Maximum likelihood estimation of phase-type distributions

Esparza, Luz Judith R; Nielsen, Bo Friis; Bladt, Mogens

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Maximum likelihood estimation of phase-type distributions

Luz Judith Rodriguez Esparza

Kongens Lyngby 2010
IMM-PHD-2010-245

DTU Informatics
Departament of Informatics and Mathematical Modeling
Technical University of Denmark

Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This work is concerned with the statistical inference of phase-type distributions and the analysis of distributions with rational Laplace transform, known as matrix-exponential distributions.

The thesis is focused on the estimation of the maximum likelihood parameters of phase-type distributions for both univariate and multivariate cases. Methods like the EM algorithm and Markov chain Monte Carlo are applied for this purpose.

Furthermore, this thesis provides explicit formulae for computing the Fisher information matrix for discrete and continuous phase-type distributions, which is needed to find confidence regions for their estimated parameters.

Finally, a new general class of distributions, called bilateral matrix-exponential distributions, is defined. These distributions have the entire real line as domain and can be used, for instance, for modelling. In addition, this class of distributions represents a generalization of the class of matrix-exponential distributions.

Resumé

Denne afhandling omhandler primært statistisk analsye af fase-type fordelinger.

Der fokuseres på estimation af parametre ved brug af maximum likelihood princippet. Både det univariate og det multivariate tilfælde behandles. Der er anvendt metoder som EM algoritmen og Markov chain Monte Carlo simulering.

Ydermere gives der formler for at beregne Fisher informationsmatrix for diskrete og kontinuerte fase-type fordelinger; denne er nødvendig for at beregne konfidensintervaller for de estimerede parametre.

Til slut introduceres en general klasse af fordelinger, der kan anvendes som modelleringsværktøj, i de tilfælde hvor den multivariate Gaussiske fordeling ikke er tilstrækkelig. Denne klasse benævnes bilaterale matrixeksponentielle fordelinger, og den har som definitionsmængde hele den reelle talakse, og repræsentersåledes en generalisering af matrixeksponentielle fordelinger.

Preface

This thesis was submitted at the Technical University of Denmark, Department of Informatics and Mathematical Modelling, in partial fulfillment of the requirements for acquiring the PhD. degree in engineering.

The thesis deals with different aspects of mathematical modelling of matrix-analytic methods. Particularly on the study of matrix-exponential distributions and phase-type distributions with special emphasis on the latter.

The PhD. project has been supervised by Associate Professor Bo Friis Nielsen and co-supervised by Professor Mogens Bladt, researcher at UNAM (Department of Statistics at the Institute for Applied Mathematics and Systems).

The thesis consists in a summary report and two research papers written during the period 2007-2010.

Lyngby, November 2010

Luz Judith Rodriguez Esparza

Papers included in the thesis

[A] Mogens Bladt, Luz Judith R. Esparza, Bo Friis Nielsen. Fisher Information and statistical inference for phase-type distributions. *Journal of Applied Probability.* Accepted, 2011.

[B] Mogens Bladt, Luz Judith R. Esparza, Bo Friis Nielsen. Bilateral matrix-exponential distributions. *Stochastic models.* Summited, 2011.

Acknowledgements

I would like to start by thanking God for being with me at every moment, for giving me the strength and the will to succeed, for being my support and my sole purpose in life.

Thanks to my supervisors Bo Friis Nielsen and Mogens Bladt. Thanks for their patience, dedication, knowledge, for their great support and assistance. I could not have taken this project forward without them.

Special thanks to DTU and MT-LAB for providing me with financial support.

Many thanks to my colleagues and friends. Thanks for supporting me, for their unconditional friendship, for their advice, and for putting up with me all this time.

I would like to thank my family, especially my nieces and nephews, they are the light of my life.

Abbreviations

AIC	Akaike information criterion
APH	Acyclic phase-type
ADPH	Acyclic discrete phase-type
BME	Bilateral matrix-exponential
BPH	Bilateral phase-type
CF	Canonical form
CDF	Cumulative distribution function
CPH	Continuous phase-type
CTMC	Continuous time Markov chain
DM	Direct method
DMC	Direct method canonical
DPH	Discrete phase-type
EM	Expectation-Maximization
EMC	Expectation-Maximization canonical
FI	Fisher information
GS	Gibbs sampler
GSC	Gibbs sampler canonical
LL	Log-likelihood
ME	Matrix-exponential
MG	Moment generating
MH	Metropolis-Hastings
MJP	Markov jump process
MLE	Maximum likelihood estimator
MPH	Multivariate phase-type
MBPH	Multivariate bilateral phase-type
MCMC	Markov chain Monte Carlo

MVME	Multivariate matrix-exponential
MVBME	Multivariate bilateral matrix-exponential
NR	Newton-Raphson
PH	Phase-type
PDF	Probability density function
RK	Runge Kutta
SD	Standard deviation

Contents

Chapter 1

Introduction

Although phase-type distributions can be traced back to the pioneering work of Erlang [29] and Jensen [36], it was not until the late seventies that Marcel F. Neuts and his co-workers established much of the modern theory ([45], [46], [47]). Most of the original applications of phase-type distributions were in the area of queueing theory (see also [4], [5], [38], [40]), still phase-type distributions have proved useful also in risk theory as we can see in the work of Asmussen [9].

Statistical inference for phase-type distributions is of more recent date, where the likelihood estimation was first proposed by Asmussen *et.al* [11] (see also [8]) using an expectation-maximization (EM) algorithm. In a companion paper Olsson [54] extended the algorithm using censored data. Moreover, a Markov chain Monte Carlo (MCMC) based approach was suggested by Bladt *et.al* [15] and later it was used by Fearnhead and Sherlock [30]. Bobbio and Telek [22] presented a maximum likelihood estimation procedure for the canonical representation of acyclic phase-type distributions (see also [19]). While Hovarth and Telek [35] presented a tool (PhFit) that allows the approximation of distributions or set of samples by phase-type distributions. Since most of the previously phase-type fitting methods were designed for fitting over the continuous phase-type class, Bobbio *et.al* [21] provided a discrete phase-type fitting method for the first time, which is restricted to the acyclic class, while the PHit algorithm (using the EM algorithm) developed by Callut and Dupont [24] can deal with general discrete phase-type distributions.

Recent applications of phase-type distributions in areas like telecommunications, civil engineering, reliability, queueing theory, finance, computer science ([49]), among others, suggested us the importance of doing a thorough statistical analysis of this class of distributions. In particular, in this work we focus on the estimation of the maximum likelihood parameters of phase-type distributions considering different optimization methods (Chapter 3). In Chapter 4 we provide a way of getting the Fisher information of these distributions.

A natural generalization of phase-type distributions is the class of multivariate phase-type distributions, which has been considered by Assaf *et.al* in [12] and by Kulkarni in [39]. Kulkarni defined this class of distributions in a restricted setting and studied some of their properties; however, neither applications nor statistical methods were proposed. In Chapter 5 we analyze in more detail this class giving an estimation of the bivariate case via the EM algorithm and via a quasi Newton-Raphson method.

Moreover, extending the domain of phase-type distributions from the positive real line to the entire line leads to the definition of bilateral phase-type distributions (see [59]). Some properties and applications of this class of distributions were studied by Ahn and Ramaswami in [2]. In Chapter 6, we study the class of multivariate bilateral phase-type distributions giving a characterization of them in terms of univariate bilateral phase-type distributions. This class of distributions turns out to be useful in areas like finance as it is showed in the work of Asmussen [7].

Many results using phase-type methodology have been generalized into the broader class of matrix-exponential distributions (distributions with rational Laplace transform), either by analytic methods (see Asmussen and Bladt [10], Bean and Nielsen [13]) or, more recently, using a flow interpretation (see Bladt and Neuts [16]). Nevertheless, the analysis of distributions with a multidimensional rational Laplace transform (also known as MVME- multivariate matrix-exponential distributions, [17]) has never been considered in its full generality. In order to generalize matrix-exponential distributions into the n-dimensional ($n \geq 1$) real space $\mathbb{R}^n$, and to unify a number of distributions, we define in Chapter 6 a new class of distributions called bilateral matrix-exponential distributions (distributions with rational moment generating function) for both univariate and multivariate cases.

The structure of the thesis is the following. First of all, we begin with some relevant background information on phase-type distributions in Chapter 2. In Chapter 3 we study their maximum likelihood estimation by different methods: EM algorithm, Markov chain Monte Carlo, Newton-Raphson method, among others. We have compared all of them taking into account the value of the log-likelihood and the execution time performed. Explicit formulae to find the

Fisher information matrix for both continuous and discrete phase-type distributions are given in Chapter 4. The multivariate case for phase-type distributions is considered in Chapter 5, and in Chapter 6 we analyze matrix-exponential distributions, giving a generalization of these. Some final remarks and perspectives are included in Chapter 7.

Chapter 2

Phase-type distributions

The embedding into a Markov process is generally referred to as the method of supplementary variables. A particular instance of the method of supplementary variables is known as the method of phases and involves ideas of remarkable simplicity which were first proposed by A. K. Erlang [29] in 1909. He observed that gamma distributions whose shape parameter is a positive integer, may be considered as the probability distributions of sums of independent, negative exponential random variables.

In the recent decades, a lot of research is carried out to handle stochastic models in which durations are phase-type distributed. Phase-type distributions were considered first by Neuts ([44],[45]). O'Cinneide [53] studied some theoretical properties of these distributions, such as their characterization.

Phase-type distributions are defined as distributions of absorption times in a Markov process with $p < \infty$ transient states (the phases) and one absorbing state. Some examples are mixtures and convolution 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 in statistical models. The most established ones come from their role as the computational

vehicle of much of applied probability because they constitute a very versatile class of distributions defined on the non-negative real numbers that lead to models which are algorithmically tractable. Their formulation also allows the Markov structure of stochastic models to be retained when they replace the familiar exponential distribution.

This Chapter is organized as follows. In Section 2.1 we provide necessary background on the theory of Markov jump processes in order to introduce the concept of phase-type distribution in Section 2.2. In Section 2.3 we introduce discrete phase-type distributions. Finally, in Section 2.4 we review the canonical form and reversed-time representation for phase-type distributions.

2.1 Markov jump process

There are several Markov processes in continuous time. In the following we shall focus on the ones which have a finite state-space. By nature, such processes are piecewise constant and transitions occur via jumps. They are often referred to as Markov jump processes (MJP) or continuous time Markov chains (CTMC).

Definition 2.1 A Markov jump process $\{X(t)\}_{t\geq 0}$, with values in the discrete state-space E, is a stochastic process with the following property

$$\mathbb{P}(X(t_n) = i_n | X(t_{n-1}) = i_{n-1}, \ldots, X(t_0) = i_0) = \mathbb{P}(X(t_n) = i_n | X(t_{n-1}) = i_{n-1}).$$

The process is called time-homogeneous if $\mathbb{P}(X(t+h) = j | X(t) = i)$ only depends on h, in which case we denote it by p_{ij}^h. We call p_{ij}^h for the transitions probabilities and define the corresponding transition matrix by $P(h) = \{p_{ij}^h\}_{i,j\in E}$.

Let $T_1, T_2, \ldots$ denote the times where $\{X(t)\}_{t\geq 0}$ jumps from one state to another, where $T_0 = 0$. Then the discrete time process $\{Y_n\}_{n\in\mathbb{N}}$, where $Y_n = X(T_n)$ is a Markov chain that keeps track of which states have been visited. Let $Q = \{q_{ij}\}_{i,j\in E}$ denote its transition matrix.

If $Y_n = i$, then $T_{n+1} - T_n$ is exponentially distributed with a certain parameter λ_i. The conditional probability that there will be a jump in the process $\{X(t)\}_{t\geq 0}$ during the infinitesimal time interval $[t, t + \mathrm{d}t)$ is $\lambda_i \mathrm{d}t$. Given a jump at time t out of state i, the probability that the jump leads to state j is by definition q_{ij}. Hence for $j \neq i$, $\lambda_i \mathrm{d}t q_{ij}$ is the probability of a jump from i to j during $[t, t + \mathrm{d}t)$. Thus for $j \neq i$,

$$\lambda_{ij} = \lambda_i q_{ij},$$

is interpreted as the intensity of jumping from state i to j. Define $\lambda_{ii} = -\sum_{j \neq i} \lambda_{ij}$, and $\mathbf{\Lambda} = \{\lambda_{ij}\}_{i,j \in E}$ be the intensity matrix or infinitesimal generator of the process. Then, we have the following important relation between $P(t)$ and $\mathbf{\Lambda}$,

$$P(t) = \exp(\mathbf{\Lambda} t),$$

where $\exp(\mathbf{A})$ denotes the exponential of a matrix $\mathbf{A}$ defined in usual way by series expansion

$$\exp(\mathbf{A}) = \sum_{n=0}^{\infty} \frac{\mathbf{A}^n}{n!}.$$

2.2 Continuous phase-type distributions

Let $\{X(t)\}_{t \geq 0}$ be a MJP on the finite state-space $E = \{1, 2, \ldots, p, p+1\}$ where the states $1, 2, \ldots, p$ are transient (i.e. given that we start in state $i \in \{1, 2, \ldots, p\}$, there is a non-zero probability that we will never return to i), and the state $p+1$ is absorbing (i.e. it is impossible to leave this state).

Then $\{X(t)\}_{t \geq 0}$ has an intensity matrix on the form

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{2.1}$$

where $\mathbf{T}$ is $(p \times p)$-dimensional matrix (satisfying $t_{ii} < 0$ and $t_{ij} \geq 0$, for $i \neq j$), $\mathbf{t}$ is a p-dimensional column vector (or $(p \times 1)$-dimensional matrix) and $\mathbf{0}$ is the p-dimensional row vector of zeros. Since the intensities of rows must sum to zero, we notice that $\mathbf{t} = -\mathbf{T}\mathbf{e}$, where $\mathbf{e}$ is a p-dimensional column vector of 1's. We suppose that absorption into the state $p+1$ from any initial state, is certain. A useful equivalent condition is given by the following lemma.

Lemma 2.1 *The states $1, \ldots, p$ are transient if and only if the matrix $\boldsymbol{T}$ is non-singular.*

Proof. See Neuts [45]. ■

The intensities t_i are the intensities by which the process jumps to the absorbing state and are known as exit rates. Let $\pi_i = \mathbb{P}(X(0) = i)$ denote the initial probabilities. Hence the initial probability vector of $\{X(t)\}_{t \geq 0}$ is given by $(\boldsymbol{\pi}, \pi_{p+1})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ and such that $\boldsymbol{\pi}\mathbf{e} + \pi_{p+1} = 1$.

Definition 2.2 The time until absorption

$$\tau = \inf\{t \geq 0 | X(t) = p + 1\}$$

is said to have a continuous phase-type (or simply phase-type (PH)) distribution, and we write

$$\tau \sim PH_p(\boldsymbol{\pi}, \mathbf{T}).$$

The set of parameters $(\boldsymbol{\pi}, \mathbf{T})$ is said to be a representation of the phase-type distribution. The dimension of $\mathbf{T}$ is said to be the order of the representation. Typically representations are non-unique and there must exist at least one representation of minimal order. Such a representation is known as minimal representation, and the order of the PH distribution itself is defined to be the order of any of its minimal representations.

Other requirement on the PH representation $(\boldsymbol{\pi}, \mathbf{T})$ is that there are no superfluous phases. That is, each phase in the Markov chain defined by $\boldsymbol{\pi}$ and $\mathbf{T}$ has a positive probability of being visited before absorption. If this is the case, then we say that the PH representation is irreducible (see [45]).

Definition 2.3 A representation $(\boldsymbol{\pi}, \mathbf{T})$ for phase-type distributions is called *irreducible* if and only if the matrix $\mathbf{T} + (1 - \pi_{p+1})^{-1}\mathbf{t}\boldsymbol{\pi}$ is irreducible.

For the definition of an irreducible matrix see [58]. If the representation is reducible, we can form an irreducible representation by simply deleting those states that are superfluous.

Note 2.4 *Throughout the thesis if we omit the subindex p in the representation, it is because we know in advance the order of the phase–type distribution.*

Now, since $\exp(\boldsymbol{\Lambda} s)$ is the transition matrix $P(s)$ of the Markov jump process $\{X(t)\}_{t\geq 0}$, we have that

$$\begin{aligned}
\exp(\boldsymbol{\Lambda} s) &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{\boldsymbol{\Lambda}^n s^n}{n!} = \mathbf{I} + \sum_{n=1}^{\infty} \frac{s^n}{n!} \begin{pmatrix} \mathbf{T}^n & -\mathbf{T}^n\mathbf{e} \\ \mathbf{0} & 0 \end{pmatrix} \\
&= \mathbf{I} + \begin{pmatrix} \sum_{n=1}^{\infty} \frac{\mathbf{T}^n s^n}{n!} & -\sum_{n=1}^{\infty} \frac{\mathbf{T}^n \mathbf{e} s^n}{n!} \\ \mathbf{0} & 0 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{T}^n s^n}{n!} & -\sum_{n=1}^{\infty} \frac{\mathbf{T}^n \mathbf{e} s^n}{n!} \\ \mathbf{0} & 1 \end{pmatrix} \\
&= \begin{pmatrix} \exp(\mathbf{T}s) & -(\exp(\mathbf{T}s)\mathbf{e} - \mathbf{I}\mathbf{e}) \\ \mathbf{0} & 1 \end{pmatrix} \\
&= \begin{pmatrix} \exp(\mathbf{T}s) & \mathbf{e} - \exp(\mathbf{T}s)\mathbf{e} \\ \mathbf{0} & 1 \end{pmatrix}.
\end{aligned}$$

The restriction of $P(s)$ to the transient states is given by $\exp(\mathbf{T}s)$. Hence we are able to compute transitions probabilities $p_{ij}^s = \mathbb{P}(X(s) = j|X(0) = i) = \exp(\mathbf{T}s)_{ij}$, for $i, j = 1, \ldots, p$.

Let f be the density of $\tau \sim PH(\boldsymbol{\pi}, \mathbf{T})$. The quantity $f(s)\mathrm{d}s$ may be interpreted as the probability $\mathbb{P}(\tau \in [s, s+\mathrm{d}s))$. If $\tau \in [s, s+\mathrm{d}s)$, then the underlying Markov jump process $\{X(t)\}_{t\geq 0}$ must be in some transient state j at time s. If the process initiates in a state i, the probability that $X(s) = j$ is $p_{ij}^s = \exp(\mathbf{T}s)_{ij}$. The probability that the process $\{X(t)\}_{t\geq 0}$ starts in state i is by definition π_i. If $X(s) = j$, the probability of a jump to the absorbing state $p+1$ during $[s, s+\mathrm{d}s)$ is $t_j\mathrm{d}s$.

Conditioning on the initial state of the process, we get that

$$\begin{aligned}
f(s)\mathrm{d}s &= \mathbb{P}(\tau \in [s, s+\mathrm{d}s)) \\
&= \sum_{j=1}^{p} \mathbb{P}(\tau \in [s, s+\mathrm{d}s)|X(s) = j)\mathbb{P}(X(s) = j) \\
&= \sum_{j=1}^{p} \mathbb{P}(\tau \in [s, s+\mathrm{d}s)|X(s) = j) \sum_{i=1}^{p} \mathbb{P}(X(s) = j|X(0) = i)\mathbb{P}(X(0) = i) \\
&= \sum_{j=1}^{p} t_j\mathrm{d}s \sum_{i=1}^{p} \exp(\mathbf{T}s)_{ij}\pi_i \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \pi_i \exp(\mathbf{T}s)_{ij}t_j\mathrm{d}s \\
&= \boldsymbol{\pi}\exp(\mathbf{T}s)\mathbf{t}\mathrm{d}s.
\end{aligned}$$

We have thus proved the following theorem:

Theorem 2.5 *If $\tau \sim PH(\boldsymbol{\pi}, \boldsymbol{T})$ its density is given by*

$$f(s) = \boldsymbol{\pi}\exp(\boldsymbol{T}s)\boldsymbol{t},$$

where $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$.

We could now obtain an expression for the distribution function by integrating the density, however, we shall retrieve this formula by an even simpler argument. If F denotes the distribution function of τ, then $1 - F(s)$ is the probability that $\{X(t)\}_{t\geq 0}$ has not yet been absorbed by time s, i.e. $\tau > s$. But the event $\{\tau > s\}$ is identical to $\{X(s) \in \{1, 2, \ldots, p\}\}$. Hence, by a similar conditioning

argument as above, we get that

$$\begin{aligned}
1 - F(s) &= \mathbb{P}(\tau > s) \\
&= \mathbb{P}(X(s) \in \{1, \ldots, p\}) \\
&= \mathbb{P}\left(\bigcup_{j=1}^{p} (X(s) = j)\right) \\
&= \sum_{j=1}^{p} \mathbb{P}(X(s) = j) \\
&= \sum_{i,j=1}^{p} \mathbb{P}(X(s) = j | X(0) = i)\mathbb{P}(X(0) = i) \\
&= \sum_{i,j=1}^{p} p_{ij}^{s} \pi_i \\
&= \sum_{i,j=1}^{p} \pi_i \exp(\mathbf{T}s)_{ij} \\
&= \boldsymbol{\pi} \exp(\mathbf{T}s)\mathbf{e}.
\end{aligned}$$

Thus we have proved:

Theorem 2.6 *If $\tau \sim PH(\boldsymbol{\pi}, \boldsymbol{T})$, the distribution function of τ is given by*

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

Example 2.1 *Exponential distribution*

Let $X \sim \exp(\lambda)$, for some $\lambda > 0$, since its density is $f(x) = \lambda e^{-\lambda x}$, its minimal PH representation is given by

$$\boldsymbol{\pi} = [1], \quad \mathbf{T} = [-\lambda], \quad \mathbf{t} = [\lambda].$$

□

Theorem 2.7 *Let $\tau \sim PH(\boldsymbol{\pi}, \boldsymbol{T})$.*

1. *The n-th moment of τ is given by $\mathbb{E}(\tau^n) = (-1)^n n! \boldsymbol{\pi} \boldsymbol{T}^{-n} \boldsymbol{e}$.*
2. *The moment generating function of τ is given by $\mathbb{E}(e^{s\tau}) = \boldsymbol{\pi}(-s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}$, where $\boldsymbol{I}$ denotes the identity matrix of the appropriate dimension.*

PROOF. We will prove the first part by induction. For $n = 1$, we have

$$\begin{aligned}
\mathbb{E}(\tau) &= \int_0^\infty s f(s)\mathrm{d}s \\
&= \int_0^\infty s\boldsymbol{\pi} e^{\mathbf{T}s}\mathbf{t}\mathrm{d}s \\
&= -\int_0^\infty \boldsymbol{\pi} e^{\mathbf{T}s}\mathbf{T}^{-1}\mathbf{t}\mathrm{d}s \\
&= \boldsymbol{\pi}\mathbf{T}^{-2}\mathbf{t} \\
&= \boldsymbol{\pi}\mathbf{T}^{-2}(-\mathbf{T}\mathbf{e}) \\
&= -\boldsymbol{\pi}\mathbf{T}^{-1}\mathbf{e}.
\end{aligned}$$

By inductive hypothesis assume that $\mathbb{E}(\tau^k) = (-1)^k k!\boldsymbol{\pi}\mathbf{T}^{-k}\mathbf{e}$ is valid for some k. Then for $k+1$,

$$\begin{aligned}
\mathbb{E}(\tau^{k+1}) &= \int_0^\infty s^{k+1} f(s)\mathrm{d}s \\
&= \int_0^\infty s^{k+1}\boldsymbol{\pi} e^{\mathbf{T}s}\mathbf{t}\mathrm{d}s \\
&= -\int_0^\infty (k+1)s^k\boldsymbol{\pi} e^{\mathbf{T}s}\mathbf{T}^{-1}\mathbf{t}\mathrm{d}s \\
&= -(k+1)\mathbf{T}^{-1}\int_0^\infty s^k\boldsymbol{\pi} e^{\mathbf{T}s}\mathbf{t}\mathrm{d}s \\
&= -(k+1)\mathbf{T}^{-1}(-1)^k k!\boldsymbol{\pi}\mathbf{T}^{-k}\mathbf{e} \\
&= (-1)^{k+1}(k+1)!\boldsymbol{\pi}\mathbf{T}^{-(k+1)}\mathbf{e}.
\end{aligned}$$

The moment generating function is given by

$$\begin{aligned}
\mathbb{E}(e^{s\tau}) &= \int_0^\infty e^{sx} f(x)\mathrm{d}x \\
&= \int_0^\infty e^{sx}\boldsymbol{\pi} e^{\mathbf{T}x}\mathbf{t}\mathrm{d}x \\
&= \int_0^\infty \boldsymbol{\pi} e^{sx}\mathbf{I} e^{\mathbf{T}x}\mathbf{t}\mathrm{d}x \\
&= \int_0^\infty \boldsymbol{\pi} e^{s\mathbf{I}x} e^{\mathbf{T}x}\mathbf{t}\mathrm{d}x \\
&= \int_0^\infty \boldsymbol{\pi} e^{(s\mathbf{I}+\mathbf{T})x}\mathbf{t}\mathrm{d}x \\
&= \boldsymbol{\pi}(-s\mathbf{I}-\mathbf{T})^{-1}\mathbf{t}.
\end{aligned}$$

■

From this theorem we can see that if $\tau \sim PH(\boldsymbol{\pi}, \mathbf{T})$, then its Laplace transform $\mathrm{L}_\tau(s) = \mathbb{E}(e^{-s\tau})$ is given by

$$\boldsymbol{\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}, \tag{2.2}$$

or $\mathrm{L}_\tau(s) = \boldsymbol{\pi}(s(-\mathbf{T})^{-1} + \mathbf{I})^{-1}\mathbf{e}$. Indeed, there is a neat probabilistic interpretation of $(-\mathbf{T})^{-1}$. Let $k \geq 0$, then

$$\begin{aligned}
\int_0^k \exp(\mathbf{T}s)\mathrm{d}s &= \int_0^k \sum_{i=0}^{\infty} \frac{(\mathbf{T}s)^i}{i!}\mathrm{d}s \\
&= \sum_{i=0}^{\infty} \mathbf{T}^i \int_0^k \frac{s^i}{i!}\mathrm{d}s \\
&= \sum_{i=0}^{\infty} \mathbf{T}^i \frac{k^{i+1}}{(i+1)!} \\
&= \mathbf{T}^{-1}(e^{\mathbf{T}k} - \mathbf{I}) \xrightarrow[k\to\infty]{} (-\mathbf{T})^{-1}.
\end{aligned}$$

Thus the element (i, j)-th of the matrix $(-\mathbf{T})^{-1}$ is the expected time spent in the phase j before absorption conditioned on the fact that the chain was started in the phase i. From this probabilistic interpretation we have that $(-\mathbf{T})^{-1} \geq 0$. Now, we get the mean time before absorption conditioning on start in i by taking row sums of $(-\mathbf{T})^{-1}$. Thus the i-th element of $(-\mathbf{T})^{-1}\mathbf{e}$ is the mean time spent in the transient states conditioning on start in i. To obtain the mean for a PH distribution with initial probability vector $\boldsymbol{\pi}$, we have to make a weighted sum of $(-\mathbf{T})^{-1}\mathbf{e}$ with $\boldsymbol{\pi}$ as weighting factors, i.e., $\mu_\tau = \boldsymbol{\pi}(-\mathbf{T})^{-1}\mathbf{e}$.

2.2.1 Properties of phase-type distributions

One of the appealing features of phase-type distributions is that the class is closed under a number of operations. The closure properties are a main contributing factor to the popularity of these distributions in probabilistic modelling of technical systems. In particular, we will see that the class is closed under addition, finite mixtures, and finite order statistics.

Let us start with some general matrix results.

Definition 2.8 For two matrices $\mathbf{A}$ and $\mathbf{B}$ of dimensions $(l \times k)$ and $(n \times m)$ respectively, we define the Kronecker product $\otimes$ as the matrix of dimension

$(ln \times km)$ written as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1k}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2k}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{l1}\mathbf{B} & a_{l2}\mathbf{B} & \dots & a_{lk}\mathbf{B} \end{pmatrix}.$$

The following rule is very convenient. If the usual matrix products $\mathbf{LU}$ and $\mathbf{MV}$ exist, then

$$(\mathbf{L} \otimes \mathbf{M})(\mathbf{U} \otimes \mathbf{V}) = \mathbf{LU} \otimes \mathbf{MV}.$$

A natural operation for continuous time phase-type distributions is $\mathbf{A}\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{B}$, as which we define as the Kronecker sum of $\mathbf{A}$ and $\mathbf{B}$, and shall be denoted by $\mathbf{A} \oplus \mathbf{B}$.

Theorem 2.9 *If $F(\cdot)$ and $G(\cdot)$ are both PH distributions with representations $(\boldsymbol{\alpha}, \boldsymbol{T})$ and $(\boldsymbol{\beta}, \boldsymbol{S})$ of orders m and n respectively, their convolution $F^*G(\cdot)$ is a PH distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{L})$, given by*

$$\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \alpha_{m+1}\boldsymbol{\beta}), \quad \boldsymbol{L} = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \cdot \boldsymbol{\beta} \\ \boldsymbol{0} & \boldsymbol{S} \end{pmatrix}, \tag{2.3}$$

where $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$.

Proof. See Neuts [45]. ■

Since the distribution of the sum of random variables is the convolution of their distributions, this shows that the family of PH distributions is closed under finite number of convolutions.

Theorem 2.10 *For $X \sim PH(\boldsymbol{\alpha}, \boldsymbol{T})$ and $Y \sim PH(\boldsymbol{\beta}, \boldsymbol{S})$ both being independent, then $Z = X + Y \sim PH(\boldsymbol{\gamma}, \boldsymbol{L})$, where $\boldsymbol{\gamma}$ and $\boldsymbol{L}$ are given in (2.3).*

Example 2.2 *Addition of exponential distributions.*

Considering the sum $Z = \sum_{i=1}^{k} X_i$ with $X_i \sim \exp(\lambda_i)$, a PH representation is given by

$$\boldsymbol{\gamma} = (1, 0, \dots, 0), \quad \mathbf{L} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_{k-1} & \lambda_{k-1} \\ 0 & 0 & 0 & \dots & 0 & -\lambda_k \end{pmatrix}.$$

This distribution is called k generalized Erlang distribution, and it can be described using a state transition diagram that has k phases in series, see Fig. 2.1. It is easy to see, without loss of generality, that the states can be ordered so that the rates $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$.

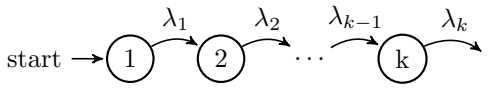

Figure 2.1: State transition diagram for an order k generalized Erlang distribution

With $\lambda_i = \lambda$ we get a sum of identically distributed exponential random variables, called an Erlang distribution (see Table 2.1). □

Table 2.1: Probability density function (PDF), cumulative distribution function (CDF), generating function (GF), and moments of the Erlang distribution

PDF	$f(x;k,\lambda)$	$\lambda\dfrac{(\lambda x)^{k-1}}{(k-1)!}e^{-\lambda x}$
CDF	$F(x;k,\lambda)$	$\displaystyle\sum_{i=k}^{\infty}\frac{(\lambda x)^i}{i!}e^{-\lambda x}$
GF	$H(x;k,\lambda)$	$\left(\dfrac{\lambda}{x+\lambda}\right)^k$
Moments	$\mu_i(k,\lambda)$	$\dfrac{(i+k-1)!}{(k-1)!\lambda^i}$

Concerning finite mixtures of phase-type random variables we have the following result.

Theorem 2.11 *Any finite convex mixture of phase-type distribution is a phase-type distribution. Let $X_i \sim PH(\boldsymbol{\alpha}_i, \boldsymbol{T}_i)$, $i = 1, \ldots, k$, such that $Z = X_i$ with probability p_i. Then $Z \in PH(\boldsymbol{\gamma}, \boldsymbol{L})$ where $\boldsymbol{\gamma} = (p_1\boldsymbol{\alpha}_1, p_2\boldsymbol{\alpha}_2, \ldots, p_k\boldsymbol{\alpha}_k)$ and*

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{T}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{T}_k \end{pmatrix}.$$

Example 2.3 *Mixture of exponential distributions.*

Consider k random variables $X_i \sim \exp(\lambda_i)$ and assume that Z takes the value of X_i with probability p_i. The distribution of Z, called hyper-exponential distribution (see Table 2.2), can be expressed as a proper mixture of the X_i's. A

PH representation is given by

$$\boldsymbol{\gamma} = (p_1, \ldots, p_k), \quad \mathbf{L} = \begin{pmatrix} -\lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -\lambda_k \end{pmatrix}.$$

This distribution can be described using a state transition diagram with k states in parallel, see Fig. 2.2. Clearly, without loss of generality, the states can be ordered so that the rates $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k$.

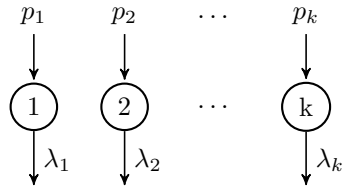

Figure 2.2: State transition diagram for an order k Hyper-exponential distribution

□

Table 2.2: Probability density function (PDF), cumulative distribution function (CDF), generating function (GF), and moments of the hyper-exponential distribution

PDF	$f(x)$	$\sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x}$
CDF	$F(x)$	$1 - \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x}$
GF	$H(x)$	$\sum_{i=1}^{k} \frac{p_i \lambda_i}{s + \lambda_i}$
Moments	μ_i	$i! \sum_{i=1}^{k} \frac{p_i}{\lambda_i^i}$

Theorem 2.12 *For $X \sim PH_k(\boldsymbol{\alpha}, \boldsymbol{T})$ and $Y \sim PH_m(\boldsymbol{\beta}, \boldsymbol{S})$, the $\min(X, Y)$ is phase-type distributed with representation $(\boldsymbol{\gamma}, \boldsymbol{L})$, where*

$$\boldsymbol{L} = \boldsymbol{T} \otimes \boldsymbol{I}_m + \boldsymbol{I}_k \otimes \boldsymbol{S},$$

and $\boldsymbol{\gamma} = \boldsymbol{\alpha} \otimes \boldsymbol{\beta}$, $\boldsymbol{I}_p$ *represents the* $(p \times p)$*-dimensional identity matrix. The* $\max(X, Y)$ *is also phase-type distributed with representation* $(\boldsymbol{\gamma}, \boldsymbol{L})$*, where*

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{T} \otimes \boldsymbol{I}_m + \boldsymbol{I}_k \otimes \boldsymbol{S} & \boldsymbol{I}_k \otimes \boldsymbol{s} & \boldsymbol{t} \otimes \boldsymbol{I}_m \\ \boldsymbol{0} & \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{S} \end{pmatrix},$$

and $\boldsymbol{\gamma} = (\boldsymbol{\alpha} \otimes \boldsymbol{\beta}, \boldsymbol{\alpha}\beta_{m+1}, \alpha_{k+1}\boldsymbol{\beta})$*. The exit vector* $\boldsymbol{l}$ *is given by*

$$\boldsymbol{l} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{t} \\ \boldsymbol{s} \end{pmatrix},$$

where $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$ *and* $\boldsymbol{s} = -\boldsymbol{S}\boldsymbol{e}$*.*

Proof. See Neuts [45]. ■

For more closure properties we refer to [40] and [42].

2.3 Discrete phase-type distributions

A discrete phase-type (DPH) distribution is the time until absorption of a discrete time Markov chain (see [26, 50, 57]). DPH distributions are defined by considering a $p+1$-state Markov chain $\mathbf{P}$ of the form

$$\mathbf{P} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix},$$

where $\mathbf{T}$ is a sub-stochastic matrix, such that $\mathbf{I} - \mathbf{T}$ is non-singular. More precisely, let $\{X(n)\}_{n \geq 0}$ denote a Markov chain with state-space $E = \{1, \ldots, p, p+1\}$, where the states $1, \ldots, p$ are transient and the state $p+1$ is absorbing. Let $\pi_i = \mathbb{P}(X(0) = i)$ denote the initial probabilities and t_{ij} the transition probabilities $\mathbb{P}(X(n+1) = j | X(n) = i)$, for $i, j = 1, \ldots, p$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ be the initial vector, $\mathbf{T} = \{t_{ij}\}_{i,j=1,\ldots,p}$ the transition matrix between transient states, and $\mathbf{t} = \mathbf{e} - \mathbf{T}\mathbf{e}$ the vector of probabilities of jumping to the absorbing state.

Definition 2.13 We say that $\tau = \inf\{n \geq 1 | X(n) = p+1\}$ has a discrete phase-type distribution with representation $(\boldsymbol{\pi}, \mathbf{T})$ and write $\tau \sim DPH_p(\boldsymbol{\pi}, \mathbf{T})$.

Sometimes it is convenient to allow for an atom at zero as well in which case we let $\pi_{p+1} > 0$ denote the initial probability of initiating in the absorbing state.

The probability density f of τ is given by

$$f(x) = \boldsymbol{\pi}\mathbf{T}^{x-1}\mathbf{t}, \quad \text{for } x \geq 1,$$

if $\pi_{p+1} > 0$ then $f(0) = \pi_{p+1}$. Let us prove this. The probability that the Markov chain is in one of the transient states $i \in \{1, \ldots, p\}$ after n steps is given by

$$p_i^{(n)} = \mathbb{P}(X(n) = i) = \sum_{k=1}^{p} \pi_k (\mathbf{T}^n)_{(k,i)}.$$

The probability of absorption of the Markov chain at time n is given by the sum over the probabilities of the Markov chain being in one of the states $\{1, \ldots, p\}$ at time $n-1$ multiplied by the probability that absorption takes place from that state. The state in the Markov chain at time $n-1$ depends on the initial state and on the $(n-1)$-step transition probability matrix $\mathbf{T}^{n-1}$. Hence we get

$$f(n) = \mathbb{P}(\tau = n) = \sum_{i=1}^{p} p_i^{(n-1)} t_i = \boldsymbol{\pi}\mathbf{T}^{n-1}\mathbf{t}, \quad n \in \mathbb{N}.$$

The distribution function can be deduced by the following probabilistic argument.

Lemma 2.2 *The distribution function of a discrete phase-type random variable is given by*

$$F(n) = 1 - \boldsymbol{\pi}\, \boldsymbol{T}^n \boldsymbol{e}.$$

Proof. We look at the probability that absorption has not yet taken place and hence the Markov chain is in one of the transient states. We get

$$\begin{aligned} 1 - F(n) &= \mathbb{P}(\tau > n) \\ &= \sum_{i=1}^{p} p_i^{(n)} \\ &= \boldsymbol{\pi}\mathbf{T}^n\mathbf{e}. \end{aligned}$$

■

The probability generating function of τ, $G_\tau(z) = \mathbb{E}(z^\tau)$, is given by

$$\begin{aligned}
\mathbb{E}(z^\tau) &= \sum_{k=0}^{\infty} z^k f(k) \\
&= \sum_{k=1}^{\infty} z^k \boldsymbol{\pi}\mathbf{T}^{k-1}\mathbf{t} \\
&= \boldsymbol{\pi}\mathbf{T}^{-1}\left(\sum_{k=1}^{\infty}(z\mathbf{T})^k\right)\mathbf{t} \\
&= \boldsymbol{\pi}\mathbf{T}^{-1}\left(\frac{z\mathbf{T}}{\mathbf{I}-z\mathbf{T}}\right)\mathbf{t} \\
&= z\boldsymbol{\pi}(\mathbf{I}-z\mathbf{T})^{-1}\mathbf{t}.
\end{aligned}$$

If $\pi_{p+1} > 0$ then $\mathbb{E}(z^\tau) = \pi_{p+1} + z\boldsymbol{\pi}(\mathbf{I}-z\mathbf{T})^{-1}\mathbf{t}$. Its factorial moments are given by

$$\begin{aligned}
G_\tau^{(k)}(1) &= \left.\frac{\mathrm{d}^k}{\mathrm{d}z^k}\right|_{z=1} G_\tau(z) \\
&= k!\boldsymbol{\pi}\mathbf{T}^{k-1}(\mathbf{I}-\mathbf{T})^{-k}\mathbf{e}.
\end{aligned}$$

A representation $(\boldsymbol{\pi}, \mathbf{T})$ for discrete phase-type distribution is called *irreducible* if every state of the Markov chain can be reached with positive probability. We can always find an irreducible representation by simply leaving out the states that cannot be reached.

Neuts [44] has given a number of elementary properties of discrete phase-type distributions, with some comments on their utility in areas like renewal theory, branching processes, and queues. He has also discussed convolution products and mixtures of these distributions.

Some properties are the following:

- Any probability density on a finite number of positive integers is discrete phase-type.
- The convolution of a finite number of densities of discrete phase-type is itself of discrete phase-type.
- Any finite mixture of probabilities densities of discrete phase-type is itself of discrete phase-type.

Example 2.4 *Geometric distribution*

$X \sim geo(p)$, with $p \in (0,1)$, i.e. $\mathbb{P}(X = x) = (1-p)^{1-x}p$ has a DPH representation given by

$$\boldsymbol{\pi} = [1], \quad \mathbf{T} = [1-p], \quad \mathbf{t} = [p].$$

□

Example 2.5 *Negative binomial distribution*

$X \sim NB(k,p)$, with $p \in (0,1)$ and $k > 0$, i.e., X is the sum of k random variables $geo(p)$-distributed, so $\mathbb{P}(X = x) = \binom{x+k-1}{k-1}(1-p)^k p^x$, for $x = 0, 1, \ldots$. X has a DPH representation given by

$$\boldsymbol{\pi} = (1, 0, \ldots, 0), \quad \mathbf{T} = \begin{pmatrix} 1-p & p & & & \\ & 1-p & p & & \\ & & \ddots & & \\ & & & 1-p & p \\ & & & & 1-p \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ p \end{pmatrix}.$$

□

2.4 On the representations of phase-type distributions

The optimization problem for general discrete phase-type (DPH) distributions is too complex to yield satisfactory results if we have a large number of phases. Bobbio and Cumani [19] have showed that the estimation problem becomes much easier if acyclic instead of general DPH distributions are used, because for this type of distributions, a canonical representation exists, which reduces the number of free parameters.

2.4.1 Canonical form

A discrete phase-type representation of a given distribution is, in general, non-unique and non-minimal. Bobbio *et.al* [21] explored a subclass of the DPH class for which the representation is an acyclic graph (ADPH). The ADPH class admits a unique minimal representation, called canonical form (CF). Cumani [27] has shown that a canonical representation for the subclass of PH distributions

with generating acyclic Markov chain (denoted by APH), is unique, minimal, and has the form of a Coxian model with real transition rates.

The use of the canonical representation for APH offers many advantages (see [20]). Some of these are shared by the whole PH class, some hold only for the APH class and, finally, some are peculiar to the CF representation.

- CF is a natural and straightforward restriction of the Coxian model obtained by forcing the transition rates to be real, but at the same time, the eigenvalue ordering ensures that the CF provides a unique representation of the whole class of APH.
- CF forms a dense set for distributions with support on $[0, \infty)$.
- APH is closed under mixture, convolution, and formation of coherent systems.

According to Bobbio *et.al* [21], one way of finding a canonical form of discrete phase-type distributions is the following.

1. Re-order the eigenvalues (diagonal elements) of the transition matrix into a decreasing sequence $q_1 \geq q_2 \geq \cdots \geq q_p$, where p is the dimension of the transition matrix. Define $d_i = 1 - q_i$, which represents the exit rate from state i.
2. Find the different paths, denoted by r_k, to reach the absorbing state.

   Any path r_k can be described as a binary vector $\mathbf{u}_k = [u_i]$ of length p defined over the ordered sequence of the q_i's. Each entry of the vector is equal to 1 if the corresponding eigenvalue q_i is present in the path, otherwise the entry is equal to 0. Hence any path r_k of length l has l ones in the vector $\mathbf{u}_k$.
3. Identify the basic paths.

   A path r_k of length l of an ADPH is called basic path if it contains the l fastest phases $q_{p-l+1}, \ldots, q_p$. The binary vector associated to a basic path is called a basic vector and it contains $(p-l)$ initial 0's and l terminal 1's.
4. Any path is assigned its characteristic binary vector. If the binary vector is not in basic form, each path is transformed into a mixture of basic paths.

   Cumani [27] has provided an algorithm which performs the transformation of any path into a mixture of basic paths in a finite number of steps.

5. Find the coefficients a_i, $i = 1, \ldots, p$, associated with $F(z, \mathbf{b}_i)$, where $\mathbf{b}_i$ is the i-th basic vector and $F(z, \mathbf{b}_i)$ is the product of the generating functions of the sojourn times spent in the consecutive states of the path (see [21] for more details).

6. Calculate the following

$$\begin{aligned} s_i &= \sum_{j=1}^{i} a_j, \qquad 1 \le i \le p, \\ e_i^* &= \frac{a_i}{s_i} d_i, \qquad 1 \le i \le p, \\ e_i &= \frac{s_{i-1}}{s_i} d_i, \qquad 2 \le i \le p. \end{aligned}$$

Definition 2.14 Canonical form CF*([21]). An ADPH is in canonical form CF* if from any phase i, $1 \le i \le p$, transitions are possible to phase i itself, $i+1$, and $p+1$. The initial probability is 1 for phase $i = 1$ and 0 for any phase $i \ne 1$.

Then the matrix representation $(\boldsymbol{\pi}, \mathbf{T})$ for the CF* is given by

$$\begin{aligned} \boldsymbol{\pi} &= (1, 0, \ldots, 0), \\ \mathbf{T} &= \begin{pmatrix} q_p & e_p & & & & \\ & q_{p-1} & e_{p-1} & & & \\ & & & \ddots & & \\ & & & & q_2 & e_2 \\ & & & & & q_1 \end{pmatrix}, \\ \mathbf{t} &= (e_p^*, e_{p-1}^*, \ldots, e_1^*)'. \end{aligned}$$

2.4.2 Reversed-time representation

Consider a PH-representation $(\boldsymbol{\pi}, \mathbf{T})$ and denote the absorption time by τ. If we are in state i of the original process at time $\tau - t$, then the process in which we are in state i at time t is called the *dual or reverse-time representation*. It can be proved that this is again a PH-representation $(\boldsymbol{\pi}^*, \mathbf{T}^*)$ (see [56]). This reversed-time representation is also valid in the discrete case, and is given by

$$\boldsymbol{\pi}^* = \mathbf{t}'\mathbf{M}, \qquad \mathbf{t}^* = \mathbf{M}^{-1}\boldsymbol{\pi}', \qquad \mathbf{T}^* = \mathbf{M}^{-1}\mathbf{T}'\mathbf{M}.$$

Here the matrix $\mathbf{M}$ is a scaling diagonal matrix

$$\mathbf{M} = diag(m_1, \ldots, m_p),$$

where the row vector $\mathbf{m} = (m_1, \ldots, m_p)$ is obtained as

$$\mathbf{m} = \boldsymbol{\pi}(\mathbf{I} - \mathbf{T})^{-1}.$$

We have the following interesting properties of the reversed-time representation:

1. The representation and its reversed-time representation rise to the same PH distribution.
2. The two representations have the same number of states and there is a one-to-one correspondence between these states.
3. The term m_i is the average time which is spent in state i before absorption. This number is finite and non-zero if the representation is irreducible ([6]).

Reversed Markov chain

If we are interested in simulating a Markov chain related to a random variable $\tau \sim DPH(\boldsymbol{\pi}, \mathbf{T})$, we have to satisfy the condition that at time τ the Markov chain is in the absorbing state. For this reason, it might be more efficient to consider a reversed Markov chain, since we can avoid rejecting Markov chains that do not satisfy these conditions.

The transition probabilities of the reversed Markov chain $\{X_i\}_{i\geq 0}$, are given by

$$\mathbb{P}(X_m = j \mid X_{m+1} = i) \quad = \quad \frac{\mathbb{P}(X_m = j)\mathbb{P}(X_{m+1} = i \mid X_m = j)}{\mathbb{P}(X_{m+1} = i)}, \quad m \geq 0,$$

where in general, if $\ell \in \{1, \ldots, p\}$, $\mathbb{P}(X_1 = \ell) = \sum_{k=1}^{p} t_{k,\ell}\pi_\ell$, and for $i \geq 2$

$$\mathbb{P}(X_i = \ell) = \sum_{k=1}^{p} t_{k,\ell}\mathbb{P}(X_{i-1} = k),$$

or simply $\mathbb{P}(X_i = \ell) = \boldsymbol{\pi}\mathbf{T}^i\mathbf{e}_\ell$.

