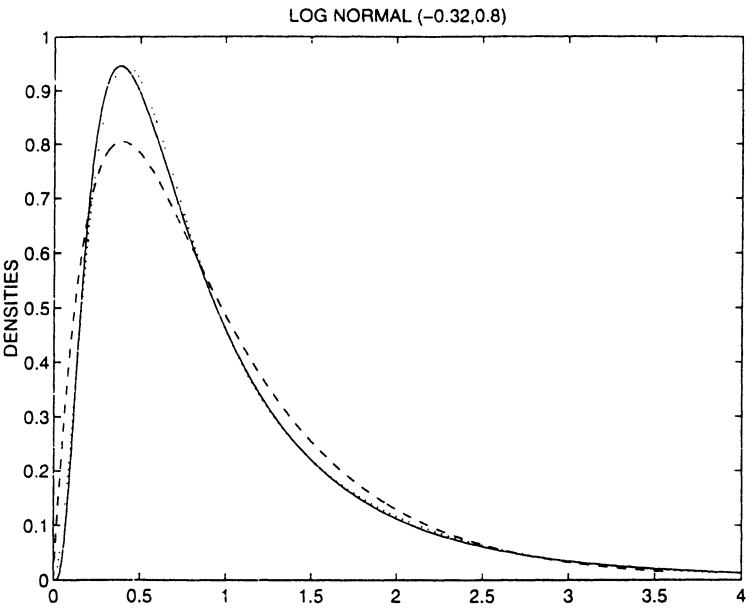


Fig. 4. Log normal (-0.32, 0.8) approximated by phase-type distributions of order $p = 2$ ---, and $p = 4$ · · ·. Both approximations are based on 2000 iterations.

5.3. Fits to an Erlang distribution with feedback

The chosen Erlang distribution with feedback has an underlying Markov process starting in state one, and from which it either jumps to the absorption state (with probability 0.4) or to state number two. From state two it can only jump to state three, thereafter to state four and so on. From the last state (number 15) it jumps to state one which gives it a so-called feedback structure.

This distribution is chosen to provide an example where it seems important to have a general phase-type structure rather than a Coxian one (it is also interesting because it exhibits wave phenomena). We have tried to approximate with both a general phase-type structure and an upper triangular structure. For lower order ($p = 5$ and 10), these fits are very poor. Therefore we tried to recover the distribution from a general phase-type distribution of the same order, and for comparison we also fitted an upper triangular phase-type structure as well as a Coxian structure of order 15. The Coxian fit was not started randomly, but initiated with parameters of decreasing values in order to try to utilize the result in Cumani (1982) and Dehon & Latouche (1982), (see section 2). Figure 5 shows clearly that the general phase-type structure succeeds much better than the upper triangular and Coxian structures to approximate this special distribution.

The reason why the general phase-type fit does not completely recover the given distribution is probably due to the fact that 10 000 iterations are not sufficient when p is as high as 15. However, the runs in large p-dimensions are very time-consuming. It might be possible to speed up the algorithm either by trying other solution methods for the differential equations in the EM step, or by using acceleration methods for the EM algorithm (see Louis, 1982; Meilijson, 1989; Jamshidian & Jennrich, 1993). We have made a first attempt to implement the method in Meilijson (1989), but so far it has not worked out well.

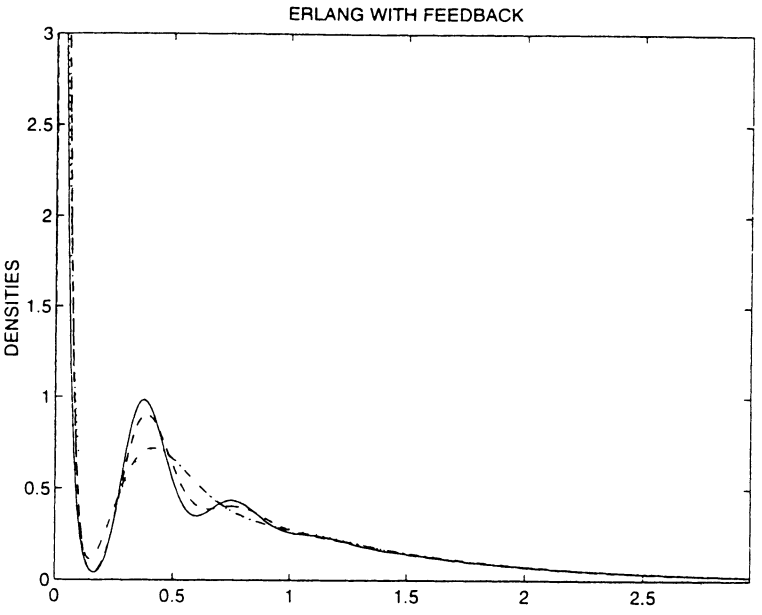


Fig. 5. Approximation of the Erlang distribution with feedback (solid line) by a general phase-type structure ---, an upper triangular structure –•–, and a Coxian structure · · ·. All fits are of order $p = 15$, and are based on 10 000 iterations.