- If $\tau = 1$

$$\mathbb{P}(X_0 = \ell \mid X_1 = p+1) \quad = \quad \frac{\pi_\ell t_\ell}{\boldsymbol{\pi}\mathbf{t}}, \quad \ell \in \{1, 2, \ldots, p\}.$$

- If $\tau \geq 2$:

1. For $\ell_{\tau-1} \in \{1, 2, \ldots, p\}$

$$\mathbb{P}(X_{\tau-1} = \ell_{\tau-1} \mid X_\tau = p+1) \quad = \quad \frac{\mathbb{P}(X_{\tau-1} = \ell_{\tau-1})}{\boldsymbol{\pi}\mathbf{T}^{\tau-1}\mathbf{t}} t_{\ell_{\tau-1}}.$$

2. If $\tau \geq 3$, from $i = \tau - 2$ to $i = 1$, $\ell_i, \ell_{i+1}, \cdots \in \{1, 2, \ldots, p\}$,

$$\begin{aligned} \mathbb{P}(X_i = \ell_i \mid X_{i+1} = \ell_{i+1}, \ldots, X_\tau = p+1) &= \mathbb{P}(X_i = \ell_i \mid X_{i+1} = \ell_{i+1}) \\ &= \frac{\mathbb{P}(X_i = \ell_i)}{\mathbb{P}(X_{i+1} = \ell_{i+1})} t_{\ell_i, \ell_{i+1}}. \end{aligned}$$

3. $i = 0$, $\ell_i, \ell_{i+1}, \cdots \in \{1, 2, \ldots, p\}$,

$$\begin{aligned} \mathbb{P}(X_0 = \ell_0 \mid X_1 = \ell_1, \ldots, X_\tau = p+1) &= \mathbb{P}(X_0 = \ell_0 \mid X_1 = \ell_1) \\ &= \frac{\pi_{\ell_0}}{\mathbb{P}(X_1 = \ell_1)} t_{\ell_0, \ell_1}. \end{aligned}$$

Chapter 3

Fitting phase-type distributions

As it is well known, the main advantage of working with phase-type distributions is the versatility that they offer in modelling.

The literature on estimation of (an approximation by) general phase-type (PH) distributions is meager and not always satisfying from a statistical point of view. The class of PH distributions has favorable computational properties, however, a PH representation is redundant and not unique ([51]), and does not appear as a good starting point for the fitting problem. One needs algorithms to determine the parameters of the applied PH distribution.

Numerical maximum likelihood methods for Coxian distributions, using non-linear constrained optimization, have been implemented in [19] and [22]; this approach appears in many ways to be one of the most satisfying developed so far, the main restriction being that only Coxian distributions are allowed. The two main classes of fitting methods differ in the kind of information they utilize: incomplete or complete information. Asmussen *et.al* [11] have given a more general estimation of phase-type distributions based on the EM algorithm for the complete class. More recently, Hovarth and Telek [35] presented a tool that allows for approximating distributions for both continuous and discrete phase-type distributions.

Bobbio *et.al* [21] have provided a discrete phase-type (DPH) fitting method that turns out to be simple and stable, but it is restricted to acyclic DPH, while the algorithm developed by Callut and Dupont [24], can deal with general DPH.

In this Chapter we present statistical approaches to estimation theory for phase-type distributions, considering both continuous and discrete cases. In Section 3.1 we introduce some methods used for finding maximum likelihood estimators. In Section 3.2 we consider the continuous case while in Section 3.3 we consider the discrete case.

3.1 Methods of finding estimators

In this Section, we will review some theory about maximum likelihood estimators. We will analyze methods such as: the Expectation-Maximization algorithm, the Gibbs sampler algorithm, and the Newton-Raphson method.

3.1.1 Maximum likelihood estimators

The method of maximum likelihood is, by far, the most popular technique for deriving estimators. Recall that if $X_1, \ldots, X_n$ are an i.i.d sample from a population with probability density function $f(x; \theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\boldsymbol{\theta}; \mathbf{x}) = L(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_k).$$

Definition 3.1 For each sample point $\mathbf{x}$, let $\hat{\boldsymbol{\theta}}(\mathbf{x})$ be a parameter value at which $L(\boldsymbol{\theta}; \mathbf{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter $\boldsymbol{\theta}$ based on a sample $\mathbf{X}$ is $\hat{\boldsymbol{\theta}}(\mathbf{X})$.

Notice that, by this construction, the range of the MLE coincides with the range of the parameter. We also use the abbreviation MLE to stand for maximum likelihood estimate when we are talking of the realized value of the estimator. Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, possessing some of the optimality properties: consistency, efficiency, and asymptotic normality.

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}; \mathbf{x}) = 0, \quad i = 1, \ldots, k. \tag{3.1}$$

Note that the solutions of (3.1) are only possible candidates for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition. Furthermore, the zeros of the first derivative locate only extreme points in the interior of the domain of a function. If the extrema occur on the boundary the first derivative may not be 0. Thus the boundary must be checked separately for extrema.

In many cases, estimation is performed using a set of independent identically distributed measurements. These may correspond to distinct elements from a random sample, repeated observations, etc. In such cases, it is of interest to determine the behavior of a given estimator as the number of measurements increases to infinity, referred to as asymptotic behavior. Under certain regularity conditions, which are listed below, the maximum likelihood estimator exhibits several characteristics which can be interpreted to mean that it is asymptotically optimal. These characteristics include:

- The MLE is asymptotically unbiased, i.e., its bias tends to zero as the number of samples increases to infinity.
- The MLE is asymptotically efficient, i.e., it achieves the Cramer-Rao lower bound when the number of samples tends to infinity. This means that, asymptotically, no unbiased estimator has lower mean squared error than the MLE.
- The MLE is asymptotically normal. As a number of samples increases, the distribution of the MLE tends to the Gaussian distribution with covariance matrix equal to the inverse of the Fisher information matrix. In addition, this property makes possible to calculate, assuming some kind of Gaussianity, confidence ranges where the true value of the parameter is confined with a given probability.

The regularity conditions required to ensure this behavior are:

1. The first and second derivatives of the log-likelihood function must be defined.
2. The Fisher information matrix must not be zero.

We let

$$I(\boldsymbol{\theta};\mathbf{y}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \tag{3.2}$$

be the matrix of negative of the second-order partial derivatives of the log-likelihood function with respect to the elements of $\boldsymbol{\theta}$, ($(')$ represents the transpose). Under regularity conditions, the expected Fisher information matrix $\mathtt{I}(\boldsymbol{\theta})$ is given by

$$\begin{aligned}\mathtt{I}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\{S(\mathbf{Y};\boldsymbol{\theta})S'(\mathbf{Y};\boldsymbol{\theta})\}\\ &= -\mathbb{E}_{\boldsymbol{\theta}}\{I(\boldsymbol{\theta};\mathbf{Y})\}\end{aligned}$$

where

$$S(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{3.3}$$

is the gradient vector of the log-likelihood function; that is, the score statistic. The operator $\mathbb{E}_{\boldsymbol{\theta}}$ denotes expectation using the parameter vector $\boldsymbol{\theta}$.

The asymptotic covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ is equal to the inverse of the expected information matrix $\mathtt{I}(\boldsymbol{\theta})$, which can be approximated by $\mathtt{I}(\hat{\boldsymbol{\theta}})$; the standard error of $\hat{\boldsymbol{\theta}}_i = (\hat{\boldsymbol{\theta}})_i$ is given by

$$SE(\hat{\boldsymbol{\theta}}_i) \approx (\mathtt{I}^{-1}(\hat{\boldsymbol{\theta}}))_{ii}^{1/2}.$$

It is common in practice to estimate the inverse of the covariance matrix of the maximum likelihood solution by the observed information matrix $I(\hat{\boldsymbol{\theta}};\mathbf{y})$, rather than the expected information matrix $\mathtt{I}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. This approach gives the approximation

$$SE(\hat{\boldsymbol{\theta}}_i) \approx (I^{-1}(\hat{\boldsymbol{\theta}};\mathbf{y}))_{ii}^{1/2},$$

also, the observed information matrix is usually more convenient to use than the expected information matrix, as it does not require an expectation to be taken.

3.1.2 Expectation-Maximization algorithm

The Expectation-Maximization (EM) (Dempster [28]) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates, useful in a variety of incomplete-data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. On each

iteration of the EM algorithm, there are two steps, called the expectation step or E-step and the maximization step or the M-step.

The situations where the EM algorithm can be applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not all natural or evident.

The basic idea of the EM algorithm is to associate with the given incomplete-data problem, a complete-data problem for which maximum likelihood estimations are computationally more tractable; for instance, the complete-data problem chosen may yield a closed-form solution to the maximum likelihood estimate. The methodology of the EM algorithm then consists in reformulating the problem in terms of this more easily solved complete-data problem, establishing a relationship between the likelihoods of these two problems. The E-step consists in manufacturing data for the complete-data problem, using the observed data set of the incomplete-data problem and the current value of the parameters, so that the simpler M-step computation can be applied to this completed data set. Starting from suitable initial parameter values, the E- and M-steps are repeated until convergence.

3.1.3 Gibbs sampler algorithm

The Gibbs sampler (GS) is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density (see [25]). The GS is a Markov chain Monte Carlo method that was introduced by German and German [32], and is a special case of the Metropolis-Hastings (MH) algorithm, developed by Metropolis *et.al* [43] and generalized by Hastings [33].

The premise of Bayesian statistics is to incorporate prior knowledge along with a given set of current observations, in order to make statistical inferences. By incorporating prior information about the parameter(s), a posterior distribution for the parameter(s) can be obtained and inferences on the model parameters and their functions can be made. The prior knowledge about the parameter(s) is expressed in terms of a pdf, called the prior distribution. The posterior distribution given the sample data, provides the updated information about the parameter(s). We can obtain the posterior distribution multiplying the prior by the likelihood function and then normalizing.

In the following, we will explain in a general way how the Gibbs sampling works. Let $\boldsymbol{\theta}$ be a vector of parameters with posterior distribution $p^*(\boldsymbol{\theta}|\mathbf{x})$, where $\mathbf{x}$ denotes the data. Suppose that $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$, where

some $\boldsymbol{\theta}_i$'s are either uni- or multidimensional and that we can simulate from the conditional posterior densities $p^*(\boldsymbol{\theta}_i|\mathbf{x}, \boldsymbol{\theta}_j, j \neq i)$. The Gibbs sampler generates a Markov chain by cycling through $p^*(\boldsymbol{\theta}_i|\mathbf{x}, \boldsymbol{\theta}_j, j \neq i)$. Starting from some $\boldsymbol{\theta}^{(0)}$, after t cycles we have a realization $\boldsymbol{\theta}^{(t)}$ that under regularity conditions, approximates a drawing from $p^*(\boldsymbol{\theta}|\mathbf{x})$.

Thus, Gibbs sampling is applicable when the joint distribution of two or more random variables, is not known explicitly, but the conditional distribution of each variable is known. The algorithm starts by drawing the initial sample from an arbitrary (possibly degenerate) prior distribution, and then, generate an instance from the distribution of each variable in turn, conditional on the current values of the other variables ([31]).

3.1.4 Newton-type method

The Newton-Raphson (NR) method was discovered by Isaac Newton and published in his book Method of Fluxions in 1736. Joseph Raphson described this method in Analysis Aequationum in 1690. The NR approximates the gradient vector $S(\mathbf{y}; \boldsymbol{\theta})$ of the log-likelihood function $\log L(\boldsymbol{\theta})$ by a linear Taylor series expansion about the current fit $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. This gives

$$S(\mathbf{y}, \boldsymbol{\theta}) \approx S(\mathbf{y}; \boldsymbol{\theta}^{(k)}) - I(\boldsymbol{\theta}^{(k)}; \mathbf{y})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}), \tag{3.4}$$

where I is given in (3.2).

A new fit $\boldsymbol{\theta}^{(k+1)}$ is obtained by solving the system of equations of (3.4) knowing $\boldsymbol{\theta}^{(k)}$. Hence

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + I^{-1}(\boldsymbol{\theta}^{(k)}; \mathbf{y}) S(\mathbf{y}; \boldsymbol{\theta}^{(k)}). \tag{3.5}$$

If the log-likelihood function is concave and unimodal, then the sequence of iterates $\{\boldsymbol{\theta}^{(k)}\}$ converges to the MLE of $\boldsymbol{\theta}$, but if the log-likelihood function is not concave, the NR method is not guaranteed to converge from an arbitrary starting value. Under reasonable assumptions on $L(\boldsymbol{\theta})$ and a sufficiently accurate starting value, the sequence of $\boldsymbol{\theta}^{(k)}$ produced by the NR method converges to a solution $\boldsymbol{\theta}^*$ of $S(\mathbf{y}; \boldsymbol{\theta}) = 0$. That is, given a norm there is a constant h such that if $\boldsymbol{\theta}^{(0)}$ is sufficiently close to $\boldsymbol{\theta}^*$, then

$$\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^* \| \leq h \| \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^* \|^2$$

holds for $k = 0, 1, 2, \ldots$. Quadratic convergence is ultimately very fast.

A broad class of methods are the so-called quasi-Newton methods, for which the solution of (3.5) takes the form

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{A}^{-1} S(\mathbf{y}; \boldsymbol{\theta}^{(k)}), \tag{3.6}$$

where $\mathbf{A}$ is an approximation to the Hessian matrix. This approximation can be maintained by doing a secant update of $\mathbf{A}$ at each iteration. Methods of this class have the advantage over the NR method of not requiring the explicit evaluation of the Hessian matrix at each iteration.

3.2 Fitting continuous phase-type distributions

Asmussen *et.al* in [11] have presented a fitting procedure for continuous phase-type (CPH) distributions via the EM algorithm. In this Section, we develop an alternative way of computing the E-step in the EM algorithm using the uniformization method (see [40]), which we call the EM unif algorithm.

A crucial part of the estimation of phase-type distributions via Markov chain Monte Carlo methods, in particular via the Gibbs sampler method (see [15]) is the simulation of the underlying Markov jump process. More precisely, for an observation from a phase-type distribution, we establish an algorithm for simulating from the conditional distribution of the underlying Markov jump process given the absorption time using the uniformization method (we denote this method by GS unif, see also [14]).

As a third method of estimation, we consider the Newton-Raphson method. In this work we refer it as the direct method (DM) (see also [48]).

3.2.1 Preliminaries

Consider $y_1, \ldots, y_M$ a realization of i.i.d random variables from $PH_p(\boldsymbol{\pi}, \mathbf{T})$. We are in a situation of incomplete information since we only have the absorption times and not the entire underlying structure is available.

Let $\mathbf{y} = (y_1, \ldots, y_M)$ and $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$, where $\mathbf{t} = -\mathbf{T}\mathbf{e}$. The incomplete data likelihood is given by

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{k=1}^{M} \boldsymbol{\pi} e^{\mathbf{T} y_k} \mathbf{t}, \tag{3.7}$$

and the log-likelihood function is

$$l(\boldsymbol{\theta};\mathbf{y}) \quad = \quad \sum_{k=1}^{M} \log f(y_k),$$

where $f(y_k) = \boldsymbol{\pi} e^{\mathbf{T}y_k}\mathbf{t}$. Substituting $\boldsymbol{\pi} = \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p'$ then

$$f(y_k) \quad = \quad \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' e^{\mathbf{T}y_k}\mathbf{t} + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p' e^{\mathbf{T}y_k}\mathbf{t}.$$

As a starting point we assume that we have got one complete observation of a Markov jump process $\{X(t)\}_{t\geq 0}$ with p states. Suppose the time until absorption is $y \in \{y_1, \ldots, y_M\}$, with n jumps to place before absorption, the sequence of states visited is $i_0, i_1, \ldots, i_n$ (here repetitions are obviously permitted), and the time spent between each of the jumps were $s_0, s_1, \ldots, s_n$, i.e., $s_0 + s_1 + \cdots + s_n = y$. In order to find the maximum likelihood estimate of $\boldsymbol{\theta}$ from the observed data, let $\mathbf{x} = \{\mathbf{x}_i\}_{i=1,\ldots,M}$ denote the full data for the M absorption times, thus the $\mathbf{x}_i$'s are trajectories of the underlying MJP. The likelihood function for the complete data is given by

$$L_f(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{p} \pi_i^{B_i} \prod_{i=1}^{p}\prod_{j\neq i}^{p} t_{ij}^{N_{ij}} e^{-t_{ij}Z_i} \prod_{i=1}^{p} t_i^{N_i} e^{-t_i Z_i}, \tag{3.8}$$

where B_i is the number of processes starting in state i, N_i the number of processes exiting from state i to the absorbing state, N_{ij} the number of jumps from state i to j among all processes, and Z_i the total time spent in state i prior to absorption for all processes.

3.2.2 The EM algorithm: CPH

Since the data $\mathbf{y} = (y_1, \ldots, y_M)$ are incomplete, in the following we shall describe a method for calculating the maximum likelihood estimators using the EM algorithm. We follow Asmussen *et.al* [11] which may be consulted for further details.

The log-likelihood function for the complete data is given by

$$\begin{aligned} l_f(\boldsymbol{\theta};\mathbf{x}) \quad = \quad & \sum_{i=1}^{p} B_i \log(\pi_i) + \sum_{i=1}^{p}\sum_{j\neq i}^{p} N_{ij} \log(t_{ij}) \\ & - \sum_{i=1}^{p}\sum_{j\neq i}^{p} t_{ij} Z_i + \sum_{i=1}^{p} N_i \log(t_i) - \sum_{i=1}^{p} t_i Z_i. \end{aligned} \tag{3.9}$$

It is immediately clear that the maximum likelihood estimators for t_{ij} and t_i are given by

$$\hat{t}_{ij} = \frac{N_{ij}}{Z_i}, \quad \hat{t}_i = \frac{N_i}{Z_i}.$$

Slightly more care has to be taken with the π_i's since they must sum to one. Applying Lagrange multipliers we get that a maximum likelihood estimator for π_i is

$$\hat{\pi}_i = \frac{B_i}{M}.$$

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0)$ denote any initial value of the parameters. The EM works as follows.

1. (E-step) Calculate the function

$$h : \boldsymbol{\theta} \to \mathbb{E}_{\boldsymbol{\theta}_0}(l_f(\boldsymbol{\theta}; \mathbf{x})|\mathbf{Y} = \mathbf{y}).$$

2. (M-step)

$$\boldsymbol{\theta}_0 = \text{argmax}_{\boldsymbol{\theta}} h(\boldsymbol{\theta}).$$

3. Goto (1).

The E-step and M-step are repeated until convergence.

Since (3.9) is a linear function of the sufficient statistics B_i, Z_i, N_i, and N_{ij}, it is enough to calculate the corresponding conditional expectations of these statistics. Let B_i^k, Z_i^k, N_i^k, and N_{ij}^k be the corresponding statistics for the k-th observation, then

$$B_i = \sum_{k=1}^{M} B_i^k, \quad Z_i = \sum_{k=1}^{M} Z_i^k, \quad N_i = \sum_{k=1}^{M} N_i^k, \quad N_{ij} = \sum_{k=1}^{M} N_{ij}^k,$$

for $i, j = 1, \ldots, p$, $i \neq j$, and hence $\mathbb{E}_{\boldsymbol{\theta}}(S|\mathbf{Y} = \mathbf{y}) = \sum_{k=1}^{M} \mathbb{E}_{\boldsymbol{\theta}}(S^k|Y_k = y_k)$, where $S \in \{B_i, Z_i, N_i, N_{ij}\}$. The main task lies in calculating $\mathbb{E}_{\boldsymbol{\theta}}(S^k|Y_k = y_k)$, if these expectations are known then we can easily calculate for more than one data point simply by summing.

The proof of the following theorem can be found in [11].

Theorem 3.2 *For $i, j = 1, \ldots, p$, $i \neq j$, we have*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}(B_i^k | Y_k = y_k) &= \frac{\pi_i \boldsymbol{e}_i' \exp(\boldsymbol{T} y_k)\boldsymbol{t}}{\boldsymbol{\pi} \exp(\boldsymbol{T} y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(Z_i^k | Y_k = y_k) &= \frac{\int_0^{y_k} \boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_i \boldsymbol{e}_i' \exp(\boldsymbol{T}(y_k - u))\boldsymbol{t} du}{\boldsymbol{\pi} \exp(\boldsymbol{T} y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_i^k | Y_k = y_k) &= \frac{t_i \boldsymbol{\pi} \exp(\boldsymbol{T} y_k)\boldsymbol{e}_i}{\boldsymbol{\pi} \exp(\boldsymbol{T} y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}^k | Y_k = y_k) &= \frac{t_{ij} \int_0^{y_k} \boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_i \boldsymbol{e}_j' \exp(\boldsymbol{T}(y_k - u))\boldsymbol{t} du}{\boldsymbol{\pi} \exp(\boldsymbol{T} y_k)\boldsymbol{t}}.
\end{aligned}$$

EM using Runge-Kutta (EM-RK)

Asmussen *et.at* [11] considered the following. Let $\mathbf{a}(y|\boldsymbol{\theta}) = \boldsymbol{\pi} \exp(\mathbf{T}y)$, $\mathbf{b}(y|\boldsymbol{\theta}) = \exp(\mathbf{T}y)\mathbf{t}$, and $\mathbf{c}(y, i|\boldsymbol{\theta}) = \int_0^y \boldsymbol{\pi} \exp(\mathbf{T}u)\mathbf{e}_i \exp(\mathbf{T}(y-u))\mathbf{t} \mathrm{d}u$, $i = 1, \ldots, p$, where $\mathbf{e}_i$ is the i-th unit vector. Then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}(B_i^k | Y_k = y_k) &= \frac{\pi_i b_i(y_k|\boldsymbol{\theta})}{\boldsymbol{\pi}\mathbf{b}(y_k|\boldsymbol{\theta})} \\
\mathbb{E}_{\boldsymbol{\theta}}(Z_i^k | Y_k = y_k) &= \frac{c_i(y_k, i|\boldsymbol{\theta})}{\boldsymbol{\pi}\mathbf{b}(y_k|\boldsymbol{\theta})} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_i^k | Y_k = y_k) &= \frac{t_i a_i(y_k|\boldsymbol{\theta})}{\boldsymbol{\pi}\mathbf{b}(y_k|\boldsymbol{\theta})} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}^k | Y_k = y_k) &= \frac{t_{ij} c_j(y_k, i|\boldsymbol{\theta})}{\boldsymbol{\pi}\mathbf{b}(y_k|\boldsymbol{\theta})}.
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{a}'(y|\boldsymbol{\theta}) &= \mathbf{a}(y|\boldsymbol{\theta})\mathbf{T} \\
\mathbf{b}'(y|\mathbf{T}) &= \mathbf{T}\mathbf{b}(y|\boldsymbol{\theta}) \\
\mathbf{c}'(y, i|\boldsymbol{\theta}) &= \mathbf{T}\mathbf{c}(y, i|\boldsymbol{\theta}) + a_i(y|\boldsymbol{\theta})\mathbf{t}, \qquad i = 1, \ldots, p.
\end{aligned}$$

By combining these equations with the initial conditions $\mathbf{a}(0|\boldsymbol{\theta}) = \boldsymbol{\pi}$, $\mathbf{b}(0|\boldsymbol{\theta}) = \mathbf{t}$, and $\mathbf{c}(0, i|\boldsymbol{\theta}) = \mathbf{0}$ for $i = 1, \ldots, p$, we can solve the system numerically, using some standard method. In the EMPHT-program, given by the authors, the Runge-Kutta method of fourth order is implemented for this purpose.

EM using uniformization

First of all, we will explain how the method of uniformization works (see [40]). Consider a Markov process $\{X(t)\}_{t\geq 0}$ with generator $\mathbf{\Lambda}$, where its diagonal elements are given by λ_{ii}, such that $|\lambda_{ii}| \leq c < \infty$ (all i) for some constant c, that automatically holds when there are only finitely many states. Then, the matrix $\mathbf{K} = \frac{1}{c}\mathbf{\Lambda} + \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, is a stochastic matrix. Now, define the stochastic process $\{Y(t)\}_{t\geq 0}$ as follows. Take a Poisson process with rate c and denote by $0 = T_0, T_1, T_2, \ldots$ the epochs of events in the process. Take a discrete time Markov chain $\{W_n\}_{n\geq 0}$ with transition matrix $\mathbf{K}$ independent of the Poisson process. Define the process $\{Y(t)\}_{t\geq 0}$ such as $Y(t) = W_n$ for $T_n \leq t < T_{n+1}$, $n \geq 0$. Not surprisingly, $\{Y(t)\}_{t\geq 0}$ happens to be a Markov process, and furthermore, its generator is equal to $\mathbf{\Lambda}$. Algebraically, if we define the transition matrix $P(t) = \{p_{ij}^t\}$ where $p_{ij}^t = \mathbb{P}(Y(t) = j | Y(0) = i)$, we obtain by a simple conditioning argument on the number of Poisson events in $(0, t]$ that

$$P(t) \quad = \quad \sum_{n=0}^{\infty} e^{-ct} \frac{(ct)^n}{n!} \mathbf{K}^n.$$

On the other hand,

$$\begin{aligned} \exp(\mathbf{\Lambda} t) &= \sum_{i=0}^{\infty} \frac{(\mathbf{\Lambda} t)^i}{i!} \\ &= \sum_{i=0}^{\infty} (ct)^i \frac{\left(\left(\frac{1}{c}\mathbf{\Lambda} + \mathbf{I}\right) - \mathbf{I}\right)^i}{i!} \\ &= \sum_{i=0}^{\infty} \frac{(ct)^i}{i!} e^{-ct} \mathbf{K}^i \\ &= P(t), \end{aligned}$$

which is the transition matrix of the process $\{Y(t)\}_{t\geq 0}$.

It allows us to interpret a continuous time Markov process as a discrete time Markov chain, for which we merely replace the constant unit of time between any two transitions by independent exponential random variables with the same parameter, hence the term uniformization.

Now, consider $y \in \{y_1, \ldots, y_M\}$ with generator $\mathbf{\Lambda}$ given in (2.1). Choosing $c = \max\{-t_{ii} : 1 \leq i \leq p\}$, the matrix $\mathbf{K} = \frac{1}{c}\mathbf{\Lambda} + \mathbf{I}$ has the form

$$\mathbf{K} = \begin{pmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{0} & 1 \end{pmatrix},$$

where $\mathbf{P} = \frac{1}{c}\mathbf{T} + \mathbf{I}$ and $\mathbf{p} = \frac{1}{c}\mathbf{t}$. Now we readily obtain that

$$\exp(\mathbf{T}x) = \sum_{i=0}^{\infty} e^{-cx} \frac{(cx)^i}{i!} \mathbf{P}^i.$$

Based on this, we calculate the integral

$$\int_0^y \boldsymbol{\pi} e^{\mathbf{T}u} \mathbf{e}_i \mathbf{e}_j' e^{\mathbf{T}(y-u)} \mathbf{t} \mathrm{d}u = \int_0^y \mathbf{e}_j' e^{\mathbf{T}(y-u)} \mathbf{t}\boldsymbol{\pi} e^{\mathbf{T}u} \mathbf{e}_i \mathrm{d}u,$$

seen as a matrix,

$$\begin{aligned}
\mathbf{J}(y) &= \int_0^y e^{\mathbf{T}(y-u)} \mathbf{t}\boldsymbol{\pi} e^{\mathbf{T}u} \mathrm{d}u \\
&= \int_0^y \left(e^{-c(y-u)} \sum_{k=0}^{\infty} \frac{(c\mathbf{K}(y-u))^k}{k!} \right) \mathbf{t}\boldsymbol{\pi} \left(e^{-cu} \sum_{j=0}^{\infty} \frac{(c\mathbf{K}u)^j}{j!} \right) \mathrm{d}u \\
&= e^{-cy} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left(\int_0^y \frac{(cu)^j}{j!} \frac{(c(y-u))^k}{k!} \mathrm{d}u \right) \mathbf{K}^j \mathbf{t}\boldsymbol{\pi} \mathbf{K}^k \\
&= e^{-cy} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{c^{j+k}}{j!k!} \left(\int_0^y u^j (y-u)^k \mathrm{d}u \right) \mathbf{K}^j \mathbf{t}\boldsymbol{\pi} \mathbf{K}^k \\
&= e^{-cy} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{c^{j+k}}{j!k!} \left(\int_0^1 (yu)^j (y-yu)^k y \mathrm{d}u \right) \mathbf{K}^j \mathbf{t}\boldsymbol{\pi} \mathbf{K}^k \\
&= e^{-cy} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{c^{j+k} y^{j+k+1}}{j!k!} \left(\int_0^1 u^j (1-u)^k \mathrm{d}u \right) \mathbf{K}^j \mathbf{t}\boldsymbol{\pi} \mathbf{K}^k.
\end{aligned}$$

Moreover, the beta function, also called the Euler integral of the first kind, is a special function defined by

$$\beta(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1} \mathrm{d}u = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where Γ is the gamma function. Then

$$\beta(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

Thus, $\mathbf{J}(y)$ can be written as

$$\begin{aligned}
\mathbf{J}(y) &= e^{-cy}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(cy)^{j+k+1}}{j!k!}\beta(j+1,k+1)\mathbf{K}^j\frac{\mathbf{t}}{c}\boldsymbol{\pi}\mathbf{K}^k \\
&= e^{-cy}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(cy)^{j+k+1}}{j!k!}\frac{j!k!}{(j+k+1)!}\mathbf{K}^j\frac{\mathbf{t}}{c}\boldsymbol{\pi}\mathbf{K}^k \\
&= e^{-cy}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(cy)^{j+k+1}}{(j+k+1)!}\mathbf{K}^j\mathbf{k}\boldsymbol{\pi}\mathbf{K}^k \\
&= e^{-cy}\sum_{m=0}^{\infty}\frac{(cy)^{m+1}}{(m+1)!}\sum_{j=0}^{m}\mathbf{K}^j\mathbf{k}\boldsymbol{\pi}\mathbf{K}^{m-j},
\end{aligned}$$

where $\mathbf{k} = \frac{1}{c}\mathbf{t}$.

The integral has the following probabilistic interpretation. The (i,j)-th entry of the matrix is the probability that a phase-type renewal process (see [10]) with interarrival distribution $\mathrm{PH}(\boldsymbol{\pi},\mathbf{T})$ starting from state i has exactly one arrival in $[0,y]$ and is in state j by time y. From this interpretation we derive the following recursive formula

$$\mathbf{J}(x+y) = e^{\mathbf{T}x}\mathbf{J}(y) + \mathbf{J}(x)e^{\mathbf{T}y}.$$

3.2.3 The Gibbs sampler algorithm: CPH

In this Section we present an alternative method for fitting phase-type distributions based on Bladt *et.al* [15].

We are interested in estimating the phase-type generator parameters given the data $\mathbf{y}$. Let $\mathbf{X} = (\{X(t)\}_{0\le t\le y_i})_{1\le i\le M}$ denote its underlying process. We shall be interested in the conditional distribution of $(\boldsymbol{\theta},\mathbf{X})$ given $\mathbf{Y} = \mathbf{y}$. We may simulate this distribution by constructing a Markov chain with a stationary distribution which coincide with this target distribution. A standard method is using a Gibbs sampler which amounts to the following scheme:

(1) Draw $\boldsymbol{\theta}$ given $\mathbf{X}$ and $\mathbf{y}$.

(2) Draw $\mathbf{X}$ given $\boldsymbol{\theta}$ and $\mathbf{y}$. Goto (1).

After a certain initial burn-in, the Markov chain will settle into stationary mode. Step (1) amounts to drawing parameters from the posterior distribution. The

second step requires the simulation of Markov jump processes which get absorbed exactly at times y_i, $i = 1, \ldots, M$.

If we choose a prior distribution with density proportional to

$$\phi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_i^{\beta_i - 1} \prod_{i=1}^{p} t_i^{\eta_i - 1} e^{-t_i \psi_i} \prod_{i=1}^{p} \prod_{j \neq i}^{p} t_{ij}^{\nu_{ij} - 1} e^{-t_{ij} \psi_i}, \tag{3.10}$$

it is easy to sample from this distribution since $\boldsymbol{\pi}$ is *Dirichlet* distributed with parameter $(\beta_1, \ldots, \beta_p)$, t_i is *Gamma* distributed with shape parameter η_i and scale parameter $1/\psi_i$, i.e. $t_i \sim Gamma(\eta_i, 1/\psi_i)$, and $t_{ij} \sim Gamma(\nu_{ij}, 1/\psi_i)$. For the choice of the prior distribution we refer to [14] and [15].

Thus, the posterior simply has the form

$$p^*(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{p} \pi_i^{B_i + \beta_i - 1} \prod_{i=1}^{p} t_i^{N_i + \eta_i - 1} e^{-t_i (Z_i + \psi_i)} \prod_{i=1}^{p} \prod_{j \neq i}^{p} t_{ij}^{N_{ij} + \nu_{ij} - 1} e^{-t_{ij} (Z_i + \psi_i)}, \tag{3.11}$$

with $\boldsymbol{\pi} \sim Dirichlet(B_1 + \beta_1, \ldots, B_p + \beta_p)$, $t_i \sim Gamma\left(N_i + \eta_i, \frac{1}{Z_i + \psi_i}\right)$, and $t_{ij} \sim Gamma\left(N_{ij} + \nu_{ij}, \frac{1}{Z_i + \psi_i}\right)$.

Drawing $\mathbf{X}$ given $(\boldsymbol{\theta}, \mathbf{y})$ is much involved. Given parameters $\boldsymbol{\theta}$ and absorption times $\mathbf{y}$ we must produced realizations of Markov jump processes with specified parameters which get absorbed exactly at times $\mathbf{y}$. Bladt *et.al* [15] applied a Metropolis-Hastings (MH) algorithm to simulate such Markov jump processes.

The Metropolis-Hastings algorithm provides a general approach for producing a correlated sequence of draws from a target density d that may be difficult to sample. The MH algorithm is defined by two steps: the first step in which a proposal value x' is drawn from the candidate generating density $q(x, x')$ and the second step in which the proposal value is accepted as the next iterate in the Markov process according to the probability

$$\min\left[1, \frac{d(x')q(x', x)}{d(x)q(x, x')}\right].$$

If the proposal value is rejected, then the next sampled value is taken to be the current value.

The MH algorithm amounts to the following simple procedure for simulating a Markov jump process j which gets absorbed exactly at time y.

ALGORITHM. *Metropolis-Hastings*

1. Draw a MJP j which is not absorbed by time y. This is done by simple rejection sampling: if a MJP is absorbed before time y it is thrown away and a new MJP is tried. We continue this way until we obtain the desired MJP.
2. Draw a new MJP j' as in step 1.
3. Draw $U \sim Unif(0,1)$.
4. If $U \leq \min(1, t_{j_{y^-}}/t_{j'_{y^-}})$ then $j = j'$, otherwise keep j.
5. Goto 2.

Here y^- denotes the limit from the left so j_{y^-} is the state just prior to exit. We iterate this procedure a number of times (burn-in) in order to get it into stationary mode. After this point and onwards, any j produced by the procedure may be considered as a draw from the desired conditional distribution and hence as a realization of a MJP which gets absorbed exactly at time y.

The full procedure Gibbs sampler is then as follows.

ALGORITHM. *Gibbs sampler with Metropolis-Hastings*

1. Draw initial parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$ from the prior distribution (3.10).
2. Draw the underlying Markov trajectories given $\boldsymbol{\theta}$ using the Metropolis-Hastings algorithm.
3. Draw the new parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$ from the posterior distribution (3.11).
4. Goto 2.

Gibbs sampler using uniformization

Our alternative algorithm for fitting phase-type distributions mainly differs on the simulation of the MJP, where we suggest to use uniformization instead of the Metropolis-Hastings algorithm (see also [30]).

The following algorithm shows how to simulate the underlying Markov jump process using uniformization.

ALGORITHM (*). *Simulation of a MJP using uniformization*

Input: $y \sim PH_p(\boldsymbol{\pi}, \mathbf{T})$.

1. Take $c = \max\{-t_{ii} : 1 \leq i \leq p\}$. Compute $\mathbf{P} = \frac{1}{c}\mathbf{T} + \mathbf{I}$.
2. Generate $N \sim Poisson(cy)$.
3. Simulate a Markov chain using the parameters $\boldsymbol{\pi}$ and $\mathbf{P}$, and the value of N as a time of absorption.
4. Find the time spent in each state s_i, $i = 0, 1, \ldots, N$, such as $\sum_{i=0}^{N} s_i = y$.

Note 3.3 *In the step 3, we can use reversed Markov chain in order to speed up the algorithm (see Section 2.4.2).*

In the following we will explain step 4 of this algorithm in more detail.

For $i = 0, 1, \ldots, N$, if $S_i \sim \exp(c)$, i.e. $S_i \sim Gamma(1, c)$, then $y = \sum_{i=0}^{N} S_i \sim Gamma(N + 1, c)$.

- If $N = 0$, then obviously $s_0 = y$.
- If $N \geq 1$, then we have that

$$f_{S_0,S_1,\ldots,S_{N-1}|\sum_{i=0}^{N} S_i}(s_0, s_1, \ldots, s_{N-1}|y) = \frac{f_{S_0,S_1,\ldots,S_{N-1},\sum_{i=0}^{N} S_i}(s_0, s_1, \ldots, s_{N-1}, y)}{f_{\sum_{i=0}^{N} S_i}(y)}.$$

If $R_0 = S_0$, $R_1 = S_1$, $\ldots$, $R_{N-1} = S_{N-1}$, and $R_N = S_0 + S_1 + \cdots + S_N$ then

$$\begin{aligned} f_{R_0,R_1,\ldots,R_N}(r_0, r_1, \ldots, r_N) &= f_{S_0,S_1,\ldots,S_N}(s_0, s_1, \ldots, s_N) \\ &= f_{S_0}(r_0) f_{S_1}(r_1) f_{S_2}(r_2) \cdots f_{S_N}\left(r_N - \sum_{j=0}^{N-1} r_j\right) \\ &= c^{N+1} e^{-cr_N}, \end{aligned}$$

since $r_N = y$, we get

$$f(s_0, \ldots, s_{N-1}, y) = f_{S_0,S_1,\ldots,S_{N-1},\sum_{i=0}^{N} S_i}(s_0, \ldots, s_{N-1}, y) = c^{N+1} e^{-cy},$$

and

$$\begin{aligned} f(s_0,\ldots,s_{N-1}|y) &= f_{S_0,S_1,\ldots,S_{N-1}|\sum_{i=0}^{N} S_i}(s_0,s_1,\ldots,s_{N-1}|y) \\ &= \frac{c^{N+1}e^{-cy}}{\frac{c}{N!}(cy)^N e^{-cy}} \\ &= \frac{N!}{y^N}. \end{aligned}$$

For $i = 0, 1, \ldots, N-1$, the general form of the conditional marginal distributions is given by

$$\begin{aligned} f(s_i|y) &= \int\cdots\int \frac{N!}{y^N}\mathrm{d}s_0\cdots\mathrm{d}s_{i-1}\mathrm{d}s_{i+1}\cdots\mathrm{d}s_{N-1} \\ &= \frac{N!}{y^N}\frac{(y-s_i)^{N-1}}{(N-1)!} \\ &= \frac{N}{y^N}(y-s_i)^{N-1}. \end{aligned} \tag{3.12}$$

Another way of getting this distribution is using the following argument which turns out to be simpler.

Consider $U_1,\ldots,U_N \sim Unif(0,y)$, and let $U_{(1)},\ldots,U_{(N)}$ be their order statistics. The joint pdf of $U_{(k)}$ and $U_{(j)}$, $1 \le k \le j \le N$, is given by

$$\begin{aligned} f_{U_{(k)},U_{(j)}}(u,v) &= \frac{N!}{(k-1)!(j-1-k)!(N-j)!} f_U(u) f_U(v) (F_U(u))^{k-1} \\ &\quad \times (F_U(v)-F_U(u))^{j-1-k}(1-F_U(v))^{N-j}, \end{aligned} \tag{3.13}$$

where $f_U(u) = \frac{1}{y}$, $F_U(u) = \frac{u}{y}$ for $u \in (0,y)$, and $U_{(0)} = 0$, $U_{(N+1)} = y$.

In general, for $i = 0, 1, \ldots, N-1$, we have

$$f_{U_{(i)},U_{(i+1)}}(u_i,u_{i+1}|y) = \frac{N!}{(i-1)!(N-i-1)!y^{i+1}} u_i^{i-1}\left(1-\frac{u_{i+1}}{y}\right)^{N-i-1}.$$

For $j = 0, 1, \ldots, N$, let $S_j = U_{(j+1)} - U_{(j)}$, then

$$f_{U_{(i)},S_i}(u,s|y) = \frac{N!}{(i-1)!(N-i-1)!y^{i+1}} u^{i-1}\left(1-\frac{s+u}{y}\right)^{N-i-1},$$

where $0 < u < y-s$. Thus, the marginal of S_i is given by

$$\begin{aligned} f_{S_i}(s|y) &= \int_0^{y-s} \frac{N!}{(i-1)!(N-i-1)!y^{i+1}} u^{i-1}\left(1-\frac{s+u}{y}\right)^{N-i-1}\mathrm{d}u \\ &= \frac{N}{y^N}(y-s)^{N-1}. \end{aligned}$$

Finally, for $N = 0$ we take $s_0 = y$, and if $N \geq 1$, $f(s_i|y) = \frac{N}{y^N}(y - s_i)^{N-1}$, for $i = 0, 1, \ldots, N-1$. Note that this density is the same as we presented in (3.12).

The following algorithm shows how to find the time spent in each state of the Markov chain (step 4 in ALGORITHM (*)).

ALGORITHM. *Time spent in each state of a Markov chain*

Input: N, y.

1. Generate N random numbers $U_1, \ldots, U_N$ from the uniform distribution, $Unif(0, y)$.
2. Find the order statistics $U_{(1)}, \ldots, U_{(N)}$.
3. For $i = 0, 1, \ldots, N$, calculate $s_i = U_{(i+1)} - U_{(i)}$, where $U_{(0)} = 0$ and $U_{(N+1)} = y$.

Hence, our algorithm to estimate PH distributions via the GS works as follows.

ALGORITHM. *Gibbs sampler using uniformization*

Input: $y_i \sim PH_p(\boldsymbol{\pi}, \mathbf{T})$; $i = 1, \ldots, M$.

1. Draw initial parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$ from the prior distribution (3.10).
2. Generate $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ where each $\mathbf{X}_i$ is a Markov jump process which gets absorbed at time y_i, obtained using uniformization (ALGORITHM (*), with $y_i \sim PH_p(\boldsymbol{\pi}, \mathbf{T})$).

   Calculate the statistics B_i, N_i, N_{ij}, Z_i; $i, j = 1, \ldots, p, i \neq j$.
3. Draw the new parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$ from the posterior distribution (3.11).
4. Goto 2.

3.2.4 Direct method: CPH

The maximum likelihood estimation of PH distributions can be interpreted as the solution of a system of non-linear equations. The most celebrated of all methods for solving a non-linear equation is the Newton-Raphson method. This is based on the idea of approximating the gradient vector, $\boldsymbol{g}$, with its linear Taylor series expansion about a working value x_k. Let $\boldsymbol{G}(x)$ be the matrix of partial derivatives of $\boldsymbol{g}(x)$ with respect to x. Using the root of the linear expansion as the new approximation gives

$$x_{k+1} = x_k - \boldsymbol{G}(x_k)^{-1}\boldsymbol{g}(x_k).$$

The same algorithm arises for minimizing $h(x)$ by approximating h with its quadratic Taylor series expansion about x_k. In the minimization case, $\boldsymbol{g}(x)$ is the derivate vector (gradient) of $h(x)$ with respect to x and the second derivate matrix $\boldsymbol{G}(x)$ is symmetric. If h is a log-likelihood function, then $\boldsymbol{g}$ is the score vector and $-\boldsymbol{G}$ is the observed information matrix. This method is not designed to work with boundary conditions. For this, we consider the unconstrained optimization given by Madsen *et.al* [41], where we have to give the explicit expression of the gradient vector with required transformations. We refer to this method as the Direct Method (DM) since it does not use the underlying probabilistic structure.

Here, we will use the log transformation, which it is the only member of the Box-Cox [23] family of transformations for which the transform of a positive-valued variable can be truly Normal, because the transformed variable is defined over the whole of the range from $-\infty$ to ∞.

For $i = 1, \ldots, p-1$, generate $-\infty < \varrho_i < \infty$, and take the following transformation

$$\pi_i = \frac{e^{\varrho_i}}{1 + \sum_{s=1}^{p-1} e^{\varrho_s}} \quad \text{and} \quad \pi_p = \frac{1}{1 + \sum_{i=1}^{p-1} e^{\varrho_i}},$$

and for $i, j = 1, \ldots, p$, generate $-\infty < \gamma_{ij} < \infty$ such that

$$t_{ij} = e^{\gamma_{ij}}, \quad i \neq j, \quad \text{and} \quad t_i = e^{\gamma_{ii}}.$$

The gradient vector is given by

$$\left(\left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \varrho_i} \right)_{i=1,\ldots,p-1}, \left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \gamma_{ij}} \right)_{i,j=1,\ldots,p} \right),$$

where

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \tau^*} = \sum_{k=1}^{M} \frac{1}{f(y_k)} \frac{\partial f(y_k)}{\partial \tau^*}, \quad \tau^* \in \{\varrho_i, \gamma_{ij}\}. \tag{3.14}$$

If $R_m(y_k) = \mathbf{e}'_m e^{\mathbf{T}y_k}\mathbf{t}$, then

$$\frac{\partial f(y_k)}{\partial \varrho_m} = \sum_{s=1}^{p-1} \frac{\partial \pi_s}{\partial \varrho_m} R_s(y_k) - \left(\sum_{s=1}^{p-1} \frac{\partial \pi_s}{\partial \varrho_m} \right) R_p(y_k), \tag{3.15}$$

where

$$\frac{\partial \pi_i}{\partial \varrho_j} = \pi_j \mathbf{1}_{\{j=i\}} - \pi_i \pi_j, \tag{3.16}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

Moreover,

$$\frac{\partial f(y_k)}{\partial \gamma_{ij}} = \sum_{s=1}^{p-1} \pi_s \frac{\partial R_s(y_k)}{\partial \gamma_{ij}} + \left(1 - \sum_{s=1}^{p-1} \pi_s \right) \frac{\partial R_p(y_k)}{\partial \gamma_{ij}}, \tag{3.17}$$

and

$$\frac{\partial R_s(y_k)}{\partial \gamma_{ij}} = \mathbf{e}'_s \frac{\partial e^{\mathbf{T}y_k}}{\partial \gamma_{ij}} \mathbf{t} + \mathbf{e}'_s e^{\mathbf{T}y_k} \frac{\partial \mathbf{t}}{\partial \gamma_{ij}},$$

where

$$\frac{\partial \mathbf{t}}{\partial \gamma_{ij}} = \mathbf{0}, \quad i \neq j, \qquad \text{and} \qquad \frac{\partial \mathbf{t}}{\partial \gamma_{ii}} = e^{\gamma_{ii}} \mathbf{e}_i.$$

In order to calculate $\frac{\partial e^{\mathbf{T}y_k}}{\partial \tau^*}$, for all τ^*, we are going to use uniformization. Let $\mathbf{K} = \mathbf{I} + \frac{1}{c}\mathbf{T}$, where $c = \max\{-t_{ii}, 1 \leq i \leq p\}$, then

$$e^{\mathbf{T}y} = \sum_{r=0}^{\infty} b_r \mathbf{K}^r,$$

where $y \in \{y_1, \ldots, y_M\}$ and $b_r = e^{-cy} \frac{(cy)^r}{r!}$. Taking the derivative we get that

$$\frac{\partial e^{\mathbf{T}y}}{\partial \tau^*} = \sum_{r=0}^{\infty} \left(b_r \frac{\partial \mathbf{K}^r}{\partial \tau^*} + \frac{\partial b_r}{\partial \tau^*} \mathbf{K}^r \right),$$

where

$$\frac{\partial b_r}{\partial \tau^*} = \frac{\partial c}{\partial \tau^*} y (b_{r-1} \mathbf{1}_{\{r>0\}} - b_r),$$

then

$$\begin{aligned}
\frac{\partial e^{\mathbf{T}y}}{\partial \tau^*} &= \sum_{r=0}^{\infty}\left(b_r\frac{\partial \mathbf{K}^r}{\partial \tau^*}+\frac{\partial c}{\partial \tau^*}y(b_{r-1}\mathbf{1}_{\{r>0\}}-b_r)\mathbf{K}^r\right)\\
&= \sum_{r=0}^{\infty}b_r\frac{\partial \mathbf{K}^r}{\partial \tau^*}+\frac{\partial c}{\partial \tau^*}y\sum_{r=0}^{\infty}b_{r-1}\mathbf{1}_{\{r>0\}}\mathbf{K}^r-\frac{\partial c}{\partial \tau^*}y\sum_{r=0}^{\infty}b_r\mathbf{K}^r\\
&= \sum_{r=0}^{\infty}b_r\frac{\partial \mathbf{K}^r}{\partial \tau^*}+\frac{\partial c}{\partial \tau^*}y\left(\sum_{r=0}^{\infty}b_r\mathbf{K}^r\right)\mathbf{K}-\frac{\partial c}{\partial \tau^*}y\sum_{r=0}^{\infty}b_r\mathbf{K}^r\\
&= \sum_{r=0}^{\infty}b_r\frac{\partial \mathbf{K}^r}{\partial \tau^*}+\frac{\partial c}{\partial \tau^*}y\left(\sum_{r=0}^{\infty}b_r\mathbf{K}^r\right)(\mathbf{K}-\mathbf{I})\\
&= \sum_{r=0}^{\infty}b_r\frac{\partial \mathbf{K}^r}{\partial \tau^*}+\frac{\partial c}{\partial \tau^*}ye^{\mathbf{T}y}(\mathbf{K}-\mathbf{I}). \qquad (3.18)
\end{aligned}$$

For $r \geq 1$ we have that

$$\frac{\partial \mathbf{K}^r}{\partial \tau^*}=\sum_{k=0}^{r-1}\mathbf{K}^k\frac{\partial \mathbf{K}}{\partial \tau^*}\mathbf{K}^{r-1-k},$$

and

$$\frac{\partial \mathbf{K}}{\partial \tau^*}=\frac{1}{c}\frac{\partial \mathbf{T}}{\partial \tau^*}-\frac{1}{c^2}\frac{\partial c}{\partial \tau^*}\mathbf{T}.$$

Assuming that the maximum of the diagonal of $-\mathbf{T}$ is given in the row k, then

$$\frac{\partial c}{\partial \gamma_{ij}}=\begin{cases}0 & \text{if } i\neq k,\forall j\neq i\\ e^{\gamma_{ij}} & \text{if } i=k,\forall j\neq i,\end{cases}\qquad \frac{\partial c}{\partial \gamma_{ii}}=\begin{cases}0 & \text{if } i\neq k\\ e^{\gamma_{ii}} & \text{if } i=k.\end{cases}$$

Finally, $\frac{\partial \mathbf{T}}{\partial \gamma_{ij}}$, $i\neq j$, is a matrix whose (r,s)-th element is given by

$$\left[\frac{\partial \mathbf{T}}{\partial \gamma_{ij}}\right]_{rs}=\begin{cases}0 & \text{if } i\neq r,\forall s,j\\ -e^{\gamma_{ij}} & \text{if } i=r,j\neq s\\ e^{\gamma_{rs}} & \text{if } i=r,j=s,\end{cases}$$

and $\frac{\partial \mathbf{T}}{\partial \gamma_{ii}}$ is a matrix whose (i,i)-th element is $-e^{\gamma_{ii}}$ and 0 otherwise.

3.2.5 Simulation results

In this Section we compare all the algorithms presented before. We ran the programs until $\frac{|LL_{i+1}-LL_i|}{|LL_i|}<10^{-15}$, where LL_i is the log-likelihood in the iteration i. For this purpose we consider the distributions given in Table 3.1.

The parameters for the Hyper-exponential distribution (see Table 2.2) are the following: $p_1 = 0.3$, $p_2 = 0.15$, $p_3 = 0.05$, $p_4 = 0.2$, $p_5 = 0.15$, $p_6 = 0.15$, and $\lambda_1 = 0.2$, $\lambda_2 = 0.8$, $\lambda_3 = 0.5$, $\lambda_4 = 0.7$, $\lambda_5 = 0.4$, $\lambda_6 = 0.3$.

Table 3.1: Distributions, number of phases, and size of data considered by the algorithms

Distribution	Phases	Observations
Exp(0.5)	3, 6, 9	200
Erlang(6,0.5)	3, 6, 9	200
Hyper-exponential	6	500
0.3*Erlang(4,0.075)+0.7*Erlang(2,0.35)	6	500

Table 3.2: Log-likelihood (LL) and execution time (time) for a Exp(0.5) distribution with 200 observations and considering dimensions 3, 6, and 9

Algorithm	3		6		9	
	LL	time	LL	time	LL	time
EM Unif	-337.324879	0.89	-337.264516	2.78	-337.211929	16.62
EM Unif Can	-337.205426	0.68	-337.149333	2.37	-337.147724	12.47
EM-RK	-337.855937	2.72	-337.701185	40.42	-337.698649	163.9
EM-RK Can	-337.201689	1.25	-337.158150	12.83	-337.144544	61.3
DM	-339.517482	235.75	-339.433725	528.64	-338.236541	612.35
DM Can	-339.461828	103.56	-338.414573	192.84	-337.126443	231.26
GS Unif	-339.653592	483.80	-339.448465	495.82	-338.826203	527.21
GS Unif Can	-339.135553	409.83	-339.025563	418.76	-337.398230	443.43
GS-MH	-339.852102	633.49	-339.614336	715.68	-338.212715	720.06
GS-MH Can	-339.482492	322.64	-339.023750	369.82	-337.065612	497.97

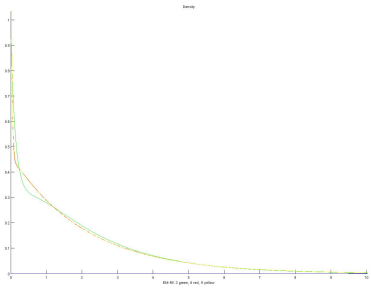

Figure 3.1: EM-RK, Exp(0.5)

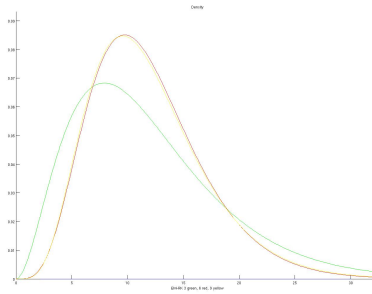

Figure 3.2: EM-RK, Erlang(6,0.5)

Table 3.3: Log-likelihood (LL) and execution time (time) for a Erlang(6,0.5) distribution with 200 observations and considering dimensions 3, 6, and 9

Algorithm	3		6		9	
	LL	time	LL	time	LL	time
EM Unif	-612.448668	0.49	-596.672830	4.56	-596.701870	12.68
EM Unif Can	-612.448668	0.26	-596.640231	4.33	-596.610579	12.41
EM-RK	-612.448517	0.81	-596.637344	5.79	-596.737192	45.46
EM-RK Can	-612.448517	0.69	-596.631987	4.62	-596.580838	16.60

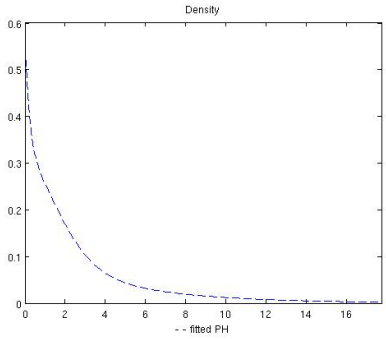

Figure 3.3: EM-RK, Hyper-exponential

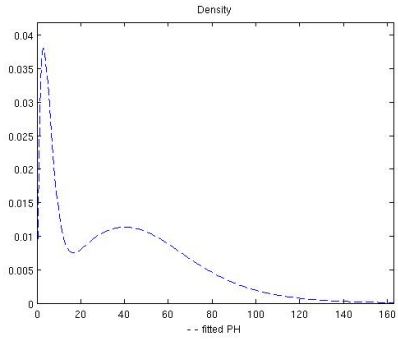

Figure 3.4: EM-RK, Mix-Erlang

Table 3.4: Log-likelihood (LL) and execution time (time) for a hyper-exponential and a mixture of Erlang distributions

Algorithm	Hyper-exponential		0.3*Erlang(4,0.075)+0.7*Erlang(2,0.35)	
	LL	time	LL	time
EM Unif	-1024.661717	9.96	-2321.917670	10.77
EM Unif Can	-1024.171364	9.55	-2286.619814	10.07
EM-RK	-1024.614153	41.17	-2316.991945	19.53
EM-RK Can	-1024.418559	17.57	-2286.542547	9.49

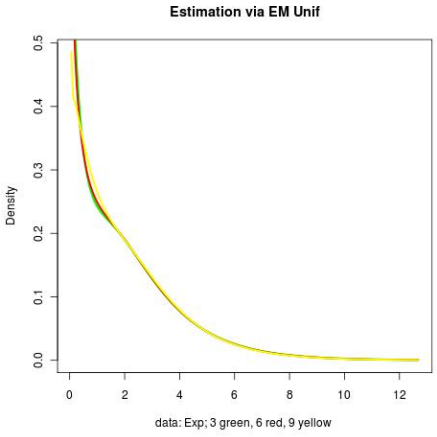

Figure 3.5: EM Unif, Exp(0.5)

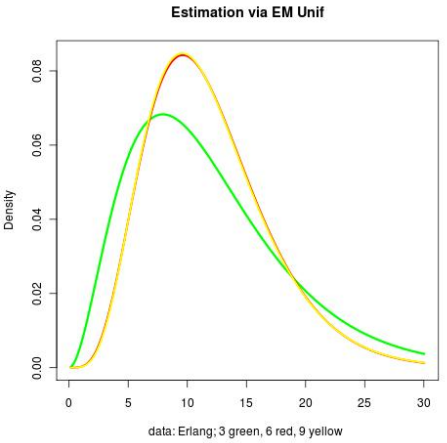

Figure 3.6: EM Unif, Erlang(6,0.5)

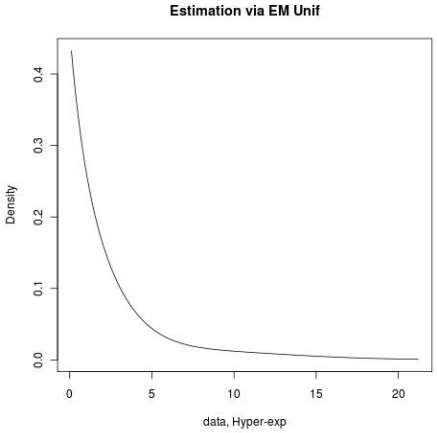

Figure 3.7: EM Unif, Hyper-exponential

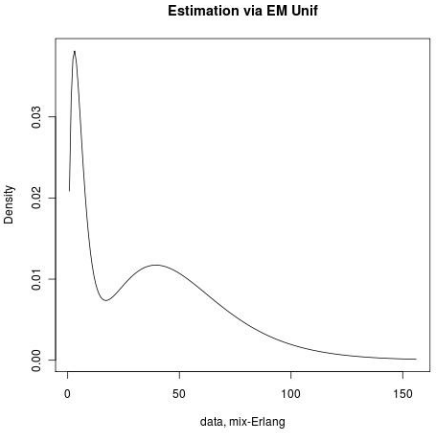

Figure 3.8: EM Unif, Mix-Erlang

3.3 Fitting discrete phase-type distributions

In this Section we apply three different methods for maximum likelihood estimation of discrete phase-type (DPH) distributions: an EM algorithm, a Gibbs sampler algorithm, and a Quasi-Newton method, where the last two methods are developed for the first time to fit DPH. We compare all of them considering

as a point of comparison their execution times. We propose some alternatives of these algorithms to accelerate them, using canonical form and reversed Markov chains.

We use an EM algorithm because of its simplicity in many applications and its desirable convergence properties. Its methodology is almost identical to the well known EM algorithm for continuous time ([11], [60]).

Nielsen and Beyer [48] presented a maximum likelihood method (Quasi-Newton method) based on counts with explicit calculation of the Fisher information matrix for an Interrupted Poisson process. Knowing this, we propose a new Quasi-Newton method, which we call direct method (DM), to estimate general and acyclic DPH.

3.3.1 Preliminaries

Consider M observations $y_1, \ldots, y_M \in \mathbb{N}$ from a $DPH_p(\boldsymbol{\pi}, \mathbf{T})$, where $\boldsymbol{\pi}$ and $\mathbf{T}$ are given as in Section 2.3. We assume that the data are independent. Initially we shall assume that $\pi_{p+1} = 0$, hence the data cannot contain zeros. Thus, y_k is the time of absorption of a Markov chain and we assume that only the absorption times are observable and not the underlying development of the Markov chains.

For each time of absorption y_k, we denote by $\mathbf{x}^{(k)} = (x_0^{(k)}, x_1^{(k)}, \ldots, x_{y_k}^{(k)})$ the sample path of the underlying Markov chain. Let $\mathbf{x} = \{\mathbf{x}^{(k)}\}_{k=1,\ldots,M}$ be the set of complete data, and let $\mathbf{y} = (y_1, \ldots, y_M)$ denote the set of incomplete observed data.

For $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$, the likelihood function is given by

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{k=1}^{M} \boldsymbol{\pi} \mathbf{T}^{y_k - 1} \mathbf{t}, \tag{3.19}$$

and the log-likelihood is

$$l(\boldsymbol{\theta}; \mathbf{y}) \quad = \quad \sum_{k=1}^{M} \log f(y_k),$$

where $f(y_k) = \boldsymbol{\pi} \mathbf{T}^{y_k - 1} \mathbf{t}$. Substituting $\boldsymbol{\pi} = \sum_{s=1}^{p-1} \pi_s \mathbf{e}_s' + \left(1 - \sum_{s=1}^{p-1} \pi_s\right) \mathbf{e}_p'$ we get

$$f(y_k) = \sum_{s=1}^{p-1} \pi_s \mathbf{e}_s' \mathbf{T}^{y_k - 1} \mathbf{t} + \left(1 - \sum_{s=1}^{p-1} \pi_s\right) \mathbf{e}_p' \mathbf{T}^{y_k - 1} \mathbf{t}.$$

If $R_m(y_k) = \mathbf{e}'_m \mathbf{T}^{y_k - 1}\mathbf{t}$, then

$$f(y_k) = \sum_{j=1}^{p-1} \pi_j R_j(y_k) + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) R_p(y_k). \tag{3.20}$$

Now consider the data from one single chain $\mathbf{x}^* \in \{\mathbf{x}^{(k)}\}_{k=1,\ldots,M}$ and suppose that y is the time of absorption. The complete likelihood function can be written in the following form

$$L_f(\boldsymbol{\theta}; \mathbf{x}^*) = \prod_{i=1}^{p} \pi_i^{B_i} \prod_{i=1}^{p} \prod_{j=1}^{p} t_{ij}^{N_{ij}} \prod_{i=1}^{p} t_i^{N_i}, \tag{3.21}$$

where B_i is equal to 1 if the Markov chain $\{X(n)\}_{n\geq 0}$ starts in the state i, and 0 otherwise, i.e., $B_i = \mathbf{1}_{\{X(0)=i\}}$; N_{ij} is the number of transitions from state i to state j, $i, j = 1, \ldots, p$; and $N_i = \mathbf{1}_{\{X(y-1)=i\}}$.

The log-likelihood function l_f is hence given by

$$l_f(\boldsymbol{\theta}; \mathbf{x}^*) = \sum_{i=1}^{p} B_i \log(\pi_i) + \sum_{i=1}^{p} \prod_{j=1}^{p} N_{ij} \log(t_{ij}) + \sum_{i=1}^{p} N_i \log(t_i). \tag{3.22}$$

Since we have M independent series of observations of the above type, then

$$B_i = \sum_{k=1}^{M} B_i^k, \quad N_i = \sum_{k=1}^{M} N_i^k, \quad N_{ij} = \sum_{k=1}^{M} N_{ij}^k,$$

where B_i^k, N_i^k, and N_{ij}^k are the corresponding statistics for the k-th observation.

3.3.2 The EM algorithm: DPH

Like in CPH, we are hence dealing with a case of incomplete information and our goal is to develop an EM algorithm for maximizing the likelihood of the data (see also [24]).

A key step in the development of the EM algorithm is to consider the full data likelihood L_f. The full log-likelihood is easily maximized by applying the method of Lagrange multipliers:

$$LA(\pi_i, t_{ij}, t_i, \lambda_1, \lambda_2) = l_f(\boldsymbol{\theta}; \mathbf{x}) + \lambda_1 \left(1 - \sum_{j=1}^{p} t_{ij} - t_i\right) + \lambda_2 \left(1 - \sum_{i=1}^{p} \pi_i\right),$$

then

$$\begin{array}{llll} \frac{\partial LA}{\partial \lambda_1} & = 1 - \sum_{j=1}^{p} t_{ij} - t_i & = 0 & \Longrightarrow \quad \sum_{j=1}^{p} t_{ij} + t_i = 1, \\ \frac{\partial LA}{\partial \lambda_2} & = 1 - \sum_{i=1}^{p} \pi_i & = 0 & \Longrightarrow \quad \sum_{j=1}^{p} \pi_i = 1, \end{array}$$

add,

$$\frac{\partial LA}{\partial \pi_i} = \frac{B_i}{\pi_i} - \lambda_2 = 0 \quad \Longrightarrow \quad \lambda_2 = \frac{B_i}{\pi_i}, \quad \Longrightarrow \quad \lambda_2 = \sum_{i=1}^{p} B_i = M,$$

and,

$$\begin{array}{llll} \frac{\partial LA}{\partial t_{ij}} & = \frac{N_{ij}}{t_{ij}} - \lambda_1 & = 0 & \Longrightarrow \quad \lambda_1 = \frac{N_{ij}}{t_{ij}}, \\ \frac{\partial LA}{\partial t_i} & = \frac{N_i}{t_i} - \lambda_1 & = 0 & \Longrightarrow \quad \lambda_1 = \frac{N_i}{t_i}, \end{array}$$

then,

$$\frac{N_{ij}}{t_{ij}} = \frac{N_i}{t_i} \quad \Longrightarrow \quad \sum_{j=1}^{p} N_{ij} t_i = \sum_{j=1}^{p} N_i t_{ij} \quad \Longrightarrow \quad t_i \sum_{j=1}^{p} N_{ij} = N_i(1 - t_i),$$

we obtain,

$$\hat{t}_i = \frac{N_i}{\sum_{j=1}^{p} N_{ij} + N_i}, \tag{3.23}$$

then

$$\lambda_1 = \sum_{j=1}^{p} N_{ij} + N_i,$$

and

$$\hat{t}_{ij} = \frac{N_{ij}}{\sum_{s=1}^{p} N_{is} + N_i}. \tag{3.24}$$

Finally,

$$\hat{\pi}_i = \frac{B_i}{\lambda_2} = \frac{B_i}{M}. \tag{3.25}$$

Since the log-likelihood function is linear in the sufficient statistics B_i, N_{ij}, and N_i, it is straightforward to calculate its conditional expectation if the corresponding conditional expectations of the sufficient statistic were known. In the following, we calculate these conditional expectations. We derive formulae for one data point y only, the general case then follows by summing the conditional expectations over all data points $y_1, \ldots, y_M$.

First we notice that $B_i = \mathbf{1}_{\{X(0)=i\}}$, then

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}(B_i | Y = y) &= \mathbb{P}(X(0) = i | Y = y) \\ &= \frac{\mathbb{P}(Y = y | X(0) = i)\mathbb{P}(X(0) = i)}{\mathbb{P}(Y = y)} \\ &= \frac{\mathbf{e}_i' \mathbf{T}^{y-1} \mathbf{t} \pi_i}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}}. \end{aligned}$$

Concerning N_{ij}, if $Y = y$ we have that

$$N_{ij} = \mathbf{1}_{\{y\geq 2\}} \sum_{k=0}^{y-2} \mathbf{1}_{\{X(k)=i, X(k+1)=j\}}.$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}|Y=y) &= \mathbf{1}_{\{y\geq 2\}} \sum_{k=0}^{y-2} \mathbb{P}(X(k)=i, X(k+1)=j|Y=y) \\
&= \mathbf{1}_{\{y\geq 2\}} \sum_{k=0}^{y-2} \frac{\mathbb{P}(Y=y|X(k+1)=j)\mathbb{P}(X(k+1)=j|X(k)=i)\mathbb{P}(X(k)=i)}{\mathbb{P}(Y=y)} \\
&= \mathbf{1}_{\{y\geq 2\}} \sum_{k=0}^{y-2} \frac{\mathbf{e}_j' \mathbf{T}^{(y-(k+1)-1)} \mathbf{t}\boldsymbol{\pi} \mathbf{T}^k \mathbf{e}_i}{\boldsymbol{\pi}\mathbf{T}^{y-1}\mathbf{t}} t_{ij}.
\end{aligned}$$

Add, $N_i = \mathbf{1}_{\{X(y-1)=i, X(y)=Y\}}$ then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}(N_i|Y=y) &= \mathbb{P}(X(y-1)=i, X(y)=Y|Y=y) \\
&= \frac{\mathbb{P}(Y=y|X(y)=Y)\mathbb{P}(X(y)=Y|X(y-1)=i)\mathbb{P}(X(y-1)=i)}{\mathbb{P}(Y=y)} \\
&= \frac{\mathbb{P}(X(y)=Y|X(y-1)=i)\mathbb{P}(X(y-1)=i)}{\mathbb{P}(Y=y)} \\
&= \frac{\boldsymbol{\pi}\mathbf{T}^{y-1}\mathbf{e}_i}{\boldsymbol{\pi}\mathbf{T}^{y-1}\mathbf{t}} t_i.
\end{aligned}$$

Then, we have proved the following theorem.

Theorem 3.4 *For $k = 1, \ldots, M$ we have the following:*

- *For $i = 1, \ldots, p$,*

$$\mathbb{E}_{\boldsymbol{\theta}}(B_i^k|Y_k = y_k) = \frac{\boldsymbol{e}_i' \boldsymbol{T}^{y_k-1}\boldsymbol{t}}{\boldsymbol{\pi}\boldsymbol{T}^{y_k-1}\boldsymbol{t}} \pi_i. \tag{3.26}$$

- *For $i, j = 1, \ldots, p$,*

$$\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}^k|Y_k = y_k) = \mathbf{1}_{\{y_k\geq 2\}} \sum_{l=0}^{y_k-2} \frac{\boldsymbol{e}_j' \boldsymbol{T}^{(y_k-(l+1)-1)} \boldsymbol{t}\boldsymbol{\pi} \boldsymbol{T}^l \boldsymbol{e}_i}{\boldsymbol{\pi}\boldsymbol{T}^{y_k-1}\boldsymbol{t}} t_{ij}. \tag{3.27}$$

- *For* $i = 1, \ldots, p$,

$$\mathbb{E}_{\boldsymbol{\theta}}(N_i^k | Y_k = y_k) = \frac{\boldsymbol{\pi} \boldsymbol{T}^{y_k - 1} \boldsymbol{e}_i}{\boldsymbol{\pi} \boldsymbol{T}^{y_k - 1} \boldsymbol{t}} t_i. \tag{3.28}$$

Finally, the EM algorithm can be performed as follows.

ALGORITHM: *EM algorithm for DPH distributions*

1. Initialize the parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0)$.
2. Find $\sum_{k=1}^{M} \mathbb{E}_{\boldsymbol{\theta}_0}(B_i^k | Y_k = y_k)$, $\sum_{k=1}^{M} \mathbb{E}_{\boldsymbol{\theta}_0}(N_{ij}^k | Y_k = y_k)$, and $\sum_{k=1}^{M} \mathbb{E}_{\boldsymbol{\theta}_0}(N_i^k | Y_k = y_k)$, (see (3.26), (3.27), (3.28)).
3. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}}, \hat{\mathbf{t}})$, (see (3.23), (3.24), (3.25)).
4. Set $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, and goto 2.

If zero is contained in the data we also need to include an atom of a certain size at zero in the specification of the phase-type distribution. Allowing for $\pi_{p+1} > 0$ we may recalculate conditional expectations and maxima as above. However, it is immediately seen that estimation procedure can be split into the following components. (1) Let $\hat{\pi}_{p+1}$ denote the proportion of zeros in the data set. (2) Eliminate the zeros from the data. (3) Fit a discrete phase-type model $DPH(\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}})$ to the remaining data. This procedure, indeed, produces a maximum likelihood estimator for the complete model (which contains an atom at zero).

Example 3.1 *Considering the distributions and phases given in Table 3.5 and using the EM algorithm we got the Figures 3.9, 3.10, 3.11, and 3.12.*

Table 3.5: Distributions, number of phases, and size of data considered by the algorithms

Distribution	Phases	Observations
Geo(0.5)	3, 6, 9	200
Bin(20,0.5)	3, 6, 9	200
0.6*Geo(0.5)+0.4*Geo(0.1)	10	100
0.5*NBin(5,0.5)+0.5*NBin(20,0.5)	10	500

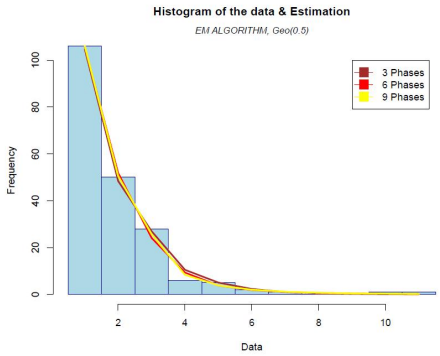

Figure 3.9: EM, Geo(0.5)

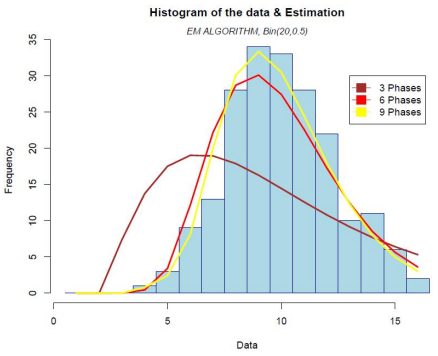

Figure 3.10: EM, Bin(20,0.5)

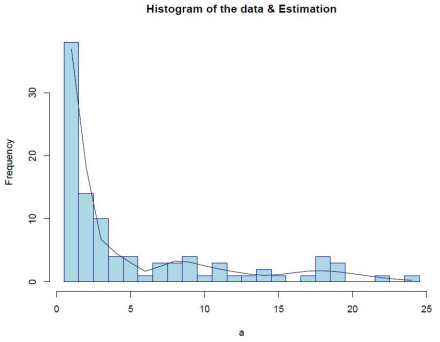

Figure 3.11: EM, 0.6*Geo(0.5)+ 0.4*Geo(0.1)

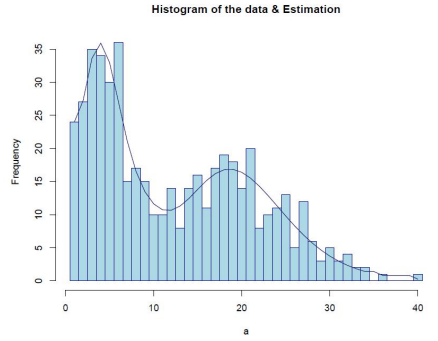

Figure 3.12: EM, 0.5*NBin(5,0.5)+ 0.5*NBin(20,0.5)

3.3.3 The Gibbs sampler algorithm: DPH

We now consider the discrete version of the Gibbs sampler presented in Section 3.2.3. We have to simulate M Markov chains with absorption times $y_1, \ldots, y_M$, using the value of the initial parameter set, and thus obtain $B_i = \mathbf{1}_{\{X(0)=i\}}$, N_{ij}, the number of transitions from state i to state j, $i, j = 1, \ldots, p$, and $N_i = \mathbf{1}_{\{X(y-1)=i\}}$, for $y \in \{y_1, \ldots, y_M\}$.

Let ϕ be the prior distribution proportional to

$$\phi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_i^{\beta_i - 1} \prod_{i=1}^{p} \prod_{j=1}^{p} t_{ij}^{\nu_{ij} - 1} \prod_{i=1}^{p} t_i^{\eta_i - 1}, \tag{3.29}$$

where β_i, ν_{ij}, and η_i are fixed parameters. Note that

$$\begin{aligned} \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_p) \sim Dirichlet(\beta_1, \ldots, \beta_p), \\ \mathbf{t}_i^* &= (t_{i1}, \ldots, t_{ip}) \sim Dirichlet(\nu_{i1}, \ldots, \nu_{ip}), \\ \mathbf{t} &= (t_1, \ldots, t_p) \sim Dirichlet(\eta_1, \ldots, \eta_p). \end{aligned}$$

The posterior distribution is then given by

$$p^*(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{i=1}^{p} \pi_i^{\beta_i + B_i - 1} \prod_{i=1}^{p} \prod_{j=1}^{p} t_{ij}^{\nu_{ij} + N_{ij} - 1} \prod_{i=1}^{p} t_i^{\eta_i + N_i - 1}, \tag{3.30}$$

with $\boldsymbol{\pi} \sim Dirichlet(\beta_1 + B_1, \ldots, \beta_p + B_p)$, $\mathbf{t}_i^* \sim Dirichlet(\nu_{i1} + N_{i1}, \ldots, \nu_{ip} + N_{ip})$, and $\mathbf{t} \sim Dirichlet(\eta_1 + N_1, \ldots, \eta_p + N_p)$.

In general, we have the following algorithm.

ALGORITHM: *Gibbs sampler algorithm for DPH distributions*

1. Draw $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0)$ from the prior distribution (3.29).
2. Simulate M Markov chains with time of absorptions $y_1, \ldots, y_M$, given $\boldsymbol{\theta}_0$. Obtain B_i, N_{ij}, N_i, $i, j = 1, \ldots, p$.
3. Draw $\boldsymbol{\theta}_1 = (\boldsymbol{\pi}_1, \mathbf{T}_1, \mathbf{t}_1)$ from the posterior distribution (3.30).
4. Set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$ and goto 2.

After a number of initial iterations (burn-in), the procedure will stabilize into a stationary mode.

Example 3.2 *Considering the distributions given in Table 3.5 and using GS with the canonical form (GSC), we obtained the Figures 3.13, 3.14, 3.15, and 3.16.*

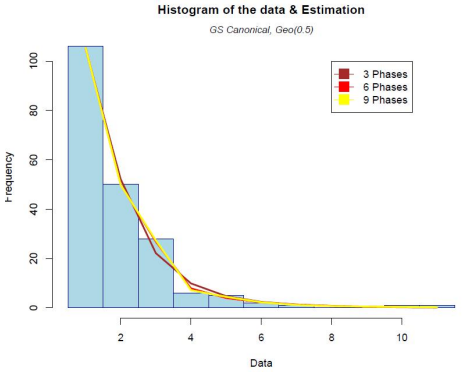

Figure 3.13: GSC, Geo(0.5)

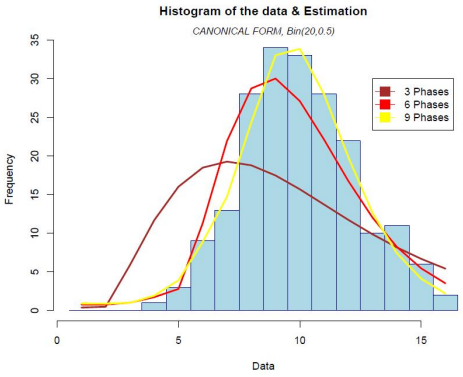

Figure 3.14: GSC, Bin(20,0.5)

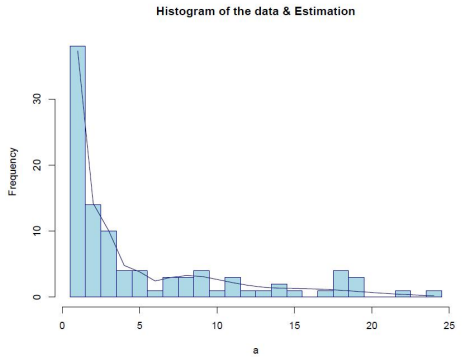

Figure 3.15: GSC, 0.6*Geo(0.5)+ 0.4*Geo(0.1)

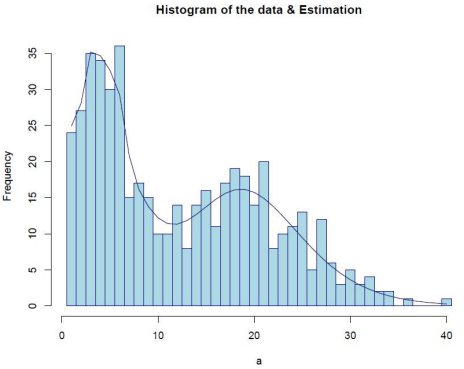

Figure 3.16: GSC, 0.5*NBin(5,0.5)+ 0.5*NBin(20,0.5)

3.3.4 Direct method: DPH

As in the CPH case, we will consider a transformation of the parameters in order to get an unconstrained optimization problem.

For $i = 1, \ldots, p-1$, generate $-\infty < \varrho_i < \infty$, and take the following transformation

$$\pi_i = \frac{e^{\varrho_i}}{1 + \sum_{s=1}^{p-1} e^{\varrho_s}}, \quad \pi_p = \frac{1}{1 + \sum_{i=1}^{p-1} e^{\varrho_i}},$$

for $i, j = 1, \ldots, p$, generate $-\infty < \gamma_{ij} < \infty$ such that

$$t_{ij} = \frac{e^{\gamma_{ij}}}{1 + \sum_{s=1}^{p} e^{\gamma_{is}}}, \quad i \neq j,$$

and

$$t_i = \frac{e^{\gamma_{ii}}}{1 + \sum_{s=1}^{p} e^{\gamma_{is}}}.$$

The gradient vector is given by

$$\left(\left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \varrho_i} \right)_{i=1,\ldots,p-1}, \left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \gamma_{ij}} \right)_{i,j=1,\ldots,p} \right),$$

where the expressions for the derivatives of $\frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial \tau^*}$, for $\tau^* \in \{\varrho_m, \gamma_{ij}\}$, $\frac{\partial f(y_k)}{\partial \varrho_m}$, and $\frac{\partial f(y_k)}{\partial \gamma_{ij}}$, are the same as (3.14), (3.15), and (3.17), respectively, considering in this case $R_m(y) = \mathbf{e}_m' \mathbf{T}^{y-1} \mathbf{t}$.

Moreover,

$$\frac{\partial R_s(y_k)}{\partial \gamma_{ij}} = \mathbf{e}_s' \frac{\partial \mathbf{T}^{y_k - 1}}{\partial \gamma_{ij}} \mathbf{t} + \mathbf{e}_s' \mathbf{T}^{y_k - 1} \frac{\partial \mathbf{t}}{\partial \gamma_{ij}}, \quad s \in \{1, \ldots, p\},$$

where for $r \geq 1$

$$\frac{\partial \mathbf{T}^r}{\partial \tau^*} = \sum_{k=0}^{r-1} \mathbf{T}^k \frac{\partial \mathbf{T}}{\partial \tau^*} \mathbf{T}^{r-1-k}, \quad \text{for all } \tau^*, \tag{3.31}$$

and particularly, $\frac{\partial \mathbf{T}}{\partial \gamma_{ij}}$ is a matrix whose (m,n)-th, $m \neq n$, element is given by

$$\frac{\partial t_{mn}}{\partial \gamma_{ij}} = t_{mn} \mathbf{1}_{\{i=m, j=n\}} - t_{mn} t_{mj} \mathbf{1}_{\{i=m\}},$$

and $\frac{\partial \mathbf{t}}{\partial \gamma_{ij}}$ is a column vector whose m-th element is

$$\frac{\partial t_m}{\partial \gamma_{ij}} = t_m \mathbf{1}_{\{i=j=m\}} - t_m t_{mj} \mathbf{1}_{\{i=m\}}.$$

Note that $\frac{\partial t_{mm}}{\partial \gamma_{ij}} = 1 - \sum_{s \neq m} \frac{\partial t_{ms}}{\partial \gamma_{ij}} - \frac{\partial t_m}{\partial \gamma_{ij}}$.

The results for the log-likelihood (LL) and the execution times of the Geo(0.5) and the Bin(20,0.5) are in Tables 3.6 and 3.7. Note that we consider the EM algorithm using also canonical form (EMC), the Gibbs sampler with the variants using canonical form (GSC) and reversed Markov chains (GS REV/GSC REV), and finally the direct method using canonical form (DMC).

We can see in Tables 3.6 and 3.7 that the EMC, the GSC REV, and the DMC methods have the lowest execution time. In Table 3.8 we present the results

Table 3.6: Log-likelihood (LL) and execution time (time) for a Geo(0.5) distribution with 200 observations and considering dimensions 3, 6, and 9

Algorithm	3		6		9	
	LL	time	LL	time	LL	time
EM	-261.50112	0.35256	-259.18334	3.249535	-259.16876	7.60085
EMC	-260.61263	0.16091	-259.06970	2.35824	-257.76120	4.55080
GSC	-263.79733	0.96749	-262.62966	2.048963	-262.98623	3.13080
GSC REV	-263.54800	0.10886	-262.35961	0.572875	-261.77490	1.27125
GS	-264.52615	1.00085	-263.21963	1.931397	-262.41165	3.77487
GS REV	-264.53692	0.12040	-263.96603	0.698154	-262.82513	1.35422
DM	-261.56296	0.23187	-261.41675	3.479794	-260.75706	10.25334
DMC	-261.04200	0.14317	-259.14148	1.418172	-258.73445	4.34394

Table 3.7: Log-likelihood (LL) and execution time (time) for a Bin(20,0.5) distribution with 200 observations and considering dimensions 3, 6, and 9

Algorithm	3		6		9	
	LL	time	LL	time	LL	time
EM	-517.20199	0.40572	-443.86606	8.083756	-437.66486	12.374817
EMC	-517.20199	0.14409	-443.86595	5.633254	-437.51209	8.045881
GSC	-525.90580	1.39880	-467.13223	3.136527	-441.48312	8.289563
GSC REV	-525.79930	0.10847	-466.73260	0.788024	-441.09375	2.233701
GS	-532.04583	1.58482	-507.31243	3.477106	-443.83995	8.387822
GS REV	-532.06085	0.18639	-507.96166	1.028186	-443.19820	2.607144
DM	-517.20199	0.25929	-444.63254	5.396215	-437.77745	10.059585
DMC	-517.12650	0.19505	-443.56324	3.216522	-437.15772	6.345557

using these methods for a mixture of geometrics $0.6 * Geo(0.5) + 0.4 * Geo(0.1)$ dimension 10, and 100 observations; and also for a mixture of negative binomials $0.5 * NBin(5, 0.5) + 0.5 * NBin(20, 0.5)$ dimension 10 and 500 observations.

Note that the worst LL was found in GSC REV, even though the time was lower than the others.

Table 3.8: Log-likelihood (LL) and execution time (time) for a mixture of geometric distributions and a mixture of negative binomials

Algorithm	0.6*Geo(0.5)+0.4*Geo(0.1)		0.5*NBin(5,0.5)+0.5*NBin(20,0.5)	
	LL	time	LL	time
EMC	-235.44286	89.37870	-1649.35240	117.53216
GSC REV	-241.68890	17.95527	-1669.59530	20.52369
DMC	-236.85436	66.22941	-1650.36523	95.23651

Chapter 4

Fisher information matrix for phase-type distributions

Fisher information (FI) is a key concept in the theory of statistical inference and essentially describes the amount of information that the data provide about unknown parameters. It has applications in finding the variance of an estimator, as well as in the asymptotic behavior of maximum likelihood estimates.

Little has been done with respect to precision and inference of phase-type distributions. Nielsen and Beyer [48] have given an explicit calculation of the FI matrix for an interrupted Poisson process. In this Chapter, we present methods for calculating the FI matrix for phase-type distributions (continuous (CPH) and discrete (DPH) cases) when the EM algorithm is applied as an optimization tool as well as for the direct Newton-Raphson method.

Consider M independent observations $y_1, \ldots, y_M$ from a $PH_p(\boldsymbol{\pi}, \mathbf{T})$ distribution, here PH denotes a phase-type distribution in general, i.e. for both DPH and CPH. We will also consider the exit vector instead of the diagonal of the matrix $\mathbf{T}$ in both distributions, i.e. the elements in the diagonal of $\mathbf{T}$ are defined as $t_{ii} = 1 - \sum_{j=1, j\neq i}^{p} t_{ij} - t_i$ in DPH, and $t_{ii} = -\sum_{j=1, j\neq i}^{p} t_{ij} - t_i$ in CPH.

Let $\boldsymbol{\theta} = (\theta_i)_{1\leq i\leq p^2+(p-1)}$ be the vector such that

$$\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{p-1}, t_1, t_{12}, \ldots, t_{1p}, t_{21}, t_2, \ldots, t_{2p}, \ldots, t_{p1}, \ldots, t_{p,p-1}, t_p).$$

The incomplete data likelihood (see (3.7) and (3.19)) function is given by

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{k=1}^{M} \boldsymbol{\pi}\boldsymbol{\Psi}(y_k)\mathbf{t}, \tag{4.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_M)$ and

$$\boldsymbol{\Psi}(y) = \begin{cases} e^{\boldsymbol{T}y} & \text{for CPH} \\ \boldsymbol{T}^{y-1} & \text{for DPH.} \end{cases}$$

The log–likelihood function is defined as $\ell(\boldsymbol{\theta};\boldsymbol{y}) = \log L(\boldsymbol{\theta};\boldsymbol{y})$.

By substituting $\boldsymbol{\pi} = \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p'$ (because $\boldsymbol{\pi}$ is a probability vector, i.e $\sum_{i=1}^{p} \pi_i = 1$, and in order to avoid the over-parameterization problem), the density of the phase–type distribution evaluated in y is given by

$$f(y) \quad = \quad \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' \boldsymbol{\Psi}(y)\mathbf{t} + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p' \boldsymbol{\Psi}(y)\mathbf{t},$$

with partial derivatives given by

$$\begin{aligned} \frac{\partial f(y)}{\partial \pi_m} &= \mathbf{e}_m' \boldsymbol{\Psi}(y)\boldsymbol{t} - \mathbf{e}_p' \boldsymbol{\Psi}(y)\boldsymbol{t} \\ \frac{\partial f(y)}{\partial t_{mn}} &= \boldsymbol{\pi} \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}} \boldsymbol{t}, \quad m \neq n \\ \frac{\partial f(y)}{\partial t_{m}} &= \boldsymbol{\pi}\boldsymbol{\Psi}(y)\mathbf{e}_m + \boldsymbol{\pi} \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{m}} \boldsymbol{t}. \end{aligned}$$

In order to compute the partial derivatives of $\boldsymbol{\Psi}$ with respect to θ_m, for $m \in \{1, \ldots, p^2 + (p-1)\}$, we shall need the derivative of $\mathbf{T}^r$ for $r \geq 1$, that it is given in (3.31), with $\left[\frac{\partial \mathbf{T}}{\partial t_{ij}}\right]_{ij} = 1$, $\left[\frac{\partial \mathbf{T}}{\partial t_{ij}}\right]_{ii} = -1$, and $\left[\frac{\partial \mathbf{T}}{\partial t_{i}}\right]_{ii} = -1$.

The derivative of $e^{\mathbf{T}y}$ is given in (3.18). Since $e^{\mathbf{T}(x+y)} = e^{\mathbf{T}x} e^{\mathbf{T}y}$, we can get a recursive version of (3.18) given by

$$\frac{\partial e^{\mathbf{T}(x+y)}}{\partial \theta_m} = e^{\mathbf{T}x} \frac{\partial e^{\mathbf{T}y}}{\partial \theta_m} + \frac{\partial e^{\mathbf{T}x}}{\partial \theta_m} e^{\mathbf{T}y}.$$

In the following sections we will give a way of getting the Fisher information matrix for PH distributions via the EM algorithm and via a direct Newton-Raphson approach.

4.1 Via the EM algorithm

The EM algorithm also allows for extracting information concerning the Fisher information matrix as noted Oakes in [52]. Considering L, the incomplete data likelihood which is maximized by the EM algorithm, the Fisher information matrix is given by

$$\frac{\partial^2 L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}^2} = \left\{ \frac{\partial^2 Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}^2} + \frac{\partial^2 Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \hat{\boldsymbol{\theta}}} \right\}_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}}, \tag{4.2}$$

where

$$Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(\ell_f(\hat{\boldsymbol{\theta}};\boldsymbol{x})\,|\boldsymbol{y})\,, \tag{4.3}$$

and $\boldsymbol{x} = (x_1, \ldots, x_M)$ denote the full data for the M absorption times.

In order to avoid notation, we define the following

$$U_i = \sum_{l=1}^{M} \frac{\mathbf{e}_i' \boldsymbol{\Psi}(y_l)\mathbf{t}}{f(y_l)}, \tag{4.4}$$

$$W_i = \sum_{l=1}^{M} \frac{\boldsymbol{\pi}\boldsymbol{\Psi}(y_l)\mathbf{e}_i}{f(y_l)}, \tag{4.5}$$

$$V_{ij} = \begin{cases} \displaystyle\sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \sum_{k=0}^{y_l-2} \frac{\mathbf{e}_j' \mathbf{T}^{y_l-k-2}\mathbf{t}\boldsymbol{\pi}\mathbf{T}^k\mathbf{e}_i}{f(y_l)}, & \text{for DPH} \\ \displaystyle\sum_{l=1}^{M} \frac{1}{f(y_l)} \int_0^{y_l} \mathbf{e}_j' e^{\mathbf{T}(y_l-u)}\mathbf{t}\boldsymbol{\pi}e^{\mathbf{T}u}\mathbf{e}_i \mathrm{d}u, & \text{for CPH.} \end{cases} \tag{4.6}$$

Then (4.3) becomes

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = & \sum_{i=1}^{p-1} \log \hat{\pi}_i U_i \pi_i + \log\left(1 - \sum_{s=1}^{p-1} \hat{\pi}_s\right) U_p \left(1 - \sum_{s=1}^{p-1} \pi_s\right) \\ & + \sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} \log \hat{t}_{ij} V_{ij} t_{ij} + \sum_{i=1}^{p} S_i V_{ii} \\ & + \sum_{i=1}^{p} \log(\hat{t}_i) W_i t_i, \end{aligned}$$

where

$$S_i = \begin{cases} \left(1 - \sum_{j=1, j\neq i}^{p} t_{ij} - t_i\right) \log\left(1 - \sum_{j=1, j\neq i}^{p} \hat{t}_{ij} - \hat{t}_i\right) & \text{for DPH} \\ -\sum_{j=1, j\neq i}^{p} \hat{t}_{ij} - \hat{t}_i & \text{for CPH.} \end{cases}$$

The Fisher information matrix is given by

$$\frac{\partial^2 L(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}^2} = \begin{pmatrix} \boldsymbol{M}_I & \boldsymbol{M}_{III} \\ \boldsymbol{M}_{II} & \boldsymbol{M}_{IV} \end{pmatrix}$$

where $\boldsymbol{M}_I$, $\boldsymbol{M}_{II}$, $\boldsymbol{M}_{III}$, and $\boldsymbol{M}_{IV}$ are themselves matrices of appropriate dimension.