5.4. *Fits to a uniform distribution*

The rate of convergence of the EM algorithm depends on the amount of missing information: the higher the order p of the phase-type distribution is, the slower the convergence rate becomes. This might be one of the reasons why we do not recover the theoretical density perfectly in the case of an Erlang distribution with feedback (see Fig. 5). Another might be that we get stuck in solutions to the likelihood equation which are local maxima or saddle points. To illustrate this phenomenon we show in Fig. 6 two fits, generated from different initial values $(\boldsymbol{\pi}, \mathbf{T})^{(0)}$, of phase-type distribution of order 10 to a uniform distribution on [0, 1]. A definite answer to which of the two phenomena is experienced would require a very large number of EM steps. However, after a quick look at Fig. 6 it seems that the local maxima hypothesis is the most plausible. Also, the difference between the maximum likelihood values is very small (log likelihood ratio ≈ 1.03).

5.5. *The geyser data-set*

Furthermore, we consider some samples. The first is a notoriously difficult example in density estimation which has been used in a number of papers, see Silverman (1986). This sample contains 107 observations of the eruption lengths (in minutes) of a famous geyser in Yellowstone National Park, USA. We fitted a general phase-type structure of order 15, (which after 10 000 iterations ended up in a Coxian structure), and a Coxian structure of order 30 (Fig. 7).

The main difficulty of finding a phase-type fit to this sample is caused by it first having a delay (the minimum observation is 1.67) and then starting off steeply. In general, it is hard to induce rapid changes of the failure rate, and it requires very high p-dimensions and a lot of "fast" states. This is especially so if the changes take place at late time points. Thus, the

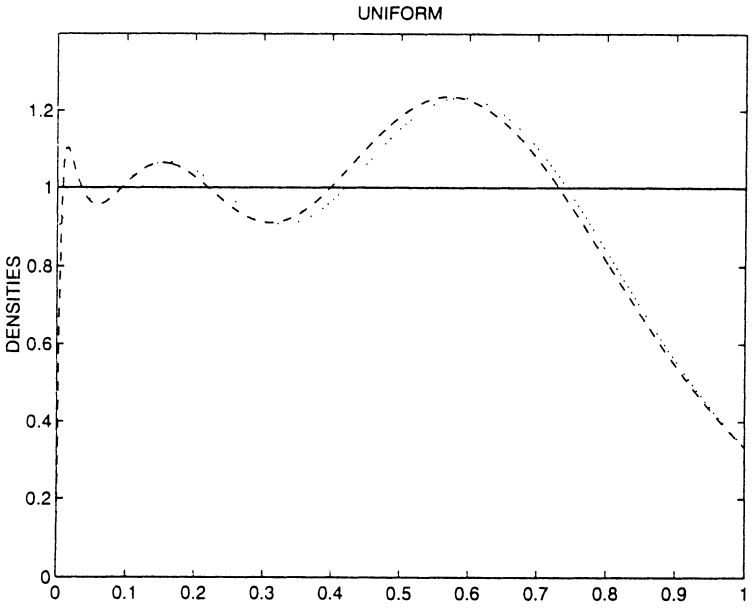


Fig. 6. Two phase-type approximations of a uniform distribution, started with different (random) initial parameters. Both fits are of order $p = 10$, and are based on 10 000 iterations.

geyser data (as well as the uniform density above) show clearly that not all positive distributions are easily approximated by phase-type distributions of moderate order.

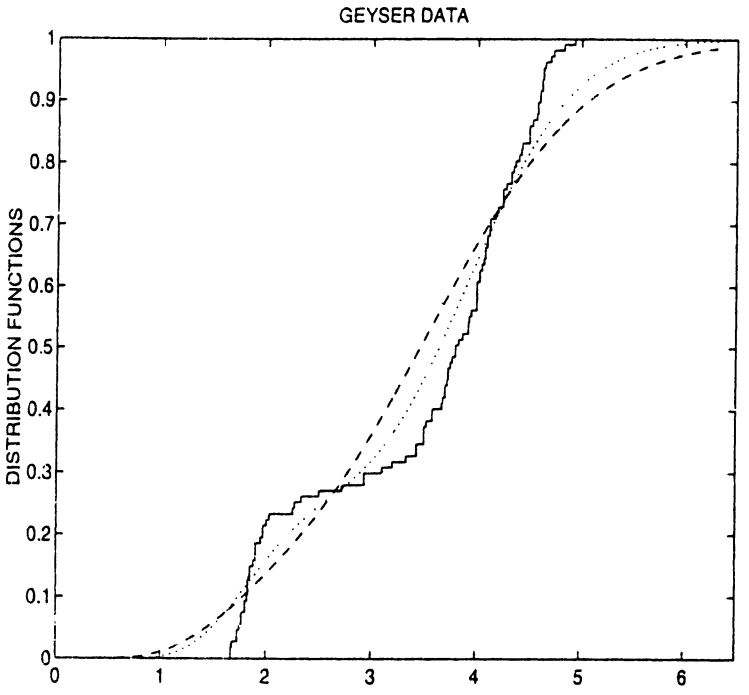
5.6. Four samples of the length of telephone calls