For $i = 1, \ldots, p-1$,

$$\left.\frac{\partial^2 Q}{\partial \hat{\pi}_i^2}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = -\frac{U_i}{\pi_i} - \frac{U_p}{\pi_p}, \quad \left.\frac{\partial^2 Q}{\partial \pi_i \partial \hat{\pi}_i}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{U_i}{\pi_i} + \frac{\partial U_i}{\partial \pi_i} + \frac{U_p}{\pi_p} - \frac{\partial U_p}{\partial \pi_p},$$

thus, the (i,i)-th element of the matrix $\boldsymbol{M}_I$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{\pi}_i^2} + \frac{\partial^2 Q}{\partial \pi_i \partial \hat{\pi}_i}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_i}{\partial \pi_i} - \frac{\partial U_p}{\partial \pi_i}.$$

For $i, j = 1, \ldots, p-1$, $i \neq j$ we have

$$\left.\frac{\partial^2 Q}{\partial \hat{\pi}_j \partial \hat{\pi}_i}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = -\frac{U_p}{\pi_p}, \quad \left.\frac{\partial^2 Q}{\partial \pi_j \partial \hat{\pi}_i}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_i}{\partial \pi_j} + \frac{U_p}{\pi_p} - \frac{\partial U_j}{\partial \pi_j},$$

thus, for $i \neq j$, the (i,j)-th element of the matrix $\boldsymbol{M}_I$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{\pi}_j \partial \hat{\pi}_i} + \frac{\partial^2 Q}{\partial \pi_j \partial \hat{\pi}_i}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_i}{\partial \pi_j} - \frac{\partial U_p}{\partial \pi_j}.$$

In general, for $i, j = 1, \ldots, p-1$ the (i,j)-th element of the matrix $\boldsymbol{M}_I$ is given by

$$\frac{\partial U_i}{\partial \pi_j} - \frac{\partial U_p}{\partial \pi_j}. \tag{4.7}$$

For $m = 1, \ldots, p-1$ and $i, j = 1, \ldots, p$, $i \neq j$,

$$\left.\frac{\partial^2 Q}{\partial \hat{t}_{ij} \partial \hat{\pi}_m}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0, \quad \left.\frac{\partial^2 Q}{\partial t_{ij} \partial \hat{\pi}_m}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_m}{\partial t_{ij}} - \frac{\partial U_p}{\partial t_{ij}},$$

thus, the $((i-1)*p+j, m)$-th element of the matrix $\boldsymbol{M}_{II}$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{t}_{ij} \partial \hat{\pi}_m} + \frac{\partial^2 Q}{\partial t_{ij} \partial \hat{\pi}_m}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_m}{\partial t_{ij}} - \frac{\partial U_p}{\partial t_{ij}}. \tag{4.8}$$

The formula remains the same considering t_i, i.e. the $((i-1)*p+i, m)$-th element of the matrix $\boldsymbol{M}_{II}$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{t}_{i} \partial \hat{\pi}_m} + \frac{\partial^2 Q}{\partial t_{i} \partial \hat{\pi}_m}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial U_m}{\partial t_{i}} - \frac{\partial U_p}{\partial t_{i}}. \tag{4.9}$$

Now, for $m = 1, \ldots, p-1$, $i, j = 1, \ldots, p$, $i \neq j$

$$\left.\frac{\partial^2 Q}{\partial \hat{\pi}_m \partial \hat{t}_{ij}}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0, \quad \left.\frac{\partial^2 Q}{\partial \pi_m \partial \hat{t}_{ij}}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial V_{ij}}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m},$$

thus, the $(m, (i-1) * p + j)$-th element of the matrix $\boldsymbol{M}_{III}$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{\pi}_m \partial \hat{t}_{ij}} + \frac{\partial^2 Q}{\partial \pi_m \partial \hat{t}_{ij}}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial V_{ij}}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m}. \tag{4.10}$$

Moreover,

$$\left.\frac{\partial^2 Q}{\partial \hat{\pi}_m \partial \hat{t}_i}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0, \quad \left.\frac{\partial^2 Q}{\partial \pi_m \partial \hat{t}_i}\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial W_i}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m},$$

thus, the $(m, (i-1) * p + i)$-th element of the matrix $\boldsymbol{M}_{III}$ is given by

$$\left.\left(\frac{\partial^2 Q}{\partial \hat{\pi}_m \partial \hat{t}_i} + \frac{\partial^2 Q}{\partial \pi_m \partial \hat{t}_i}\right)\right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \frac{\partial W_i}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m}. \tag{4.11}$$

Finally, for $i, j, m, n = 1, \ldots, p$, the $(ip - 1 + j, mp - 1 + n)$-th element of the matrix $\boldsymbol{M}_{IV}$ is given by

$$\begin{cases} \dfrac{\partial V_{ij}}{\partial t_{mn}} - \dfrac{\partial V_{ii}}{\partial t_{mn}} & \text{if } i \neq j, m \neq n \\ \dfrac{\partial V_{ij}}{\partial t_m} - \dfrac{\partial V_{ii}}{\partial t_m} & \text{if } i \neq j, m = n \\ \dfrac{\partial W_i}{\partial t_{mn}} - \dfrac{\partial V_{ii}}{\partial t_{mn}} & \text{if } i = j, m \neq n \\ \dfrac{\partial W_i}{\partial t_m} - \dfrac{\partial V_{ii}}{\partial t_m} & \text{if } i = j, m = n. \end{cases}$$

Let $R_i(u) = \boldsymbol{\pi}\boldsymbol{\Psi}(u)\mathbf{e}_i$ and $Q_i(u) = \mathbf{e}_i'\boldsymbol{\Psi}(u)\mathbf{t}$. Then, their derivatives are given by

$$\begin{array}{ll} \dfrac{\partial R_i(u)}{\partial \pi_m} = \mathbf{e}_m'\boldsymbol{\Psi}(u)\mathbf{e}_i - \mathbf{e}_p'\boldsymbol{\Psi}(u)\mathbf{e}_i, & \dfrac{\partial Q_i(u)}{\partial \pi_m} = 0, \\ \dfrac{\partial R_i(u)}{\partial t_{mn}} = \boldsymbol{\pi}\dfrac{\partial \boldsymbol{\Psi}(u)}{\partial t_{mn}}\mathbf{e}_i, \quad m \neq n, & \dfrac{\partial Q_i(u)}{\partial t_{mn}} = \mathbf{e}_i'\dfrac{\partial \boldsymbol{\Psi}(u)}{\partial t_{mn}}\mathbf{t}, \quad m \neq n, \\ \dfrac{\partial R_i(u)}{\partial t_m} = \boldsymbol{\pi}\dfrac{\partial \boldsymbol{\Psi}(u)}{\partial t_m}\mathbf{e}_i, & \dfrac{\partial Q_i(u)}{\partial t_m} = \mathbf{e}_i'\boldsymbol{\Psi}(u)\mathbf{e}_m + \mathbf{e}_i'\dfrac{\partial \boldsymbol{\Psi}(u)}{\partial t_m}\mathbf{t}. \end{array}$$

Thus, U_i, W_i, and V_{ij} (see (4.4), (4.5), and (4.6)) become

$$\begin{aligned}
U_i &= \sum_{l=1}^{M} \frac{Q_i(y_l)}{f(y_l)}, \\
W_i &= \sum_{l=1}^{M} \frac{R_i(y_l)}{f(y_l)}, \\
V_{ij} &= \begin{cases} \displaystyle\sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \sum_{k=0}^{y_l-2} \frac{Q_j(y_l-k-1)R_i(k+1)}{f(y_l)}, & \text{for DPH} \\ \displaystyle\sum_{l=1}^{M} \frac{1}{f(y_l)} \int_0^{y_l} Q_j(y_l-u)R_i(u)\mathrm{d}u, & \text{for CPH.} \end{cases}
\end{aligned}$$

Hence, for $n \in \{1, \ldots, p^2 + (p-1)\}$, the derivatives w.r.t. θ_n are given by

$$\begin{aligned}
\frac{\partial U_i}{\partial \theta_n} &= \sum_{l=1}^{M} \frac{1}{f(y_l)^2} \left(f(y_l) \frac{\partial Q_i(y_l)}{\partial \theta_n} - Q_i(y_l) \frac{\partial f(y_l)}{\partial \theta_n} \right), \\
\frac{\partial W_i}{\partial \theta_n} &= \sum_{l=1}^{M} \frac{1}{f(y_l)^2} \left(f(y_l) \frac{\partial R_i(y_l)}{\partial \theta_n} - R_i(y_l) \frac{\partial f(y_l)}{\partial \theta_n} \right),
\end{aligned}$$

the derivative of V_{ij} for DPH is given by

$$\begin{aligned}
\frac{\partial V_{ij}}{\partial \theta_n} &= \sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \sum_{k=0}^{y_l-2} \frac{1}{f(y_l)^2} \left[f(y_l) \left(Q_j(y_l-k-1) \frac{\partial R_i(k+1)}{\partial \theta_n} \right.\right. \\
&\quad \left.\left. + \frac{\partial Q_j(y_l-k-1)}{\partial \theta_n} R_i(k+1) \right) - \left(\frac{\partial f(y_l)}{\partial \theta_n} \right) Q_j(y_l-k-1)R_i(k+1) \right].
\end{aligned}$$

Concerning the computation of $\frac{\partial V_{ij}}{\partial \theta_n}$ for CPH we have that

$$\begin{aligned}
\frac{\partial V_{ij}}{\partial \theta_n} &= \sum_{l=1}^{M} \frac{1}{f(y_l)^2} \left[f(y_l) \int_0^{y_l} Q_j(y_l-u) \left(\frac{\partial R_i(u)}{\partial \theta_n} \right) + \left(\frac{\partial Q_j(y_l-u)}{\partial \theta_n} \right) R_i(u)\mathrm{d}u \right. \\
&\quad \left. - \left(\frac{\partial f(y_l)}{\partial \theta_n} \right) \int_0^{y_l} Q_j(y_l-u)R_i(u)\mathrm{d}u \right].
\end{aligned}$$

Define the following integrals

$$\begin{aligned}
\mathbf{J}_1(y;\mathbf{M}) &= \int_0^y e^{\mathbf{T}(y-u)}\mathbf{M}e^{\mathbf{T}u}\mathrm{d}u = e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+1}}{(s+1)!}\mathbf{D}_{\mathbf{J}_1}(s),\\
\mathbf{J}_2(y;\theta_n,\mathbf{M}) &= \int_0^y e^{\mathbf{T}(y-u)}\mathbf{M}\frac{\partial e^{\mathbf{T}u}}{\partial\theta_n}\mathrm{d}u\\
&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+2}}{(s+2)!}(\mathbf{D}_{\mathbf{J}_2,1}(s,\theta_n)+\mathbf{D}_{\mathbf{J}_2,2}(s,\theta_n)),\\
\mathbf{J}_3(y;\theta_n,\mathbf{M}) &= \int_0^y \frac{\partial e^{\mathbf{T}(y-u)}}{\partial\theta_n}\mathbf{M}e^{\mathbf{T}u}\mathrm{d}u\\
&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+2}}{(s+2)!}(\mathbf{D}_{\mathbf{J}_3,1}(s,\theta_n)+\mathbf{D}_{\mathbf{J}_3,2}(s,\theta_n)),
\end{aligned}$$

where $\mathbf{M}$ is a $(p\times p)$-dimensional matrix and

$$\begin{aligned}
\mathbf{D}_{\mathbf{J}_1}(s) &= \sum_{j=0}^{s}\mathbf{K}^j\frac{1}{c}\mathbf{M}\mathbf{K}^{s-j},\\
\mathbf{D}_{\mathbf{J}_2,1}(s,\theta_n) &= \sum_{j=0}^{s}\mathbf{K}^j\frac{1}{c}\mathbf{M}\frac{\partial\mathbf{K}^{s-j+1}}{\partial\theta_n},\\
\mathbf{D}_{\mathbf{J}_2,2}(s,\theta_n) &= \sum_{j=0}^{s}\mathbf{K}^j(s+1-j)\frac{1}{c^2}\frac{\partial c}{\partial\theta_n}\mathbf{M}(\mathbf{K}-\mathbf{I})\mathbf{K}^{s-j},\\
\mathbf{D}_{\mathbf{J}_3,1}(s,\theta_n) &= \sum_{j=0}^{s}\frac{\partial\mathbf{K}^{s-j+1}}{\partial\theta_n}\frac{1}{c}\mathbf{M}\mathbf{K}^j,\\
\mathbf{D}_{\mathbf{J}_3,2}(s,\theta_n) &= \sum_{j=0}^{s}\mathbf{K}^j(j+1)\frac{1}{c^2}\frac{\partial c}{\partial\theta_n}(\mathbf{K}-\mathbf{I})\mathbf{M}\mathbf{K}^{s-j}.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial V_{ij}}{\partial \pi_m} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \left(\mathbf{e}_j' \, \mathbf{J}_1(y_k; \mathbf{t}\mathbf{e}_m')\mathbf{e}_i - \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\mathbf{e}_p')\mathbf{e}_i \right) \right. \\
&\quad \left. - \frac{\partial f(y_k)}{\partial \pi_m} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i \right], \\
\frac{\partial V_{ij}}{\partial t_{mn}} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \left(\mathbf{e}_j' \mathbf{J}_2(y_k; t_{mn}, \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i + \mathbf{e}_j' \mathbf{J}_3 \left(y_k; t_{mn}, \mathbf{t}\boldsymbol{\pi}\right) \mathbf{e}_i \right) \right. \\
&\quad \left. - \frac{\partial f(y_k)}{\partial t_{mn}} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i \right], \\
\frac{\partial V_{ij}}{\partial t_m} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \left(\mathbf{e}_j' \mathbf{J}_2(y_k; t_m, \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i + \mathbf{e}_j' \mathbf{J}_1 \left(x_k; \mathbf{e}_m \boldsymbol{\pi}\right) \mathbf{e}_i \right. \right. \\
&\quad \left. \left. + \, \mathbf{e}_j' \mathbf{J}_3(y_k; t_m, \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i \right) - \frac{\partial f(y_k)}{\partial t_m} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi})\mathbf{e}_i \right].
\end{aligned}$$

A proper truncation of the infinite sums involved in $\mathbf{J}_i$, $i = 1, 2, 3$, can be obtained. Since we are working with stochastic (sub-stochastic) matrices, then we can see that $\mathbf{D}_{\mathbf{J}_1}(s)$ is bounded by $s + 1$, thus

$$\begin{aligned}
\sum_{s=0}^{\infty} e^{-cy} \frac{(cy)^{s+1}}{(s+1)!} \cdot (s+1) &= \sum_{s=1}^{\infty} e^{-cy} \frac{(cy)^s}{s!} \cdot s \\
&= cy \left[\frac{\sum_{s=1}^{\infty} e^{-cy} \frac{(cy)^s}{s!} \cdot s}{(cy)} \right] \\
&= cy \left[\sum_{s=1}^{\infty} \frac{f(i, cy) \cdot s}{(cy)} \right] \\
&= cy \left[\sum_{s=1}^{\infty} f_1(i, cy) \right],
\end{aligned}$$

where $f(x, \lambda)$ is the Poisson density with parameter λ and $f_1(x, \lambda)$ is the first order moment distribution of f. Hence, the truncation is the standard uniformization level plus 1.

In the same manner, we have that $\mathbf{D}_{\mathbf{J}_2,1}(s, \cdot)$, $\mathbf{D}_{\mathbf{J}_2,2}(s, \cdot)$, $\mathbf{D}_{\mathbf{J}_3,1}(s, \cdot)$, and $\mathbf{D}_{\mathbf{J}_3,2}(s, \cdot)$, are bounded by $\frac{1}{2}(s+1)(s+2)$, thus

$$\sum_{s=0}^{\infty} e^{-cy} \frac{(cy)^{s+2}}{(s+2)!} \cdot \frac{1}{2}(s+1)(s+2) = \sum_{s=2}^{\infty} e^{-cy} \frac{(cy)^s}{s!} \cdot \frac{1}{2} s(s-1)$$

$$= \frac{1}{2}\sum_{s=2}^{\infty} e^{-cy}\frac{(cy)^s}{s!}\cdot s^2 - \frac{1}{2}\sum_{s=2}^{\infty} e^{-cy}\frac{(cy)^s}{s!}\cdot s.$$

The right hand side we have already analyzed, and

$$\begin{aligned}\sum_{s=2}^{\infty} e^{-cy}\frac{(cy)^s}{s!}\cdot s^2 &= \mu_2\left[\frac{\sum_{s=2}^{\infty} e^{-cy}\frac{(cy)^s}{(s)!}\cdot s^2}{\mu_2}\right] \\ &= \mu_2\left[\sum_{s=2}^{\infty}\frac{f(i,cy)\cdot s^2}{\mu_2}\right] \\ &= \mu_2\left[\sum_{s=2}^{\infty} f_2(i,cy)\right],\end{aligned}$$

where $\mu_2 = \int_0^\infty x^2 f(x,cy)\mathrm{d}x$ and $f_2(x,\lambda)$ is the second order moment distribution of f. The truncation level is the standard uniformization plus 2.

Finally, in Table 4.1, we can see the general way to compute the Fisher information matrix of a PH distribution.

Table 4.1: Fisher Information matrix for a phase-type distribution with set of parameters $\boldsymbol{\theta}$

	π_m	t_{ij}	t_i
$\boldsymbol{\theta}$	$\frac{\partial U_m}{\partial \boldsymbol{\theta}} - \frac{\partial U_p}{\partial \boldsymbol{\theta}}$	$\frac{\partial V_{ij}}{\partial \boldsymbol{\theta}} - \frac{\partial V_{ii}}{\partial \boldsymbol{\theta}}$	$\frac{\partial W_i}{\partial \boldsymbol{\theta}} - \frac{\partial V_{ii}}{\partial \boldsymbol{\theta}}$

4.2 Newton–Raphson estimation

The Newton–Raphson method is based on the idea of approximating a function with its first or second order Taylor expansion. Thus, we need to calculate the gradient vector of the log–likelihood function. This is computationally demanding, particularly if the dimension is large. However, the cost of calculating the gradient could be compensated for by fewer iterations. The method is not designed to work with boundary conditions. While the calculation of the gradient is rather straightforward, the task of making an efficient numerical implementation of the formulae is by no means trivial.

Using the idea given by B. F. Nielsen, *et.al* [48], we want to work with an unconstrained system, and use a package for unconstrained optimization written

by K. Madsen, *et.al* [41]. Their program, as well as many other standard routines available for unconstrained optimization, find the maximum of a given function using the gradient vector. Since we want to find the maximum of the log-likelihood function, we calculate the gradient vector based on the parameter transformation which provides the unconstrained optimization problem. We shall again refer to this method as the Direct Method (DM).

The direct method we employ assumes that the parameters are unbounded. This is obviously not the case for the phase–type intensities so we consider a re–parametrisation $\boldsymbol{\tau}$ of the parameters. We also need to provide the gradient at a given point of the transformed parameters.

Considering the transformations of the parameters given in sections 3.2.4 and 3.3.4, we need to provide the gradient at a given point of the transformed parameters, denoted by $\boldsymbol{\tau}$, in order to find the maximum,

$$\mathbf{g} = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\tau}} = \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \tau_m} \right)_{m=1,\ldots,p^2+(p-1)}.$$

By the chain rule this vector can be obtained as

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\tau}} = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}, \tag{4.12}$$

where $\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}}$ is a $p^2 + (p-1)$-dimensional row vector and $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}$ is the Jacobian matrix. Taking the derivative of the log-likelihood function w.r.t $\boldsymbol{\theta}$ we get that

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{M} \frac{1}{f(y_k)} \frac{\partial f(y_k)}{\partial \boldsymbol{\theta}},$$

where f is the density of the phase–type distribution parameterized by $\boldsymbol{\theta}$.

To obtain the Fisher information matrix using the direct method, we take the second derivative of (4.12), which at the optimum gives

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\tau}} \partial \boldsymbol{\tau}} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}} \frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} \frac{\partial \bar{\boldsymbol{\theta}}}{\partial \bar{\boldsymbol{\tau}}} \tag{4.13}$$

where $\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}}$ is a square matrix of second-order partial derivatives given by

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} = \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \frac{\partial^2 f(y_k)}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} - \frac{\partial f(y_k)}{\partial \bar{\boldsymbol{\theta}}} \frac{\partial f(y_k)}{\partial \boldsymbol{\theta}} \right].$$

The second derivatives of the density with respect to the initial probabilities are 0, i.e.

$$\frac{\partial^2 f(y)}{\partial \pi_n \partial \pi_m} = 0,$$

while with respect to the elements of the matrix $\mathbf{T}$, the second derivatives are given by

$$\frac{\partial^2 f(y)}{\partial t_{mn}\partial t_{ij}} = \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_{mn}\partial t_{ij}}\mathbf{t}, \quad m \neq n, i \neq j,$$

and w.r.t the exit probabilities

$$\frac{\partial^2 f(y)}{\partial t_m \partial t_i} = \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_m}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{e}_m + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_m \partial t_i}\mathbf{t}.$$

Finally,

$$\begin{aligned}
\frac{\partial^2 f(y)}{\partial \pi_m \partial t_{ij}} = \frac{\partial^2 f(y)}{\partial t_{ij}\partial \pi_m} &= \mathbf{e}_m' \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{ij}}\mathbf{t} - \mathbf{e}_p' \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{ij}}\mathbf{t}, \quad i \neq j \\
\frac{\partial^2 f(y)}{\partial \pi_m \partial t_i} = \frac{\partial^2 f(y)}{\partial t_i \partial \pi_m} &= \mathbf{e}_m' \boldsymbol{\Psi}(y)\mathbf{e}_i - \mathbf{e}_p' \boldsymbol{\Psi}(y)\mathbf{e}_i + \mathbf{e}_m' \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{t} - \mathbf{e}_p' \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{t} \\
\frac{\partial^2 f(y)}{\partial t_{mn}\partial t_i} &= \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_{mn}\partial t_i}\mathbf{t}, \quad m \neq n \\
\frac{\partial^2 f(y)}{\partial t_i \partial t_{mn}} &= \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_i \partial t_{imn}}\mathbf{t}, \quad m \neq n.
\end{aligned}$$

For $m, n \in \{1, \ldots, p^2 + (p-1)\}$, and taking the second derivative of (3.31) we get

$$\frac{\partial^2 \mathbf{T}^r}{\partial \theta_n \partial \theta_m} = \sum_{k=0}^{r-1} \mathbf{T}^k \frac{\partial \mathbf{T}}{\partial \theta_m}\frac{\partial \mathbf{T}^{r-1-k}}{\partial \theta_n} + \frac{\partial \mathbf{T}^k}{\partial \theta_n}\frac{\partial \mathbf{T}}{\partial \theta_m}\mathbf{T}^{r-1-k}. \tag{4.14}$$

In the same way from (3.18), we have that

$$\frac{\partial^2 e^{\mathbf{T}y}}{\partial \theta_n \partial \theta_m} = e^{-cy}\sum_{k=0}^{\infty}\frac{(cy)^{k+1}}{(k+1)!}\frac{\partial^2 \mathbf{K}^{k+1}}{\partial \theta_n \partial \theta_m} + \frac{\partial c}{\partial \theta_m} y\left(e^{\mathbf{T}y}\frac{\partial \mathbf{K}}{\partial \theta_n} + \frac{\partial e^{\mathbf{T}y}}{\partial \theta_n}(\mathbf{K} - \mathbf{I})\right), \tag{4.15}$$

where $\frac{\partial^2 \mathbf{K}^r}{\partial \theta_n \partial \theta_m}$ is calculated like (4.14).

The quasi-Newton method presented in [48] gives an approximate value of the Hessian matrix for the transformed parameters $\boldsymbol{\tau}$ used in the optimization. This can be transformed into an approximation for the inverse Fisher information matrix using

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\theta}}\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \bar{\boldsymbol{\tau}} \partial \boldsymbol{\tau}}\frac{\partial \bar{\boldsymbol{\tau}}}{\partial \bar{\boldsymbol{\theta}}}.$$

4.3 Experimental results

In order to make the Fisher information meaningful, we consider the canonical form representation given by

$$\boldsymbol{\pi} = (1, 0, \ldots, 0), \quad \mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & & & \\ & t_{22} & t_{23} & & \\ & & \ddots & \ddots & \\ & & & t_{p-1,p-1} & t_{p-1,p} \\ & & & & t_{pp} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_{p-1} \\ t_p \end{pmatrix}. \tag{4.16}$$

In this Section we present the results of an estimation study considering simulated data from discrete and continuous phase-type distributions. First of all, we consider a shifted Negative binomial distribution $NB(r = 3, p_1 = 0.2)$, which canonical form is given by

$$\boldsymbol{\pi} = (1, 0, 0), \quad \mathbf{T} = \begin{pmatrix} 1 - p_1 & (1 - p_1^2)p_1 & 0 \\ 0 & 1 - p_1 & p_1 - \frac{2p_1^2}{1+p_1} \\ 0 & 0 & 1 - p_1 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} p_1^3 \\ \frac{2p_1^2}{1+p_1} \\ p_1 \end{pmatrix}.$$

For the continuous case, we considered a mixture of three exponential distributions: $\exp(\lambda_1 = 1.0)$, $\exp(\lambda_2 = 0.1)$, and $\exp(\lambda_3 = 0.01)$. This distribution is also called Hyper-exponential (HE), and has a canonical representation (see Cumani [27]) given by

$$\boldsymbol{\pi} = (1, 0, 0), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 - t_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 - t_2 \\ 0 & 0 & -\lambda_3 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \pi_1\lambda_1 + \pi_2\lambda_2 + \pi_3\lambda_3 \\ \frac{\pi_2\lambda_2(\lambda_1-\lambda_2)+\pi_3\lambda_3(\lambda_1-\lambda_3)}{\pi_2(\lambda_1-\lambda_2)+\pi_3(\lambda_1-\lambda_3)} \\ \lambda_3 \end{pmatrix},$$

where $\pi_1 = 0.9$, $\pi_2 = 0.09$, and $\pi_3 = 0.01$.

We generated samples of different sizes. Then, we ran the algorithms for different dimensions of the PH generator. We found the maximum likelihood estimators (MLE) not only via EM-algorithm but also for the DM. The maxima were generally the same.

In order to check the order of the distribution for the simulated data, we calculated the Akaike Information Criterion (AIC) [3], which is calculated by

$$AIC = 2k - 2\log(\hat{L})$$

where k is the number of parameters in the statistical model, and $\hat{L}$ is the maximized value of the likelihood function for the estimated model.

In Tables 4.2 and 4.3 we can see the corresponding AIC. As we expected, considering a sufficient amount of data, we found the correct order in both distributions.

Table 4.2: AIC of the shifted Negative binomial(3,0.2)

dim	Size of data			
	500	1000	5000	10000
1	3528.581819	7020.810118	35213.462481	70413.184709
2	3388.114359	6656.243035	33451.875728	66951.502089
3	**3387.396123**	**6635.565645**	**33325.139564**	**66690.026858**
4	3390.728760	6635.645945	33328.647163	66693.147736
5	3393.884284	6639.629079	33331.568906	66694.314071

Table 4.3: AIC of the Hyper-exponential distribution

dim	Size of data			
	500	1000	5000	20000
1	1503.015560	3657.734462	19550.565214	83008.788613
2	**1386.046313**	**2893.167184**	14281.181393	59627.389521
3	1390.042883	2894.846373	**14077.871083**	**59006.193568**
4	1393.997920	2898.832350	14080.991236	59010.065036
5	1397.408475	2902.826303	14084.990615	59014.071250

After finding the MLE, the FI matrix was obtained considering only the non-zero parameters. Since the inverse of the FI matrix is the empirical variance-covariance matrix, we could obtain the standard deviation (SD) of the parameters (see Tables 4.4 and 4.5).

Table 4.4: Maximum likelihood estimators (MLE) and standard deviations (SD) of the shifted Negative binomial(3,0.2), considering 10000 observations

Parameter	true value	EM		DM	
		MLE	SD	MLE	SD
$\hat{t}_1$	8.0E-3	9.375498E-3	9.3409303E-4	9.390321E-3	9.354187E-4
$\hat{t}_{12}$	0.192000	0.193870	0.042604	0.193925	0.045525
$\hat{t}_2$	0.066667	0.059214	0.011803	0.059187	0.012520
$\hat{t}_{23}$	0.133333	0.144035	0.038740	0.144065	0.040769
$\hat{t}_3$	0.200000	0.203261	0.042618	0.203243	0.044996

The corresponding correlations are given in Tables 4.6 and 4.7.

Table 4.5: Maximum likelihood estimators (MLE) and standard deviations (SD) of the Hyper-exponential, considering 20000 observations

Parameter	true value	EM		DM	
		MLE	SD	MLE	SD
$\hat{t}_1$	0.909100	0.915968	7.966296E-3	0.924815	8.026784E-3
$\hat{t}_{12}$	0.090900	0.093476	3.7044226E-3	0.092316	3.685568E-3
$\hat{t}_2$	0.090198	0.092183	4.012385E-3	0.092144	4.028796E-3
$\hat{t}_{23}$	0.009802	0.013576	1.5451314E-3	0.015261	1.6862311E-3
$\hat{t}_3$	0.010000	0.011547	9.307426E-4	0.012154	9.596845E-4

Table 4.6: Correlations of the shifted Negative binomial(3,0.2)

	$\hat{t}_1$	$\hat{t}_{12}$	$\hat{t}_2$	$\hat{t}_{23}$	$\hat{t}_3$
$\hat{t}_1$	1.000000	-0.011774	-0.185519	0.067717	0.010276
$\hat{t}_{12}$	-0.011774	1.000000	-0.933626	-0.262275	-0.497264
$\hat{t}_2$	-0.185519	-0.933626	1.000000	0.191576	0.451236
$\hat{t}_{23}$	0.067717	-0.262275	0.191576	1.000000	-0.684175
$\hat{t}_3$	0.010276	-0.497264	0.451236	-0.684175	1.000000

Table 4.7: Correlations of the Hyper-exponential

	$\hat{t}_1$	$\hat{t}_{12}$	$\hat{t}_2$	$\hat{t}_{23}$	$\hat{t}_3$
$\hat{t}_1$	1.000000	0.345050	0.241808	0.059097	0.042929
$\hat{t}_{12}$	0.345050	1.000000	0.577651	0.187375	0.114814
$\hat{t}_2$	0.241808	0.577651	1.000000	0.417144	0.229993
$\hat{t}_{23}$	0.059097	0.187375	0.417144	1.000000	0.488739
$\hat{t}_3$	0.042929	0.114814	0.229993	0.488739	1.000000

Chapter 5

Multivariate phase-type distributions

There exist a vast amount of definitions concerning multivariate distributions of either exponential or gamma type in the literature (see e.g. Kotz *et.al* [37]). Such distributions either have exponentially or gamma distributed marginals. This has resulted in a rather extensive amount of distributions, many of which are related or only differ from each other vaguely. Also, the class of phase-type distributions, which generalize certain gamma type distributions, has been extended to a multivariate setting, first by Assaf *et.al* [12] and later by Kulkarni [39] (Section 5.1). The latter class, which contains the former as a special case, provides an elegant construction of multivariate phase-type distributions in terms of a single underlying Markov jump process.

In Section 5.2 we will consider the estimation of bivariate phase-type distributions by different methods.

5.1 Two classes of multivariate phase-type distributions

Let $\{X(t)\}_{t\geq 0}$ be a MJP on the finite state-space $\{1, 2, \ldots, p, p+1\}$ with intensity matrix on the form

$$\begin{pmatrix} \mathbf{T} & -\mathbf{Te} \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\mathbf{T}$ is a $(p \times p)$-dimensional invertible matrix. Suppose the initial distribution is $(\boldsymbol{\pi}, \pi_{p+1})$ and define

$$\tau = \inf\{t \geq 0 | X(t) = p + 1\}. \tag{5.1}$$

Assaf *et.al* [12] introduced a class of multivariate phase-type distributions (denoted by MPH) by considering the hitting times to different (possibly overlapping) subsets of the state-space. More specifically, let Γ_i, $i = 1, \ldots, k$, denote absorbing subsets of the state-space and let Y_i be the first hitting time of $\{X(t)\}_{t\geq 0}$ to Γ_i. Then, the k-dimensional vector $\mathbf{Y} = (Y_1, \ldots, Y_k)$ is said to have a phase-type distribution in the class MPH. A rephrasing of the definition of this class says that the reward for Y_i is accumulated with rate 1 in the states belonging to Γ_i^c, where Γ_i^c is the complement of Γ_i. Based on this interpretation Kulkarni [39] introduced the class MPH*, which is a generalization of the MPH class, where rates can be any non-negative real constants.

Let $\mathbf{r} = (r(1), r(2), \ldots, r(p))'$ be a non-negative p-dimensional vector, where $r(i)$ is the rate at which a reward is obtained when the system is in the state i. Now, define

$$Y = \int_0^\tau r(X(t))\mathrm{d}t,$$

where Y is the total reward obtained until absorption in the state $p + 1$. Note that if $r(i) = 1$ for all $1 \leq i \leq p$, then $Y = \tau$. Kulkarni proved in [39] that Y has a phase-type distribution, i.e. when the reward rates are non-negative, the accumulated reward until absorption has a phase-type distribution. This provides a natural way of defining a class of multivariate phase-type distributions that is explained in the following analysis.

For $i = 1, \ldots, k$, let $\mathbf{r}_i = (r_i(1), r_i(2), \ldots, r_i(p))'$ be k non-negative reward vectors. Define $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_k)$ as the $(p \times k)$-dimensional reward matrix. Let $Y_i = \int_0^\tau r_i(X(t))\mathrm{d}t$, the random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ is said to have a multivariate phase-type (MPH*) distribution, with representation $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$, where $\boldsymbol{\pi}$ is the vector of initial probabilities, and $\mathbf{T}$ is the intensity matrix.

Example 5.1 *Let* $\boldsymbol{Y} = (Y_1, Y_2) \sim MPH^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ *with* $\boldsymbol{\pi} = [1]$, $\boldsymbol{T} = [-\lambda]$, $\boldsymbol{r}_1 = [c_1]$ *and* $\boldsymbol{r}_2 = [c_2]$, *where* $\lambda > 0, c_1 \geq 0, c_2 \geq 0$. *Thus, the continuous time Markov chain (CTMC) starts in state 1, spends an* $\exp(\lambda)$ *amount of time there and then gets absorbed in state 2. Let* H *be the time spent in state 1. Then,* $Y_i = c_i H$ *for* $i = 1, 2$. *We have*

$$\begin{aligned} \bar{F}(y_1, y_2) &= \mathbb{P}(Y_1 > y_1, Y_2 > y_2) \\ &= \exp(-\lambda \max(y_1/c_1, y_2/c_2)). \end{aligned}$$

The mass of $\boldsymbol{Y}$ *is concentrated along the line* $L = \{(c_1 y, c_2 y) : y \geq 0\}$ *with the following density*

$$f_{Y_1,Y_2}(c_1 y, c_2 y) = \lambda e^{-\lambda y}.$$

□

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k) \sim MPH^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. Kulkarni [39] has presented three computational techniques for finding the joint distribution of $\mathbf{Y}$.

1) **Partial differential equations**.

For $1 \leq i \leq p$, define

$$\bar{F}_i(y_1, \ldots, y_k) = \mathbb{P}(Y_1 > y_1, \ldots, Y_k > y_k | X(0) = i).$$

Then,

$$\begin{aligned} \bar{F}(y_1, \ldots, y_k) &= \mathbb{P}(Y_1 > y_1, \ldots, Y_k > y_k) \\ &= \sum_{i=1}^{p} \pi_i \bar{F}_i(y_1, \ldots, y_k). \end{aligned}$$

Theorem 5.1 *For* $1 \leq i \leq p$, *the functions* $\bar{F}_i(y_1, \ldots, y_k)$ *satisfy the following system of simultaneous linear partial differential equations*

$$\sum_{j=1}^{k} r_j(i) \frac{\partial \bar{F}_i}{\partial y_j} = \sum_{j=1}^{p} t_{ij} \bar{F}_j. \tag{5.2}$$

2) **Laplace transforms**.

The Laplace Stieltjes transform of $\mathbf{Y}$ is given by

$$\begin{aligned} \phi(s_1, \ldots, s_k) &= \mathbb{E}(\exp(-s_1 Y_1 - s_2 Y_2 - \cdots - s_k Y_k)) \\ &= \sum_{i=1}^{p} \pi_i \phi_i(s_1, \ldots, s_k) + \pi_{p+1}. \end{aligned}$$

where $\phi_i(s_1, \ldots, s_k) = \mathbb{E}(\exp(-s_1 Y_1 - s_2 Y_2 - \cdots - s_k Y_k | X(0) = i)$ and s_i are complex with $Re(s_i) > 0$.

Theorem 5.2 *The conditional Laplace Stieltjes transforms are given by the unique solution*

$$(\boldsymbol{D} - \boldsymbol{T})\boldsymbol{\phi}'(\boldsymbol{s}) = -\boldsymbol{T}\boldsymbol{e}, \tag{5.3}$$

where $\boldsymbol{s} = (s_1, \ldots, s_k)$, $\boldsymbol{\phi}(\boldsymbol{s}) = (\phi_1(\boldsymbol{s}), \ldots, \phi_k(\boldsymbol{s}))$, *and* $\boldsymbol{D}$ *is a diagonal matrix with* $D_{ii} = \sum_{j=1}^{k} s_j r_j(i)$.

Since $\mathbf{T}$ is invertible, $\mathbf{D} - \mathbf{T}$ is invertible in a non-empty neighborhood of $\mathbf{s} = \mathbf{0}$. Hence (5.3) has a unique solution. Thus the Laplace Stieltjes transforms are given by

$$\boldsymbol{\phi}' = -(\mathbf{D} - \mathbf{T})^{-1}\mathbf{T}\mathbf{e}.$$

We could use (5.3) to compute joint moments by taking appropriate derivatives. Even more, Bladt and Nielsen [17] presented a result of how to compute the variance-covariance matrix of $\mathbf{Y}$.

3) **Occupation times**.

Let Z_i be the total time spent by the MJP $\{X(t)\}_{t\geq 0}$ in state i.

Theorem 5.3 *The conditional distribution of* $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ *is given by*

$$\mathbb{P}(Z_1 \leq z_1, Z_2 \leq z_2, \ldots, Z_p \leq z_p | X(0) = i) =$$

$$\sum_{(a_1,\ldots,a_p)} \psi_i(a_1, \ldots, a_p) \prod_{l=1}^{p} \left(1 - \sum_{j=1}^{a_l} e^{-t_l z_l} \frac{(t_l z_l)^j}{j!}\right),$$

where ψ_i *satisfies the following recursive equations*

$$\begin{aligned}
\psi_i(0, \ldots, 0) &= 0, \\
\psi_i(a_1, \ldots, a_p) &= p_{i,p+1}, \quad \text{if } a_j = \delta_{ij}, \\
\psi_i(a_1, \ldots, a_p) &= \sum_{j=1}^{p} p_{ij}\psi_j(a_1, \ldots, a_{j-1}, a_j - 1, a_{j+1}, \ldots, a_p),
\end{aligned}$$

where $\delta_{ij} = 1$ *if* $i = j$ *and 0 otherwise, and* p_{ij} *are the transitions probabilities of the embedded discrete time Markov chain, i.e.,*

$$p_{ij} = \begin{cases} t_{ij}/(-t_{ii}) & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

Now, it is clear that

$$Y_j = \sum_{i=1}^{p} r_j(i) Z_i, \quad 1 \leq j \leq k.$$

We can thus compute the joint distribution of $\mathbf{Y}$ from that of $\mathbf{Z}$ by using standard methods.

The first two methods both face numerical difficulties, since solving a system of simultaneous partial differential equations and inverting multidimensional Laplace transforms is difficult. The third method which is based upon occupations times in CTMCs, turns out to be simpler.

Finally, some properties of the MPH* class are the following:

- Let $Y_1, Y_2, \ldots, Y_k$ be independent random variables in PH. Then $(Y_1, Y_2, \ldots, Y_k)$ is in MPH*.
- Let $(Y_1, Y_2, \ldots, Y_k)$ be in MPH*, then Y_i is in PH for $1 \leq i \leq k$.
- Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ be in MPH* and let $\mathbf{A}$ be a $j \times k$ matrix of non-negative real numbers, for some $j \in \mathbb{N}$. Let $\mathbf{Z}' = \mathbf{A}\mathbf{Y}'$, then $\mathbf{Z}$ is in MPH*.
- The class MPH* is closed under finite convolutions. In particular if (Y_1, Y_2) is in MPH* then $Y_1 + Y_2$ is in PH.
- Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_j)$ be two independent random vectors in MPH*. Then so is $(\mathbf{Y}, \mathbf{Z}) = (Y_1, Y_2, \ldots, Y_k, Z_1, Z_2, \ldots, Z_j)$.
- The class MPH* is closed under finite mixtures.
- Let MPH_n^* be the set of all n-dimensional distributions in MPH*. The set MPH_n^* is dense in the set of all distributions $\mathbb{R}_+^n$.

Finally, let $\mathbf{Y} = (Y_1, \ldots, Y_k)$ such that $\mathbf{Y} \sim MPH^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. Then (see Bladt and Nielsen [17])

$$\langle \mathbf{a}, \mathbf{Y} \rangle \sim PH(\boldsymbol{\pi}, \Delta(\mathbf{Ra})^{-1}\mathbf{T}),$$

for all non-zero $\mathbf{a} \in \mathbb{R}_+^k$, where $\Delta(\mathbf{v})$ is the diagonal matrix with vector $\mathbf{v}$ as diagonal and $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^k$. Thus the Laplace transform of $\langle \mathbf{a}, \mathbf{Y} \rangle$ is given by

$$\begin{aligned} L_{\langle \mathbf{a}, \mathbf{Y} \rangle}(s) &= \boldsymbol{\pi}(s\mathbf{I} - \Delta(\mathbf{Ra})^{-1}\mathbf{T})^{-1}(-\Delta(\mathbf{Ra})^{-1}\mathbf{T})\mathbf{e} \\ &= \boldsymbol{\pi}(s\mathbf{I} - \Delta(\mathbf{Ra})^{-1}\mathbf{T})^{-1}\Delta(\mathbf{Ra})^{-1}(-\mathbf{Te}) \\ &= \boldsymbol{\pi}(\Delta(\mathbf{Ra})(s\mathbf{I} - \Delta(\mathbf{Ra})^{-1}\mathbf{T}))^{-1}\mathbf{t}, \qquad (\text{where } \mathbf{t} = -\mathbf{Te}) \\ &= \boldsymbol{\pi}(s\Delta(\mathbf{Ra}) - \mathbf{T})^{-1}\mathbf{t}, \end{aligned}$$

or,

$$\begin{aligned} L_{\langle \mathbf{a}, \mathbf{Y} \rangle}(s) &= \boldsymbol{\pi}(s\mathbf{I} - \Delta(\mathbf{Ra})^{-1}\mathbf{T})^{-1}(-\Delta(\mathbf{Ra})^{-1}\mathbf{T})\mathbf{e} \\ &= \boldsymbol{\pi}\left((-\mathbf{T})^{-1}\Delta(\mathbf{Ra})(s\mathbf{I} - \Delta(\mathbf{Ra})^{-1}\mathbf{T})\right)^{-1}\mathbf{e} \\ &= \boldsymbol{\pi}\left(s(-\mathbf{T})^{-1}\Delta(\mathbf{Ra}) + \mathbf{I}\right)^{-1}\mathbf{e}, \end{aligned}$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension. For $s = 1$, $L_{\langle \mathbf{a}, \mathbf{Y} \rangle}(s)$ is the joint Laplace transform of $\mathbf{Y}$ at $\mathbf{a}$.

5.2 Estimation of bivariate phase-type distributions

Consider two phase-type random variables $Y_1 \sim PH_{p_1}(\boldsymbol{\pi}_1, \mathbf{T}_{11})$ and $Y_2 \sim PH_{p_2}(\boldsymbol{\pi}_2, \mathbf{T}_{22})$, and let $\{X(t)\}_{t \geq 0}$ be the Markov jump process with the set of transient states $\{1, \ldots, p\}$ split into two sets: $E_1 = \{1, \ldots, p_1\}$ and $E_2 = \{p_1 + 1, \ldots, p\}$. Let p_2 denotes the number of states in E_2, i.e. $p_2 = p - p_1$. The state space is thus $E = E_1 \cup E_2 \cup \{p + 1\}$, where the state $p + 1$ is absorbing. Suppose the Markov process is only allowed to start in a state belonging to E_1, which imposes the following structure on its initial distribution

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{p_1}, 0, \ldots, 0) = (\boldsymbol{\pi}_1, \mathbf{0}).$$

That is, only the first p_1 elements of $\boldsymbol{\pi}$, which are collected in the vector $\boldsymbol{\pi}_1$, are allowed to be larger than zero.

The generator of the Markov process is partitioned with a $(p \times p)$-dimensional matrix $\mathbf{T}$ as its largest part. Now, due to the partition of the p non-absorbing states into two groups, E_1 and E_2, and for simplicity (see [1]), the subintensity matrix $\mathbf{T}$ can partitioned into four blocks

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix},$$

where $\mathbf{T}_{11}$ is a $(p_1 \times p_1)$-dimensional matrix containing the transitions intensities for jumps within E_1, $\mathbf{T}_{12}$ is of dimension $(p_1 \times p_2)$ and contains the intensities for transitions from E_1 to E_2, and $\mathbf{T}_{22}$ is a $(p_2 \times p_2)$-dimensional matrix containing the intensities for transitions within E_2. The elements in the fourth sub-matrix are all set to zero, which implies that the process cannot return to a state in E_1 once it has entered E_2.

In order to make absorption possible only from the states in the second group E_2, we fix the first p_1 states in the exit vector $\mathbf{t}$ to zero, i.e.

$$\mathbf{t} = (0, \ldots, 0, t_{p_1+1}, \ldots, t_p)' = (\mathbf{0}, \mathbf{t}_2)'.$$

Consider the reward matrix given by

$$\mathbf{R} = \begin{pmatrix} \mathbf{e} & \mathbf{0} \\ \mathbf{0} & \mathbf{e} \end{pmatrix}. \tag{5.4}$$

The joint density of (Y_1, Y_2) is given by

$$\begin{aligned} f_{(Y_1,Y_2)}(y_1,y_2) &= \sum_j \mathbb{P}(Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2 | X(0)=j)\mathbb{P}(X(0)=j) \\ &= \sum_j \mathbb{P}(Y_2 = y_2 | Y_1 = y_1)\mathbb{P}(Y_1 = y_1 | X(0)=j)\mathbb{P}(X(0)=j) \\ &= \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2, \end{aligned} \tag{5.5}$$

the corresponding joint Laplace transform is thus given by

$$L_{(Y_1,Y_2)}(s_1,s_2) = \boldsymbol{\pi}_1(s_1\mathbf{I}-\mathbf{T}_{11})^{-1})\mathbf{T}_{12}(s_2\mathbf{I}-\mathbf{T}_{22})^{-1}\mathbf{t}_2.$$

The marginal distribution of Y_2 is given by

$$\begin{aligned} f_{Y_2}(y_2) &= \int_0^\infty \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2 \mathrm{d}y_1 \\ &= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2. \end{aligned}$$

A sufficient condition for $Y_2 \sim PH(\boldsymbol{\pi}_2, \mathbf{T}_{22})$ is hence $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}$. Now, we can obtain the other marginal

$$\begin{aligned} f_{Y_1}(y_1) &= \int_0^\infty \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2 \mathrm{d}y_2 \\ &= \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-1}\mathbf{t}_2 \\ &= \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-1}(-\mathbf{T}_{22}\mathbf{e}) \\ &= \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}\mathbf{e}. \end{aligned}$$

Since $Y_1 \sim PH(\boldsymbol{\pi}_1, \mathbf{T}_{11})$, then $\mathbf{t}_1 = \mathbf{T}_{12}\mathbf{e}$. Thus, the matrix $\mathbf{T}_{12}$ has to satisfy the following system of equations

$$\begin{cases} \boldsymbol{\pi}_2 &= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{T}_{12} \\ \mathbf{t}_1 &= \mathbf{T}_{12}\mathbf{e}. \end{cases} \tag{5.6}$$

Suppose this system can be written in the matrix form $\mathbf{Ax} = \mathbf{b}$, where the entries of the vector $\mathbf{x}$ are the elements of the matrix $\mathbf{T}_{12}$, i.e. if

$$\mathbf{T}_{12} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1p_2} \\ d_{21} & d_{22} & \dots & d_{2p_2} \\ \vdots & \vdots & \dots & \vdots \\ d_{p_1 1} & d_{p_1 2} & \dots & d_{p_1 p_2} \end{pmatrix},$$

then $\mathbf{x} = (d_{11}, d_{12}, \dots, d_{1p_2}, d_{21}, d_{22}, \dots, d_{2p_2}, \dots, d_{p_1 1}, d_{p_1 2}, \dots, d_{p_1 p_2})'$. The

matrix $\mathbf{A}$ is the coefficient matrix, and has the following general form

$$\begin{pmatrix} u_1 & 0 & \dots & 0 & \dots & u_{p-1} & 0 & \dots & 0 & u_{p_1} & 0 & \dots & 0 \\ 0 & u_1 & \dots & 0 & \dots & 0 & u_{p-1} & \dots & 0 & 0 & u_{p_1} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & u_1 & \dots & 0 & 0 & \dots & u_{p-1} & \dots & 0 & 0 & u_{p_1} \\ 1 & 1 & \dots & 1 & \dots & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{pmatrix},$$

where u_i represents the i-th element of $\boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}$. Note that the matrix $\mathbf{A}$ is of dimension $(p_2 + p_1) \times (p_1 p_2)$, which turns out to be singular, with rank equals to $p_1 + p_2 - 1$, this means that we are in the case with infinite solutions. In order to avoid it, we choose $\mathbf{T}_{12}$ with degree of freedom equals to $p_1 p_2 - (p_1 + p_2 - 1)$, i.e. fixing the values of $d_{11}, d_{12}, \dots, d_{1,p_2-1}, d_{21}, d_{22}, \dots, d_{2,p_2-1}$, $\dots, d_{p_1-1,1}, d_{p_1-1,2}, \dots, d_{p_1-1,p_2-1}$, we can re-write the matrix $\mathbf{A}$ in the following form

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & \dots & 0 & u_{p_1} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & u_{p_1} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ u_1 & u_2 & \dots & u_{p_1-1} & 0 & 0 & \dots & u_{p_1} \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix},$$

and the vector $\mathbf{x}$ as $(d_{1p_2}, d_{2p_2}, \dots, d_{p_1-1,p_2}, d_{p_1 1}, d_{p_1 2}, \dots, d_{p_1,p_2})'$. Finally, the vector $\mathbf{b}$ is given by

$$\mathbf{b} = \begin{pmatrix} v_1 - \sum_{i=1}^{p_1-1} u_i d_{i1} \\ v_2 - \sum_{i=1}^{p_1-1} u_i d_{i1} \\ \vdots \\ v_{p_2-1} - \sum_{i=1}^{p_1-1} u_i d_{i1} \\ v_{p_2} \\ w_1 - (p_2 - 1) \\ w_2 - (p_2 - 1) \\ \vdots \\ w_{p_1-1} - (p_2 - 1) \end{pmatrix},$$

where v_i is the i-th element of $\boldsymbol{\pi}_2$ and w_i is the i-th element of $\mathbf{t}_1$.

In order to get the joint moments and the variance-covariance matrix of (Y_1, Y_2) we present the following result (see also [18]).