The data file underlying Figs 8–10 was kindly supplied by Professor O. Kella, Hebrew University, Jerusalem, and Professor A. Mandelbaum, Technion, Haifa. In the file, lengths of incoming telephone calls to the service centre of one of Israel's major television cable companies are recorded and the calls are classified into types 0–10. We took the four types, 1, 3, 4, and 7, of incoming calls having the largest number N of observations. The types have the following meaning ($\bar{X}$ is the empirical mean in minutes):

(i) type 1: "home services", receiving notices from subscribers on problems, and transferring the information to technicians, here $\bar{X} = 2.69$, $N = 2039$;
(ii) type 3: "sales", notices on sales actions, including seeking help on prices, times, clarifications with sales people, etc, here $\bar{X} = 2.40$, $N = 472$;
(iii) type 4: "billings", providing information to customers on payments' procedures, here $\bar{X} = 3.18$, $N = 904$;
(iv) type 7: "general information", including change of address, private calls, here $\bar{X} = 2.15$, $N = 3189$.

To all four samples we have fitted both a general phase-type structure and a Coxian structure. For all samples but one (type 3) it has not been possible to distinguish the fitted Coxian density from the fitted general phase-type density in the graphs, (even though the estimates of $\boldsymbol{\pi}$ and $\mathbf{T}$ are very different in all fits). The unit of scale on the x-axis in Figs 8–10 is minutes.

(a)



(b)

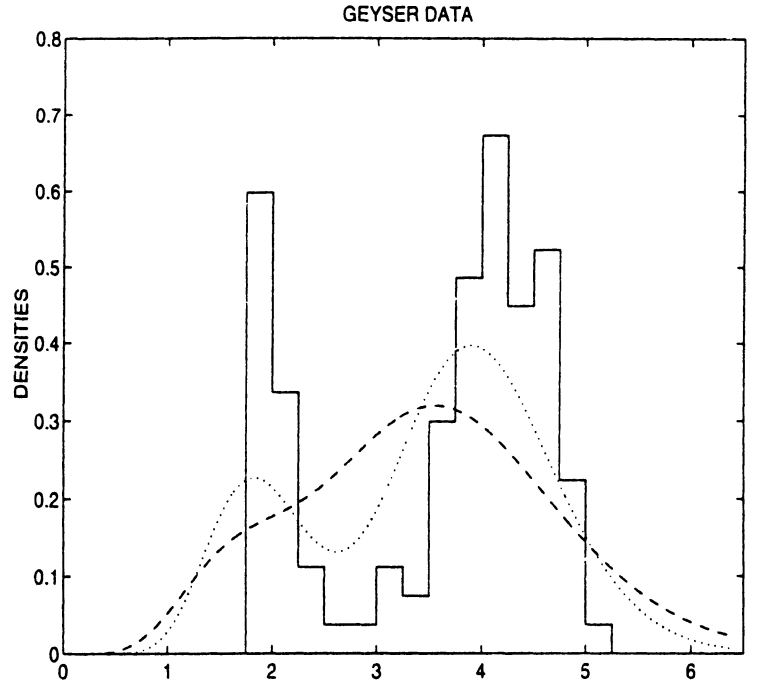


Fig. 7. Phase-type fits to the geyser data by a general structure of order $p = 15$ --- (based on 10 000 iterations), and a Coxian structure of order $p = 30$ · · · (based on 3000 iterations). In (a) the solid line is the empirical distribution function, and in (b) a histogram of the relative frequencies is given as a comparison to the fitted densities.

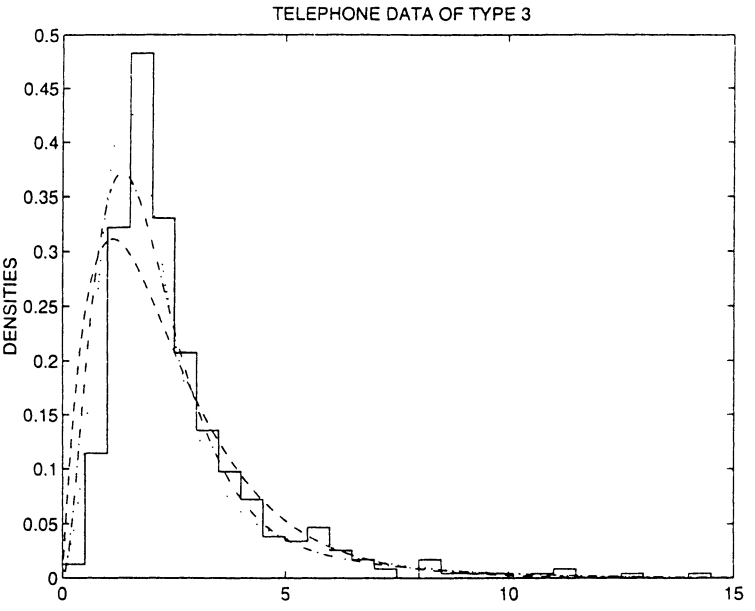


Fig. 8. Fits to the telephone data of type 3. Phase-type fits of order $p = 2$ ---, $p = 4$ - · -, and $p = 6$ · · ·.