Theorem 5.4 *The joint moments of Y_1 and Y_2 are given by*

$$\mathbb{E}(Y_1^{n_1} Y_2^{n_2}) = n_1! n_2! \boldsymbol{\pi}_1 (-\boldsymbol{T}_{11})^{-n_1-1} \boldsymbol{T}_{12} (-\boldsymbol{T}_{22})^{-n_2-1} \boldsymbol{t}_2, \quad n_1, n_2 \in \mathbb{N}.$$

In particular, the covariance between Y_1 and Y_2 is given by

$$\mathrm{Cov}(\mathrm{Y}_1, \mathrm{Y}_2) = \boldsymbol{\pi}_1 (-\boldsymbol{T}_{11})^{-1} ((-\boldsymbol{T}_{11})^{-1} \boldsymbol{T}_{12} - \boldsymbol{e}\boldsymbol{\pi}_2)(-\boldsymbol{T}_{22})^{-1} \boldsymbol{e}.$$

Proof. Follows immediately differentiating the Laplace transform. In deriving the formula for the covariance we assume that $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}$. We know that $\mathbb{E}(Y_1) = \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{e}$, $\mathbb{E}(Y_2) = \boldsymbol{\pi}_2(-\mathbf{T}_{22})^{-1}\mathbf{e}$, $\mathbb{E}(Y_1 Y_2) = \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-2}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-2}\mathbf{t}_2$ then

$$\begin{aligned}
\mathrm{Cov}(\mathrm{Y}_1, \mathrm{Y}_2) &= \mathbb{E}(Y_1, Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2) \\
&= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-2}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-2}\mathbf{t}_2 - \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{e}\boldsymbol{\pi}_2(-\mathbf{T}_{22})^{-1}\mathbf{e} \\
&= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}((-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-2}\mathbf{t}_2 - \mathbf{e}\boldsymbol{\pi}_2(-\mathbf{T}_{22})^{-1}\mathbf{e}) \\
&= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}((-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}(-\mathbf{T}_{22})^{-1}\mathbf{e} - \mathbf{e}\boldsymbol{\pi}_2(-\mathbf{T}_{22})^{-1}\mathbf{e}) \\
&= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}((-\mathbf{T}_{11})^{-1}\mathbf{T}_{12} - \mathbf{e}\boldsymbol{\pi}_2)(-\mathbf{T}_{22})^{-1}\mathbf{e}.
\end{aligned}$$

■

Example 5.2 *Let $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = (1, 0, 0, 0)$ and*

$$\boldsymbol{T}_{11} = \boldsymbol{T}_{22} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

And let $\boldsymbol{T}$ and $\boldsymbol{R}$ defined as before. Then the conditions $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_1(-\boldsymbol{T}_{11})^{-1}\boldsymbol{T}_{12}$ and $\boldsymbol{T}_{12}\boldsymbol{e} = \boldsymbol{t}_1$ implies that the only possible form of $\boldsymbol{T}_{12}$ is the matrix

$$\boldsymbol{T}_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

□

5.2.1 Via the EM algorithm

Let $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ denote the full data. The likelihood function for the complete data is given by

$$\begin{aligned} L_f(\boldsymbol{\theta}; \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \prod_{i=1}^{p_1} \pi_i^{B_i} \prod_{i=1}^{p_1} \prod_{j=1, j\neq i}^{p_1} t_{ij}^{N_{ij}} e^{-t_{ij} Z_i} \prod_{i=1}^{p_1} \prod_{j=p_1+1}^{p} t_{ij}^{N_{ij}} e^{-t_{ij} Z_i} \\ &\quad \prod_{i=p_1+1}^{p} \prod_{j=p_1+1, j\neq i}^{p} t_{ij}^{N_{ij}} e^{-t_{ij} Z_i} \prod_{i=p_1+1}^{p} t_i^{N_i} e^{-t_i Z_i}, \end{aligned}$$

where B_i is the number of processes starting in state i, N_i the number of processes exiting from state i to the absorbing state, N_{ij} the number of jumps from state i to j among all processes, and Z_i the total time spent in state i.

Like the analysis given in Section 3.2.2 for the univariate case, we will find the conditional expectations of the sufficient statistics: B_i's, N_i's, Z_i's, and N_{ij}'s (see Asmussen's notation in [11]).

In the following analysis we present the formulas of the conditional expectations considering the reward matrix $\mathbf{R}$ given in (5.4).

Formula for B_i's

For $i = 1, \ldots, p_1$,

$$\begin{aligned} \mathbb{E}(B_i | Y_1 = y_1, Y_2 = y_2) &= \mathbb{P}(X(0) = i | Y_1 = y_1, Y_2 = y_2) \\ &= \frac{\mathbb{P}(X(0) = i, Y_1 = y_1, Y_2 = y_2)}{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)} \\ &= \frac{\mathbb{P}(X(0) = i)\mathbb{P}(Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2 | X(0) = i)}{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)} \\ &= \frac{\mathbb{P}(X(0) = i)\mathbb{P}(Y_1 \in \mathrm{d}y_1 | X(0) = i)\mathbb{P}(Y_2 \in \mathrm{d}y_2 | Y_1 \in \mathrm{d}y_1)}{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)} \\ &= \frac{\pi_i \boldsymbol{e}_i' e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2}, \end{aligned}$$

where π_i is the i-th element of the vector $\boldsymbol{\pi}_1$.

Formula for N_i's

For $i = p_1 + 1, \ldots, p$, let $i^* = i - p_1$ then,

$$\begin{aligned}
&\mathbb{E}(X(Y_2 - \epsilon) = i^* | Y_1 = y_1, Y_2 = y_2) \\
&= \frac{\mathbb{P}(X(Y - \epsilon) = i^*, Y_1 = y_1, Y_2 = y_2)}{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)} \\
&= \frac{\mathbb{P}(Y_1 \in \mathrm{d}y_1)\mathbb{P}(X(Y - \epsilon) = i^* | Y_1 \in \mathrm{d}y_1)\mathbb{P}(Y_2 \in \mathrm{d}y_2 | X(Y_2 - \epsilon) = i^*)}{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)} \\
&= \frac{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22}(y_2 - \epsilon)} \mathbf{e}_{i^*} t_{i^*}}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2},
\end{aligned}$$

taking $\epsilon \to 0$, we get

$$\mathbb{E}(N_i | Y_1 = y_1, Y_2 = y_2) = \frac{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{e}_{i^*} t_{i^*}}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2},$$

where t_{i^*} is the i^*-th element of the vector $\mathbf{t}_2$.

Formula for Z_i's

For $i = 1, \ldots, p$,

$$\begin{aligned}
\mathbb{E}(Z_i | Y_1 = y_1, Y_2 = y_2) &= \int \mathbb{P}(X(u) = i | Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)\mathrm{d}u \\
&= \frac{\int \mathbb{P}(X(u) = i, Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)\mathrm{d}u}{\mathbb{P}(Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)}.
\end{aligned}$$

(a) $1 \leq i \leq p_1$

Analyzing the numerator, we have

$$\begin{aligned}
&\int_0^{y_1} \mathbb{P}(X(u) = i, Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)\mathrm{d}u \\
&= \int_0^{y_1} \mathbb{P}(X(u) = i)\mathbb{P}(Y_1 \in \mathrm{d}y_1 | X(u) = i)\mathbb{P}(Y_2 \in \mathrm{d}y_2 | Y_1 \in \mathrm{d}y_1)\mathrm{d}u \\
&= \int_0^{y_1} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11} u} \mathbf{e}_i \mathbf{e}_i' e^{\mathbf{T}_{11}(y_1 - u)} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2 \mathrm{d}u,
\end{aligned}$$

thus,

$$\mathbb{E}(Z_i | Y_1 = y_1, Y_2 = y_2) = \frac{\int_0^{y_1} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11} u} \mathbf{e}_i \mathbf{e}_i' e^{\mathbf{T}_{11}(y_1 - u)} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2 \mathrm{d}u}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_1} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_2} \mathbf{t}_2}.$$

(b) $\underline{p_1+1 \le i \le p}$

Let $i^* = i - p_1$,

$$\begin{aligned}
&\int_0^{y_2} \mathbb{P}(X(u) = i^*, Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)\mathrm{d}u \\
&\quad = \int_0^{y_2} \mathbb{P}(Y_1 \in \mathrm{d}y_1)\mathbb{P}(X(u) = i^* | Y_1 \in \mathrm{d}y_1)\mathbb{P}(Y_2 \in \mathrm{d}y_2 | X(u) = i^*)\mathrm{d}u \\
&\quad = \int_0^{y_2} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}u}\mathbf{e}_{i^*}\mathbf{e}'_{i^*}e^{\mathbf{T}_{22}(y_2-u)}\mathbf{t}_2\mathrm{d}u,
\end{aligned}$$

thus,

$$\mathbb{E}(Z_i | Y_1 = y_1, Y_2 = y_2) = \frac{\int_0^{y_2} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}u}\mathbf{e}_{i^*}\mathbf{e}'_{i^*}e^{\mathbf{T}_{22}(y_2-u)}\mathbf{t}_2\mathrm{d}u}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2}.$$

Formula for N_{ij}'s

We have that $N_{ij}^{\epsilon} = \sum_u \mathbf{1}_{\{X(u\epsilon)=i, X((u+1)\epsilon)=j\}}$ for $\epsilon > 0$ and $i \neq j$. Then,

$$\mathbb{E}(N_{ij}^{\epsilon} | Y_1 = y_1, Y_2 = y_2) = \frac{\sum \mathbb{P}(X(u\epsilon) = i, X((u+1)\epsilon) = j, Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)}{\mathbb{P}(Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2)}.$$

We will analyze only the numerator.

(a) $\underline{1 \le i \le p_1 \text{ and } 1 \le j \le p}$

(a.1) $\underline{1 \le j \le p_1}$

$$\begin{aligned}
&\sum \mathbb{P}(X(u\epsilon) = i, X((u+1)\epsilon) = j, Y_1 \in \mathrm{d}y_1, Y_2 \in \mathrm{d}y_2) \\
&\quad = \sum \mathbb{P}(X(u\epsilon) = i)\mathbb{P}(X((u+1)\epsilon) = j | X(u\epsilon) = i) \\
&\qquad \times \mathbb{P}(Y_1 \in \mathrm{d}y_1 | X((u+1)\epsilon) = j)\mathbb{P}(Y_2 \in \mathrm{d}y_2 | Y_1 \in \mathrm{d}y_1) \\
&\quad = \sum \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}u\epsilon}\mathbf{e}_i t_{ij}\mathbf{e}'_j e^{\mathbf{T}_{11}(y_1-(u+1)\epsilon)}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2,
\end{aligned}$$

if $\epsilon \to 0$ then

$$\mathbb{E}(N_{ij} | Y_1 = y_1, Y_2 = y_2) = t_{ij}\frac{\int_0^{y_1} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}u}\mathbf{e}_i\mathbf{e}'_j e^{\mathbf{T}_{11}(y_1-u)}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2\mathrm{d}u}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2},$$

where t_{ij} is the (i,j)-th element of $\mathbf{T}_{11}$.

(a.2) $\underline{p_1+1 \le j \le p}$

Let $j^* = j - p_1$,

$$\begin{aligned}
&\mathbb{P}(X(u)=i, X(u+1)=j^*, Y_1=y_1, Y_2=y_2) \\
&\quad = \mathbb{P}(Y_1 \in \mathrm{d}y_1)\mathbb{P}(X(u)=i|Y_1 \in \mathrm{d}y_1) \\
&\qquad \times \mathbb{P}(X(u+1)=j^*|X(u)=i)\mathbb{P}(Y_2 \in \mathrm{d}y_2|X(u+1)=j^*) \\
&\quad = \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{e}_i t_{ij^*}\mathbf{e}'_{j^*}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2,
\end{aligned}$$

then,

$$\mathbb{E}(N_{ij}|Y_1=y_1, Y_2=y_2) = t_{ij^*}\frac{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{e}_i\mathbf{e}'_{j^*}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2},$$

where t_{ij^*} is the (i,j^*)-th element of $\mathbf{T}_{12}$.

(b) $\underline{p_1+1 \le i \le p \text{ and } p_1+1 \le j \le p}$

Let $i^* = i - p_1$ and $j^* = j - p_1$,

$$\begin{aligned}
&\mathbb{P}(X(u)=i^*, X(u+1)=j^*, Y_1=y_1, Y_2=y_2) \\
&\quad = \mathbb{P}(Y_1 \in \mathrm{d}y_1)\mathbb{P}(X(u)=i^*|Y_1 \in \mathrm{d}y_1) \\
&\qquad \times \mathbb{P}(X(u+1)=j^*|X(u)=i^*)\mathbb{P}(Y_2 \in \mathrm{d}y_2|X(u+1)=j^*) \\
&\quad = \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}u}\mathbf{e}_{i^*}t_{i^*j^*}\mathbf{e}'_{j^*}e^{\mathbf{T}_{22}(y_2-u)}\mathbf{t}_2,
\end{aligned}$$

then

$$\mathbb{E}(N_{ij}|Y_1=y_1, Y_2=y_2) = t_{i^*j^*}\frac{\int_0^{y_2}\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}u}\mathbf{e}_{i^*}\mathbf{e}'_{j^*}e^{\mathbf{T}_{22}(y_2-u)}\mathbf{t}_2\mathrm{d}u}{\boldsymbol{\pi}_1 e^{\mathbf{T}_{11}y_1}\mathbf{T}_{12}e^{\mathbf{T}_{22}y_2}\mathbf{t}_2},$$

where $t_{i^*j^*}$ is the (i^*,j^*)-th element of $\mathbf{T}_{22}$.

Thus, we have proved the following theorem.

Theorem 5.5 *The conditional expectations of the sufficient statistics for bivariate phase-type distributions are given by:*

- *For* $i = 1, \ldots, p_1$,

$$\mathbb{E}(B_i|Y_1=y_1, Y_2=y_2) = \frac{\pi_i \boldsymbol{e}'_i e^{\boldsymbol{T}_{11}y_1}\boldsymbol{T}_{12}e^{\boldsymbol{T}_{22}y_2}\boldsymbol{t}_2}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11}y_1}\boldsymbol{T}_{12}e^{\boldsymbol{T}_{22}y_2}\boldsymbol{t}_2}, \tag{5.7}$$

where π_i *is the* i*-th element of the vector* $\boldsymbol{\pi}_1$*.*

- *For* $i = p_1 + 1, \ldots, p$,

$$\mathbb{E}(N_i | Y_1 = y_1, Y_2 = y_2) = \frac{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{e}_{i-p_1} t_{i-p_1}}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, \tag{5.8}$$

where t_{i-p_1} *is the* $i - p_1$*-th element of the vector* $\boldsymbol{t}_2$.

- $\mathbb{E}(Z_i | Y_1 = y_1, Y_2 = y_2) =$

$$\begin{cases} \dfrac{\int_0^{y_1} \boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} u} \boldsymbol{e}_i \boldsymbol{e}_i' e^{\boldsymbol{T}_{11}(y_1 - u)} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2 \, du}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, & i = 1 \ldots, p_1, \\ \dfrac{\int_0^{y_2} \boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} u} \boldsymbol{e}_{i-p_1} \boldsymbol{e}_{i-p_1}' e^{\boldsymbol{T}_{22}(y_2 - u)} \boldsymbol{t}_2 \, du}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, & i = p_1 + 1 \ldots, p. \end{cases} \tag{5.9}$$

- $\mathbb{E}(N_{ij} | Y_1 = y_1, Y_2 = y_2) =$

$$\begin{cases} t_{ij} \dfrac{\int_0^{y_1} \boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} u} \boldsymbol{e}_i \boldsymbol{e}_j' e^{\boldsymbol{T}_{11}(y_1 - u)} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2 \, du}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, \quad i, j = 1, \ldots, p_1, \\ \quad \text{where } t_{ij} \text{ is the } (i,j)\text{-th element of } \boldsymbol{T}_{11}. \\ \\ t_{i,j-p_1} \dfrac{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{e}_i \boldsymbol{e}_{j-p_1}' e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, \quad i = 1, \ldots, p_1, j = p_1 + 1, \ldots, p \\ \quad \text{where } t_{i,j-p_1} \text{ is the } (i, j-p_1)\text{-th element of } \boldsymbol{T}_{12}. \\ \\ t_{i-p_1,j-p_1} \dfrac{\int_0^{y_2} \pi_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} u} \boldsymbol{e}_{i-p_1} \boldsymbol{e}_{j-p_1}' e^{\boldsymbol{T}_{22}(y_2 - u)} \boldsymbol{t}_2 \, du}{\boldsymbol{\pi}_1 e^{\boldsymbol{T}_{11} y_1} \boldsymbol{T}_{12} e^{\boldsymbol{T}_{22} y_2} \boldsymbol{t}_2}, \\ \quad i, j = p_1 + 1, \ldots, p, \text{ where } t_{i-p_1,j-p_1} \text{ is the } (i-p_1, j-p_1)\text{-th} \\ \quad \text{element of } \boldsymbol{T}_{22}. \end{cases} \tag{5.10}$$

In general, if the matrix $\mathbf{R}$ has the following form

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 \end{pmatrix},$$

where $\mathbf{R}_1$ is a column vector of dimension p_1, $\mathbf{R}_2$ is a p_2-dimensional column vector, and $\mathbf{0}$ is zero vector of the required dimension, the new joint density of (Y_1, Y_2) is given by

$$f_{(Y_1,Y_2)}(y_1, y_2) = \boldsymbol{\pi}_1 e^{\Delta(\mathbf{R}_1)^{-1} \mathbf{T}_{11} y_1} \Delta(\mathbf{R}_1)^{-1} \mathbf{T}_{12} e^{\Delta(\mathbf{R}_2)^{-1} \mathbf{T}_{22} y_2} \Delta(\mathbf{R}_2)^{-1} \mathbf{t}_2, \tag{5.11}$$

where $\Delta(\mathbf{a})$ is the diagonal matrix of the vector $\mathbf{a}$.

Taking

$$\begin{aligned}\mathbf{T}_{11}^{\bullet} &= \Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{11}\\ \mathbf{T}_{12}^{\bullet} &= \Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{12}\\ \mathbf{T}_{22}^{\bullet} &= \Delta(\mathbf{R}_2)^{-1}\mathbf{T}_{22}\\ \mathbf{t}_{2}^{\bullet} &= \Delta(\mathbf{R}_2)^{-1}\mathbf{t}_{2},\end{aligned}$$

the density given in (5.11) becomes

$$f_{(Y_1,Y_2)}(y_1,y_2) = \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}^{\bullet} y_1}\mathbf{T}_{12}^{\bullet} e^{\mathbf{T}_{22}^{\bullet} y_2}\mathbf{t}_{2}^{\bullet}. \tag{5.12}$$

We can see that the densities given in (5.5) and (5.12) have the same form. Actually, the conditions given in (5.6) remain the same, since

1.

$$\begin{aligned}\boldsymbol{\pi}_2 &= \boldsymbol{\pi}_1(-\mathbf{T}_{11}^{\bullet})^{-1}\mathbf{T}_{12}^{\bullet}\\ &= \boldsymbol{\pi}_1(-\Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{11})^{-1}\Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{12}\\ &= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\Delta(\mathbf{R}_1)\Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{12}\\ &= \boldsymbol{\pi}_1(-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}.\end{aligned}$$

2.

$$\begin{aligned}\mathbf{t}_{1}^{\bullet} &= \mathbf{T}_{12}^{\bullet}\mathbf{e}\\ -\Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{11}\mathbf{e} &= \Delta(\mathbf{R}_1)^{-1}\mathbf{T}_{12}\mathbf{e}\\ -\mathbf{T}_{11}\mathbf{e} &= \mathbf{T}_{12}\mathbf{e}\\ \mathbf{t}_1 &= \mathbf{T}_{12}\mathbf{e}.\end{aligned}$$

Hence, given the vectors $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, the matrices $\mathbf{T}_{11}$, $\mathbf{T}_{22}$, $\mathbf{T}_{12}$ (found it above), and the general form of $\mathbf{R}$, we can obtain $\mathbf{T}_{11}^{\bullet}$, $\mathbf{T}_{12}^{\bullet}$, $\mathbf{T}_{22}^{\bullet}$, and, $\mathbf{t}_{2}^{\bullet}$. The new transition matrix, denoted by $\mathbf{T}^{\bullet}$, will be given by

$$\mathbf{T}^{\bullet} = \begin{pmatrix}\mathbf{T}_{11}^{\bullet} & \mathbf{T}_{12}^{\bullet}\\ \mathbf{0} & \mathbf{T}_{22}^{\bullet}\end{pmatrix},$$

and the formulae for the estimation via the EM algorithm given in (5.7)-(5.10) remain the same.

L. Ahlstrom *et.al* in [1] presented the estimation of bivariate phase-type distributions via the EM algorithm considering the matrix of rewards $\mathbf{R}$ in the following form

$$\mathbf{R} = \begin{pmatrix}\mathbf{e} & \mathbf{0}\\ \mathbf{e} & \mathbf{e}\end{pmatrix}. \tag{5.13}$$

Example 5.1 *We generate a sample of 100 observations from a bivariate distribution with gamma marginals:* $Y_1 \sim Gamma(shape = 2, scale = 1)$, $Y_2 \sim Gamma(shape = 3, scale = 2)$ *and a normal copula.*

Table 5.1: Log-likelihood (LL) of bivariate copula of $Gamma(2,1)$ and $Gamma(3,2)$

LL	time
-404.9689	1944.0126

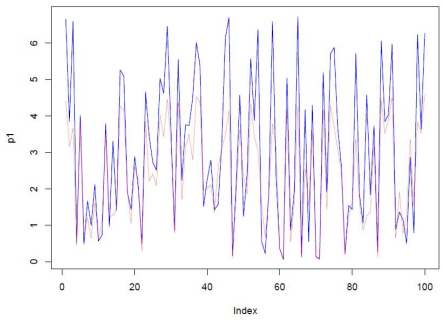

Figure 5.1: Estimation of bivariate gamma using the EM-algorithm

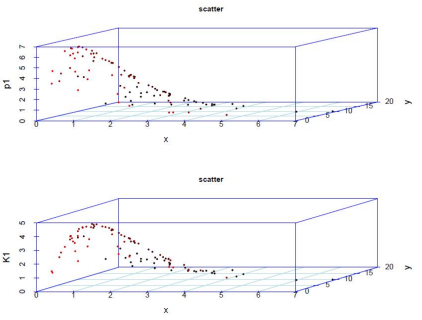

Figure 5.2: Scatterplot of bivariate gamma

5.2.2 Via direct method

Based on the idea of the estimation of univariate PH distributions by the DM (see Section 3.2.4), we will consider the estimation of bivariate PH distributions via the DM.

Let $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}) = \{(y_1^{(1)}, y_1^{(2)}), (y_2^{(1)}, y_2^{(2)}), \ldots, (y_M^{(1)}, y_M^{(2)})\}$ be a bivariate sample from the density (5.5) of size M.

The likelihood function is given by

$$L(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) = \prod_{k=1}^{M} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2,$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$. Hence the log-likelihood function is

$$l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) = \sum_{k=1}^{M} \log f(y_k^{(1)}, y_k^{(2)}),$$

where $f(y_k^{(1)}, y_k^{(2)}) = \boldsymbol{\pi}_1 e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2$. Let π_1^j be the j-th element of the vector $\boldsymbol{\pi}_1$, then substituting $\boldsymbol{\pi}_1 = \sum_{j=1}^{p_1-1} \pi_1^j \mathbf{e}_j' + \left(1 - \sum_{j=1}^{p_1-1} \pi_1^j\right) \mathbf{e}_{p_1}'$ we get

$$\begin{aligned} f(y_k^{(1)}, y_k^{(2)}) &= \left(\sum_{j=1}^{p_1-1} \pi_1^j \mathbf{e}_j' \right) e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2 \\ &\quad + \left(1 - \sum_{j=1}^{p_1-1} \pi_1^j \right) \mathbf{e}_{p_1}' e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2. \end{aligned}$$

If $R_m(y_k^{(1)}, y_k^{(2)}) = \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2$, then

$$f(y_k^{(1)}, y_k^{(2)}) = \sum_{j=1}^{p_1-1} \pi_1^j R_j(y_k^{(1)}, y_k^{(2)}) + \left(1 - \sum_{j=1}^{p_1-1} \pi_1^j \right) R_{p_1}(y_k^{(1)}, y_k^{(2)}).$$

Like in the univariate case we have to consider a transformation of the parameters. For $r = 1, 2$, (representing both variables), and for $i = 1, \ldots, p_r - 1$, generate $-\infty < \varrho_i, \xi_i < \infty$. Let us take the following transformations

$$\pi_r^i = \frac{e^{\tau_i}}{1 + \sum_{s=1}^{p_r-1} e^{\tau_s}}, \quad \text{and} \quad \pi_r^{p_r} = \frac{1}{1 + \sum_{s=1}^{p_r-1} e^{\tau_s}},$$

where

$$\tau_i = \begin{cases} \varrho_i & \text{if } r = 1 \\ \xi_i & \text{if } r = 2. \end{cases}$$

For $i, j = 1, \ldots, p_r$, generate $-\infty < \gamma_{ij}, \eta_{ij} < \infty$, such as

$$t_{ij} = e^{\tau_{ij}}, \quad i \neq j, \quad t_i = e^{\tau_{ii}},$$

where

$$\tau_{ij} = \begin{cases} \gamma_{ij} & \text{if } r = 1 \\ \eta_{ij} & \text{if } r = 2. \end{cases}$$

The gradient vector is given by

$$\left(\left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)})}{\partial \varrho_i} \right)_{i=1,\ldots,p_1-1}, \left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)})}{\partial \gamma_{ij}} \right)_{i,j=1,\ldots,p_1}, \right.$$

$$\left. \left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)})}{\partial \xi_i} \right)_{i=1,\ldots,p_2-1}, \left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)})}{\partial \eta_{ij}} \right)_{i,j=1,\ldots,p_2} \right).$$

Suppose $\tau^* \in \{\varrho_i, \xi_i, \gamma_{ij}, \eta_{ij}\}$, then

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y}^{(1)}, \mathbf{y}^{(2)})}{\partial \tau^*} = \sum_{k=1}^{M} \frac{1}{f(y_k^{(1)}, y_k^{(2)})} \frac{\partial f(y_k^{(1)}, y_k^{(2)})}{\partial \tau^*},$$

where

$$\begin{aligned}\frac{\partial f(y_k^{(1)}, y_k^{(2)})}{\partial \varrho_m} &= \sum_{s=1}^{p_1-1} \pi_1^s \left(\frac{\partial R_s(y_k^{(1)}, y_k^{(2)})}{\partial \varrho_m} - \frac{\partial R_{p_1}(y_k^{(1)}, y_k^{(2)})}{\partial \varrho_m} \right) \\ &\quad + \frac{\partial \pi_1^s}{\partial \varrho_m} (R_s(y_k^{(1)}, y_k^{(2)}) - R_{p_1}(y_k^{(1)}, y_k^{(2)})) + \frac{\partial R_{p_1}(y_k^{(1)}, y_k^{(2)})}{\partial \varrho_m}.\end{aligned}$$

For $s \in \{1, \ldots, p_1 - 1\}$, the derivative $\frac{\partial \pi_1^s}{\partial \varrho_m}$ is the same as in the univariate case, (see (3.16)), and

$$\frac{\partial R_s(y_k^{(1)}, y_k^{(2)})}{\partial \varrho_m} = \mathbf{e}_s' e^{\mathbf{T}_{11} y_k^{(1)}} \frac{\partial \mathbf{T}_{12}}{\partial \varrho_m} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2.$$

For $\tau^* \in \{\xi_i, \gamma_{ij}, \eta_{ij}\}$ we have

$$\frac{\partial f(y_k^{(1)}, y_k^{(2)})}{\partial \tau^*} = \sum_{s=1}^{p_1-1} \pi_1^s \frac{\partial R_s(y_k^{(1)}, y_k^{(2)})}{\partial \tau^*} + \left(1 - \sum_{s=1}^{p_1-1} \pi_1^s\right) \frac{\partial R_p(y_k^{(1)}, y_k^{(2)})}{\partial \tau^*},$$

and

$$\begin{aligned}
\tfrac{\partial R_m(y_k^{(1)}, y_k^{(2)})}{\partial \gamma_{ij}} &= \mathbf{e}_m' \tfrac{\partial e^{\mathbf{T}_{11} y_k^{(1)}}}{\partial \gamma_{ij}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2 + \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \tfrac{\partial \mathbf{T}_{12}}{\partial \gamma_{ij}} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2, \\
\tfrac{\partial R_m(y_k^{(1)}, y_k^{(2)})}{\partial \xi_i} &= \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \tfrac{\partial \mathbf{T}_{12}}{\partial \xi_i} e^{\mathbf{T}_{22} y_k^{(2)}} \mathbf{t}_2, \\
\tfrac{\partial R_m(y_k^{(1)}, y_k^{(2)})}{\partial \eta_{ij}} &= \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} \tfrac{\partial e^{\mathbf{T}_{22} y_k^{(2)}}}{\partial \eta_{ij}} \mathbf{t}_2, \quad i \neq j, \\
\tfrac{\partial R_m(y_k^{(1)}, y_k^{(2)})}{\partial \eta_{ii}} &= \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} e^{\mathbf{T}_{22} y_k^{(2)}} \tfrac{\partial \mathbf{t}_2}{\partial \eta_{ii}} + \mathbf{e}_m' e^{\mathbf{T}_{11} y_k^{(1)}} \mathbf{T}_{12} \tfrac{\partial e^{\mathbf{T}_{22} y_k^{(2)}}}{\partial \eta_{ii}} \mathbf{t}_2.
\end{aligned}$$

Since the matrix $\mathbf{T}_{12}$ has to solve the system of equations given in (5.6), i.e. it is in terms of ϱ_i, γ_{ij}, and ξ_i, then $\frac{\partial \mathbf{T}_{12}}{\partial \eta_{ij}} = 0$. And, for the other parameters, the analytic form of this derivative is not straightforward.

Let $c = \max\{-t_{ii} : 1 \leq i \leq p\}$ and suppose the maximum of the diagonal of $-\mathbf{T}$ is given in the row k. If $1 \leq k \leq p_1$, then

$$\frac{\partial c}{\partial \gamma_{ij}} = \begin{cases} 0 & \text{if } i \neq k, \forall j \neq i, \\ e^{\gamma_{ij}} & \text{if } i = k, \forall j \neq i, \end{cases} \qquad \frac{\partial c}{\partial \gamma_{ii}} = \begin{cases} 0 & \text{if } i \neq k, \\ e^{\gamma_{ii}} & \text{if } i = k, \end{cases}$$

and $\frac{\partial \mathbf{T}_{11}}{\partial \gamma_{ij}}$, $i \neq j$, is a matrix whose (i,i)-th element is $-e^{\gamma_{ij}}$, the (i,j)-th element is $e^{\gamma_{ij}}$, and 0 otherwise. Moreover, $\frac{\partial \mathbf{T}_{11}}{\partial \gamma_{ii}}$ is a matrix whose (i,i)-th element is $-e^{\gamma_{ii}}$ and 0 otherwise.

If $p_1 + 1 \leq k \leq p$ the formulae remain the same, replacing η_{ij} instead of γ_{ij}.

Finally,

$$\frac{\partial \mathbf{t}_2}{\partial \eta_{ii}} = e^{\eta_{ii}} \mathbf{e}_i.$$

Chapter 6

Matrix-exponential distributions

We will study the distributions of non-negative random vectors with a joint rational Laplace transform, i.e., a fraction between two multi-dimensional polynomials. These distributions are in the univariate case known as matrix-exponential (ME) distributions, since their densities can be written as linear combinations of the elements in the exponential of a matrix.

Matrix-exponential distributions were studied in [10] and [16], and these are a generalization of phase-type (PH) distributions, in which the probabilistic interpretation is a priori less clear. For every ME distribution there exist many matrix representations called ME representations. ME distributions and ME representations deserve attention from researchers for a number of reasons. First, ME distributions are useful in the analysis of stochastic models and, as was demonstrated in [10] and [13] can be used in the analysis of renewal processes and queueing systems. Second, the class of ME distributions includes all PH distributions and all Coxian distributions.

The literature on ME distributions is limited. Asmussen and Bladt [10] identified some necessary and sufficient conditions for an ME representation to be minimal and developed a method for computing a minimal ME representation. Bladt and Neuts [16] studied the class of ME distributions and they related ME renewal processes through a randomly stopped deterministic flow model.

More recently, Bladt and Nielsen [17] have given a characterization of the multivariate class, stating that a vector follows a multivariate matrix-exponential (MVME) distribution if and only if all non-negative, non-null linear combinations of its coordinates have a univariate matrix-exponential distribution.

This Chapter is organized as follows. The first two sections will provide the necessary background on matrix-exponential distributions for the univariate case (Section 6.1) and the multivariate case (Section 6.2). In order to generalize these distributions, in Section 6.3 we defined a new class of distributions called bilateral matrix-exponential distributions including both the univariate and the multivariate cases.

6.1 Univariate matrix-exponential distributions

We already know from (2.2) that the Laplace transform from $Y \sim PH(\boldsymbol{\pi}, \mathbf{T})$ is given by $\boldsymbol{\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$, which is essentially a linear combination of the terms in the inverse of $s\mathbf{I} - \mathbf{T}$. The matrix $s\mathbf{I} - \mathbf{T}$ may be inverted by the method of elementary operations which amounts to scaling and subtracting rows. The resulting inverse matrix consists of elements which are rationals functions in s and hence the conclusion that the Laplace transform of Y is also a rational function in s.

Consider a non-negative random variable Y with rational Laplace transform, i.e.,

$$L_Y(s) = \mathbb{E}(e^{-sY}) = \frac{p(s)}{q(s)},$$

where $p(s)$ and $q(s)$ are polynomials. First, we notice that since $Y \geq 0$, then either $L_Y(s) \to 0$ as $s \to \infty$ if the distribution of Y is absolutely continuous or $L_Y(s) \to c$ as $s \to 0$ for some constant c if Y has an atom at zero. Thus we conclude that the $degree(p(s)) \leq degree(q(s))$. If Y has no atom at zero, then $degree(p(s)) < degree(q(s))$.

Thus the general form of a rational Laplace transform of an absolutely continuous non-negative random variable Y is given by

$$L_Y(s) = \mathbb{E}(\exp(-sY)) = \frac{f_1 s^{m-1} + f_2 s^{m-2} + \cdots + f_m}{s^m + g_1 s^{m-1} + \cdots + g_m}. \tag{6.1}$$

Since $L_Y(0) = 1$ we notice that $f_m = g_m$.

Definition 6.1 A non-negative random variable Y is said to have a matrix-exponential distribution if the Laplace transform $L(s) = \mathbb{E}(\exp(-sY))$ is a rational function in s.

Another characterization of the class of matrix-exponential distributions is given by Asmussen and Bladt [10].

Definition 6.2 A random variable is matrix-exponential distributed if and only if there exists a triple $(\boldsymbol{\beta}, \mathbf{D}, \mathbf{d})$ such that the density $f(y)$ of Y can be expressed as

$$f(y) = \boldsymbol{\beta} e^{\mathbf{D}y}\mathbf{d},$$

and we write $Y \sim ME(\boldsymbol{\beta}, \mathbf{D}, \mathbf{d})$. Here, $\boldsymbol{\beta}$ is a row vector of dimension some m, $\mathbf{d}$ is a column vector of the same dimension, and $\mathbf{D}$ is an $m \times m$ matrix, possibly with complex elements.

Indeed, the distribution function of Y is given by $F(y) = 1 + \boldsymbol{\beta} e^{\mathbf{D}y}\mathbf{D}^{-1}\mathbf{d}$, since

$$\begin{aligned} F(y) &= \int_0^y f(u)\mathrm{d}u = \int_0^y \boldsymbol{\beta} e^{\mathbf{D}u}\mathbf{d}\mathrm{d}u \\ &= \left(\boldsymbol{\beta} e^{\mathbf{D}u}\mathbf{D}^{-1}\mathbf{d}\big|_0^y\right) = \boldsymbol{\beta} e^{\mathbf{D}y}\mathbf{D}^{-1}\mathbf{d} - \boldsymbol{\beta}\mathbf{D}^{-1}\mathbf{d}, \end{aligned}$$

but $-\boldsymbol{\beta}\mathbf{D}^{-1}\mathbf{d} = \int_0^\infty \boldsymbol{\beta} e^{\mathbf{D}u}\mathbf{d}\mathrm{d}u = 1$.

The triple $(\boldsymbol{\beta}, \mathbf{D}, \mathbf{d})$ is called a representation of the matrix-exponential distribution. Any matrix-exponential distribution has infinitely many representations. The dimension of $\mathbf{D}$ is called the order of the representation. A representation is called minimal if it is not possible to find another representation of lower dimension.

The Laplace transform of Y can be determined from a representation $(\boldsymbol{\beta}, \mathbf{D}, \mathbf{d})$ as

$$L(s) = \boldsymbol{\beta}(s\mathbf{I} - \mathbf{D})^{-1}\mathbf{d}, \tag{6.2}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimension.

Theorem 6.3 *Let $Y \sim ME(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{d})$. Then its non-centralized moments are given by*

$$M_i = \mathbb{E}(Y^i) = i!\boldsymbol{\beta}(-\boldsymbol{D})^{-(i+1)}\boldsymbol{d}, \quad i = 0, 1, 2, \ldots.$$

PROOF.

$$\begin{aligned}\mathbb{E}(Y^i) &= \int_0^\infty y^i \boldsymbol{\beta} e^{\mathbf{D}y}\mathbf{d}\mathrm{d}y \\ &= -\int_0^\infty i y^{i-1}\boldsymbol{\beta}\mathbf{D}^{-1}e^{\mathbf{D}y}\mathbf{d}\mathrm{d}y \\ &= \cdots \\ &= (-1)^i i!\int_0^\infty \boldsymbol{\beta}\mathbf{D}^{-i}e^{\mathbf{D}y}\mathbf{d}\mathrm{d}y \\ &= (-1)^{i+1}i!\boldsymbol{\beta}\mathbf{D}^{-(i+1)}\mathbf{d}.\end{aligned}$$

■

Let $Y \sim ME(\boldsymbol{\beta}, \mathbf{D}, \mathbf{d})$ with rational Laplace transform given in (6.1). Then, its reduced moments

$$\mu_i = \frac{M_i}{i!}, \quad i = 0, 1, 2\ldots, \tag{6.3}$$

satisfy (see [17])

$$\mu_{m+j} = \sum_{i=0}^{m-1}\frac{g_i}{g_m}(-1)^{m+i+1}\mu_{i+j}, \quad \text{for } j \geq 0.$$

6.1.1 Order of matrix-exponential distributions

By a continued fraction we understand an expression on the form

$$d_0 + \frac{c_1|}{|d_1} + \frac{c_2|}{|d_2} + \frac{c_3|}{|d_3} + \cdots = d_0 + \cfrac{c_1}{d_1 + \cfrac{c_2}{d_2 + \cfrac{c_3}{d_3 + \cdots}}}.$$

A continued fraction is said to be finite if the sum above contains a finite number of terms. Of particular interest for our analysis are the C-continued fractions, which are expressions on the form

$$1 + \frac{c_1 s^{r_1}|}{|1} + \frac{c_2 s^{r_2}|}{|1} + \frac{c_3 s^{r_3}|}{|1} + \cdots.$$

The moment generating function $M(s)$ of a matrix-exponential distributed random variable has a power series expansion

$$M(s) = 1 + \mu_1 s + \mu_2 s^2 + \cdots,$$

where the μ_i's are defined in (6.3). Note that $\mu_i > 0$ for all i.

According to Perron [55], any power series (Taylor series) with constant term 1 corresponds uniquely to a C-continued fraction. If furthermore the series is a power series expansion of a rational function, then the corresponding continued fraction is finite. Particularly tractable are the regular C-continued fraction, where $r_i = 1$ for all i.

Hence, the power series expansion of the moment generating function $M(s)$ of a matrix-exponentially distributed random variable, corresponds uniquely to a regular C-continued fraction (see [17]).

Theorem 6.4 *Consider a matrix-exponential distributed random variable Y with reduced moments $\mu_i = \mathbb{E}(Y^i)/i!$. Then, the rational moment generating function of Y can be written as a finite and regular C-continued fraction*

$$1 + \frac{c_1 s|}{|1} + \frac{c_2 s}{|1} + \frac{c_3 s|}{|1} + \cdots + \frac{c_{2n} s|}{|1}.$$

The coefficients c_i can be calculated in terms of the Hankel determinants

$$\phi_n = \begin{vmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \\ \mu_2 & \mu_3 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n-1} \end{vmatrix} \text{ and } \psi_n = \begin{vmatrix} \mu_2 & \mu_3 & \cdots & \mu_n \\ \mu_3 & \mu_4 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n-2} \end{vmatrix}$$

(for all $n = 1, 2, \ldots$ for ϕ_n and for $n = 2, 3 \ldots$ for ψ_n) as follows:

$$c_1 = \phi_1, \qquad c_{2n} = -\frac{\psi_{n+1}\phi_{n-1}}{\psi_n \phi_n}, \qquad c_{2n+1} = -\frac{\psi_{n+1}\phi_n}{\psi_{n+1}\phi_n},$$

where $\phi_0 = 1$. The Hankel determinants $\phi_m = 0$ for $m > n$ and $\psi_m = 0$ for $m > n + 1$.

Proof. See [17]. ■

Van de Liefvoort [61], He and Zhang [34], and Bladt and Nielsen [17] have showed how to find the minimal order of matrix-exponential distributions. Let l be the order of the rational moment-generating function, then the following Hankel determinant

$$H_l = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_l \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{l+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_l & \mu_{l+1} & \mu_{l+2} & \cdots & \mu_{2l} \end{vmatrix}$$

is 0. Indeed, $H_{l-1} \neq 0$ and $H_i = 0$, for $i \geq l$.

6.1.2 Properties of matrix-exponential distributions

Let $Y_1 \sim ME(\boldsymbol{\beta}, \mathbf{S}, \mathbf{s})$ and $Y_2 \sim ME(\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$, with densities f_{Y_1} and f_{Y_2}, respectively. The convolution of two independent matrix-exponential distributions is obviously again matrix-exponential since the product of two rational functions (their Laplace transform) is again a rational function.

Theorem 6.5 *Y_1+Y_2 has a matrix-exponential distribution with representation*

$$\left((\boldsymbol{\beta}, \mathbf{0}), \begin{pmatrix} \boldsymbol{S} & \boldsymbol{s\pi} \\ \mathbf{0} & \boldsymbol{T} \end{pmatrix}, \begin{pmatrix} \mathbf{0} \\ \boldsymbol{t} \end{pmatrix}\right).$$

PROOF. The Laplace transform corresponding to the proposal is given by

$$L(s) = (\boldsymbol{\beta}, \mathbf{0}) \left(s\mathbf{I} - \begin{pmatrix} \mathbf{S} & \mathbf{s\pi} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}\right)^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{t} \end{pmatrix} = (\boldsymbol{\beta}, \mathbf{0}) \begin{pmatrix} s\mathbf{I} - \mathbf{S} & -\mathbf{s\pi} \\ \mathbf{0} & s\mathbf{I} - \mathbf{T} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{t} \end{pmatrix}$$

$$\begin{aligned} &= (\boldsymbol{\beta}, \mathbf{0}) \begin{pmatrix} (s\mathbf{I} - \mathbf{S})^{-1} & (s\mathbf{I} - \mathbf{S})^{-1}\mathbf{s\pi}(s\mathbf{I} - \mathbf{T})^{-1} \\ \mathbf{0} & (s\mathbf{I} - \mathbf{T})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{t} \end{pmatrix} \\ &= \boldsymbol{\beta}(s\mathbf{I} - \mathbf{S})^{-1}\mathbf{s\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}, \end{aligned}$$

which is a rational function. ■

Theorem 6.6 *Let $p \in (0,1)$. Then the mixture of $f = pf_{Y_1} + (1-p)f_{Y_2}$ is again a matrix-exponential distribution with representation*

$$\left((p\boldsymbol{\beta}, (1-p)\boldsymbol{\pi}), \begin{pmatrix} \boldsymbol{S} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T} \end{pmatrix}, \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{t} \end{pmatrix}\right).$$

PROOF. The Laplace transform $L(s)$ of f is given by

$$\begin{aligned} L(s) &= pL_{Y_1}(s) + (1-p)L_{Y_2}(s) \\ &= p\boldsymbol{\beta}(s\mathbf{I} - \mathbf{S})^{-1}\mathbf{s} + (1-p)\boldsymbol{\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t} \\ &= (p\boldsymbol{\beta}, (1-p)\boldsymbol{\pi}) \begin{pmatrix} s\mathbf{I} - \mathbf{S} & \mathbf{0} \\ \mathbf{0} & s\mathbf{I} - \mathbf{T} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{s} \\ \mathbf{t} \end{pmatrix} \\ &= (p\boldsymbol{\beta}, (1-p)\boldsymbol{\pi}) \left(s\mathbf{I} - \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}\right)^{-1} \begin{pmatrix} \mathbf{s} \\ \mathbf{t} \end{pmatrix}. \end{aligned}$$

■

Theorem 6.7 *The minimum* $\min(Y_1, Y_2)$ *is matrix-exponential with representation*

$$(\boldsymbol{\beta} \otimes \boldsymbol{\pi}, \boldsymbol{S} \oplus \boldsymbol{T}, \boldsymbol{s} \otimes (-\boldsymbol{T})^{-1}\boldsymbol{t} + (-\boldsymbol{S})^{-1}\boldsymbol{s} \otimes \boldsymbol{t}).$$

If $\boldsymbol{s} = -\boldsymbol{S}\boldsymbol{e}$ *and* $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$*, then*

$$\min(Y_1, Y_2) \sim ME(\boldsymbol{\beta} \otimes \boldsymbol{\pi}, \boldsymbol{S} \oplus \boldsymbol{T}).$$

PROOF. Let $\bar{F}_{Y_i}$, $i = 1, 2$, denote their survival functions. Since $\bar{F}(u) = \mathbb{P}(Y_1 > u, Y_2 > u) = \bar{F}_{Y_1}(u)\bar{F}_{Y_2}(u)$, the density of the minimum, f is given by

$$\begin{aligned} f(x) &= -f_{Y_1}(x)\bar{F}_{Y_2}(x) - \bar{F}_{Y_1}(X)f_{Y_2}(x) \\ &= -\boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{s}\boldsymbol{\pi}e^{\mathbf{T}x}\mathbf{T}^{-1}\mathbf{t} - \boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{S}^{-1}\mathbf{s}\boldsymbol{\pi}e^{\mathbf{T}x}\mathbf{t} \\ &= (\boldsymbol{\beta} \otimes \boldsymbol{\pi})e^{(\mathbf{S}\oplus\mathbf{T})x}\left[\mathbf{s} \otimes (-\mathbf{T})^{-1}\mathbf{t} + (-\mathbf{S})^{-1}\mathbf{s} \otimes \mathbf{t}\right]. \end{aligned}$$

If $\mathbf{s} = -\mathbf{S}\mathbf{e}$ and $\mathbf{t} = -\mathbf{T}\mathbf{e}$, then

$$\begin{aligned} \mathbf{s} \otimes (-\mathbf{T})^{-1}\mathbf{t} + (-\mathbf{S})^{-1}\mathbf{s} \otimes \mathbf{t} &= \mathbf{s} \otimes \mathbf{e} + \mathbf{e} \otimes \mathbf{t} \\ &= (-\mathbf{S}\mathbf{e}) \otimes \mathbf{I}\mathbf{e} + \mathbf{I}\mathbf{e} \otimes (-\mathbf{T}\mathbf{e}) \\ &= (-\mathbf{S} \otimes \mathbf{I})(\mathbf{e} \otimes \mathbf{e}) + (\mathbf{I} \otimes (-\mathbf{T}))(\mathbf{e} \otimes \mathbf{e}) \\ &= -(\mathbf{S} \oplus \mathbf{T})\mathbf{e}. \end{aligned}$$

■

Corollary 6.8 *Let* Y_1 *and* Y_2 *be two independent matrix-exponentially distributed random variables with representation* $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{t})$ *such that* $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$ *and such that there exist dual representations. Then* $\min(Y_1, Y_2)$ *is matrix-exponentially distributed with representation*

$$(\boldsymbol{\pi} \otimes \boldsymbol{\pi}, \boldsymbol{T} \oplus \boldsymbol{T}, (\boldsymbol{t} \oplus \boldsymbol{t})\boldsymbol{e}),$$

and $\max(Y_1, Y_2)$ *is matrix-exponentially distributed with representation*

$$\left((\boldsymbol{\pi} \otimes \boldsymbol{\pi}, \boldsymbol{0}), \begin{pmatrix} \boldsymbol{T} \oplus \boldsymbol{T} & \boldsymbol{t} \oplus \boldsymbol{t} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix}, \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{t} \end{pmatrix}\right).$$

Lemma 6.9 *Let* $(\boldsymbol{\pi}, \boldsymbol{T}, -\boldsymbol{T}\boldsymbol{e})$ *be a representation for a matrix-exponential distribution. Further let* $\boldsymbol{\alpha} = \frac{1}{\mu}\boldsymbol{\pi}(-\boldsymbol{T})^{-1}$ *be non-zero solution to the equation system* $\boldsymbol{\alpha}(\boldsymbol{T} + \boldsymbol{t}\boldsymbol{\pi}) = \boldsymbol{0}$ *with* $\mu = \boldsymbol{\pi}(-\boldsymbol{T})^{-1}\boldsymbol{e}$*. Assume that all elements of* $\boldsymbol{\alpha}$ *are different from zero. Then the distribution has an alternatively representation* $(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{T}}, -\tilde{\boldsymbol{T}}\boldsymbol{e})$*. Here*

$$\tilde{\boldsymbol{\pi}} = \frac{\boldsymbol{\alpha} \bullet \boldsymbol{t}}{\boldsymbol{\alpha}\boldsymbol{t}}, \qquad \tilde{\boldsymbol{T}} = \Delta(\boldsymbol{\alpha})^{-1}\boldsymbol{T}'\Delta(\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha} \bullet \boldsymbol{t} = (\alpha_1 t_1, \ldots, \alpha_p t_p)$*, with* p *the order of the distribution.*

PROOF. The matrix $\mathbf{T}$ is non-singular while $\mathbf{T}+\mathbf{t}\boldsymbol{\pi}=\mathbf{T}-\mathbf{T}\mathbf{e}\boldsymbol{\pi}=\mathbf{T}(\mathbf{I}-\mathbf{e}\boldsymbol{\pi})$ obviously has a single zero eigenvalue with corresponding left eigenvector $\boldsymbol{\alpha}$. Define $\Delta(\boldsymbol{\alpha})$ as the diagonal matrix which has the elements of $\boldsymbol{\alpha}$ as entries. Then

$$\begin{aligned}\boldsymbol{\pi}e^{\mathbf{T}y}\mathbf{t} &= \boldsymbol{\pi}\Delta(\boldsymbol{\alpha})^{-1}e^{\Delta(\boldsymbol{\alpha})\mathbf{T}\Delta(\boldsymbol{\alpha})^{-1}y}\Delta(\boldsymbol{\alpha})\mathbf{t}\\ &= \mathbf{t}'\Delta(\boldsymbol{\alpha})e^{\Delta(\boldsymbol{\alpha})^{-1}\mathbf{T}'\Delta(\boldsymbol{\alpha})y}\Delta(\boldsymbol{\alpha})^{-1}\boldsymbol{\pi}'\\ &= \frac{\mathbf{t}'\Delta(\boldsymbol{\alpha})}{\boldsymbol{\alpha}\mathbf{t}}e^{\Delta(\boldsymbol{\alpha})^{-1}\mathbf{T}'\Delta(\boldsymbol{\alpha})y}\Delta(\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}'.\end{aligned}$$

And we obtained $\tilde{\boldsymbol{\pi}}=\frac{\mathbf{t}'\Delta(\boldsymbol{\alpha})}{\boldsymbol{\alpha}\mathbf{t}}=\frac{\boldsymbol{\alpha}\bullet\mathbf{t}}{\boldsymbol{\alpha}\mathbf{t}}$ and $\tilde{\mathbf{T}}=\Delta(\boldsymbol{\alpha})^{-1}\mathbf{T}'\Delta(\boldsymbol{\alpha})$.