When fitting phase-type distributions of order 3 and 4 to the telephone data of type 3, we discovered that the general phase-type structure seemed to converge to a structure with feedback. In these cases the general phase-type approximations gave better fits (according to the log likelihood) than the Coxian structure, although the difference is hard to see in plots of the densities. For the approximations of order 6, the fits of the general phase-type and the

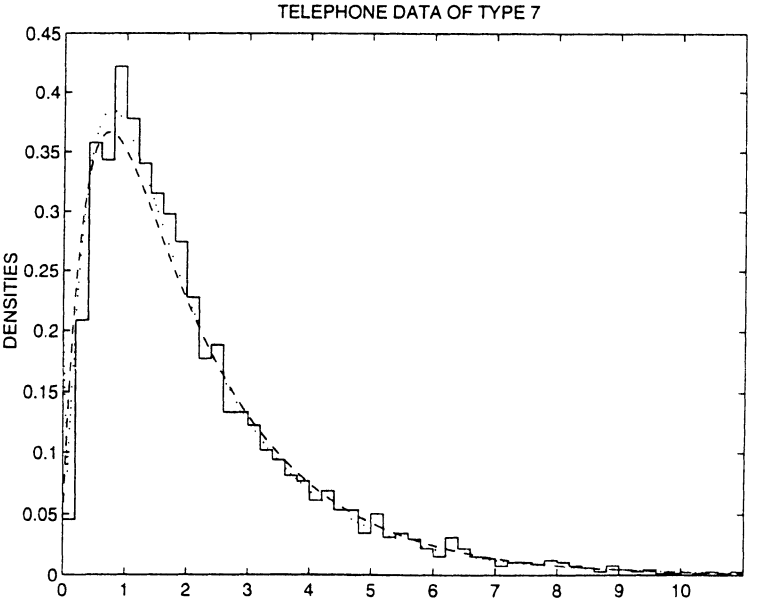


Fig. 9. Phase-type fits of order $p = 2$ ---, and $p = 4$ · · ·, to the telephone data of type 7.

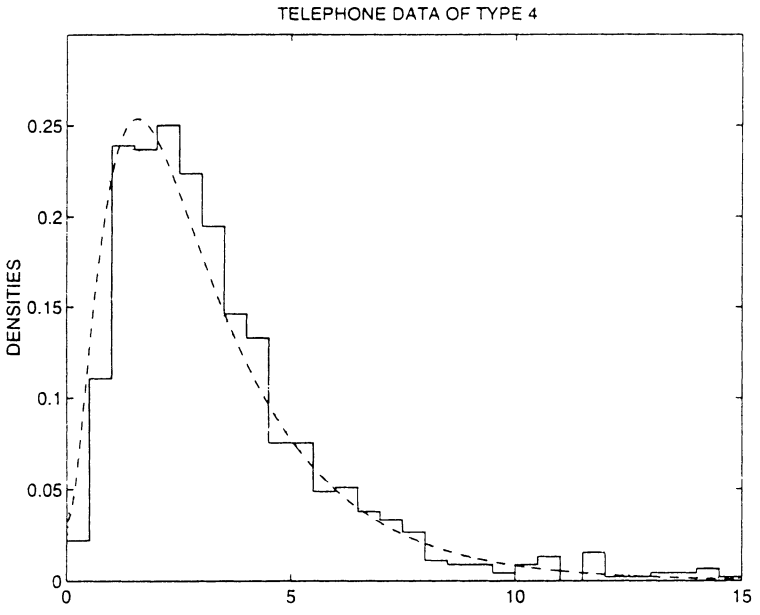


Fig. 10. Fits to the telephone data of type 4. Comparison of two fitted Coxian distributions (both of order $p = 3$), started with different initial values of the parameters.

Coxian structures gave the same log likelihood value. Of course, the Coxian structure has the advantage of being much faster to fit.

Another example of the algorithm converging to different local maxima (probably) when started from different initial values of the parameters, is shown in Fig. 10. Here a Coxian structure of order 3 has been fitted; in both cases using 5000 iterations.