On other hand,

$$\begin{aligned}\boldsymbol{\alpha}(\mathbf{T}+\mathbf{t}\boldsymbol{\pi}) &= \mathbf{0}\\ \boldsymbol{\alpha}\mathbf{T}+\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi} &= \mathbf{0}\\ \boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi} &= -\boldsymbol{\alpha}\mathbf{T}\\ \boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}' &= -\mathbf{T}'\boldsymbol{\alpha}', \qquad (\boldsymbol{\alpha}\mathbf{t}\text{ is a constant})\\ \boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}' &= -\mathbf{T}'\Delta(\boldsymbol{\alpha})\mathbf{e}, \qquad (\text{since }\boldsymbol{\alpha}'=\Delta(\boldsymbol{\alpha})\mathbf{e})\\ \Delta(\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}' &= -\Delta(\boldsymbol{\alpha})^{-1}\mathbf{T}'\Delta(\boldsymbol{\alpha})\mathbf{e}\\ &= -\tilde{\mathbf{T}}\mathbf{e},\end{aligned}$$

if we take $\tilde{\mathbf{t}}=-\tilde{\mathbf{T}}\mathbf{e}$ we have proven the lemma. ■

The alternative representation of Lemma 6.9 is known as the reversed-time representation in the phase-type case.

Definition 6.10 For a conservative representation of a matrix-exponential distribution the representation of Lemma 6.9 is called the dual of the representation.

We can always find a dual pair representation for a matrix-exponential distribution.

Lemma 6.11 *Let Y_1 and Y_2 be two independent matrix-exponentially distributed random variables with conservative representation $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{t})$ such that $\boldsymbol{t}=-\boldsymbol{T}\boldsymbol{e}$. The dual of the conservative representations of $\min(Y_1,Y_2)$ and $\max(Y_1,Y_2)$ given in Corollary 6.8 are $(\tilde{\boldsymbol{\pi}}_m, \tilde{\boldsymbol{T}}_m, \tilde{\boldsymbol{t}}_m)$ and $(\tilde{\boldsymbol{\pi}}_M, \tilde{\boldsymbol{T}}_M, \tilde{\boldsymbol{t}}_M)$ with*

$$\begin{aligned}\tilde{\boldsymbol{\pi}}_m &= \mu_m\boldsymbol{\alpha}_m\bullet(\boldsymbol{t}\oplus\boldsymbol{t})\boldsymbol{e}, & \tilde{\boldsymbol{T}}_m &= \Delta(\boldsymbol{\alpha}_m)^{-1}(\boldsymbol{T}\oplus\boldsymbol{T})'\Delta(\boldsymbol{\alpha}_m),\\ \tilde{\boldsymbol{\pi}}_M &= (\mu_m+\mu_r)\boldsymbol{\alpha}_r\bullet\boldsymbol{t}, & \tilde{\boldsymbol{T}}_M &= \begin{pmatrix}\Delta(\boldsymbol{\alpha}_r)^{-1}\boldsymbol{T}'\Delta(\boldsymbol{\alpha}_r) & \Delta(\boldsymbol{\alpha}_r)^{-1}(\boldsymbol{t}\oplus\boldsymbol{t})'\Delta(\boldsymbol{\alpha}_m^*)\\ \boldsymbol{0} & \tilde{\boldsymbol{T}}_m\end{pmatrix}.\end{aligned}$$

Here

$$\begin{aligned}
\mu_m &= (\boldsymbol{\pi}\otimes\boldsymbol{\pi})(-\boldsymbol{T}\oplus\boldsymbol{T})^{-1}\boldsymbol{e}\\
\boldsymbol{\alpha}_m &= \mu_m^{-1}(\boldsymbol{\pi}\otimes\boldsymbol{\pi})(-\boldsymbol{T}\oplus\boldsymbol{T})^{-1}\\
\mu_r &= (\boldsymbol{\pi}\otimes\boldsymbol{\pi})(\boldsymbol{T}\oplus\boldsymbol{T})^{-1}(\boldsymbol{t}\oplus\boldsymbol{t})\boldsymbol{T}^{-1}\boldsymbol{e}\\
\boldsymbol{\alpha}_r &= (\mu_m+\mu_r)^{-1}(\boldsymbol{\pi}\otimes\boldsymbol{\pi})(\boldsymbol{T}\oplus\boldsymbol{T})^{-1}(\boldsymbol{t}\oplus\boldsymbol{t})(\boldsymbol{T})^{-1}\\
\boldsymbol{\alpha}_m^* &= \tfrac{\mu_m}{\mu_m+\mu_r}\boldsymbol{\alpha}_m.
\end{aligned}$$

PROOF. Taking $\mu = \boldsymbol{\pi}(-\mathbf{T})^{-1}\mathbf{e}$, we have

$$\begin{aligned}
\boldsymbol{\alpha}(\mathbf{T}+\mathbf{t}\boldsymbol{\pi}) &= \mathbf{0}\\
\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi} &= -\boldsymbol{\alpha}\mathbf{T}\\
\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}(-\mathbf{T})^{-1} &= \boldsymbol{\alpha}\\
\boldsymbol{\alpha}\mathbf{t}\boldsymbol{\pi}(-\mathbf{T})^{-1}\mathbf{e} &= \boldsymbol{\alpha}\mathbf{e}\\
\boldsymbol{\alpha}\mathbf{t}\mu &= 1\\
\mu &= \frac{1}{\boldsymbol{\alpha}\mathbf{t}},
\end{aligned}$$

then

$$\mu_m = \frac{1}{\boldsymbol{\alpha}_m(\mathbf{t}\oplus\mathbf{t})\mathbf{e}}.$$

Note that $\boldsymbol{\alpha}_M = \left(\frac{\mu_m}{\mu_m+\mu_r}\boldsymbol{\alpha}_m, \boldsymbol{\alpha}_r\right)$ and $\mu_M = \mu_m + \mu_r$.

$$\tilde{\mathbf{T}}_M = \begin{pmatrix} \Delta(\boldsymbol{\alpha}_m^*)^{-1}(\mathbf{T}\oplus\mathbf{T})'\Delta(\boldsymbol{\alpha}_m^*) & \mathbf{0} \\ \Delta(\boldsymbol{\alpha}_r)^{-1}(\mathbf{t}\oplus\mathbf{t})'\Delta(\boldsymbol{\alpha}_m^*) & \Delta(\boldsymbol{\alpha}_r)^{-1}\mathbf{T}'\Delta(\boldsymbol{\alpha}_r)\end{pmatrix}.$$

The initial vector is similarly found to be $(\mathbf{0}, (\mu_m+\mu_r)\boldsymbol{\alpha}_r \bullet \mathbf{t})$. By swapping the blocks we obtain the expressions for $\tilde{\boldsymbol{\pi}}_M$ and $\tilde{\mathbf{T}}_M$. ■

6.2 Multivariate matrix-exponential distributions

The definition of multivariate matrix-exponential distributions is a natural extension of the univariate case (see [17]).

Definition 6.12 A non-negative random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ of dimension n is said to have multivariate matrix-exponential distribution if the joint Laplace transform $L(\mathbf{s}) = \mathbb{E}(\exp(-\langle\mathbf{Y}, \mathbf{s}\rangle))$ is a multi-dimensional rational function, that is, a fraction between two multi-dimensional polynomials. Here $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^n$ and $\mathbf{s} = (s_1, \ldots, s_n)$. This class of distributions is denoted by MVME.

We will consider distributions with no point mass at zero, i.e. if $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a multivariate matrix-exponential distribution, then the Laplace transform of $\mathbf{Y}$ is of the form

$$L_{\mathbf{Y}}(\mathbf{s}) = \mathbb{E}(e^{\langle \mathbf{Y}, \mathbf{s} \rangle}) = \frac{p(\mathbf{s})}{q(\mathbf{s})}, \quad \text{for } \mathbf{s} = (s_1, \ldots, s_n),$$

where p and q are polynomials in n variables, and the degree of p is always lower than the degree of q. Here the *degree* of the multi-dimensional polynomial can be obtained as the largest sum of exponents of its monomials.

Bladt and Nielsen [17] have given a characterization of this class of distributions. In order to get it, they proved the following three lemmas.

Lemma 6.1 *Assume that $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle$ has a matrix-exponential distribution for all $\boldsymbol{a} \geq 0$. Then the (minimal) order of the univariate matrix-exponential distribution for $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle$ is a bounded function of $\boldsymbol{a}$.*

Lemma 6.2 *Assume that $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle$ has a univariate matrix-exponential distribution for all $\boldsymbol{a} \geq 0$. Then exists a set N of n-dimensional Lebesgue measure zeros such that the Laplace transform $L_{\langle \boldsymbol{Y}, \boldsymbol{a} \rangle}(s)$ of $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle$ is a rational function in s for all $\boldsymbol{a} \in [0, \infty)^n \setminus N$.*

Lemma 6.3 *If $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle$ satisfies the conditions of Lemma 6.2 then we may write its Laplace transform as*

$$\frac{\hat{f}_1(\boldsymbol{a}) s^{m-1} + \hat{f}_2(\boldsymbol{a}) s^{m-2} + \cdots + \hat{f}_{m-1}(\boldsymbol{a}) s + 1}{\hat{g}_0(\boldsymbol{a}) s^m + \hat{g}_1(\boldsymbol{a}) s^{m-1} + \cdots + \hat{g}_{m-1}(\boldsymbol{a}) s + 1}$$

where the terms $\hat{f}_i(\boldsymbol{a})$ and $\hat{g}_i(\boldsymbol{a})$ are sums of monomials in $\boldsymbol{a}$ of order $m - i$ and m is the common order except a set of measure zero.

Now, their main theorem which characterizes the class of MVME is the following.

Theorem 6.13 *A vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ follows a multivariate matrix-exponential distribution if and only if $\langle \boldsymbol{Y}, \boldsymbol{a} \rangle = \sum_{i=1}^n a_i Y_i$ has a univariate matrix-exponential distribution for all non-negative vectors $\boldsymbol{a} \neq \mathbf{0}$.*

Moreover, if $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a MVME distribution and $\mathbf{A}$ is a non-negative $m \times n$ matrix, then $\mathbf{Z}' = \mathbf{A}\mathbf{Y}'$ has a MVME distribution. In particular, all marginals distributions are again matrix-exponentially distributed.

Inspired by this analysis, Bladt and Nielsen [17] proposed the following definition of a multivariate phase-type distributions.

Definition 6.14 A vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a multivariate phase-type (MVPH) distribution if $\langle \mathbf{Y}, \mathbf{a} \rangle$ has a (univariate) phase-type distribution for all non-negative $\mathbf{a} \neq \mathbf{0}$.

Certainly if Y_i are n independent PH distributed random variables then $\langle \mathbf{Y}, \mathbf{a} \rangle$ will be phase-type distributed. Indeed, the MPH* class (see Section 5.1) belongs to the MVPH class.

The following definition is a natural extension of the MPH* structure to matrix-exponential distributions.

Definition 6.15 Let MME* be the subclass of MVME, such as $\langle \mathbf{Y}, \mathbf{a} \rangle$ has representation $(\boldsymbol{\gamma}, \Delta(\mathbf{Ra})^{-1}\mathbf{T}, \mathbf{t})$, where $\mathbf{t} = -\Delta(\mathbf{Ra})^{-1}\mathbf{Te}$. We say that the triple $(\boldsymbol{\gamma}, \mathbf{T}, \mathbf{R})$ is a MME* representation of the multivariate distribution.

There exists MVME distributions where the MVME order is strictly less than the MME* order (see [17]).

Theorem 6.16 *The cross–moments* $\mathbb{E}\left(\prod_{i=1}^{n} Y_i^{a_i}\right)$, *where* $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \sim MME^*(\boldsymbol{\gamma}, \boldsymbol{T}, \boldsymbol{R})$ *and* $a_i \in \mathbb{N}$, *are given by*

$$\boldsymbol{\gamma} \sum_{l=1}^{a!} \left(\prod_{i=1}^{a} (-\boldsymbol{T})^{-1} \Delta(\boldsymbol{R}_{\sigma_l(i)}) \right) \boldsymbol{e},$$

where $a = \sum_{i=1}^{n} a_i$, $\boldsymbol{R}_i$ *is the i-th column of* $\boldsymbol{R}$, *and* $\sigma_1, \ldots, \sigma_{a!}$ *are the ordered permutations of a-tuples of derivatives, within* $\sigma_l(i)$ *being the value among* $1, \ldots, n$ *at the i-th position of the permutation* σ_l.

Proof. The joint Laplace transform of $\mathbf{Y}$ (see (2.2)) is given by

$$H(\mathbf{s}) = \boldsymbol{\gamma}((-\mathbf{T})^{-1}\Delta(\mathbf{Rs}) + \mathbf{I})^{-1}\mathbf{e}.$$

Thus, we can obtain the cross–moments by

$$\mathbb{E}\left(\prod_{i=1}^{n} Y_i^{a_i} \right) = \left. \frac{\mathrm{d}^a H(\mathbf{s})}{\mathrm{d}s_1^{a_1} \mathrm{d}s_2^{a_2} \ldots \mathrm{d}s_k^{a_n}} \right|_{\mathbf{s}=\mathbf{0}}.$$

For more details of the demonstration we refer the reader to [49]. ■

6.3 Bilateral matrix-exponential distributions

The class of phase-type (PH) distributions has become quite popular in the last decades. Previous work on PH distributions ([44], [45]) has been extended into the real line as bilateral phase-type distributions in [2] and [59]. Ahn and Ramaswami [2] defined the class of bilateral phase-type distributions, denoted by BPH*, in terms of rewards $r(i)$, which can be negative (see also Section 5.1). Being m the order of the representation, let $\Delta(\mathbf{r})$ be the diagonal matrix composed of the reward rates of the transient states $\mathbf{r} = (r(1), \ldots, r(m))'$.

Definition 6.17 [2] Let X denote the total accumulated reward until absorption. X is said to be bilaterally phase–type distributed random variable with initial probability vector $\boldsymbol{\alpha}$, transient generator $\mathbf{T}$ and reward matrix $\Delta(\mathbf{r})$. We denote this by $X \sim BPH^*(\boldsymbol{\alpha}, \mathbf{T}, \Delta(\mathbf{r}))$.

It is clear from the construction of the BPH* class that it has an atom at 0 if and only if $\alpha_{m+1} = 1 - \boldsymbol{\alpha}\mathbf{e} > 0$.

Indeed, the MG function of X is given by

$$M_X(s) = \alpha_{m+1} + \boldsymbol{\alpha}(s\mathbf{T}^{-1}\Delta(\mathbf{r}) + \mathbf{I})^{-1}\mathbf{e}. \tag{6.4}$$

We will define the class of multivariate bilateral phase-type (MBPH*) distributions as a generalization for both the multivariate case of PH distributions proposed by Kulkarni [39] (denoted by MPH*), and the BPH* class.

Definition 6.18 Let $X_j \sim BPH^*(\boldsymbol{\alpha}, \mathbf{T}, \Delta(\mathbf{r}_j))$, for $j = 1, \ldots, n$, where the column vectors $\mathbf{r}_j$ are the rewards associated with the variable X_j. Note that now the rewards can be negative. Then, for $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$, we say that $\mathbf{X} = (X_1, \ldots, X_n)$ is multivariate bilateral phase–type (denoted by MBPH*) distributed with representation $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$.

Inspired by Definition 6.14 we propose the following lemma.

Lemma 6.4 *$\boldsymbol{X} = (X_1, \ldots, X_n) \sim MBPH^*(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{R})$ if and only if $\langle \boldsymbol{X}, \boldsymbol{a} \rangle \sim BPH^*(\boldsymbol{\alpha}, \boldsymbol{T}, \Delta(\boldsymbol{Ra}))$, for all n-dimensional real vector $\boldsymbol{a}$.*

Proof. See Appendix B. ■

Another generalization of PH distributions is considering the class of matrix-exponential (ME) distributions, that have been studied in [10] and [16], and in the multivariate case (MVME) in [17].

In order to generalize matrix-exponential (univariate and multivariate) distributions into the real line, we define the class of bilateral ME distributions for both the univariate and multivariate cases, as an extension of the ME and MVME distributions, respectively.

Definition 6.19 We say that X is univariate bilateral matrix-exponential or simply bilateral matrix-exponential (BME) distributed, if it has rational moment generating (MG) function, i.e. if $\mathbb{E}(e^{sX})$ is rational in s. We denote by $X \sim BME(\boldsymbol{\alpha}_+, \mathbf{T}_+, \mathbf{t}_+, \boldsymbol{\alpha}_-, \mathbf{T}_-, \mathbf{t}_-)$ if X has the following density

$$f_X(x) = \boldsymbol{\alpha}_+ e^{\mathbf{T}_+ x}\mathbf{t}_+ \mathbf{1}_{\{x>0\}} + \boldsymbol{\alpha}_- e^{\mathbf{T}_- |x|}\mathbf{t}_- \mathbf{1}_{\{x<0\}}, \tag{6.5}$$

where $\boldsymbol{\alpha}_+$ is a row vector of some dimension m_+, $\mathbf{T}_+$ is a matrix of dimension $m_+ \times m_+$, and $\mathbf{t}_+$ is an m_+-dimensional column vector. Similarly, both the vectors $\boldsymbol{\alpha}_-$,$\mathbf{t}_-$, and the matrix $\mathbf{T}_-$, are defined by some dimension m_-.

The study of BME distributions and their representations is also important to consider. Having a BME-representation $(\boldsymbol{\alpha}_+, \mathbf{T}_+, \mathbf{t}_+, \boldsymbol{\alpha}_-, \mathbf{T}_-, \mathbf{t}_-)$, we know the dimension of the matrix $\mathbf{T}_+$ plus the dimension of the matrix $\mathbf{T}_-$ is called the *order* of the representation, the smallest order from among the equivalent representations is called the *degree* (see [61]).

A representation whose order is equal to the degree is said to be of *minimal order*, this is called the order of the distribution ([34]).

Asmussen and Bladt [10] identified some necessary and sufficient conditions for an ME representation to be minimal and developed a method for computing a minimal ME representation from an ME distribution. The ME order can play an important role in finding minimal representations. Qi-Ming and Hanqin [34] introduced certain Hankel matrices that can be used to compute the ME order of ME distributions.

Here we establish a relationship between the MG function and the minimal BME representation using Hankel matrices. Having X with density given in (6.5) and being m the minimal order of the distribution, the Hankel determinant given by

$$H_l = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_l \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{l+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_l & \mu_{l+1} & \mu_{l+2} & \cdots & \mu_{2l} \end{vmatrix}$$

is 0 if $l \geq m$ and $H_{m-1} \neq 0$, where μ_j are the reduced moments given by

$$\mu_j = \boldsymbol{\alpha}_+(-\mathbf{T}_+)^{-(j+1)}\mathbf{t}_+ + (-1)^j \boldsymbol{\alpha}_-(-\mathbf{T}_-)^{-(j+1)}\mathbf{t}_-.$$

We define multivariate bilateral matrix-exponential (MVBME) distributions as a natural extension of the univariate case.

Definition 6.20 A random vector $\mathbf{X} \in \mathbb{R}^n$ of dimension n is multivariate bilateral matrix-exponential (MVBME) distributed if the joint MG function $\mathbb{E}(e^{\langle \mathbf{X},\mathbf{s}\rangle})$, $\mathbf{s} \in \mathbb{R}^n$, is a multi-dimensional rational function.

The marginal distributions are hence univariate bilateral matrix-exponential distributions.

In order to give a characterization of these distributions we present the following lemmas (their proofs are given in Appendix B).

Lemma 6.5 *Assume that $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ has a BME distribution for all $\boldsymbol{a} \in \mathbb{R}^n \setminus \boldsymbol{0}$. Then the (minimal) order $m(\boldsymbol{a})$ of the univariate BME distribution for $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ is a bounded function of $\boldsymbol{a}$.*

Lemma 6.6 *Assume that $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ has a univariate bilateral matrix-exponential distribution for all $\boldsymbol{a} \in \mathbb{R}^n \setminus \boldsymbol{0}$, and suppose the order of the distribution of $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ is bounded by some m. Then, we may write the MG function of $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ as*

$$\frac{\tilde{b}_m(\boldsymbol{a})s^m + \tilde{b}_{m-1}(\boldsymbol{a})s^{m-1} + \cdots + \tilde{b}_1(\boldsymbol{a})s + 1}{\tilde{a}_m(\boldsymbol{a})s^m + \tilde{a}_{m-1}(\boldsymbol{a})s^{m-1} + \cdots + \tilde{a}_1(\boldsymbol{a})s + 1},$$

where the terms $\tilde{b}_j(\boldsymbol{a})$ and $\tilde{a}_j(\boldsymbol{a})$ are sums of n-dimensional monomials in $\boldsymbol{a}$ of degree j.

Our theorem which characterizes the class of MVBME is the following.

Theorem 6.21 *A vector $\boldsymbol{X}$ follows a multivariate bilateral matrix-exponential distribution, i.e. $\boldsymbol{X} \sim MVBME$, if and only if $\langle \boldsymbol{X}, \boldsymbol{a}\rangle \sim BME$ for all $\boldsymbol{a} \in \mathbb{R}^n \setminus \boldsymbol{0}$.*

Proof. Let $\mathbf{X} \sim MVBME$, then $\mathbb{E}(e^{\langle \mathbf{X}, s\mathbf{a}\rangle})$ is rational in $s\mathbf{a}$ for $s \in \mathbb{R}$ and $\mathbf{a} \in \mathbb{R}^n \setminus \mathbf{0}$. Since

$$\mathbb{E}(e^{\langle \mathbf{X}, s\mathbf{a}\rangle}) = \mathbb{E}(e^{s\langle \mathbf{X}, \mathbf{a}\rangle}),$$

then $\mathbb{E}(e^{s\langle \mathbf{X}, \mathbf{a}\rangle})$ is rational in s, i.e. $\langle \mathbf{X}, \mathbf{a}\rangle \sim BME$.

On the other hand, suppose that $\langle \mathbf{X}, \mathbf{a}\rangle$ has rational MG function, for all $\mathbf{a} \in \mathbb{R}^n \setminus \mathbf{0}$. Then we know that the MG function can be expressed in the form of Lemma 6.6. By setting $s = 1$ this rational function coincides with the multidimensional MG function of $\mathbf{X}$ at $\mathbf{a}$. ■

Chapter 7

Conclusion and Outlook

This work is focused on statistical analysis and estimation of phase–type (PH) distributions. The initial idea of this work was the estimation and analysis of multivariate matrix–exponential distributions. However, we considered it prudent to analyze firstly the sub-class of PH distributions due to their importance in the recent decades in different areas of applied probability. Many results for PH distributions can be generalized to matrix–exponential (ME) distributions, by providing proofs that do not depend on underlying Markov chains or, possibly, by using a continuation argument.

Asmussen *et.al* [11] provided the statistical framework for obtaining maximum likelihood estimates of continuous PH distributions using the EM algorithm, while Bladt *et.al* [15] used Markov chain Monte Carlo for doing so. In this work we proposed some alternatives of these algorithms using uniformization, canonical form, and reversed-time Markov chains. Furthermore, we proposed an up-to-date (quasi) Newton-Raphson method (called Direct method (DM)) for obtaining maximum likelihood estimates of continuous as well as for discrete PH distributions.

We noticed that there is no significant difference between the maximum likelihoods using the EM and the DM algorithms, however, the maximum likelihoods using the Gibbs sampler algorithm depended on the prior distributions.

With respect to the execution times performed by the algorithms, there were quite some differences. Definitely, using canonical form and reversed-time Markov chains for discrete PH, and using uniformization for continuous PH, contributed to accelerate the algorithms. Since many matrix-matrix and matrix-vector products are used, it might thus be possible to optimize our implementations further with different strategies for calculating and storing intermediate results.

As a natural extension, we also considered the estimation of bivariate PH distributions, via the EM algorithm and the DM method.

Moreover, we realized that little has been done on statistical inference of PH (discrete and continuous) distributions. For this reason, we decided to dig deeper into the subject, considering the Fisher information matrix of PH distributions, since the inverse of this matrix provides the variance and covariance of the estimated parameters. We discussed two different ways of analytically obtaining the Fisher information matrix, one of these methods is based on a direct calculation of second derivatives of the log–likelihood function while the other is based on a paper by Oakes [52] where the partial derivatives are made using a split of the log–likelihood function as in the EM algorithm. There are still serious issues concerning identifiability and over-parameterization.

Finally, we analyzed matrix-exponential distributions for both univariate and multivariate cases. Since ME distributions (distributions with rational Laplace transform) have support $[0, \infty)$, we considered prudent before pursuing multivariate estimation, to introduce a new class of distributions whose support is $(-\infty, \infty)$, obviously with similar features to the ME distributions. We called this type of distributions *bilateral ME distributions*, and they are defined as distributions with rational moment generating function. It should be pointed out that a more comprehensive multivariate analysis of these distributions is needed as well as the estimation of their parameters.

Appendix A

Fisher information and statistical inference for phase-type distributions

---

Paper for the Conference in Honor of Soren Asmussen - New Frontiers in Applied Probability. August 2011.

Accepted

Applied Probability Trust (3 February 2011)

FISHER INFORMATION AND STATISTICAL INFERENCE FOR PHASE–TYPE DISTRIBUTIONS

MOGENS BLADT,* *Universidad Nacional Autonoma de Mexico*

JUDITH R. ESPARZA,** *Technical University of Denmark*

BO F. NIELSEN,** *Technical University of Denmark*

Abstract

This paper is concerned with statistical inference for both continuous and discrete phase–type distributions. We consider maximum likelihood estimation, where traditionally the EM algorithm has been employed. Certain numerical aspects of this method is revised and we provide an alternative method for dealing with the E–step. We also compare the EM algorithm to a direct Newton–Raphson optimization of the likelihood function. As one of the main contributions of the paper, we provide formulae for calculating the Fisher information matrix both for the EM algorithm and Newton–Raphson approach. The inverse of the Fisher information matrix provides the variances and covariances of the estimated parameters.

Keywords: Phase–type distributions, Fisher information, EM–algorithm, Newton–Raphson

2000 Mathematics Subject Classification: Primary 62F25
Secondary 60J10, 60J27, 60J75

1. Introduction

Phase–type distributions have played an important role in the modeling of complex stochastic phenomena in the recent decades. They are mathematically tractable and often allow for exact solutions to functionals of interest such as e.g. the ruin probability in risk theory or waiting time distributions in queueing theory. Such solutions are typically explicit or given in terms of some deterministic equations which may require some standard numerical procedure for its evaluation.

Phase–type distributions [8] can be defined for both discrete and continuous distributions. A continuous (discrete) phase–type distribution is the time until absorption of a Markov jump process (Markov chain) with finitely many states, one of which is absorbing and the remaining being transient. It is the Markov jump (Markov chain) structure underlying the absorption times that makes the phase–type distributions tractable, and most manipulations with phase–type distributions use this underlying structure directly in establishing probabilistic arguments.

Estimation and statistical inference for phase–type distributions is of considerable importance when consolidating its role in applications. The paper by Asmussen, *et. al.* [2] was the first to establish a general approach to maximum likelihood estimation of continuous phase–type distributions. In spite of being mathematically tractable

* Postal address: IIMAS-UNAM, A.P. 20-726, 01000 Mexico, D.F., Mexico
** Postal address: Richard Petersens Plads, Building 305, 2800 Kgs. Lyngby, Denmark

due to their probabilistic interpretation, this very interpretability complicates the estimation and inference for phase–type distributions considerably: there are serious issues concerning identifiability and over–parameterization.

One of the main reasons for using phase–type distributions is their tractability in many areas of applied probability. Many of the key functionals of interest such as ruin probabilities in insurance risk and the waiting time distributions in queueing theory are invariant under different equivalent representations of the same phase–type distribution.

The main contributions of this paper are methods for calculating the Fisher information matrix for discrete and continuous phase–type distributions, and we provide formulae which relate to both the EM algorithm and the Newton Raphson approach. The Fisher information matrix is then employed to find confidence regions for the estimated parameters. We also review some necessary background concerning the EM algorithms for the discrete and continuous cases, and we shall suggest an alternative method for calculating matrix–exponentials and related integrals appearing in the E–step, where originally (see [2]) a Runge–Kutta method was employed. Our method will speed up the execution of the EM algorithm considerably for small and medium sized data sets, while the Runge–Kutta method may outperform our method for large amounts of data.

While the problem concerning over–parameterization in general persists, we shall only consider distributions which have a unique representation. Confidence regions for parameters in models which are over–parameterized or non–unique are not well defined.

The remainder of this paper is organized as follows. In Section 2, we provide some relevant background on phase–type distributions, while in Section 3 we analyze the maximum likelihood estimation of these distributions via the EM algorithm and a Newton–Raphson method. In Section 4 we present methods for obtaining the Fisher information matrix. A simulation study is provided in Section 5. Finally, the work is summarized in Section 6.

2. Some basic properties of phase–type distributions

Let $\{X_t\}_{t\in I}$ be a Markov chain (Markov jump process) with $I = \{0, 1, 2, \dots\}$ ($I = [0, \infty)$) and state space $E = \{1, \dots, p, p+1\}$, where the states $1, \dots, p$ are transient and the state $p + 1$ is absorbing. Let $\pi_i = \mathbb{P}(X_0 = i)$ be the initial probabilities and define the row vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$. Let t_{ij} denote the transition probabilities (transition rates) between the transient states. The transition rates for continuous time processes are the entries of the intensity matrix. Let $\mathbf{T} = \{t_{ij}\}_{i,j=1,\dots,p}$ denote the transition matrix (intensity matrix) restricted to the transient states. Finally, let $\mathbf{t} = (t_1, \dots, t_p)'$ be the vector of exit probabilities (exit rates). With $\mathbf{e}$ being a p-dimensional column vector of 1's, we have $\mathbf{t} = \mathbf{e} - \mathbf{Te}$ ($\mathbf{t} = -\mathbf{Te}$). We say that $\tau = \inf\{t \in I | X_t = p + 1\}$ has a phase–type distribution with representation $(\boldsymbol{\pi}, \mathbf{T})$, and write $\tau \sim \mathrm{PH}_p(\boldsymbol{\pi}, \mathbf{T})$. In this paper we will refer to the discrete case by DPH and to the continuous case by CPH.

Sometimes it is convenient to allow for an atom at zero as well in which case we let $\pi_{p+1} > 0$ denote the probability of initiating in the absorbing state. If $\pi_{p+1} = 0$, the probability mass (density) function of τ is $f(x) = \boldsymbol{\pi}\mathbf{T}^{x-1}\mathbf{t}$, $(\boldsymbol{\pi}e^{\mathbf{T}x}\mathbf{t})$, $x > 0$. We shall initially assume that the phase–type distributions under consideration do not have an

atom at zero.

3. Maximum likelihood estimation of phase–type distributions

Consider M independent observations $y_1, \ldots, y_M$ from a $\mathrm{PH}_p(\boldsymbol{\pi}, \mathbf{T})$ distribution, where throughout the paper the order of the distribution p, is assumed to be known. One may use the Akaike information criterion [1] for estimating p, but this matter will not be pursued here. Let $\boldsymbol{y} = (y_1, \ldots, y_M)$. We only observe the times until absorption and have no information about the underlying Markov chains (jump processes). We may consider this as a situation of incomplete data since ideally we would be able to observe all the underlying Markov chains (jump processes) which generate the absorption times.

Let $\boldsymbol{\theta}$ denote a vector containing the parameters $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$. The incomplete data likelihood function is given by

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{k=1}^{M} \boldsymbol{\pi} \mathbf{T}^{y_k - 1} \mathbf{t} \quad \text{(DPH)}, \quad L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{k=1}^{M} \boldsymbol{\pi} e^{\mathbf{T} y_k} \mathbf{t} \quad \text{(CPH)}. \tag{1}$$

The log–likelihood function is defined as $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \log L(\boldsymbol{\theta}; \boldsymbol{y})$.

3.1. EM algorithm

One approach to maximizing the incomplete likelihood function is via the EM algorithm (Expectation–Maximization) [5] for which we shall need the full data or complete likelihood function, L_f. Let $\boldsymbol{x} = (x_1, \ldots, x_M)$ denote the full data for the M absorption times. Thus the x_i's are trajectories of the underlying Markov chains (Markov jump processes) up to the time of absorption. The full data likelihood is given in terms of sufficient statistics,

$$L_f(\boldsymbol{\theta}; \boldsymbol{x}) = \begin{cases} \prod_{i=1}^{p} \pi_i^{B_i} \prod_{i,j=1}^{p} t_{ij}^{N_{ij}} \prod_{i=1}^{p} t_i^{N_i} & \text{(DPH)} \\ \prod_{i=1}^{p} \pi_i^{B_i} \prod_{\substack{i,j=1 \\ i \neq j}}^{p} t_{ij}^{N_{ij}} e^{-t_{ij} Z_i} \prod_{i=1}^{p} t_i^{N_i} e^{-t_i Z_i}, & \text{(CPH)} \end{cases} \tag{2}$$

where B_i is the number of Markov chains (Markov jump processes) initiating in state i, N_{ij} the number of transitions from state i to state j, N_i the number of chains (processes) jumping from state i to the absorbing state, and Z_i the total time the processes spent in state i.

The full log–likelihood function ℓ_f is hence given by

$$\ell_f(\boldsymbol{\theta}; \boldsymbol{x}) = \begin{cases} \sum_{i=1}^{p} B_i \log(\pi_i) + \sum_{i,j=1}^{p} N_{ij} \log(t_{ij}) + \sum_{i=1}^{p} N_i \log(t_i) & \text{(DPH)} \\ \sum_{i=1}^{p} B_i \log(\pi_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{p} N_{ij} \log(t_{ij}) + \sum_{i=1}^{p} N_i \log(t_i) & \\ \quad - \sum_{\substack{i,j=1 \\ i \neq j}}^{p} t_{ij} Z_i - \sum_{i=1}^{p} t_i Z_i. & \text{(CPH)} \end{cases} \tag{3}$$

The full likelihood is easily maximized applying e.g. the method of Lagrange multipliers, attending the constraints. We get that the maximum likelihood estimators of

$\boldsymbol{\pi}, \mathbf{T}$, and $\mathbf{t}$, are given by

$$\hat{\pi}_i = \frac{B_i}{M}, \quad \hat{t}_{ij} = \frac{N_{ij}}{J_i}, \quad \hat{t}_i = \frac{N_i}{J_i}, \qquad \text{(DPH)}$$

$$\hat{\pi}_i = \frac{B_i}{M}, \quad \hat{t}_{ij} = \frac{N_{ij}}{Z_i}, \quad \hat{t}_i = \frac{N_i}{Z_i}, \qquad \text{(CPH)}$$

where J_i is the total number of jumps out of state i (DPH), whereas $\hat{t}_{ii} = 1 - \sum_{j \neq i} \hat{t}_{ij} - \hat{t}_i$ (DPH) and $\hat{t}_{ii} = -\sum_{j \neq i} \hat{t}_{ij} - \hat{t}_i$ (CPH).

The EM algorithm works as follows. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0)$ be (in principle) any choice of initial parameters.

1. Calculate $h : \boldsymbol{\theta} \longmapsto \mathbb{E}_{\boldsymbol{\theta}_0}(\ell_f(\boldsymbol{\theta}; \boldsymbol{x})|\boldsymbol{y})$.
2. Maximize h. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}}, \hat{\mathbf{t}})$ denote the point which maximizes h.
3. Set $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$ and goto 1.

Since the log–likelihood function is linear in the sufficient statistics B_i, N_{ij}, and N_i, it is straightforward to calculate its conditional expectation if the corresponding conditional expectations of the sufficient statistics are known. To this end, consider one data point y (time until absorption). The general case with more than one data then follows by summing up the conditional expectations over all data points $y_1, \ldots, y_M$.

First, we consider the discrete case (see also [3]). We notice that $B_i = \mathbf{1}_{\{X_0 = i\}}$ and hence

$$\begin{aligned} \mathbb{E}(B_i|\tau = y) &= \mathbb{P}(X_0 = i|\tau = y) \\ &= \frac{\mathbb{P}(\tau = y|X_0 = i)\mathbb{P}(X_0 = i)}{\mathbb{P}(\tau = y)} \\ &= \frac{\mathbf{e}_i' \mathbf{T}^{y-1} \mathbf{t}}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}} \pi_i. \end{aligned}$$

Here $\mathbf{e}_i$ denotes a p-dimensional column vector with 1 in the i-th entry and 0 otherwise.

Concerning N_{ij}, if $\tau = y$ we have that

$$N_{ij} = \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \mathbf{1}_{\{X_k = i, X_{k+1} = j\}}.$$

Thus

$$\begin{aligned} \mathbb{E}(N_{ij}|\tau = y) &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \mathbb{P}(X_k = i, X_{k+1} = j|\tau = y) \\ &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \frac{\mathbb{P}(\tau = y|X_{k+1} = j)\mathbb{P}(X_{k+1} = j|X_k = i)\mathbb{P}(X_k = i)}{\mathbb{P}(\tau = y)} \\ &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \frac{\mathbf{e}_j' \mathbf{T}^{(y-(k+1)-1)} \mathbf{t} \boldsymbol{\pi} \mathbf{T}^k \mathbf{e}_i}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}} t_{ij}. \end{aligned}$$

Similar calculations yields

$$\mathbb{E}(N_i|\tau = y) = \frac{\boldsymbol{\pi}\mathbf{T}^{y-1}\mathbf{e}_i}{\boldsymbol{\pi}\mathbf{T}^{y-1}\mathbf{t}} t_i.$$

Finally, $\mathbb{E}(J_i) = \sum_{j=1}^{p} \mathbb{E}(N_{ij}) + \mathbb{E}(N_i)$. The final EM algorithm for the discrete case then translates into

0. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0)$.
1. Under $\boldsymbol{\theta}_0$, calculate conditional expectations $\mathbb{E}_{\boldsymbol{\theta}_0}(B_i|\boldsymbol{y})$, $\mathbb{E}_{\boldsymbol{\theta}_0}(N_{ij}|\boldsymbol{y})$, and $\mathbb{E}_{\boldsymbol{\theta}_0}(N_i|\boldsymbol{y})$. Let $\mathbb{E}(J_i|\boldsymbol{y}) = \sum_{j=1}^{p} \mathbb{E}(N_{ij}|\boldsymbol{y}) + \mathbb{E}(N_i|\boldsymbol{y})$.
2. Let $\hat{\pi}_i = \dfrac{\mathbb{E}_{\boldsymbol{\theta}_0}(B_i|\boldsymbol{y})}{M}$, $\hat{t}_{ij} = \dfrac{\mathbb{E}_{\boldsymbol{\theta}_0}(N_{ij}|\boldsymbol{y})}{\mathbb{E}_{\boldsymbol{\theta}_0}(J_i|\boldsymbol{y})}$, and $\hat{t}_i = \dfrac{\mathbb{E}_{\boldsymbol{\theta}_0}(N_i|\boldsymbol{y})}{\mathbb{E}_{\boldsymbol{\theta}_0}(J_i|\boldsymbol{y})}$.
3. Set $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0, \mathbf{T}_0, \mathbf{t}_0) = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}}, \hat{\mathbf{t}})$ and goto 1.

The EM algorithm for the CPH is similar, only changing the formulae for the conditional expectations which can be found in [2]. As a curiosity, in the derivation of the conditional expectation of N_{ij} given discrete data in continuous time, [2] uses a discretization argument where they approximate the continuous process by the corresponding Markov chain formula derived above.

The corresponding formulae for the CPH (see [2]) are given by

$$\begin{aligned}
\mathbb{E}(B_i|\tau = y) &= \frac{\mathbf{e}_i' e^{\mathbf{T}y}\mathbf{t}}{\boldsymbol{\pi} e^{\mathbf{T}y}\mathbf{t}}\pi_i \\
\mathbb{E}(N_{ij}|\tau = y) &= \frac{\int_0^y \boldsymbol{\pi} e^{\mathbf{T}u}\mathbf{e}_i\mathbf{e}_j' e^{\mathbf{T}(y-u)}\mathbf{t}\,du}{\boldsymbol{\pi} e^{\mathbf{T}y}\mathbf{t}} t_{ij} \\
\mathbb{E}(N_i|\tau = y) &= \frac{\boldsymbol{\pi} e^{\mathbf{T}y}\mathbf{e}_i}{\boldsymbol{\pi} e^{\mathbf{T}y}\mathbf{t}} t_i \\
\mathbb{E}(Z_i|\tau = y) &= \frac{\int_0^y \boldsymbol{\pi} e^{\mathbf{T}u}\mathbf{e}_i\mathbf{e}_i' e^{\mathbf{T}(y-u)}\mathbf{t}\,du}{\boldsymbol{\pi} e^{\mathbf{T}y}\mathbf{t}}.
\end{aligned}$$

If zero is contained in the data we also need to include an atom of a certain size at zero in the specification of the phase–type distribution. Allowing for $\pi_{p+1} > 0$ we may recalculate conditional expectations and maxima as above. However, it is immediately seen that the estimation procedure can be split into the following components. (1) Let $\hat{\pi}_{p+1}$ denote the proportion of zeros in the data set. (2) Eliminate the zeros from the data. (3) Fit a phase–type distribution $\mathrm{PH}_p(\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}})$ to the remaining data. This procedure, indeed, produces a maximum likelihood estimator for the full model which contains an atom at zero.

The EM algorithm always converges to a (possibly local) maximum. The convergence is known to be quite slow. Various random initiations of the algorithm may be needed in order to support the hypothesis that the local maxima reached represents a global maximum. Also it is important to initiate the algorithm with a representation of full dimension. If we, for example in the discrete case, decided to initiate with $t_{ij} = 1/(p+1)$ and $\pi_i = 1/p$, p being the dimension of the representation, then this is equivalent to a geometric distribution and it is not difficult to see that all

subsequent iterations will again give geometric distributions. Hence the maximum likelihood estimator will also satisfy that all elements of the transition matrix are equal. If some parameter t_{ij} is set to zero initially, then all subsequent values of t_{ij} through the iterations will remain zero. This makes it possible to estimate subclasses of general discrete phase–type distribution by adequately specifying zeros of certain transition probabilities from the beginning. If other subclasses or re–parameterizations like e.g. letting all remaining t_{ij} only depend on i are to be considered, then we need to intervene directly into the likelihood function and calculate new expressions for the maximum likelihood estimators. The conditional expectations, however, still remain valid.

The evaluation of the E-step in the CPH version of the EM algorithm can be numerically challenging. In [2] the authors propose to use a Runge-Kutta method. Another, and by now standard, method for the evaluation of the matrix exponential is uniformization. This method can also be applied in the evaluation of $\mathbb{E}(N_{ij}|\tau = y)$. The advantage of uniformization is the higher numerical precision. In most cases we found uniformization to be superior in terms of efficiency too, albeit for a very high number of observations our implementation was outperformed with respect to speed by the Runge–Kutta implementation of [2].

EM for CPH using uniformization. In standard uniformization (see [6]) we let $\mathbf{K} = \frac{1}{c}\mathbf{T} + \mathbf{I}$, where $c = \max\{-t_{ii} : 1 \le i \le p\}$ and $\mathbf{I}$ is the identity matrix of appropriate dimension $(p \times p)$, we have that

$$e^{\mathbf{T}y} = \sum_{r=0}^{\infty} e^{-cy}\frac{(cy)^r}{r!}\mathbf{K}^r.$$

Also, for $y \in \{y_1, \ldots, y_M\}$ we have to evaluate the integral $\mathbf{J}(y) = \int_0^y e^{\mathbf{T}(y-u)}\mathbf{t}\boldsymbol{\pi}e^{\mathbf{T}u}\mathrm{d}u$ for which we shall use uniformization. Here

$$\begin{aligned}
\mathbf{J}(y) &= \int_0^y \left(e^{-c(y-u)}\sum_{k=0}^{\infty}\frac{(c\mathbf{K}(y-u))^k}{k!}\right)\mathbf{t}\boldsymbol{\pi}\left(e^{-cu}\sum_{j=0}^{\infty}\frac{(c\mathbf{K}u)^j}{j!}\right)\mathrm{d}u \\
&= e^{-cy}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\left(\int_0^y \frac{(cu)^j}{j!}\frac{(c(y-u))^k}{k!}\mathrm{d}u\right)\mathbf{K}^j\mathbf{t}\boldsymbol{\pi}\mathbf{K}^k \\
&= e^{-cy}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(cy)^{j+k+1}}{j!k!}\frac{j!k!}{(j+k+1)!}\mathbf{K}^j\frac{1}{c}\mathbf{t}\boldsymbol{\pi}\mathbf{K}^k \\
&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+1}}{(s+1)!}\mathbf{D}_{\mathbf{J}}(s), \qquad (4)
\end{aligned}$$

where $\mathbf{D}_{\mathbf{J}}(s) = \sum_{j=0}^{s}\mathbf{K}^j\frac{1}{c}\mathbf{t}\boldsymbol{\pi}\mathbf{K}^{s-j}$, which may be calculated recursively. The matrix $\mathbf{J}(y)$ has the following probabilistic interpretation. The (i,j)-th entry of the matrix is the probability that a phase–type renewal process with inter-arrival distribution $\mathrm{PH}_p(\boldsymbol{\pi}, \mathbf{T})$ (CPH) starting from state i has exactly one arrival in $[0, y]$ and is in state j by time y. From this interpretation we derive the following recursive formula

$$\mathbf{J}(x+y) = e^{\mathbf{T}x}\mathbf{J}(y) + \mathbf{J}(x)e^{\mathbf{T}y}.$$

By this formula we may calculate $\mathbf{J}(x + \Delta x)$, using previous terms, improving the efficiency considerably.