5.7. Some final remarks

A theoretical property of a phase-type distribution is that it always has an Erlang-like tail:

$$F(u) = 1 - \boldsymbol{\pi} e^{\mathbf{T}u}\mathbf{e} \sim cu^{k-1}e^{\rho u}$$

where ρ is the (real) eigenvalue of $\mathbf{T}$ with the largest real part. This tail behaviour of course implies that the failure rate $r(u)$ converges to a constant, as $u \to \infty$, a fact that can be seen in our failure rate plots. Probabilistically one can think of the Markov process, conditioned on non-absorption, approaching a quasi-stationary distribution, which makes the failure rate approach the corresponding weighted sum of state dependent instantaneous failure rates.

A deeper discussion of the theoretical properties of the maximum likelihood estimates (which we hope to find via the EM algorithm) has been postponed for several reasons. One is that due to the over-parameterization the situation is somewhat non-standard, although usual asymptotic distribution properties concerning estimable quantities such as mean, median, and other quantiles, as well as distribution-, density- and failure rate functions, should be derivable from knowledge of the existence of a sufficiently regular unique parameterization. (Candidates for such a parameterization are either the zeros and poles of the Laplace transform, or maybe a sequence of moments, see section 2). Another reason is that asymptotic theory tells us nothing when we fit theoretical distributions. Also, the relevant asymptotic is quite hard to derive when the phase-type assumption is only an

approximation and not a model assumption. For a recent discussion of the latter topic see Hjort (1992).

Acknowledgement

We would like to thank Mogens Bladt for performing the phase-type fits to the theoretical densities shown in section 5.

References

Aalen, O. O. (1993). Phase-type distributions: computer algebra and a simple mixing model. Manuscript, University of Oslo.

Aalen, O. O. (1995). On phase type distributions in survival analysis. *Scand. J. Statist.* **22,** 447–463.

Alberg, A. (1961). Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Ann. Math. Statist.* **38,** 727–753.

Asmussen, S. (1992). Phase-type representations in random walls and queueing problems. *Ann. Probab.* **20,** 772–789.

Asmussen, S. & Bladt, M. (1996). Renewal theory and queueing algorithms for matrix-exponential distributions. In *Matrix-analytic methods in stochastic models* (eds A. S. Alfa & S. Chakravarty), Marcel Dekker, New York (to appear).

Asmussen, S. & Nerman, O. (1991). Fitting phase-type distributions via the EM algorithm. *Symposium i Anvendt Statistik, Copenhagen, January 21–23, 1991* (ed. K. Vest Nielsen), 335–346. UNI-C, Copenhagen.

Asmussen, S. & Rolski, T. (1991). Computational methods in risk theory: a matrix-algorithmic approach. *Insurance: Math. Econom.* **10,** 259–274.

Basawa, I. V. & Rao, B. L. S. (1980). *Statistical inference for stochastic processes.* Academic Press, London.

Baum, L. E., Pertrif, T., Soules, G. & Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41,** 164–171.

Bobbio, A., Cumani, A., Premoli A. & Saracco (1980). Modelling and identification of non-exponential distributions by homogeneous Markov processes. *Proc. 6th Adv. Reliab. Symp., Bradford,* pp. 373–392.

Bobbio, A. & Cumani, A. (1990). ML estimation fo the parameters of a PH distributions in triangular canonical form. In *Computer performance evaluation* (eds G. Balbo & G. Serazzi), 33–46, Elsevier, Amsterdam.

Bobbio, A. & Telek, M. (1994). A benchmark for PH estimation algorithms: results for acyclic PH. *Commun. Statist. Stochastic Models* **10,** 661–667.

Bux, W. & Herzog, U. (1977). The phase concept: approximation of measured data and performance analysis. In *Computer performance* (eds K. M. Chandy & M. Reiser), 23–38. North-Holland, Amsterdam.

Cox, D. R., (1953). A use of complex probabilities in the theory of stochastic processes. *Proc. Camb. Philos. Soc.* **51,** 313–319.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectron. Reliab.* **22,** 583–602.

Dehon, M. & Latouch, G. (1982). A geometric interpretation of the relations between the exponential and the generalized Erlang distributions. *Adv. Appl. Probab.* **14,** 885–897.

Dempster, A. P., Laird, N. M. & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B.* **39,** 1–38.

Faddy, M. J. (1990). Compartmental models with phase-type residence time distributions. *Appl. Stochastic Models Data Anal.* **6,** 121–127.

Faddy, M. J. (1993). A structured compartmental model for drug kinetics. *Biometrics* **49,** 243–248.

Faddy, M. J. (1994). Examples of fitting structured phase type distributions. *Appl. Stochastic Models Data Anal.* **10,** 247–256.

Harris, C. M, & Sykes, E. A. (1987). Likelihood estimation for generalized mixed exponential distributions. *Naval Res. Logist. Quart.* **34,** 251–279.

Hjort, N. L. (1992). On inference in parametric survival data models. *Int. Statist. Rev.* **60,** 355–387.

Hoem, J. M. (1969). Purged and partial Markov chains. *Skand. Aktuarietidskr.* **56,** 147–155.

Häggström, O., Asmussen, S. & Nerman, O. (1992). EMPHT—A program for fitting phase-type distributions. Technical report, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden.

Jamshidian, M. & Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *J. Amer. Statist. Assoc.* **88,** 221–228.

Johnson, M. A., (1990). Selecting parameters of phase distributions: combining nonlinear programming, heuristics and Erlang distributions. *Technical Report.*

Johnson, M. A. & Taaffe, M. R. (1989). Matching moments to phase distributions: mixtures of Erlang distributions of common order. *Commun. Statist. Stochastic Models* **5,** 711–743.

Johnson, M. A. & Taaffe, M. R., (1990a). Matching moments to phase distributions: nonlinear programming approaches. *Commun. Statist. Stochastic Models* **6,** 259–281.

Johnson, M. A. & Taaffe, M. R. (1990b). Matching moments to phase distributions: density function shapes. *Commun. Statist. Stochastic Models* **6,** 283–306.

Jonsson, E., Andersson, M. & Asmussen, S. (1994). A practical dependability measure for degradable computer systems with non-exponential degradation. To appear in the proceedings of Safeprocess '94, Helsinki, June 1994.

Kao, E. P. C. (1988). Computing the phase-type renewal and related functions. *Technometrics* **30,** 87–93.

Kullback, S. (1978). *Information theory and statistics.* Peter Smith, Gloucester, MA.

Lang, A. & Arthur J. L. (1994). Parameter approximation for phase-type distributions. Technical Report, Oregon State University, Corvallis.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov chains. *Stochastic Process. Applic.* **40,** 127–143.

Lipsky, L. (1992). *Queueing theory—a linear algebraic approach.* Macmillan, New York.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44,** 226–233.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. Ser. B* **51,** 127–138.

Moler, C. & Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.* **20,** 801–836.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models.* Johns Hopkins University Press, Baltimore, MD.

Ó Cinnéide, C. A. (1987). On non-uniqueness of representations of phase-type distributions. *Commun. Statist. Stochastic Models* **5,** 247–259.

Olsson, M. (1996). Estimation of phase type distributions from censored data. *Scand. J. Statist.* **23,** 443–460.

Orchard, T. & Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proc. 6th Berkeley Symposium on Math. Statist. and Probab.* **1,** 697–715.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and EM algorithm. *SIAM Rev.* **26,** 195–237.

Ruhe, A. (1980). Fitting empirical data by positive sums of exponentials. *Siam. J. Sci. Statist. Comput.* **1,** 481–498.

Ryden, T. (1993). Parameter estimation for Markov modulated Poisson processes and overload control of spc switches. PhD thesis, Department of Mathematical Statistics, Lund Institute of Technology, Sweden.

Sengupta, B. (1989). Markov processes whose steady-state distribution is matrix-exponential with an application to the GI/G/1 queue. *Adv. Appl. Probab.* **21,** 159–180.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall, New York.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1,** 49–58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Commun. Statist. Simulation Computation* **B5,** 55–64.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11,** 95–103.

Received June 1994, in final form October 1995

Olle Nerman, Department of Mathematics, University of Göteborg, S-412 96 Göteborg, Sweden.

Appendix

A.1. Derivation of the conditional expectation

We shall motivate the conditional expectations of the three groups of random variables: B_i, Z_i and N_{ij} $i = 1, \ldots, p$, $j = 0, 1, \ldots, p$, used in section 3.2. To simplify the notation we assume that $n = 1$ and consider a single Markov process J_u corresponding to the phase-type parameters $(\boldsymbol{\pi}, \mathbf{T})$ with absorption time Y.

It is elementary to derive the conditional expectations of the initial state indicators

$$\begin{aligned}
\mathbb{E}[B_i \mid Y = y] &= \frac{\mathbb{P}(J_0 = i,\ Y \in dy)}{\mathbb{P}(Y \in dy)} \\
&= \frac{\mathbb{P}(J_0 = i)\mathbb{P}(Y \in dy \mid J_0 = i)}{\mathbb{P}(Y \in dy)} \\
&= \frac{\pi_i \mathbf{e}_i' \exp\{\mathbf{T}y\}\mathbf{t}}{\boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{t}} \\
&= \frac{\pi_i b_i(y \mid \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y \mid \mathbf{T})} \quad i = 1, \ldots, p.
\end{aligned}$$

(See section 3.2 for the definition of **a**, **b** and **c**.)