One of the strengths of the uniformization method is the exact upper bound that can be given on the absolute truncation error, due to the role of the weighting factors as the terms in the Poisson probability mass function.

A similar exact upper bound can be given when determining an upper limit for the truncation of the sum involved in calculating $\mathbf{J}(y)$. To see this, we will consider the distribution

$$q_i = \frac{i \cdot \lambda^i}{\lambda i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots,$$

or

$$q_i = \frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda}, \quad i = 1, 2, \ldots,$$

that is the size biased distribution derived from the Poisson distribution with the probabilistic interpretation that it tells what is the fraction of the mean contributed by observations of exactly size i. It is a nice property to see, that in a sense the Poisson distribution is closed under size biasing albeit a shift to the right. If we consider the factors $\mathbf{D_J}(s)$ in the expression for $\mathbf{J}(y)$, we see that all row sums of $\mathbf{D_J}(s)$ are bounded by $s+1$ and thus we can obtain the upper bound for the truncation from the size biased distribution of the Poisson distribution which happens to be the truncation limit for the standard uniformization factor plus 1.

3.2. Newton–Raphson maximization

The EM algorithm is a numerical method for optimizing the incomplete likelihood function. It uses the underlying probabilistic structure of the model and convergence is guaranteed. As an alternative we shall explore a state–of–the–art Newton–Raphson algorithm, and compare its performance to the EM algorithm.

The Newton–Raphson method is based on the idea of approximating a function with its first or second order Taylor expansion. Thus, we need to calculate the gradient vector of the log–likelihood function. This is computationally demanding, particularly if the dimension is large. However, the cost of calculating the gradient could be compensated for by fewer iterations. The method is not designed to work with boundary conditions. While the calculation of the gradient is rather straightforward, the task of making an efficient numerical implementation of the formulae is by no means trivial.

Using the idea given by B. F. Nielsen, *et. al.* [9], we want to work with an unconstrained system, and use a package for unconstrained optimization written by K. Madsen, *et. al.* [7]. Their program, as well as many other standard routines available for unconstrained optimization, find the maximum of a given function using the gradient vector. Since we want to find the maximum of the log-likelihood function, we calculate the gradient vector based on the parameter transformation which provides the unconstrained optimization problem. We shall refer to this method as the Direct Method (DM) since it does not use the underlying probabilistic structure.

The direct method we employ assumes that the parameters are unbounded. This is obviously not the case for the phase–type intensities so we consider a re–parameterization $\boldsymbol{\tau}$ of the parameters. We also need to provide the gradient at a given point of the

transformed parameters,

$$\mathbf{g} = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\tau}} = \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \tau_m} \right)_{m=1,\ldots,p^2+(p-1)}.$$

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p^2+(p-1)})$. By the chain rule this vector can be obtained as

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\tau}} = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}, \tag{5}$$

where $\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}}$ is a $p^2 + (p-1)$-dimensional row vector and $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}}$ is the Jacobian matrix. Taking the derivative of the log-likelihood function w.r.t $\boldsymbol{\theta}$ we get that

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{M} \frac{1}{f(y_k)} \frac{\partial f(y_k)}{\partial \boldsymbol{\theta}},$$

where f is the density of the phase–type distribution parameterized by $\boldsymbol{\theta}$. Thus, the problem reduces to finding the derivative of f with respect to the original parameters. To do this, we introduce

$$\boldsymbol{\Psi}(y) = \begin{cases} \mathbf{T}^{y-1} & \text{for the discrete case (DPH)} \\ e^{\mathbf{T}y} & \text{for the continuous case (CPH).} \end{cases}$$

By substituting $\boldsymbol{\pi} = \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p'$, the density of the phase–type distribution evaluated in y is given by

$$f(y) = \sum_{j=1}^{p-1} \pi_j \mathbf{e}_j' \boldsymbol{\Psi}(y) \mathbf{t} + \left(1 - \sum_{j=1}^{p-1} \pi_j\right) \mathbf{e}_p' \boldsymbol{\Psi}(y) \mathbf{t},$$

and its partial derivatives w.r.t the original parameters are given by

$$\begin{aligned} \frac{\partial f(y)}{\partial \pi_m} &= \mathbf{e}_m' \boldsymbol{\Psi}(y) \mathbf{t} - \mathbf{e}_p' \boldsymbol{\Psi}(y) \mathbf{t} \\ \frac{\partial f(y)}{\partial t_{mn}} &= \boldsymbol{\pi} \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}} \mathbf{t}, \quad m \neq n \\ \frac{\partial f(y)}{\partial t_m} &= \boldsymbol{\pi} \boldsymbol{\Psi}(y) \mathbf{e}_m + \boldsymbol{\pi} \frac{\partial \boldsymbol{\Psi}(y)}{\partial t_m} \mathbf{t}. \end{aligned}$$

In order to compute the partial derivatives of $\boldsymbol{\Psi}$ with respect to θ_m, for $m \in \{1, \ldots, p^2 + (p-1)\}$, we shall need the derivatives of $\mathbf{T}^r$ for $r \geq 1$, and the derivative of $e^{\mathbf{T}y}$. In general, we have that

$$\frac{\partial \mathbf{T}^r}{\partial \theta_m} = \sum_{k=0}^{r-1} \mathbf{T}^k \frac{\partial \mathbf{T}}{\partial \theta_m} \mathbf{T}^{r-1-k}, \quad r \geq 1, \tag{6}$$

where $\left[\frac{\partial \mathbf{T}}{\partial t_{ij}}\right]_{ij} = 1$, $\left[\frac{\partial \mathbf{T}}{\partial t_{ij}}\right]_{ii} = -1$, and $\left[\frac{\partial \mathbf{T}}{\partial t_i}\right]_{ii} = -1$.

Concerning the derivative of $e^{\mathbf{T}y}$, we shall use a uniformization argument similar to (4). We get that

$$\frac{\partial e^{\mathbf{T}y}}{\partial \theta_m} = e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+1}}{(s+1)!}\mathbf{D}_m(s) + \frac{\partial c}{\partial \theta_m}ye^{\mathbf{T}y}(\mathbf{K}-\mathbf{I}), \tag{7}$$

where $\mathbf{D}_m(s) = \dfrac{\partial \mathbf{K}^{s+1}}{\partial \theta_m}$ which is calculated like (6). Since $e^{\mathbf{T}(x+y)} = e^{\mathbf{T}x}e^{\mathbf{T}y}$, we can get a recursive version of (7) given by

$$\frac{\partial e^{\mathbf{T}(x+y)}}{\partial \theta_m} = e^{\mathbf{T}x}\frac{\partial e^{\mathbf{T}y}}{\partial \theta_m} + \frac{\partial e^{\mathbf{T}x}}{\partial \theta_m}e^{\mathbf{T}y}.$$

In order to deal with unconstrained parameters in the optimization we propose the following transformation. For $m = 1, \ldots, p^2 + (p-1)$, let $-\infty < \tau_m < \infty$ be such that

$$\pi_i = \frac{\exp(\tau_i)}{1+\sum_{s=1}^{p-1}\exp(\tau_s)}, i = 1, \ldots, p-1, \quad \pi_p = \frac{1}{1+\sum_{i=1}^{p-1}\exp(\tau_i)},$$

and for $i, j = 1, \ldots, p$, $i \neq j$,

$$t_{ij} = \frac{\exp(\tau_{ip+(j-1)})}{1+\sum_{s=1}^{p}\exp(\tau_{ip+(s-1)})}, \quad t_i = \frac{\exp(\tau_{ip+(i-1)})}{1+\sum_{s=1}^{p}\exp(\tau_{ip+(s-1)})} \qquad \text{(DPH)}$$

$$t_{ij} = \exp(\tau_{ip+(j-1)}), \qquad t_i = \exp(\tau_{ip+(i-1)}). \qquad \text{(CPH)}$$

The elements in the diagonal of $\mathbf{T}$ are defined as $t_{ii} = 1 - \sum_{j=1, j\neq i}^{p} t_{ij} - t_i$ in DPH, and $t_{ii} = -\sum_{j=1, j\neq i}^{p} t_{ij} - t_i$ in CPH. Note that zeros for π_i and t_{ij} are not a possibility in this re–parameterization. However, we can choose to bound t_{ij} or t_i to 0 with obvious changes for the τ_m's.

The Jacobian matrix is constructed as follows. For $i, j = 1, \ldots, p-1$ the (i,j)-th element of this matrix is given by

$$\frac{\partial \pi_i}{\partial \tau_j} = \pi_j \mathbf{1}_{\{j=i\}} - \pi_i\pi_j.$$

For $i, j = 1, \ldots, p$, and $m = p, \ldots, p^2 + (p-1)$, the $(ip + (j-1), m)$-th element of the matrix is given by $\frac{\partial t_{ij}}{\partial \tau_m}$ if $i \neq j$ and $\frac{\partial t_i}{\partial \tau_m}$ if $i = j$, where

$$\frac{\partial t_{ij}}{\partial \tau_m} = \begin{cases} t_{ij}\mathbf{1}_{\{m=ip+(j-1)\}} - t_{ij}\sum_{r=1}^{p}(t_i\mathbf{1}_{\{i=r\}} + t_{ir}\mathbf{1}_{\{i\neq r\}})\mathbf{1}_{\{m=ip+(r-1)\}} & \text{(DPH)} \\ t_{ij}\mathbf{1}_{\{m=ip+(j-1)\}} & \text{(CPH)} \end{cases}$$

$$\frac{\partial t_i}{\partial \tau_m} = \begin{cases} t_i\mathbf{1}_{\{m=ip+(i-1)\}} - t_i\sum_{r=1}^{p}(t_i\mathbf{1}_{\{i=r\}} + t_{ir}\mathbf{1}_{\{i\neq r\}})\mathbf{1}_{\{m=ip+(r-1)\}} & \text{(DPH)} \\ t_i\mathbf{1}_{\{m=ip+(i-1)\}}. & \text{(CPH)} \end{cases}$$

4. Fisher information

Fisher information is a key concept in the theory of statistical inference and essentially describes the amount of information data provide about unknown parameters.

It has applications to finding the variance of an estimator, as well as in the asymptotic behavior of maximum likelihood estimates, and in Bayesian inference.

We present formulae for the Fisher information matrix for a general phase–type distribution. Frequently we may consider sub–classes such as generalized Erlang or Hyper–exponential distributions, where several intensities are assumed to be zero. The corresponding Fisher information is then calculated with the same formulae but summing over indices where the parameters are different from zero. We present methods for calculating the Fisher information matrix for both the EM algorithm and the Newton–Raphson method.

Throughout, we shall assume that the parameters are freely varying and not linked to each other through some common parameters or formulae.

4.1. Fisher information via the EM algorithm

The EM algorithm also allows for extracting information concerning the Fisher information matrix as noted by D. Oakes in [10]. Considering L, the incomplete data likelihood which is maximized by the EM algorithm, the Fisher information matrix is given by

$$\frac{\partial^2 L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}^2} = \left\{ \frac{\partial^2 Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}^2} + \frac{\partial^2 Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \hat{\boldsymbol{\theta}}} \right\}_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}}, \tag{8}$$

where

$$Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(\ell_f(\hat{\boldsymbol{\theta}};\boldsymbol{x})\,|\boldsymbol{y})\,. \tag{9}$$

Define

$$U_i = \sum_{l=1}^{M} \frac{\mathbf{e}_i' \boldsymbol{\Psi}(y_l)\mathbf{t}}{f(y_l)}, \tag{10}$$

$$W_i = \sum_{l=1}^{M} \frac{\boldsymbol{\pi}\boldsymbol{\Psi}(y_l)\mathbf{e}_i}{f(y_l)}, \tag{11}$$

$$V_{ij} = \begin{cases} \displaystyle\sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \frac{1}{f(y_l)} \sum_{k=0}^{y_l-2} \mathbf{e}_j' \mathbf{T}^{y_l-k-2} \mathbf{t}\boldsymbol{\pi}\mathbf{T}^k \mathbf{e}_i & \text{(DPH)} \\ \displaystyle\sum_{l=1}^{M} \frac{1}{f(y_l)} \int_0^{y_l} \mathbf{e}_j' e^{\mathbf{T}(y_l-u)} \mathbf{t}\boldsymbol{\pi} e^{\mathbf{T}u} \mathbf{e}_i \mathrm{d}u. & \text{(CPH)} \end{cases} \tag{12}$$

Then (9) becomes

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) &= \sum_{i=1}^{p-1} \log \hat{\pi}_i U_i \pi_i + \log\left(1 - \sum_{s=1}^{p-1} \hat{\pi}_s\right) U_p \left(1 - \sum_{s=1}^{p-1} \pi_s\right) \\ &\quad + \sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} \log \hat{t}_{ij} V_{ij} t_{ij} + \sum_{i=1}^{p} S_i V_{ii} + \sum_{i=1}^{p} \log(\hat{t}_i) W_i t_i, \end{aligned}$$

where

$$S_i = \begin{cases} \left(1 - \sum_{j=1, j\neq i}^{p} t_{ij} - t_i\right) \log\left(1 - \sum_{j=1, j\neq i}^{p} \hat{t}_{ij} - \hat{t}_i\right) & \text{(DPH)} \\ -\sum_{j=1, j\neq i}^{p} \hat{t}_{ij} - \hat{t}_i. & \text{(CPH)} \end{cases}$$

The elements of the Fisher information matrix (8) are given as follows. For $i, j = 1, \ldots, p-1$, the (i,j)-th element is given by

$$\frac{\partial U_i}{\partial \pi_j} - \frac{\partial U_p}{\partial \pi_j},$$

for $m = 1, \ldots, p-1$ and $i, j = 1, \ldots, p$, the $(ip-1+j, m)$-th element is given by

$$\frac{\partial U_m}{\partial t_{ij}} - \frac{\partial U_p}{\partial t_{ij}} \quad \text{if } i \neq j, \qquad \frac{\partial U_m}{\partial t_i} - \frac{\partial U_p}{\partial t_i} \quad \text{if } i = j,$$

the $(m, ip-1+j)$-th element is given by

$$\frac{\partial V_{ij}}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m} \quad \text{if } i \neq j, \qquad \frac{\partial W_i}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m} \quad \text{if } i = j,$$

and finally, for $i, j, m, n = 1, \ldots, p$, the $(ip-1+j, mp-1+n)$-th element is given by

$$\begin{aligned}
&\frac{\partial V_{ij}}{\partial t_{mn}} - \frac{\partial V_{ii}}{\partial t_{mn}} && \text{if } i \neq j, m \neq n \\
&\frac{\partial V_{ij}}{\partial t_m} - \frac{\partial V_{ii}}{\partial t_m} && \text{if } i \neq j, m = n \\
&\frac{\partial W_i}{\partial t_{mn}} - \frac{\partial V_{ii}}{\partial t_{mn}} && \text{if } i = j, m \neq n \\
&\frac{\partial W_i}{\partial t_m} - \frac{\partial V_{ii}}{\partial t_m} && \text{if } i = j, m = n.
\end{aligned}$$

The explicit formulae of the above derivatives are given in Appendix A.

4.2. Newton–Raphson estimation

To obtain the Fisher information matrix using the direct method, we take the second derivative of (5), which at the optimum gives

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\tau}} \partial \boldsymbol{\tau}} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\tau}} \frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} \frac{\partial \bar{\boldsymbol{\theta}}}{\partial \bar{\boldsymbol{\tau}}}, \tag{13}$$

where $\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}}$ is a square matrix of second–order partial derivatives. For this, we need the second derivatives of the density f w.r.t the original parameters (see Appendix B).

For $m, n \in \{1, \ldots, p^2 + (p-1)\}$, and taking the second derivative of (6) we get

$$\frac{\partial^2 \mathbf{T}^r}{\partial \theta_n \partial \theta_m} = \sum_{k=0}^{r-1} \mathbf{T}^k \frac{\partial \mathbf{T}}{\partial \theta_m} \frac{\partial \mathbf{T}^{r-1-k}}{\partial \theta_n} + \frac{\partial \mathbf{T}^k}{\partial \theta_n} \frac{\partial \mathbf{T}}{\partial \theta_m} \mathbf{T}^{r-1-k}. \tag{14}$$

In the same way from (7), we have that

$$\frac{\partial^2 e^{\mathbf{T}y}}{\partial \theta_n \partial \theta_m} = e^{-cy} \sum_{k=0}^{\infty} \frac{(cy)^{k+1}}{(k+1)!} \frac{\partial^2 \mathbf{K}^{k+1}}{\partial \theta_n \partial \theta_m} + \frac{\partial c}{\partial \theta_m} y \left(e^{\mathbf{T}y} \frac{\partial \mathbf{K}}{\partial \theta_n} + \frac{\partial e^{\mathbf{T}y}}{\partial \theta_n} (\mathbf{K} - \mathbf{I}) \right), \tag{15}$$

where $\frac{\partial^2 \mathbf{K}^r}{\partial \theta_n \partial \theta_m}$, can be calculated like (14).

The quasi Newton method presented in [9] gives an approximate value of the Hessian matrix for the transformed parameters $\boldsymbol{\tau}$ used in the optimization. This can be transformed into an approximation for the inverse Fisher information matrix using

$$\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\bar{\boldsymbol{\theta}}\partial\boldsymbol{\theta}} = \frac{\partial\boldsymbol{\tau}}{\partial\boldsymbol{\theta}}\frac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{x})}{\partial\bar{\boldsymbol{\tau}}\partial\boldsymbol{\tau}}\frac{\partial\bar{\boldsymbol{\tau}}}{\partial\bar{\boldsymbol{\theta}}}.$$

5. Simulation results

The phase–type representation of a given distribution is, in general, non–unique and non–minimal. Hence, we explore a subclass of PH distributions for which the representation is an acyclic graph (APH). A. Cumani [4] has shown that a canonical representation for the APH subclass exists, and this representation is unique, minimal and has the form of a Coxian model with real transition rates. This representation is called a canonical form.

The canonical form representation is given by

$$\boldsymbol{\pi} = (1,0,\ldots,0), \quad \mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & & & & \\ & t_{22} & t_{23} & & & \\ & & \ddots & \ddots & & \\ & & & t_{p-1,p-1} & t_{p-1,p} \\ & & & & t_{pp} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_{p-1} \\ t_p \end{pmatrix}. \tag{16}$$

In this section we present the results of an estimation study considering simulated data from discrete and continuous phase–type distributions. The discrete phase–type distribution has the distribution of a shifted negative binomial random variable, $1 + N$, where N is negative binomially distributed with parameters $(3, 0.2)$. The PH–representation is given by

$$\boldsymbol{\pi} = (1,0,0), \quad \mathbf{T} = \begin{pmatrix} 1-p_1 & (1-p_1)p_1 & (1-p_1)p_1^2 \\ 0 & 1-p_1 & (1-p_1)p_1 \\ 0 & 0 & 1-p_1 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} p_1^3 \\ p_1^2 \\ p_1 \end{pmatrix}.$$

Its equivalent canonical representation is given by

$$\boldsymbol{\pi} = (1,0,0), \quad \mathbf{T} = \begin{pmatrix} 1-p_1 & (1-p_1^2)p_1 & 0 \\ 0 & 1-p_1 & p_1 - \frac{2p_1^2}{1+p_1} \\ 0 & 0 & 1-p_1 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} p_1^3 \\ \frac{2p_1^2}{1+p_1} \\ p_1 \end{pmatrix}.$$

For the continuous case, we consider a mixture of three exponential distributions with parameters $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$. This distribution is also called Hyper–exponential, and has a PH–representation given by

$$\boldsymbol{\pi} = (\pi_1,\pi_2,\pi_3), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & 0 \\ 0 & 0 & -\lambda_3 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix},$$

where $\pi_1 = 0.9$, $\pi_2 = 0.09$, and $\pi_3 = 0.01$. Its equivalent canonical form is given by

$$\boldsymbol{\pi} = (1,0,0), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 - t_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 - t_2 \\ 0 & 0 & -\lambda_3 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \pi_1\lambda_1 + \pi_2\lambda_2 + \pi_3\lambda_3 \\ \frac{\pi_2\lambda_2(\lambda_1-\lambda_2)+\pi_3\lambda_3(\lambda_1-\lambda_3)}{\pi_2(\lambda_1-\lambda_2)+\pi_3(\lambda_1-\lambda_3)} \\ \lambda_3 \end{pmatrix}.$$

The way to obtain the canonical form is given in A. Cumani [4]. All estimation is performed using the canonical form.

After finding the MLE, the Fisher Information (FI) matrix was obtained considering only the non–zero parameters. As the inverse of the FI is the empirical variance–covariance matrix, we could obtain the standard deviation (SD) of the parameters (see tables 1 and 2). The corresponding correlations are given in tables 3 and 4.

TABLE 1: Maximum likelihood estimators (MLE) and standard deviations (SD) of the shifted Negative binomial(3,0.2), considering 10000 observations

Parameter	true value	EM		DM	
		MLE	SD	MLE	SD
$\hat{t}_1$	0.0080	0.0094	0.0009	0.0094	0.0009
$\hat{t}_{12}$	0.1920	0.1939	0.0426	0.1939	0.0455
$\hat{t}_2$	0.0667	0.0592	0.0118	0.0591	0.0125
$\hat{t}_{23}$	0.1333	0.1440	0.0387	0.1441	0.0408
$\hat{t}_3$	0.2000	0.2033	0.0426	0.2032	0.0450

TABLE 2: Maximum likelihood estimators (MLE) and standard deviations (SD) of the Hyper-exponential, considering 20000 observations

Parameter	true value	EM		DM	
		MLE	SD	MLE	SD
$\hat{t}_1$	0.9091	0.9160	0.0080	0.9248	0.0080
$\hat{t}_{12}$	0.0909	0.0934	0.0037	0.0923	0.0037
$\hat{t}_2$	0.0902	0.0922	0.0040	0.0921	0.0040
$\hat{t}_{23}$	0.0098	0.0136	0.0015	0.0152	0.0017
$\hat{t}_3$	0.0100	0.0115	0.0009	0.0121	0.0010

TABLE 3: Correlations of the shifted Negative binomial(3,0.2)

	$\hat{t}_1$	$\hat{t}_{12}$	$\hat{t}_2$	$\hat{t}_{23}$	$\hat{t}_3$
$\hat{t}_1$	1.0000	-0.0118	-0.1855	0.0677	0.0103
$\hat{t}_{12}$	-0.0118	1.0000	-0.9336	-0.2623	-0.4973
$\hat{t}_2$	-0.1855	-0.9336	1.0000	0.1916	0.4512
$\hat{t}_{23}$	0.0677	-0.2623	0.1916	1.0000	-0.6842
$\hat{t}_3$	0.0103	-0.4973	0.4512	-0.6842	1.0000

6. Concluding remarks

The paper by Asmussen *et. al.* [2] provided the statistical framework for obtaining maximum likelihood estimates of continuous PH distributions using the EM algorithm. In this paper we have demonstrated how one can obtain uncertainty estimates of

Table 4: Correlations of the Hyper–exponential

	$\hat{t}_1$	$\hat{t}_{12}$	$\hat{t}_2$	$\hat{t}_{23}$	$\hat{t}_3$
$\hat{t}_1$	1.0000	0.3451	0.2418	0.0591	0.0429
$\hat{t}_{12}$	0.3451	1.0000	0.5777	0.1874	0.1148
$\hat{t}_2$	0.2418	0.5777	1.0000	0.4171	0.2300
$\hat{t}_{23}$	0.0591	0.1874	0.4171	1.0000	0.4887
$\hat{t}_3$	0.0429	0.1148	0.2300	0.4887	1.0000

the parameters in cases where the PH distribution is not over–parameterized. The development is done for discrete as well as for continuous PH distributions. We have discussed two different ways of analytically obtaining the Fisher Information matrix in such cases. One of these methods is based on a direct calculation of second derivatives of the log–likelihood function while the other is based on a paper by Oakes [10] where the partial derivatives are made using a split of the log–likelihood function as in the EM algorithm. The methods are quite similar with respect to the actual analytical and numerical calculations. In particular, the truncation error of the algorithm can in both cases be controlled exactly in the same way as for the uniformization method. In turn we suggest a technical alternative based on uniformization for the calculation of matrix–exponentials and certain integrals in the continuous version of the EM algorithm. The main advantage of using the uniformization based approach is the possibility of controlling the numerical error during the successive iterations. We also demonstrate how one could alternatively obtain maximum likelihood estimates by a direct approach using an up–to–date (quasi) Newton–Raphson method.

We have demonstrated our results using a couple of numerical examples, one for the discrete case and one for the continuous case. The two algorithms gave the same result for the Fisher information, a result that was verified by the approximate information on the Hessian matrix provided by the quasi Newton–Raphson method.

Our implementations did not provide significant evidence that one of the two optimization methods should be preferred over the other. In most cases our implementations were competitive with the Runge–Kutta based approach also in terms of efficiency.

In the future we will modify our approach to be able to handle cases with fewer free parameters in the PH representations. For example, we may consider phase–type distributions in arbitrary dimensions where certain transition rates are equal or proportional to each other. In this case we need to provide alternative formulae for the EM algorithm and the Fisher information.

Another topic for future study is to improve the efficiency of the algorithms. Many matrix–matrix and matrix–vector products are used a number of times throughout. It might thus be possible to optimize our implementations further with different strategies for calculating and storing intermediate results.

Appendix A. Fisher information matrix, EM

Let $R_i(u) = \boldsymbol{\pi}\boldsymbol{\Psi}(u)\mathbf{e}_i$ and $Q_i(u) = \mathbf{e}_i'\boldsymbol{\Psi}(u)\mathbf{t}$. Then, their derivatives are given by

$$\begin{array}{ll}
\frac{\partial R_i(u)}{\partial \pi_m} = \mathbf{e}_m'\boldsymbol{\Psi}(u)\mathbf{e}_i - \mathbf{e}_p'\boldsymbol{\Psi}(u)\mathbf{e}_i, & \frac{\partial Q_i(u)}{\partial \pi_m} = 0, \\
\frac{\partial R_i(u)}{\partial t_{mn}} = \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(u)}{\partial t_{mn}}\mathbf{e}_i, \quad m \neq n, & \frac{\partial Q_i(u)}{\partial t_{mn}} = \mathbf{e}_i'\frac{\partial \boldsymbol{\Psi}(u)}{\partial t_{mn}}\mathbf{t}, \quad m \neq n, \\
\frac{\partial R_i(u)}{\partial t_m} = \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(u)}{\partial t_m}\mathbf{e}_i, & \frac{\partial Q_i(u)}{\partial t_m} = \mathbf{e}_i'\boldsymbol{\Psi}(u)\mathbf{e}_m + \mathbf{e}_i'\frac{\partial \boldsymbol{\Psi}(u)}{\partial t_m}\mathbf{t}.
\end{array}$$

Then U_i, W_i, and V_{ij} (see (10), (11), and (12)) become

$$\begin{aligned}
U_i &= \sum_{l=1}^{M} \frac{Q_i(y_l)}{f(y_l)}, \\
W_i &= \sum_{l=1}^{M} \frac{R_i(y_l)}{f(y_l)}, \\
V_{ij} &= \begin{cases} \sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \frac{1}{f(y_l)} \sum_{k=0}^{y_l-2} Q_j(y_l-k-1)R_i(k+1) & \text{(DPH)} \\ \sum_{l=1}^{M} \frac{1}{f(y_l)} \int_0^{y_l} Q_j(y_l-u)R_i(u)\mathrm{d}u. & \text{(CPH)} \end{cases}
\end{aligned}$$

Hence, for $n \in \{1, \ldots, p^2 + (p-1)\}$, the derivatives w.r.t θ_n are given by

$$\begin{aligned}
\frac{\partial U_i}{\partial \theta_n} &= \sum_{l=1}^{M} \frac{1}{f(y_l)^2}\left(f(y_l)\frac{\partial Q_i(y_l)}{\partial \theta_n} - Q_i(y_l)\frac{\partial f(y_l)}{\partial \theta_n}\right), \\
\frac{\partial W_i}{\partial \theta_n} &= \sum_{l=1}^{M} \frac{1}{f(y_l)^2}\left(f(y_l)\frac{\partial R_i(y_l)}{\partial \theta_n} - R_i(y_l)\frac{\partial f(y_l)}{\partial \theta_n}\right), \\
\frac{\partial V_{ij}}{\partial \theta_n} &= \begin{cases} \sum_{l=1}^{M} \mathbf{1}_{\{y_l \geq 2\}} \sum_{k=0}^{y_l-2} \frac{1}{f(y_l)^2}\Big[f(y_l)\left(Q_j(y_l-k-1)\frac{\partial R_i(k+1)}{\partial \theta_n} + \frac{\partial Q_j(y_l-k-1)}{\partial \theta_n}R_i(k+1)\right) \\ -\left(\frac{\partial f(y_l)}{\partial \theta_n}\right)Q_j(y_l-k-1)R_i(k+1)\Big] & \text{(DPH)} \\ \sum_{l=1}^{M} \frac{1}{f(y_l)^2}\Big[f(y_l)\int_0^{y_l} Q_j(y_l-u)\left(\frac{\partial R_i(u)}{\partial \theta_n}\right) + \left(\frac{\partial Q_j(y_l-u)}{\partial \theta_n}\right)R_i(u)\mathrm{d}u \\ -\left(\frac{\partial f(y_l)}{\partial \theta_n}\right)\int_0^{y_l} Q_j(y_l-u)R_i(u)\mathrm{d}u\Big]. & \text{(CPH)} \end{cases}
\end{aligned}$$

Concerning the computation of $\frac{\partial V_{ij}}{\partial \theta_n}$ for CPH, we define the following integrals

$$\begin{aligned}
\mathbf{J}_1(y;\mathbf{M}) &= \int_0^y e^{\mathbf{T}(y-u)}\mathbf{M}e^{\mathbf{T}u}\mathrm{d}u &&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+1}}{(s+1)!}\mathbf{D}_{\mathbf{J}_1}(s), \\
\mathbf{J}_2(y;\theta_n,\mathbf{M}) &= \int_0^y e^{\mathbf{T}(y-u)}\mathbf{M}\tfrac{\partial e^{\mathbf{T}u}}{\partial \theta_n}\mathrm{d}u &&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+2}}{(s+2)!}\left(\mathbf{D}_{\mathbf{J}_2,1}(s,\theta_n) + \mathbf{D}_{\mathbf{J}_2,2}(s,\theta_n)\right), \\
\mathbf{J}_3(y;\theta_n,\mathbf{M}) &= \int_0^y \tfrac{\partial e^{\mathbf{T}(y-u)}}{\partial \theta_n}\mathbf{M}e^{\mathbf{T}u}\mathrm{d}u &&= e^{-cy}\sum_{s=0}^{\infty}\frac{(cy)^{s+2}}{(s+2)!}\left(\mathbf{D}_{\mathbf{J}_3,1}(s,\theta_n) + \mathbf{D}_{\mathbf{J}_3,2}(s,\theta_n)\right),
\end{aligned}$$

where $\mathbf{M}$ is a $(p \times p)$–matrix and

$$\begin{aligned}
\mathbf{D}_{\mathbf{J}_1}(s) &= \sum_{j=0}^{s} \mathbf{K}^j \frac{1}{c} \mathbf{M} \mathbf{K}^{s-j}, \\
\mathbf{D}_{\mathbf{J}_2,1}(s,\theta_n) &= \sum_{j=0}^{s} \mathbf{K}^j \frac{1}{c} \mathbf{M} \frac{\partial \mathbf{K}^{s-j+1}}{\partial \theta_n}, \quad \mathbf{D}_{\mathbf{J}_2,2}(s,\theta_n) = \sum_{j=0}^{s} \mathbf{K}^j (s+1-j) \frac{1}{c^2} \frac{\partial c}{\partial \theta_n} \mathbf{M}(\mathbf{K}-\mathbf{I}) \mathbf{K}^{s-j}, \\
\mathbf{D}_{\mathbf{J}_3,1}(s,\theta_n) &= \textstyle\sum_{j=0}^{s} \frac{\partial \mathbf{K}^{s-j+1}}{\partial \theta_n} \frac{1}{c} \mathbf{M} \mathbf{K}^j, \quad \mathbf{D}_{\mathbf{J}_3,2}(s,\theta_n) = \sum_{j=0}^{s} \mathbf{K}^j (j+1) \frac{1}{c^2} \frac{\partial c}{\partial \theta_n} (\mathbf{K}-\mathbf{I}) \mathbf{M} \mathbf{K}^{s-j}.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial V_{ij}}{\partial \pi_m} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \left(\mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\mathbf{e}_m') \mathbf{e}_i - \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\mathbf{e}_p') \mathbf{e}_i \right) - \frac{\partial f(y_k)}{\partial \pi_m} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i \right], \\
\frac{\partial V_{ij}}{\partial t_{mn}} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \left(\mathbf{e}_j' \mathbf{J}_2(y_k; t_{mn}, \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j' \mathbf{J}_3(y_k; t_{mn}, \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i \right) - \frac{\partial f(y_k)}{\partial t_{mn}} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i \right], \\
\frac{\partial V_{ij}}{\partial t_m} &= \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[-f(y_k) \left(\mathbf{e}_j' \mathbf{J}_2(y_k; t_m, \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j' \mathbf{J}_1(x_k; \mathbf{e}_m \boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j' \mathbf{J}_3(y_k; t_m, \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i \right) \right. \\
&\qquad \left. - \frac{\partial f(y_k)}{\partial t_m} \mathbf{e}_j' \mathbf{J}_1(y_k; \mathbf{t}\boldsymbol{\pi}) \mathbf{e}_i \right].
\end{aligned}$$

A proper truncation of the infinite sums involved in $\mathbf{J}_i$, $i = 1, 2, 3$, can be obtained using the same approach as for $\mathbf{J}$ discussed in Section 3.1. The row sums of the matrix $\mathbf{D}_{\mathbf{J}_1}(s)$ are like the ones for $\mathbf{D}_{\mathbf{J}}(s)$ bounded by $s+1$, while the row sums of $\mathbf{D}_{\mathbf{J}_2,1}(s,\cdot)$, $\mathbf{D}_{\mathbf{J}_2,2}(s,\cdot)$, $\mathbf{D}_{\mathbf{J}_3,1}(s,\cdot)$, and $\mathbf{D}_{\mathbf{J}_3,2}(s,\cdot)$ are bounded by $\frac{1}{2}(s+1)(s+2)$. Thus to find a proper level for truncation we can restrict ourselves to the scalar sum

$$\sum_{s=0}^{\infty} e^{-cy} \frac{(cy)^{s+2}}{(s+2)!} \cdot \frac{1}{2}(s+1)(s+2) = -\frac{1}{2} \sum_{s=2}^{\infty} e^{-cy} \frac{(cy)^s}{s!} \cdot s + \frac{1}{2} \sum_{s=2}^{\infty} e^{-cy} \frac{(cy)^s}{s!} \cdot s^2,$$

which represents the summation of the first and second order moment distribution of the Poisson distribution.

As in Section 3.1 the truncation level is thus the standard uniformization level plus 1 and plus 2, respectively.

Appendix B. Hessian matrix for the Newton–Raphson method

Taking the second derivative of the log–likelihood function we get

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} = \sum_{k=1}^{M} \frac{1}{f(y_k)^2} \left[f(y_k) \frac{\partial^2 f(y_k)}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}} - \frac{\partial f(y_k)}{\partial \bar{\boldsymbol{\theta}}} \frac{\partial f(y_k)}{\partial \boldsymbol{\theta}} \right],$$

where the second derivatives of the density with respect to the initial probabilities are 0, i.e.

$$\frac{\partial^2 f(y)}{\partial \pi_n \partial \pi_m} = 0.$$

While with respect to the elements of the matrix $\mathbf{T}$, the second derivatives are given by

$$\frac{\partial^2 f(y)}{\partial t_{mn}\partial t_{ij}} = \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_{mn}\partial t_{ij}}\mathbf{t}, \quad m \neq n, i \neq j,$$

and w.r.t the exit probabilities

$$\frac{\partial^2 f(y)}{\partial t_m \partial t_i} = \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_m}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{e}_m + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_m \partial t_i}\mathbf{t}.$$

Finally,

$$\begin{aligned}
\frac{\partial^2 f(y)}{\partial \pi_m \partial t_{ij}} = \frac{\partial^2 f(y)}{\partial t_{ij}\partial \pi_m} &= \mathbf{e}_m'\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{ij}}\mathbf{t} - \mathbf{e}_p'\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{ij}}\mathbf{t}, \quad i \neq j \\
\frac{\partial^2 f(y)}{\partial \pi_m \partial t_i} = \frac{\partial^2 f(y)}{\partial t_i \partial \pi_m} &= \mathbf{e}_m'\boldsymbol{\Psi}(y)\mathbf{e}_i - \mathbf{e}_p'\boldsymbol{\Psi}(y)\mathbf{e}_i + \mathbf{e}_m'\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{t} - \mathbf{e}_p'\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_i}\mathbf{t} \\
\frac{\partial^2 f(y)}{\partial t_{mn}\partial t_i} &= \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_{mn}\partial t_i}\mathbf{t}, \quad m \neq n \\
\frac{\partial^2 f(y)}{\partial t_i \partial t_{mn}} &= \boldsymbol{\pi}\frac{\partial \boldsymbol{\Psi}(y)}{\partial t_{mn}}\mathbf{e}_i + \boldsymbol{\pi}\frac{\partial^2 \boldsymbol{\Psi}(y)}{\partial t_i \partial t_{mn}}\mathbf{t}, \quad m \neq n.
\end{aligned}$$

Acknowledgements

Mogens Bladt would like to acknowledge the support of research grant 15945 from Sistema Nacional de Investigadores and grant 48538-F of CONACYT.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *EEE Transactions on Automatic Control*, **19**, 716–723.

[2] Asmussen, S., Nerman, O. and Olsson, M. (1996). Fitting Phase–type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, **23**, 419–441.

[3] Callut, J. and Dupont, P. (2006). Sequence Discrimination using Phase–type Distributions. *Springer-Verlag Berlin Heidelberg*, Lecture Notes in Artificial Intelligence, **4212**, 78–89.

[4] Cumani, A. (1982). On the canonical representation of homogeneous Markov Process modelling failure–time distributions. *Microelectronics and Reliability*, **22**, 583–602.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38.

[6] Latouche, G. and Ramaswami, V. (1999). Introduction to Matrix Analytic Methods in Stochastic Modeling, ASA–SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[7] Madsen, K., Nielsen, H. and Sondergaard, J. (2002). Robust subroutines for non-linear optimization. *Technical University of Denmark*. Technical Report. IMM-REP-2002-02.

[8] Neuts, M. F. (1981). *Matrix–geometric solutions in stochastic models*, volume 2 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, Md.

[9] Nielsen, B. F. and Beyer, J. E. (2005). Estimation of Interrupted Poisson Process Parameters from Counts. *Institute Mittag–Leffler*. Technical Report. IML-R- -21-04/05- -SE+fall.

[10] Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society*, **61**, 479–482.

Appendix B

Bilateral matrix-exponential distributions

Paper for The Seventh International Conference on Matrix-Analytic Methods in Stochastic Models (MAM7). June 2011.

Summited

BILATERAL MATRIX–EXPONENTIAL DISTRIBUTIONS

MOGENS BLADT, LUZ JUDITH R. ESPARZA, AND BO FRIIS NIELSEN

Abstract. In this article we consider the classes of bilateral phase–type distributions and univariate matrix–exponential distributions in order to define a new general class of probability distributions called bilateral matrix–exponential (BME) distributions, whose support is the entire real line and whose moment–generating function is a rational function. Moreover, this is extended to the multivariate case (MVBME) where the distributions have multidimensional rational moment–generating function. We prove a characterization that states that a random variable is MVBME distributed if and only if all non-null linear combinations of the coordinates are univariate BME distributed.

Primary 62H05; Secondary 60E10

Bilateral phase–type distribution, matrix–exponential distribution, multivariate phase–type distribution, moment–generating function.

1. Introduction

Phase–type (PH) distributions (Neuts [12, 13]) have become quite popular in the last decades as they have been used in a wide range of applications of stochastic models. A PH distributed random variable can be interpreted as the time till absorption in an absorbing Markov chain. This class of distributions has enjoyed such popularity because it forms a dense subset in the space of all distributions defined on the non-negative real numbers. That is, they can approximate any non-negative distribution arbitrarily closely (see Asmussen [3]).

In the multivariate case, this class of distributions has been considered initially by Assaf *et.al* [6] and later by Kulkarni [11]. Previous work on PH distributions has been extended into the real line by Shanthikumar [16] and by Ahn and Ramaswami [1], defining the class of bilateral phase–type distributions.

Another generalization of PH distributions is the class of matrix–exponential (ME) distributions (distributions with rational Laplace transform), that have been studied, for instance, by Asmussen and Bladt [5], Bladt and Neuts [7], and in the multivariate case (denoted by MVME) by Bladt and Nielsen [8].

Asmussen and Bladt [5] have studied the class of ME distributions applying them in the study of a class of queueing systems. Also identifying some necessary and sufficient conditions for an ME representation to be minimal. Liefvoort [17]

Date: March 31, 2011.

proposed a method that provides insight into the minimal representation problem for phase–type distributions. Indeed, this method characterizes completely the ME distributions with a finite order. In 2007, Qi-Ming and Hanqin [10] established some relationships between the Laplace transforms, the distribution functions, and the minimal ME representations of ME distributions.

Moreover, Bladt and Neuts [7] have studied the class of ME distributions and they related ME renewal process through a randomly stopped deterministic flow model. More recently, Bodrog *et.al* [9] have given a characterization of ME processes, presenting an algorithm to compute their finite dimensional moments based on a set of required (low order) moments.

The main purpose of this paper is to generalize the class of matrix–exponential (univariate and multivariate) distributions into the real space, in order to unify a number of distributions and use them, for instance, for modelling whenever the multivariate Gaussian is not sufficient. For this goal, we introduce the class of bilateral ME distributions (distributions with rational moment–generating (MG) function) for both univariate and multivariate cases, as a natural extension of the ME and MVME distributions, respectively.

The remainder of this paper is organized as follows. In Section 2 we provide necessary background on PH and ME distributions. Bilateral ME distributions are defined in Section 3. In Section 4 we give a generalization of Bilateral PH distributions considering the multivariate case. The minimal order of Bilateral ME distributions is analyzed in Section 5. The multivariate case of Bilateral ME is considered in Section 6. In Section 7 as an application, we study terminal distributions of Markov additive processes with absorption. Finally, the article is concluded in Section 8.

2. Background

Let $J = \{J(t)\}_{t\geq 0}$ be a continuous time Markov chain with state space composed by m transient states $\{1, 2, \ldots, m\}$ and an absorbing one $\{m+1\}$. Suppose that J has an initial probability vector $(\boldsymbol{\alpha}, \alpha_{m+1})$, where $\boldsymbol{\alpha}$ in turn denotes a vector of dimension m; and a generator matrix given by

$$\begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{1}$$

where $\mathbf{T}$ is an invertible $(m \times m)$-dimensional matrix satisfying $t_{ii} < 0$ and $t_{ij} \geq 0$, for $i \neq j$; and $\mathbf{t}$ is an m-dimensional column vector such that $\mathbf{t} = -\mathbf{T}\mathbf{e}$, where $\mathbf{e}$ denotes a column vector with 1 at all entries. Then the time to absorption of J, $\tau = \inf\{t \geq 0 : J(t) = m+1\}$, is said to be phase–type distributed with initial probability vector $\boldsymbol{\alpha}$ and sub-generator matrix $\mathbf{T}$. Since the dimension of $\mathbf{T}$ is m, we may write this by $\tau \sim \mathrm{PH}_m(\boldsymbol{\alpha}, \mathbf{T})$. The probability density function of τ, for $x > 0$, is given by $f(x) = \boldsymbol{\alpha} e^{\mathbf{T}x}\mathbf{t}$ (see Neuts [12, 13]).

Sometimes it is convenient to allow for an atom at zero as well, in which case we let $\alpha_{m+1} > 0$ denote the probability of initiating the process in the absorbing state.

In general, let X be a non-negative random variable (r.v.) with density function $b(x) = \boldsymbol{\alpha} e^{\mathbf{T}x}\mathbf{t}$, where $\boldsymbol{\alpha}$ is a row vector, $\mathbf{t}$ is a column vector, and $\mathbf{T}$ is a matrix, then we say that X is matrix–exponentially distributed. The triple $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{t})$ is called a representation for the distribution of X, and we write $X \sim \mathrm{ME}(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{t})$. The Laplace–Stieltjes transform of X, its moments, and reduced moments can be computed as:

$$\begin{array}{lll} L_X(s) & = \mathbb{E}(e^{-sX}) & = \alpha_{m+1} + \boldsymbol{\alpha}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}, \\ M_i & = \mathbb{E}(X^i) & = i!\boldsymbol{\alpha}(-\mathbf{T})^{-(i+1)}\mathbf{t}, \\ \mu_i & = \frac{\mathbb{E}(X^i)}{i!} & = \boldsymbol{\alpha}(-\mathbf{T})^{-(i+1)}\mathbf{t}, \end{array}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimension. Indeed, if X is a strictly continuous r.v., it has no probability mass at zero, i.e., $\boldsymbol{\alpha}(-\mathbf{T})^{-1}\mathbf{t} = 1$.

The Laplace–Stieltjes transform of an ME distributed r.v. is thus rational. It is immediate that any r.v. with rational Laplace–Stieltjes transform is also ME, see [5] for details.

It is also clear that a phase–type distribution is matrix–exponential with representation $(\boldsymbol{\alpha}, \mathbf{T}, -\mathbf{T}\mathbf{e})$. Without loss of generality we can take $0 \leq \boldsymbol{\alpha}\mathbf{e} \leq 1$ and $\mathbf{T}\mathbf{e} + \mathbf{t} = 0$ also in the ME case.

3. Univariate bilateral matrix–exponential distributions

We aim to generalize the class of matrix–exponential distributions to a class that we shall call bilateral matrix–exponential distributions on the entire line $(-\infty, \infty)$.

If a random variable X has a rational moment–generating function, then that function can be expressed as the fraction of two polynomials $A(s)$ and $B(s)$ as

$$M_X(s) = \frac{B(s)}{A(s)}. \tag{2}$$

Theorem 3.1. *X has a rational moment–generating function if and only if the continuous part of its density function can be written as follows*

$$f_X(x) = \boldsymbol{\alpha}_+ e^{\boldsymbol{T}_+ x}\boldsymbol{t}_+ \mathbf{1}_{\{x>0\}} + \boldsymbol{\alpha}_- e^{\boldsymbol{T}_- |x|}\boldsymbol{t}_- \mathbf{1}_{\{x<0\}}, \tag{3}$$

where $\boldsymbol{\alpha}_+$ is a row vector of some dimension m_+, $\boldsymbol{T}_+$ is a matrix of dimension $m_+ \times m_+$, and $\boldsymbol{t}_+$ is an m_+-dimensional column vector. Similarly, both the vectors $\boldsymbol{\alpha}_-$,$\boldsymbol{t}_-$, and the matrix $\boldsymbol{T}_-$, are defined by some dimension m_-.

Without loss of generality, we can take $\boldsymbol{\alpha}_+$, $\boldsymbol{\alpha}_-$, $\boldsymbol{T}_+$, and $\boldsymbol{T}_-$ real valued such that $0 \leq \boldsymbol{\alpha}_+\boldsymbol{e} + \boldsymbol{\alpha}_-\boldsymbol{e} \leq 1$, and $\boldsymbol{T}_+\boldsymbol{e} + \boldsymbol{t}_+ = \boldsymbol{T}_-\boldsymbol{e} + \boldsymbol{t}_- = 0$.

Proof. Let X be a random variable with density given by (3), then its MG function, $M_X(s) = \mathbb{E}(e^{sX})$, is given by

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{\infty} e^{sx} \mathrm{d}F(x) \\ &= (1 - \boldsymbol{\alpha}_+\mathbf{e} - \boldsymbol{\alpha}_-\mathbf{e}) + \int_{-\infty}^{\infty} e^{sx} \left(\boldsymbol{\alpha}_+ e^{\mathbf{T}_+ x}\mathbf{t}_+ \mathbf{1}_{\{x>0\}} + \boldsymbol{\alpha}_- e^{\mathbf{T}_- |x|}\mathbf{t}_- \mathbf{1}_{\{x<0\}} \right) \mathrm{d}x \\ &= (1 - \boldsymbol{\alpha}_+\mathbf{e} - \boldsymbol{\alpha}_-\mathbf{e}) + \int_{0}^{\infty} e^{sx} \boldsymbol{\alpha}_+ e^{\mathbf{T}_+ x}\mathbf{t}_+ \mathrm{d}x + \int_{-\infty}^{0} e^{sx} \boldsymbol{\alpha}_- e^{\mathbf{T}_- |x|}\mathbf{t}_- \mathrm{d}x \\ &= (1 - \boldsymbol{\alpha}_+\mathbf{e} - \boldsymbol{\alpha}_-\mathbf{e}) + \boldsymbol{\alpha}_+(-s\mathbf{I} - \mathbf{T}_+)^{-1}\mathbf{t}_+ + \boldsymbol{\alpha}_-(s\mathbf{I} - \mathbf{T}_-)^{-1}\mathbf{t}_-, \end{aligned}$$

where both terms $\boldsymbol{\alpha}_+(-s\mathbf{I} - \mathbf{T}_+)^{-1}\mathbf{t}_+$ and $\boldsymbol{\alpha}_-(s\mathbf{I} - \mathbf{T}_-)^{-1}\mathbf{t}_-$ are rational ([5]). Thus $M_X(s)$ is the sum of rational functions in s, and then rational.