Almost as simple is the derivation of the conditional expectations of the occupation times

$$\begin{aligned}
\mathbb{E}[Z_i \mid Y = y] &= \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{J_u = i\}}\, du \mid Y = y\right] \\
&= \int_0^\infty \mathbb{P}(J_u = i \mid Y = y)\, du \\
&= \int_0^\infty \frac{\mathbb{P}(J_u = i,\ Y \in dy)}{\mathbb{P}(Y \in dy)}\, du \\
&= \frac{\int_0^y \mathbb{P}(J_u = i)\mathbb{P}(Y \in dy \mid J_u = i)\, du}{\mathbb{P}(Y \in dy)} \\
&= \frac{\int_0^y \boldsymbol{\pi} \exp\{\mathbf{T}u\}\mathbf{e}_i\mathbf{e}_i' \exp\{\mathbf{T}(y-u)\}\mathbf{t}\, du}{\boldsymbol{\pi} \exp\{\mathbf{T}y\}\mathbf{t}} \\
&= \frac{c_i(y; i \mid \boldsymbol{\pi}, \mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y \mid \mathbf{T})} \quad i = 1, \ldots, p.
\end{aligned}$$

Here the exchange of the order of integration and conditional expectation is motivated by positivity of the integrand. (The fact that $J_u = 0$ for $u > y$ motivates the change of the upper integration bound from ∞ to y in the third equality).

To derive the (intuitively natural) formula for the conditional expectation of the number of jumps between two non-absorbing states with reasonable rigour, is slightly more complicated. First observe that the expectation of the total number of jumps $\mathbb{E}[\Sigma_{i \neq j} N_{ij}]$ is finite (a fact which follows from straightforward arguments on the level of the embedded jump chain). Second, observe that the discrete approximations of N_{ij}

$$N_{ij}^\varepsilon = \sum_{k=0}^{\infty} \mathbf{1}_{\{J_{k\varepsilon} = i,\ J_{(k+1)\varepsilon} = j\}} \quad \varepsilon > 0,\ i \neq j$$

are all dominated by $\Sigma_{i\neq j} N_{ij}$ and converge to N_{ij} as $\varepsilon \downarrow 0$. Furthermore

$$\begin{aligned}\mathbb{E}[N_{ij}^{\varepsilon} \mid Y=y] &= \sum_{k=0}^{[y/\varepsilon]-1} \frac{\mathbb{P}(J_{k\varepsilon}=i, J_{(k+1)\varepsilon}=j,\ Y\in dy)}{\mathbb{P}(Y\in dy)} \\ &= \sum_{k=0}^{[y/\varepsilon]-1} \frac{\mathbb{P}(J_{k\varepsilon}=i)\mathbb{P}(J_{(k+1)\varepsilon}=j \mid J_{k\varepsilon}=i)\mathbb{P}(Y\in dy \mid J_{(k+1)\varepsilon}=j)}{\mathbb{P}(Y\in dy)} \\ &= \frac{\sum_{k=0}^{[y/\varepsilon]-1} (\boldsymbol{\pi}\exp\{\mathbf{T}k\varepsilon\}\mathbf{e}_i)(\mathbf{e}_i'\exp\{\mathbf{T}\varepsilon\}\mathbf{e}_j)(\mathbf{e}_j'\exp\{\mathbf{T}(y-(k+1)\varepsilon)\}\mathbf{t})}{\boldsymbol{\pi}\exp\{\mathbf{T}y\}t} \\ &\to \frac{\int_0^y \boldsymbol{\pi}\exp\{\mathbf{T}u\}\mathbf{e}_i t_{ij}\mathbf{e}_j'\exp\{\mathbf{T}(y-u)\}\mathbf{t}\,du}{\boldsymbol{\pi}\exp\{\mathbf{T}y\}\mathbf{t}} \quad \text{as } \varepsilon\downarrow 0, \text{ for } i,j=1,\ldots,p,\ i\neq j,\end{aligned}$$

follows from the continuity of $\exp\{\mathbf{T}u\}$ and the fact that

$$\frac{\exp\{\mathbf{T}\varepsilon\}-I}{\varepsilon}\to\mathbf{T} \quad \text{as } \varepsilon\downarrow 0.$$

Now dominated convergence (for conditional expectations) yields

$$\begin{aligned}\mathbb{E}[N_{ij} \mid Y=y] &= \frac{\int_0^y \boldsymbol{\pi}\exp\{\mathbf{T}u\}\mathbf{e}_i t_{ij}\mathbf{e}_j'\exp\{\mathbf{T}(y-u)\}\mathbf{t}\,du}{\boldsymbol{\pi}\exp\{\mathbf{T}y\}\mathbf{t}} \\ &= \frac{t_{ij}c_j(y;\, i\mid \boldsymbol{\pi},\mathbf{T})}{\boldsymbol{\pi}\mathbf{b}(y\mid\mathbf{T})} \quad i,j=1,\ldots,p,\ i\neq j.\end{aligned}$$

Finally, the conditional expectations of the number of jumps from the non-absorbing states to 0 can be interpreted as the conditional probability that the final absorbing jump at time y came from state i. Again we can make an ε-argument:

$$\begin{aligned}\mathbb{P}(J_{Y-\varepsilon}=i \mid Y=y) &= \frac{\mathbb{P}(J_{y-\varepsilon}=i)\mathbb{P}(Y\in dy \mid J_{y-\varepsilon}=i)}{\mathbb{P}(Y\in dy)} \\ &= \frac{\boldsymbol{\pi}\exp\{\mathbf{T}(y-\varepsilon)\}\mathbf{e}_i\mathbf{e}_i'\exp\{\mathbf{T}\varepsilon\}\mathbf{t}}{\boldsymbol{\pi}\exp\{\mathbf{T}y\}t} \quad i=1,\ldots,p,\quad y>\varepsilon>0.\end{aligned}$$

As $\varepsilon\downarrow 0$ this relation becomes

$$\begin{aligned}\mathbb{E}[N_{i0} \mid Y=y] &= \frac{\boldsymbol{\pi}\exp\{\mathbf{T}y\}\mathbf{e}_i t_i}{\boldsymbol{\pi}\exp\{\mathbf{T}y\}\mathbf{t}} \\ &= \frac{a_i(y\mid\boldsymbol{\pi},\mathbf{T})t_i}{\boldsymbol{\pi}\mathbf{b}(y\mid\mathbf{T})} \quad i=1,\ldots,p.\end{aligned}$$

A.2. EM minimization of information divergence

The information divergence (Kullback–Leibler information or relative entropy) of the probability density f with respect to the probability density h is defined as Kullback (1978),

$$\begin{aligned}I(f,h) &= \int \log\frac{h(x)}{f(x)}h(x)\mu(dx) \\ &= \int \log(h(x))h(x)\mu(dx) - \int \log(f(x))h(x)\mu(dx),\end{aligned}$$

where both densities are assumed to be with respect to the measure μ. From Jensen's

inequality we get that $I(f, h) \geq 0$, with equality iff $f = h$ μ – almost everywhere. It is also easy to see that I does not depend on the choice of the μ-measure.

To find $\hat{f}$ that minimizes this I-divergence over some class of densities for a given h can naturally be thought of as a maximum likelihood problem of infinite sample size.

Now consider an I-divergence minimization where we wish to fit to h a density $g_\gamma(y)$ which can be thought of as the density of $Y = u(X)$, a partial observation of a random variable (or vector) X with density $f_\gamma(x)$, say. The density of Y is supposed to be with respect to $\nu = \mu u^{-1}$. Denote the conditional density of X given $Y = y$ by $k_\gamma(x \mid y)$ and define the density $h_\gamma(x)$ by

$$h_\gamma(x) = h(u(x))k_\gamma(x \mid u(x)).$$

Then it is straightforward to see that

$$I(g_\gamma, h) = I(f_\gamma, h_\gamma).$$

Also, the basic additivity property of information divergence as a sum of the marginal and expected conditional information divergence (sometimes called the chain rule for relative entropy), together with the non-negativity of the I-divergence yields

$$I(f_\gamma, h'_\gamma) \geq I(f_\gamma, h_\gamma).$$

This shows that if γ_1 minimizes $\gamma \to I(f_\gamma, h_{\gamma 0})$ then

$$I(g_{\gamma 1}, h) = I(f_{\gamma 1}, h_{\gamma 1}) \leq I(f_{\gamma 1}, h_{\gamma 0}) \leq I(f_{\gamma 0}, h_{\gamma 0}) = I(g_{\gamma 0}, h).$$

We can characterize γ_1 as the value of γ that maximizes

$$L(\gamma_0, \gamma) = \int \log (f_\gamma(x))h_{\gamma 0}(x)\mu(dx)$$

$$= \int \mathbb{E}_{\gamma 0}[\log f_\gamma(X) \mid u(X) = y]h(y)\nu(dy).$$

This motivates an algorithm for the minimization which is completely analogous to the EM algorithm. In the E-step we calculate

$$L(\gamma_n, \gamma) = \int \mathbb{E}_{\gamma_n}[\log f_\gamma(X) \mid u(X) = y]h(y)\nu(dy)$$

and in the M-step we find γ_{n+1} that maximizes $\gamma \to L(\gamma_n, \gamma)$.

Certainly many of the properties and problems associated with the ordinary EM algorithm carry over, including problems with convergence to local minima or saddle points.

Now suppose f_γ belongs to a (possibly curved) multi-dimensional exponential family with density

$$f_\gamma(x) = \exp \{\theta(\gamma)'\mathbf{S}(x) + d(\theta(\gamma))\}.$$

Then the E-step gives

$$L(\gamma_n, \gamma) = \theta(\gamma)'\bar{\mathbf{S}}_n + d(\theta(\gamma)),$$

where

$$\bar{\mathbf{S}}_n = \int \mathbb{E}\gamma_n[\mathbf{S}(X) \mid u(\mathbf{X}) = y]h(y)\nu(dy).$$

In the M-step we must find γ_{n+1} that maximizes

$$\gamma \to L(\gamma_n, \gamma) = \theta(\gamma)'\bar{\mathbf{S}}_n + d(\theta(\gamma)),$$

just as if $\bar{\mathbf{S}}_n$ was a sample average and we tried to find a maximum likelihood estimate.