On the other hand, let $M_X(s)$ be the MG function of X given by (2). We can write $A(s) = A_+(s)A_-(s)$ where $A_+(s)$ is the polynomial which has roots in the positive half plane and $A_-(s)$ the one which has roots in the negative half plane. Now define $B_+(s)$ and $B_-(s)$ (see Appendix A), such that

$$B(s) \quad = \quad (1 - \boldsymbol{\alpha}_+\mathbf{e} - \boldsymbol{\alpha}_-\mathbf{e})(A_+(s)A_-(s)) + A_+(s)B_-(s) + A_-(s)B_+(s),$$

then the MG function becomes

$$M_X(s) = (1 - \boldsymbol{\alpha}_+\mathbf{e} - \boldsymbol{\alpha}_-\mathbf{e}) + \frac{B_+(s)}{A_+(s)} + \frac{B_-(s)}{A_-(s)}, \tag{4}$$

where the functions related to $\frac{B_+(s)}{A_+(s)}$ and $\frac{B_-(s)}{A_-(s)}$ are non-negative, having support on the positive and negative reals, respectively.

If we are in the case where there are no positive (negative) roots, then we define $A_+(s) = 1$ and $B_+(s) = 0$ ($A_-(s) = 1$ and $B_-(s) = 0$).

Then using Lemma 2.1 from Asmussen and Bladt [5] with the appropriate notation, we get that $M_X(s) = (1-\boldsymbol{\alpha}_+\mathbf{e}-\boldsymbol{\alpha}_-\mathbf{e})+\boldsymbol{\alpha}_+(-s\mathbf{I}-\mathbf{T}_+)^{-1}\mathbf{t}_+ + \boldsymbol{\alpha}_-(s\mathbf{I}-\mathbf{T}^-)^{-1}\mathbf{t}_-$, which represents the MG function of a r.v. with density given by (3). □

Definition 3.1. *We say that X is univariate bilateral matrix–exponential or simply bilateral matrix–exponential (BME) distributed, if it has rational moment–generating function, i.e. if $\mathbb{E}(e^{sX})$ is rational in s.*

We denote by $X \sim \mathrm{BME}(\boldsymbol{\alpha}_+, \mathbf{T}_+, \mathbf{t}_+, \boldsymbol{\alpha}_-, \mathbf{T}_-, \mathbf{t}_-)$ when X has the density given by (3).

Observation 3.1. *We have seen that if $X \sim BME(\boldsymbol{\alpha}_+, \boldsymbol{T}_+, \boldsymbol{t}_+, \boldsymbol{\alpha}_-, \boldsymbol{T}_-, \boldsymbol{t}_-)$ its MG function can be written as (4), where the degree of $A_+(s) = \det(-s\boldsymbol{I} - \boldsymbol{T}_+)$ is the dimension of $\boldsymbol{T}_+$, let us say m_+, and in the same way the degree of $A_-(s) = \det(s\boldsymbol{I} - \boldsymbol{T}_-)$ is m_-, then*

$$\begin{aligned} M_X(s) &= (1 - \boldsymbol{\alpha}_+\boldsymbol{e} - \boldsymbol{\alpha}_-\boldsymbol{e}) + \frac{B_+(s)}{A_+(s)} + \frac{B_-(s)}{A_-(s)} \\ &= \frac{(1 - \boldsymbol{\alpha}_+\boldsymbol{e} - \boldsymbol{\alpha}_-\boldsymbol{e})(A_+(s)A_-(s))}{A_+(s)A_-(s)} + \frac{B_+(s)A_-(s) + B_-(s)A_+(s)}{A_+(s)A_-(s)} \end{aligned}$$

has degree $m_+ + m_-$. If $B_+(s)$ and $A_+(s)$ ($B_-(s)$ and $A_-(s)$) have no common factors, then $\boldsymbol{T}_+$ ($\boldsymbol{T}_-$) has the lowest dimension possible. When both $\boldsymbol{T}_+$ and $\boldsymbol{T}_-$ have the lowest dimension, we say that the representation is of minimal order. The number $m = m_+ + m_-$ is the order of the distribution. We will analyze this issue in more detail in Section 5.

4. A generalization of phase–type distributions

Consider a phase–type distributed random variable τ with an order-m representation $(\boldsymbol{\alpha}, \mathbf{T})$. We can interpret τ as resulting from a simple reward structure on a finite Markov jump process $\{J(t)\}_{t\geq 0}$. If the reward rate is 1 in each state, then the total reward is phase–type distributed.

Now, with the help of Markov reward models, we have the following analysis of phase–type distributions. We assign a real valued constant $r(i)$, referred to as the reward rate to each state, and a real valued reward function $W(t)$, to $J(t)$ such that $W(t)$ describes the reward accumulated by $J(t)$ in the interval $(0, t)$. We assume $W(0) = 0$, during the sojourn in state i the amount of accumulated reward increases at rate $r(i)$, i.e. $\mathrm{d}W(t)/\mathrm{d}t = r(i)$, when $J(t) = i$. If $r(i)$ is negative $W(t)$ decreases during the sojourn in i. The amount of reward accumulated during the interval $(0, t)$ is

$$W(t) = \int_0^t r(J(s))\mathrm{d}s. \tag{5}$$

If the rewards are different in each state and strictly positive, we obtain a phase–type distributed random variable, $X \sim \mathrm{PH}_m(\boldsymbol{\alpha}, \boldsymbol{\Delta}(\mathbf{r})^{-1}\mathbf{T})$, where $\boldsymbol{\Delta}(\mathbf{r})$ is the diagonal matrix composed of the reward rates of the transient states $\mathbf{r} = (r(1), \ldots, r(m))'$. Its MG function is given by

$$M_X(s) = \alpha_{m+1} + \boldsymbol{\alpha}(s\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{r}) + \mathbf{I})^{-1}\mathbf{e}, \tag{6}$$

see Ahn and Ramaswami [1].

When the reward vector $\mathbf{r}$ is a non-zero real vector, we obtain the class of bilateral phase–type distributions, which was introduced by Ahn and Ramaswami [1] and it is denoted in this paper by BPH*.

Definition 4.1. [1] *Let X denote the total accumulated reward until absorption, that is, $X = W(\tau)$. X is said to be a bilaterally phase–type distributed random variable with initial probability vector $\boldsymbol{\alpha}$, transient generator $\boldsymbol{T}$, and reward matrix $\boldsymbol{\Delta}(\boldsymbol{r})$. We denote this by $X \sim BPH^*(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\Delta}(\boldsymbol{r}))$.*

It is clear from the construction of the BPH* class that it has an atom at zero if and only if $\alpha_{m+1} = 1 - \boldsymbol{\alpha}\mathbf{e} > 0$.

Note that the moment–generating function (6) is rational (see Theorem 3.1), i.e. X is BME distributed with representation given in the Theorem 4.1 by Ahn and

Ramaswami [1]. And, even more, it turns out to be also valid for some i such as $r(i) = 0$ (see Bladt and Nielsen [8]).

Kulkarni [11] used a similar construction as (5) to define a multivariate phase–type (MPH*) distributed random variable. In order to define the class of multivariate bilateral phase–type distributions, in the following analysis we will present the construction of the MPH* class and its characterization given in [8].

For $j = 1, \ldots, k$, let $\mathbf{r}_j = (r_j(1), \ldots, r_j(m))'$ be k non-negative m-column reward vectors. And define $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_k)$ the $(m \times k)$-dimensional reward matrix. Now, like (5) we have

$$X_j = \int_0^\tau r_j(J(t))\mathrm{d}t, \quad 1 \leq j \leq k,$$

and the vector $\mathbf{X} = (X_1, \ldots, X_k)$ is said to have MPH* distribution with representation $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$.

From Theorem 2.3.2 by Bladt and Nielsen [8], we get that if $\mathbf{X} \sim \mathrm{MPH}^*(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$, then $\langle \mathbf{X}, \mathbf{a} \rangle \sim \mathrm{PH}_m(\boldsymbol{\alpha}, \boldsymbol{\Delta}(\mathbf{Ra})^{-1}\mathbf{T})$, for all k-dimensional column vectors $\mathbf{a}$ such that $\mathbf{Ra} > \mathbf{0}$. In addition, the MG function of $\langle \mathbf{X}, \mathbf{a} \rangle$ is given by

$$M_{\langle \mathbf{X}, \mathbf{a} \rangle}(s) = \alpha_{m+1} + \boldsymbol{\alpha}(s\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{Ra}) + \mathbf{I})^{-1}\mathbf{e}. \tag{7}$$

Now, let $X_j \sim \mathrm{BPH}^*(\boldsymbol{\alpha}, \mathbf{T}, \boldsymbol{\Delta}(\mathbf{r}_j))$, where the m-dimensional column vectors $\mathbf{r}_j$ are the rewards associated with the variable X_j. Note that, now the rewards can be negative. Then, for $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_k)$, we say that $\mathbf{X} = (X_1, \ldots, X_k)$ is multivariate bilateral phase–type (denoted by MBPH*) distributed with representation $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$.

Theorem 4.1. *$\boldsymbol{X} = (X_1, \ldots, X_k) \sim$ MBPH$^*(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{R})$ if and only if $\langle \boldsymbol{X}, \boldsymbol{a} \rangle \sim$ BPH$^*(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\Delta}(\boldsymbol{Ra}))$, for all k-dimensional real vector $\boldsymbol{a}$.*

Proof. From (6) we can get the MG function of $\langle \mathbf{X}, \mathbf{a} \rangle$, which turns out to be the same as (7).

On the other hand, since the MG function of $\mathbf{X}$ is given in (7) with $s = 1$ and $\mathbf{a} = \mathbf{s}$, then we get the result. □

For more details of the proof see [8].

Partial differential equations. A computational technique for the distributions in the MBPH* class is using partial differential equations. Let $\mathbf{X} = (X_1, \ldots, X_k)$ be in MBPH* with representation $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$. For $1 \leq i \leq m$, define

$$\bar{F}_i(x_1, \ldots, x_k) = \mathbb{P}(X_1 > x_1, \ldots, X_k > x_k | J(0) = i),$$

then

$$\begin{aligned} \bar{F}(x_1, \ldots, x_k) &= \mathbb{P}(X_1 > x_1, \ldots, X_k > x_k) \\ &= \sum_{i=1}^{m} \alpha_i \bar{F}_i(x_1, \ldots, x_k). \end{aligned}$$

Theorem 4.2. *The functions $\bar{F}_i(x_1,\ldots,x_k)$, $1 \le i \le m$, satisfy the following system of simultaneous linear partial differential equations*

$$\sum_{j=1}^{k} r_j(i)\frac{\partial \bar{F}_i}{\partial x_j} = \sum_{l=1}^{m} t_{il}\bar{F}_l, \quad 1 \le i \le m.$$

See Kulkarni [11] for a proof.

In order to generalize the MBPH* class, we define the following.

Definition 4.2. *For $\boldsymbol{X} = (X_1,\ldots,X_k)$, let MVBME* be the class of distributions such that the MG function of $\boldsymbol{X}$ is given by*

$$M_{\boldsymbol{X}}(\boldsymbol{s}) = \alpha_{m+1} + \boldsymbol{\alpha}(\boldsymbol{T}^{-1}\boldsymbol{\Delta}(\boldsymbol{R}\boldsymbol{s}) + \boldsymbol{I})^{-1}\boldsymbol{e}, \tag{8}$$

then we say that the vector $\boldsymbol{X}$ is MVBME with representation $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{R})$.*

The partial differential equations are still valid when generalizing MBPH* to MVBME*. In addition, if $\mathbf{T}$ is such that $(\mathbf{e}_i', \mathbf{T}, -\mathbf{T}\mathbf{e})$, where $\mathbf{e}_i$ denotes the i-th column unit vector, is a distribution for all $1 \le i \le m$, then we also have a corresponding probabilistic interpretation.

The following theorem gives an explicit formula for calculating cross–moments of the components of a MVBME* distributed random variable. Bladt and Nielsen [8] have proved a similar result for a class which generalizes MPH* distributions.

Theorem 4.3. *The cross–moments $\mathbb{E}\left(\prod_{i=1}^{k} X_i^{a_i}\right)$, where $\boldsymbol{X} = (X_1,\ldots,X_k) \sim MVBME^*(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{R})$ and $a_i \in \mathbb{N}$, are given by*

$$\boldsymbol{\alpha}\sum_{l=1}^{a!}\left(\prod_{i=1}^{a}\boldsymbol{T}^{-1}\boldsymbol{\Delta}(\boldsymbol{R}_{\sigma_l(i)})\right)\boldsymbol{e},$$

where $a = \sum_{i=1}^{k} a_i$, $\boldsymbol{R}_i$ is the i-th column of $\boldsymbol{R}$, and $\sigma_1,\ldots,\sigma_{a!}$ are the ordered permutations of a-tuples of derivatives, within $\sigma_l(i)$ being the value among $1,\ldots,k$ at the i-th position of the permutation σ_l.

Proof. We can obtain the cross–moments by

$$\mathbb{E}\left(\prod_{i=1}^{k} X_i^{a_i}\right) = M^a_{\mathbf{X}}(\mathbf{0}) = \left.\frac{\mathrm{d}^a M_{\mathbf{X}}(\mathbf{s})}{\mathrm{d}s_1^{a_1}\mathrm{d}s_2^{a_2}\ldots\mathrm{d}s_k^{a_k}}\right|_{\mathbf{s}=\mathbf{0}},$$

where $M_{\mathbf{X}}(\mathbf{s})$ is given in (8).

Since

$$\tfrac{\mathrm{d}}{\mathrm{d}s_i}\left(\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{R}\mathbf{s}) + \mathbf{I}\right)^{-1} = \left(\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{R}\mathbf{s}) + \mathbf{I}\right)^{-1}\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{R}_i)\left(\mathbf{T}^{-1}\boldsymbol{\Delta}(\mathbf{R}\mathbf{s}) + \mathbf{I}\right)^{-1},$$

then by induction and substituting $\mathbf{s} = \mathbf{0}$, we get the result. □

For more details of the demonstration we refer the reader to Nielsen *et.al* [14].

One example of using bilateral phase–type distributions is considering the Ladder process, as Asmussen has shown in [2]. The same example but considering matrix–exponential distributions as ladder height distributions is given by Asmussen and Bladt [5].

5. Order of bilateral matrix–exponential distributions

The study of matrix–exponential distributions and their representations have been deeply studied in the last decades. This is due to their important role in applications in areas like queuing theory, insurance, stochastic modeling, among others. In this section we will extend this analysis considering BME distributions.

We know in advance that the *order* of the ME-representation $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{t})$ is given by the dimension of $\mathbf{T}$, and the smallest order from among the equivalent representations is called the *degree* (see Liefvoort [17]). Also, a representation whose order is equal to the degree is said to be of *minimal order*, and this is called the order of the distribution (He and Zhang [10]). Having these concepts, we will translate them directly to BME-representations, where the *order* of the representation $(\boldsymbol{\alpha}_+, \mathbf{T}_+, \mathbf{t}_+, \boldsymbol{\alpha}_-, \mathbf{T}_-, \mathbf{t}_-)$ is given by the dimension of the matrix $\mathbf{T}_+$ plus the dimension of the matrix $\mathbf{T}_-$.

Asmussen and Bladt [5] identified some necessary and sufficient conditions for an ME representation to be minimal and developed a method for computing a minimal ME representation from an ME distribution. The ME order can play an important role in finding minimal representations, He and Zhang [10] introduced certain Hankel matrices that can be used to compute the ME order of ME distributions.

In this section, we will establish a relationship between the MG function and the minimal BME representation of BME distributions using Hankel matrices.

For $j \geq 0$ the non-centralized moments of $X \sim \text{BME}(\boldsymbol{\alpha}_+, \mathbf{T}_+, \mathbf{t}_+, \boldsymbol{\alpha}_-, \mathbf{T}_-, \mathbf{t}_-)$ are given by

$$\begin{aligned} M_j &= \mathbb{E}(x^j) \\ &= M_j^+ + M_j^-, \end{aligned}$$

where $M_j^+ = j!\boldsymbol{\alpha}_+(-\mathbf{T}_+)^{-(j+1)}\mathbf{t}_+$ and $M_j^- = (-1)^j j!\boldsymbol{\alpha}_-(-\mathbf{T}_-)^{-(j+1)}\mathbf{t}_-$. The reduced moments are given by

$$\mu_j = \frac{M_j}{j!} = \frac{M_j^+}{j!} + \frac{M_j^-}{j!} =: \mu_j^+ + \mu_j^-, \tag{9}$$

where $\mu_j^+ > 0$, for all j, and $\mu_j^- > 0$ if j is even and $\mu_j^- < 0$ if j is odd.

Moreover, the MG function of X is rational and has a power series expansion of the form $M_X(s) = \sum_j \mu_j s^j$, where μ_j is the j-th reduced moment. Then by (9), we get $M_X(s) = \sum_j \mu_j^+ s^j + \sum_j \mu_j^- s^j$.

Let m the minimal order of the distribution, then its MG function can be written as follows

$$M_X(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + 1}{a_m s^m + a_{m-1}s^{m-1} + \cdots + a_1 s + 1} = \frac{B(s)}{A(s)},$$

that is well defined in a strip containing the imaginary axis.

Since $\mu_0 = 1$, we get that

$$\frac{B(s)}{A(s)} = 1 + \sum_{j=1}^{\infty} \mu_j s^j, \tag{10}$$

multiplying (10) by $A(s)$ and equating coefficients, we get that the equations corresponding to powers $m+1, \ldots, 2m$ of s satisfy the following system

$$-\boldsymbol{\mu}_m = \boldsymbol{H}_m \mathbf{a}_m,$$

where $\boldsymbol{\mu}_m = (\mu_{m+1}, \ldots, \mu_{2m})'$, $\mathbf{a}_m = (a_m, \ldots, a_1)'$, and $\boldsymbol{H}_m$ is the $(m \times m)$-dimensional Hankel matrix given by

$$\boldsymbol{H}_m = \begin{pmatrix} \mu_1 & \mu_2 & & \mu_m \\ \mu_2 & \mu_3 & & \mu_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_m & \mu_{m+1} & & \mu_{2m-1} \end{pmatrix}, \tag{11}$$

with Hankel determinant defined by

$$\phi_m = \det(\boldsymbol{H}_m). \tag{12}$$

The equation system must have a unique solution due to the uniqueness of the MG function and the assumption of minimality, then $\boldsymbol{H}_m$ must have full rank, i.e. $\phi_m \neq 0$. On the other hand, considering the equations corresponding to powers $m+1, \ldots, 2m+1$ of s, we get that

$$\mathbf{0} = \boldsymbol{H}_{m+1} \mathbf{a}_m^*$$

where $\mathbf{0}$ is the $(m+1)$-dimensional column vector of zeros and $\mathbf{a}_m^* = (a_m, \ldots, a_1, 1)'$. Note that $\boldsymbol{H}_{m+1}$ has rank m, since the determinant of the lower-left $m \times m$ submatrix is different from zero. Hence $\phi_{m+1} = 0$.

By continuation of the argument we see that $\operatorname{rank}(\boldsymbol{H}_l) = m$ for $l \geq m$, this means that $\phi_l = 0$ for $l > m$.

Thus the minimal order of the BME distribution can be checked through the verification of the determinants to be the highest index of the determinant for which it is different from zero. Note that some determinants ϕ_l for $l < m$ could be zero or non-zero. See also Liefvoort [17] and He and Zhang [10].

Example 5.1. *Suppose X is a random variable with density given by*

$$f_X(x) = pe^{-x}\mathbf{1}_{\{x>0\}} + (1-p)e^{x}\mathbf{1}_{\{x<0\}}, \quad p \in (0,1).$$

With the notation presented before, we have that $m_+ = 1$ and $m_- = 1$. The MG function is given by

$$M_X(s) = \frac{(1-2p)s - 1}{s^2 - 1},$$

and the Hankel determinants are given by

$$\phi_1 = 2p - 1, \quad \phi_2 = 4p^2 - 4p, \quad \phi_l = 0, \text{ for } l > 2.$$

Example 5.2. *Suppose X has the following density*

$$f_X(x) = p\left(\frac{2}{3}e^{-x}(1+\cos(x))\right)\mathbf{1}_{\{x>0\}} + (1-p)e^{x}\mathbf{1}_{\{x<0\}},$$

with $p \in (0,1)$.

Then we have that $m_+ = 3$ and $m_- = 1$. The MG function of X is given by

$$M_X(s) = \frac{1}{3}\frac{(-7p+3)s^3 + (13p-9)s^2 + (-10p+12)s - 6}{s^4 - 2s^3 + s^2 + 2s - 2},$$

and the Hankel determinants are given by

$$\begin{aligned}
\phi_1 &= (5/3)p - 1 \\
\phi_2 &= (9/4)p^2 - (13/6)p \\
\phi_3 &= (307/432)p^3 - (103/144)p^2 \\
\phi_4 &= -(25/216)p^4 + (25/216)p^3 \\
\phi_l &= 0, \text{ for } l > 4.
\end{aligned}$$

6. Multivariate bilateral matrix–exponential distributions

We will define the class of multivariate bilateral matrix–exponential distributions as a natural extension of the univariate case.

Definition 6.1. *A random vector $\boldsymbol{X} \in \mathbb{R}^k$ of dimension k is multivariate bilateral matrix–exponential (MVBME) distributed if the joint moment–generating function $\mathbb{E}(e^{\langle \boldsymbol{X}, \boldsymbol{s}\rangle})$, $\boldsymbol{s} \in \mathbb{R}^k$, is a multidimensional rational function.*

Let $\mathbf{X} \in \mathbb{R}^k$. In order to prove our main characterization we proceed by deriving the following two lemmas.

Lemma 6.1. *Assume that $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ has a BME distribution for all $\boldsymbol{a} \in \mathbb{R}^k \setminus \boldsymbol{0}$. Then the (minimal) order $m(\boldsymbol{a})$ of the univariate BME distribution for $\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ is a bounded function of $\boldsymbol{a}$.*

Proof. Let $\phi_i(\mathbf{a})$ denote the ith-order Hankel determinant (see (12)) corresponding to $\langle \mathbf{X}, \mathbf{a}\rangle$, and let $C_i = \{\mathbf{a} \in \mathbb{R}^k \setminus \mathbf{0} : \phi_j(\mathbf{a}) = 0, j \geq i\}$. For $\mathbf{a}_1 \in \mathbb{R}^k \setminus \mathbf{0}$ we let $m_1 = m(\mathbf{a}_1)$, then $\phi_i(\mathbf{a}) = 0$ for $i > m_1$.

The ith-order Hankel determinant is a sum of monomials of order $i(i+1)$ and hence a continuous function. Thus there exists a neighborhood B around $\mathbf{a}_1$ for which $\phi_m(\mathbf{b}) \neq 0$ and $\mathbf{b} \in B$. Hence the order of the BME distribution of $\langle \mathbf{X}, \mathbf{b}\rangle$ is at least the order of $\langle \mathbf{X}, \mathbf{a}\rangle$ for $\mathbf{b} \in B$.

Since ϕ_{m_1} is a non-vanishing k-dimensional polynomial, then C_{m_1} has k-dimensional Lebesgue measure zero. Suppose there exists $\mathbf{a}_2 \in \mathbb{R}^k \setminus \mathbf{0}$ such that $m_2 = m(\mathbf{a}_2) > m_1$, then $\mathbf{a}_1 \in C_{m_2}$, and $C_{m_1} \subseteq C_{m_2}$.

If the order of the MG function for $\langle \mathbf{X}, \mathbf{a} \rangle$ is unbounded, then there exists a sequence $\mathbf{a}_i$ with $m_i = m(\mathbf{a}_i)$ such that $m_i \uparrow \infty$, and the set $C = \cup_{i=1}^{\infty} C_{m_i}$ has k-dimensional Lebesgue measure zero, contradicting the assumption of $\langle \mathbf{X}, \mathbf{a} \rangle$ being BME distributed (of finite order). □

The next lemma shows that the rational MG function is of a particularly simple form.

Lemma 6.2. *Assume that $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ has a univariate bilateral matrix–exponential distribution for all $\boldsymbol{a} \in \mathbb{R}^k \setminus \boldsymbol{0}$, and suppose the order of the distribution of $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ is bounded by some m. Then, we may write the MG function of $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ as*

$$\frac{\tilde{b}_m(\boldsymbol{a})s^m + \tilde{b}_{m-1}(\boldsymbol{a})s^{m-1} + \cdots + \tilde{b}_1(\boldsymbol{a})s + 1}{\tilde{a}_m(\boldsymbol{a})s^m + \tilde{a}_{m-1}(\boldsymbol{a})s^{m-1} + \cdots + \tilde{a}_1(\boldsymbol{a})s + 1},$$

where the terms $\tilde{b}_j(\boldsymbol{a})$ and $\tilde{a}_j(\boldsymbol{a})$ are sums of k-dimensional monomials in $\boldsymbol{a}$ of degree j.

Proof. Since $\langle \mathbf{X}, \mathbf{a} \rangle \sim$ BME its MG function can be written as

$$\frac{\tilde{b}_m(\mathbf{a})s^m + \tilde{b}_{m-1}(\mathbf{a})s^{m-1} + \cdots + \tilde{b}_1(\mathbf{a})s + 1}{\tilde{a}_m(\mathbf{a})s^m + \tilde{a}_{m-1}(\mathbf{a})s^{m-1} + \cdots + \tilde{a}_1(\mathbf{a})s + 1}, \tag{13}$$

where $\tilde{b}_i(\mathbf{a})$ and $\tilde{a}_i(\mathbf{a})$ ($\tilde{a}_m(\mathbf{a}) \neq 0$) are functions in $\mathbf{a}$.

Let $\tilde{a}_i(\mathbf{a}) = P_i(\mathbf{a}) + E_i(\mathbf{a})$, where $P_i(\mathbf{a})$ is a sum of all, if any, i-th order monomials appearing in the expression for $\tilde{a}_i(\mathbf{a})$, while $E_i(\mathbf{a}) = \tilde{a}_i(\mathbf{a}) - P_i(\mathbf{a})$.

Let $\boldsymbol{\mu}_m(\mathbf{a}) = (\mu_{m+1}(\mathbf{a}), \ldots, \mu_{2m}(\mathbf{a}))'$ and $\boldsymbol{H}_m(\mathbf{a})$ the Hankel matrix (11) now depending on $\mathbf{a}$.

Now, since

$$\frac{\tilde{b}_m(\mathbf{a})s^m + \tilde{b}_{m-1}(\mathbf{a})s^{m-1} + \cdots + \tilde{b}_1(\mathbf{a})s + 1}{\tilde{a}_m(\mathbf{a})s^m + \tilde{a}_{m-1}(\mathbf{a})s^{m-1} + \cdots + \tilde{a}_1(\mathbf{a})s + 1} = 1 + \sum_{j=1}^{\infty} \mu_j(\mathbf{a})s^j, \tag{14}$$

we get the following system of equations

$$-\boldsymbol{\mu}_m(\mathbf{a}) = \boldsymbol{H}_m(\mathbf{a})\mathbf{P}_m(\mathbf{a}) + \boldsymbol{H}_m(\mathbf{a})\mathbf{E}_m(\mathbf{a}),$$

where $\mathbf{P}_m(\mathbf{a}) = (P_m(\mathbf{a}), \ldots, P_1(\mathbf{a}))'$ and $\mathbf{E}_m(\mathbf{a}) = (E_m(\mathbf{a}),$
$\ldots, E_1(\mathbf{a}))'$.

For $1 \leq j \leq m$, $\mu_{m+j}(\mathbf{a})$ is a sum of monomials of order $m + j$ as the corresponding terms of $\boldsymbol{H}_m(\mathbf{a})\mathbf{P}_m(\mathbf{a})$. Note that we can re-write $E_j(\mathbf{a})$ as $E_{>j}(\mathbf{a}) + E^j_{\text{irra}}(\mathbf{a}) + E^j_{\text{rat}}(\mathbf{a})$, where $E_{>j}$ represents the sum of monomials with order greater than j, and E^j_{irra} (respectively E^j_{rat}) is the sum of irrational (rational) monomials in the j-th equation. Then, we get that $\mathbf{E}_m(\mathbf{a}) = \mathbf{E}_{>m}(\mathbf{a}) + \mathbf{E}_{\text{irra}}(\mathbf{a}) + \mathbf{E}_{\text{rat}}(\mathbf{a})$.

It is easy to see that $\boldsymbol{H}_m(\mathbf{a})\mathbf{E}_m(\mathbf{a})$ does not contain monomials of order $m+j$, since:

- $\boldsymbol{H}_m(\mathbf{a})\mathbf{E}_{>m}(\mathbf{a})$ has monomials of order greater than $m+j$,
- $\boldsymbol{H}_m(\mathbf{a})\mathbf{E}_{\text{irra}}(\mathbf{a})$ has irrational monomials,

and for the rational case, i.e. $\boldsymbol{H}_m(\mathbf{a})\mathbf{E}_{\text{rat}}(\mathbf{a})$ we refer the reader to [8] to see a proof that does not have monomials of order $m+j$. Then, by coefficient matching we get that

$$\boldsymbol{H}_m(\mathbf{a})\mathbf{E}_m(\mathbf{a}) = \mathbf{0}.$$

This implies that $\mathbf{E}_m(\mathbf{a}) = \mathbf{0}$, since $\boldsymbol{H}_m(\mathbf{a})$ is non-singular. Hence all $\tilde{a}_i(\mathbf{a})$ are sums of monomials of order i. From (14) we can also see that $\tilde{b}_i(\mathbf{a})$ are sums of monomials of order i. □

Our theorem which characterizes the class of MVBME distributions is the following.

Theorem 6.1. *A vector $\boldsymbol{X}$ follows a multivariate bilateral matrix–exponential distribution, i.e. $\boldsymbol{X} \sim MVBME$, if and only if $\langle \boldsymbol{X}, \boldsymbol{a}\rangle \sim BME$ for all $\boldsymbol{a} \in \mathbb{R}^k \setminus \boldsymbol{0}$.*

Proof. Let $\mathbf{X} \sim MVBME$, then $\mathbb{E}(e^{\langle \mathbf{X}, s\mathbf{a}\rangle})$ is rational in $s\mathbf{a}$ for $s \in \mathbb{R}$ and $\mathbf{a} \in \mathbb{R}^k \setminus \mathbf{0}$. Since

$$\mathbb{E}(e^{\langle \mathbf{X}, s\mathbf{a}\rangle}) = \mathbb{E}(e^{s\langle \mathbf{X}, \mathbf{a}\rangle}),$$

then $\mathbb{E}(e^{s\langle \mathbf{X}, \mathbf{a}\rangle})$ is rational in s, i.e. $\langle \mathbf{X}, \mathbf{a}\rangle \sim BME$.

On the other hand, suppose that $\langle \mathbf{X}, \mathbf{a}\rangle$ has rational MG function, for all $\mathbf{a} \in \mathbb{R}^k \setminus \mathbf{0}$. Then we know that the MG function can be expressed in the form of Lemma 6.2. By setting $s = 1$ this rational function coincide with the multidimensional MG function of $\mathbf{X}$ at $\mathbf{a}$. □

Example 6.1. *Wishart distribution.*

The Wishart distribution was formulated by John Wishart in 1928, [18]. Let $\mathbf{X}_1 = (x_{i1})_{1\le i\le p}$, $\mathbf{X}_2 = (x_{i2})_{1\le i\le p}$, …, $\mathbf{X}_\nu = (x_{i\nu})_{1\le i\le p}$ be p-dimensional random column vectors distributed independently according to the p-dimensional Normal distributions $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \ldots, \mathcal{N}_p(\boldsymbol{\mu}_\nu, \boldsymbol{\Sigma})$ with mean vectors $\boldsymbol{\mu}_1 = (\mu_{i1})_{1\le i\le p}$, …, $\boldsymbol{\mu}_\nu = (\mu_{i\nu})_{1\le i\le p}$ (respectively), and a common variance–covariance matrix $\boldsymbol{\Sigma}$. The distribution of a $(p \times p)$ symmetric random matrix $\boldsymbol{W} = (w_{ij})_{1\le i,j\le p}$ defined by $w_{ij} = \sum_{t=1}^{\nu} x_{it}x_{jt}$ is the real non-central Wishart distribution $W_p(\nu, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = (\lambda_{ij})_{1\le i,j\le p}$ is the mean square matrix defined by $\lambda_{ij} = \sum_{t=1}^{\nu} \mu_{it}\mu_{jt}$. The Wishart distribution for $\boldsymbol{\Lambda} = \mathbf{0}$ is said to be central and is denoted by $W_p(\nu, \boldsymbol{\Sigma})$.

The moment–generating function of the central Wishart distribution (Numata and Kuriki [15]) is given by

$$M_{\boldsymbol{W}}(\boldsymbol{\Theta}) = \mathbb{E}[e^{\text{tr}(\boldsymbol{\Theta}\boldsymbol{W})}] = \det(\ \mathbf{I} - 2\boldsymbol{\Theta}\boldsymbol{\Sigma})^{-\frac{\nu}{2}}, \tag{15}$$

where $\mathbf{\Theta} = (\theta_{ij})_{1\leq i,j\leq p}$ is a symmetric parameter matrix, and $\mathrm{tr}(\cdot)$ is the trace of a matrix.

If we define the following vectors in $\mathbb{R}^{p^2}$

$$\begin{aligned} \boldsymbol{s} &= ((\theta_{i1})_{1\leq i\leq p}, (\theta_{i2})_{1\leq i\leq p}, \ldots, (\theta_{ip})_{1\leq i\leq p}), \\ \mathbf{X} &= ((w_{i1})_{1\leq i\leq p}, (w_{i2})_{1\leq i\leq p}, \ldots, (w_{ip})_{1\leq i\leq p}), \end{aligned}$$

then $\mathbb{E}(e^{\langle \mathbf{s},\mathbf{X}\rangle})$ is given by (15), which is a rational function whenever ν is an even integer number. This means that $\mathbf{X} \sim$ MVBME.

7. Markov additive processes with absorption

Let $\mathbf{Y} = (Y_1, \ldots, Y_m) \sim \text{MVBME}^*(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{I})$, where $\mathbf{T}$ is of dimension m. When the distribution is in the MBPH* class, then it can be interpreted as the joint distribution of the sojourn times in each of the transient phases before absorption.

Now we consider a multidimensional reward structure $\mathbf{X} = (X_1, \ldots, X_k)$ such that

$$X_j = \sum_{i=1}^{m} B_{ij}, \quad j = 1, \ldots, k$$

where $\mathbf{B}_i = (B_{i1}, \ldots, B_{ik}) \sim \mathcal{N}_k(Y_i\mathbf{r}(i), Y_i\mathbf{\Sigma}(i))$, with $\mathbf{\Sigma}(i) = \boldsymbol{\sigma}(i)\boldsymbol{\sigma}(i)'$ for some $\boldsymbol{\sigma}(i)$, $i = 1, \ldots, m$.

The joint moment–generating function of $\mathbf{X}$ is given by

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{s}) &= \mathbb{E}(e^{\langle \mathbf{s},\mathbf{X}\rangle}) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{m} \exp\left(y_i \mathbf{s}\mathbf{r}(i)' + y_i \frac{1}{2}\mathbf{s}\mathbf{\Sigma}(i)\mathbf{s}' \right) \mathrm{d}F(\boldsymbol{y}), \end{aligned}$$

which is the moment–generating function of $\mathbf{Y}$ evaluated in $\theta_i = \mathbf{s}\mathbf{r}(i)' + \frac{1}{2}\mathbf{s}\mathbf{\Sigma}(i)\mathbf{s}'$, i.e. (see (8))

$$M_{\mathbf{X}}(\mathbf{s}) = \boldsymbol{\alpha}\left(\mathbf{T}^{-1}\mathbf{\Delta}(\boldsymbol{\theta}) + \mathbf{I}\right)^{-1}\mathbf{e},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$.

Thus, the moment–generating function is obviously rational so $\mathbf{X}$ is MVBME distributed.

Note that all these arguments are also valid for the MBPH* class (see (7)), where the probabilistic interpretation is easier. In the following analysis we will present an application of this class considering Markov additive processes.

Analysis of terminal distributions with added multidimensional Brownian components. Let $J = \{J(t)\}_{t\geq 0}$ be an irreducible Markov (jump) process as given in Section 2, i.e. with finite state space $\{1, \ldots, m, m+1\}$, where J eventually get absorbed at state $\{m+1\}$. The infinitesimal generator matrix of J according to $E = \{1, \ldots, m\}$ and $\{m+1\}$ is given by (1).

Now, define the real-valued process $W = \{W(t)\}_{t\geq 0}$ as evolving a Brownian motion with parameters $r(i)$ (drift) and $\sigma^2(i)$ (variation) during intervals when the phase equals $i \in E$, such as

$$W(t) = \int_0^t r(J(s))\mathrm{d}s + \int_0^t \sigma(J(s))\mathrm{d}B(s), \tag{16}$$

where B is a standard Brownian motion. This is in fact known to be the most general Markov additive process on J with skipfree (continuous) paths ([4]). We assume that all states in E communicate and that the absorption time, say τ, is finite a.s. Then, W gets absorbed as well at $W(\tau)$, the terminal value.

The case of all σ^2 being equal to 0 corresponds to a fluid model, and when $\sigma^2(i) > 0$ for all i, we say that corresponds to the Brownian case. Asmussen [4] has proved that the class of terminal distributions for both cases fluid and Brownian, is a natural way to approach the class BPH*. Indeed, in Corollaries 1 and 3 of [4], we can find phase–type representations.

For the multivariate case, let us define $\mathbf{X} = (X_1, \ldots, X_k)$ where each X_n, $1 \leq n \leq k$, is given by

$$X_n = \int_0^\tau r_n(J(t))\mathrm{d}t + \int_0^\tau \sigma_n(J(t))\mathrm{d}B(t),$$

then we can write $\mathbf{X}$ as follows

$$\mathbf{X} = \int_0^\tau \mathbf{r}(J(t))\mathrm{d}t + \int_0^\tau \boldsymbol{\sigma}(J(t))\mathrm{d}\mathbf{B}(t), \tag{17}$$

where $\mathbf{B}$ is a k-dimensional standard Brownian motion. Thus, $\mathbf{X}$ evolves a k-dimensional Brownian motion with drift vector $\mathbf{r}(i)$ and diffusion matrix $\boldsymbol{\Sigma}(i) = \boldsymbol{\sigma}(i)\boldsymbol{\sigma}(i)'$.

In Section 4, we analyzed the phase–type representation of $\mathbf{X}$ for the fluid model, i.e. when $\boldsymbol{\Sigma}(i) = \mathbf{0}$ for all i. And when $\boldsymbol{\Sigma}(i)$ is positive definite (Brownian model), the phase–type representation was already given in this section.

8. Conclusion

In this article we have generalized the class of matrix–exponential distributions considering a general support. For this purpose we defined the new class called Bilateral ME distributions (distributions with rational moment–generating function). We also analyzed the multivariate case, which domain is the real space. Our main characterization of this is based on the one presented in [8] for multivariate ME distributions.

Moreover, we have analyzed and used the theory already written about bilateral phase–type distributions ([1]) in order to give a generalization of them for the multivariate case. Indeed, we have applied this into Markov additive processes. We believe that these distributions have high use in areas like statistics, finance, and computer science, where general reward rates may have advantages.

A more comprehensive multivariate analysis of this class is needed as well as the estimation of their parameters.

9. Acknowledgements

Luz Judith Rodriguez Esparza and Bo Friis Nielsen would like to thank the Villum Kann Rasmussen Foundation and The Danish Council for Strategic Research for their support through the MTlab a VKR Centre of Excellence and the UNITE project under grant no 2140-08-0011.

References

[1] S. Ahn and V. Ramaswami. Bilateral Phase-type distributions. *Stochastic Models*, 21:239–259, 2005.

[2] S. Asmussen. Phase-type representations in random walk and queueing problems. *The Annals of Probability*, 20:772–789, 1992.

[3] S. Asmussen. *Applied probability and queues*. Springer-Verlag, New York, 2003.

[4] S. Asmussen. Terminal distributions of skipfree Markov additive processes with absorption. Technical Report 14, MaPhySto, 2004.

[5] S. Asmussen and M. Bladt. Renewal theory and queueing algorithms for Matrix-Exponential distributions. *Matrix-analytic methods in stochastic models*, pages 313–341, 1997.

[6] D. Assaf, N. A. Langberg, T. H. Savits, and M. Shaked. Multivariate phase-type distributions. *Operations Research*, 32(3):688–702, 1984.

[7] M. Bladt and M. F. Neuts. Matrix-exponential distributions: calculus and interpretations via flows. *Stochastic Models*, 19:113–124, 2003.

[8] M. Bladt and B. F. Nielsen. Multivariate matrix-exponential distributions. *Stochastic Models*, 26:1–26, 2010.

[9] L. Bodrog, A. Hovarth, and M. Telek. Moment characterization of matrix-exponential and Markovian arrival processes. *Annals of Operations Research*, 160:51–68, 2008.

[10] Q. He and H. Zhang. On matrix-exponential distributions. *Advances in Applied Probability*, 39:271–292, 2007.

[11] V. G. Kulkarni. A new class of multivariate phase-type distributions. *Operations Research*, 37:151–158, 1989.

[12] M. F. Neuts. Probability distributions of phase-type. *In Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206, 1975.

[13] M. F. Neuts. *Matrix Geometric solutions in stochastic models*, volume 2. Jhons Hopkins University Press, Baltimore, Md., 1981.

[14] B. F. Nielsen, F. Nielson, and H. R. Nielson. Model checking multivariate state rewards. *IEEE Computer Society, QEST*, 17:7–16, 2010.

[15] Y. Numata and S. Kuriki. On formulas for moments of the Wishart distributions as weighted generating functions of matchings. *Discrete mathematics and Theoretical computer science*, pages 821–832, 2010.

[16] J. G. Shanthikumar. Bilateral phase type distributions. *Naval Res. Log. Quarterly*, 32:119–136, 1985.

[17] A. Van de Liefvoort. The moment problem for continuous distributions. Technical report, University of Missouri, 1990.

[18] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A:32–52, 1928.

Appendix A. Existence of B_+ and B_-

In the following, we will give an analysis of the existence of B_+ and B_- assuming that we do not have an atom at zero.

Suppose that the polynomial $A(s)$ can be written as $A(s) = \prod_{j=1}^{r}(s-\lambda_j)^{\nu_j}$, for some r such as $\sum_{j=1}^{r}\nu_j = \deg(A)$, and whose poles are given by λ_j. Then, for

$$A_k(s) = \prod_{j\neq k}(s-\lambda_j)^{\nu_j} = \frac{A(s)}{(s-\lambda_k)^{\nu_k}}, \quad k = 1,\ldots,r,$$

we get that

$$\frac{B(s)}{A(s)} = \sum_{j=1}^{r}\frac{C_j(s)}{(s-\lambda_j)^{\nu_j}}, \tag{18}$$

where the polynomial $C_j(s)$ is the Taylor polynomial of $\frac{B(s)}{A_j(s)}$ of order $\nu_j - 1$ at the point λ_j, i.e.

$$C_j(s) := \sum_{k=0}^{\nu_j-1}\frac{1}{k!}\left(\frac{B(s)}{A_j(s)}\right)^{k}\lambda_j(s-\lambda_j)^k.$$

Taylor's theorem (in the real or complex case) provides a proof of the existence and uniqueness of the partial fraction decomposition, and a characterization of the coefficients. If we define

$$A_+(s) := \prod_{j=1}^{r}(s-\lambda_j)^{\nu_j}\mathbf{1}_{\{\lambda_j>0\}}, \quad A_-(s) := \prod_{j=1}^{r}(s-\lambda_j)^{\nu_j}\mathbf{1}_{\{\lambda_j<0\}},$$

From (18) we get

$$\begin{aligned}\frac{B(s)}{A(s)} &= \sum_{j=1}^{r}\frac{C_j(s)}{(s-\lambda_j)^{\nu_j}}\mathbf{1}_{\{\lambda_j>0\}} + \sum_{j=1}^{r}\frac{C_j(s)}{(s-\lambda_j)^{\nu_j}}\mathbf{1}_{\{\lambda_j<0\}}\\ &= \frac{B_+(s)}{A_+(s)} + \frac{B_-(s)}{A_-(s)},\end{aligned}$$

where

$$\begin{aligned}B_+(s) &:= \sum_{j=1}^{r}C_j(s)\mathbf{1}_{\{\lambda_j>0\}}\prod_{k\neq j}^{r}(s-\lambda_k)^{\nu_k}\mathbf{1}_{\{\lambda_k>0\}},\\ B_-(s) &:= \sum_{j=1}^{r}C_j(s)\mathbf{1}_{\{\lambda_j<0\}}\prod_{k\neq j}^{r}(s-\lambda_k)^{\nu_k}\mathbf{1}_{\{\lambda_k<0\}}.\end{aligned}$$

Bibliography

[1] L. Ahlstrom, M. Olsson, and O. Nerman. A parametric estimation procedure for relapse time distributions. *Lifetime Data Analysis*, 5:113–132, 1999.

[2] S. Ahn and V. Ramaswami. Bilateral Phase-type distributions. *Stochastic Models*, 21:239–259, 2005.

[3] H. Akaike. A new look at the statistical model identification. *EEE Transactions on Automatic Control*, 19:716–723, 1974.

[4] A. S. Alfa. Matrix-geometric solution of discrete time MAP/PH/1 priority queue. *Naval Research Logistics*, 45:23–50, 1998.

[5] A. S. Alfa, B. Liu, and Q. M. He. Discrete-time analysis of MAP/PH/1 multiclass general preemptive priority queue. *Naval Research Logistics*, 50:662–682, 2003.

[6] A. T. Andersen, M. F. Neuts, and B. F. Nielsen. On the Time Reversal of Markovian Arrival Processes. *Stochastic Models*, 20(2):237–260, 2004.

[7] S. Asmussen. Phase-type representations in random walk and queueing problems. *The Annals of Probability*, 20:772–789, 1992.

[8] S. Asmussen. *Phase-type distributions and related point process: fitting and recent advances*, volume 183, pages 137–149. Matrix-analytic methods in stochastic models, 1997.

[9] S. Asmussen. *Ruin Probability*, volume 2 of *Advance Series on Statistical Science & Applied probability*. World Scientific Publishing Co. Inc., N. J, 2000.

[10] S. Asmussen and M. Bladt. Renewal theory and queueing algorithms for Matrix-Exponential distributions. *Matrix-analytic methods in stochastic models*, pages 313–341, 1997.

[11] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorihtm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[12] D. Assaf, N. A. Langberg, T. H. Savits, and M. Shaked. Multivariate phase-type distributions. *Operations Research*, 32(3):688–702, 1984.

[13] N. G. Bean and B. F. Nielsen. Quasi-Birth-and-Death Process with Rational Arrival Process Components. *Stochastic Models*, 26:309–334, 2010.

[14] M. Bladt. A review on phase-type distributions and their use in risk theory. *International Actuarial Association*, 35:145–161, 2005.

[15] M. Bladt, A. Gonzalez, and S. Lauritzen. The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavia Actuarial*, 4:280–300, 2003.

[16] M. Bladt and M. F. Neuts. Matrix-exponential distributions: calculus and interpretations via flows. *Stochastic Models*, 19:113–124, 2003.

[17] M. Bladt and B. F. Nielsen. Multivariate matrix-exponential distributions. *Stochastic Models*, 26:1–26, 2010.

[18] M. Bladt and B. F. Nielsen. On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models*, 26:295–308, 2010.

[19] A. Bobbio and A. Cumani. ML estimation of the parameters of a PH distributions in triangular canonical form. *Computer performance evaluations*, pages 33–46, 1992.

[20] A. Bobbio, A. Cumani, A. Premoli, and O. Saracco. Modelling and identification of non-exponential distributions by homogeneous Markov process. *Proceedings of the 6th Advances in Reliability Technology Symposium*, pages 373–392, 1980.

[21] A. Bobbio, A. Horvath, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions: properties and a parameter estimation algorithm. *Performance evaluation An international Journal*, 54:1–32, 2003.

[22] A. Bobbio and M. Telek. A benchmark of Ph estimation algorithms: results for acyclic-PH. *Stochastic models*, 10:661–677, 1994.

[23] G. E. Box and D. R. Cox. An analysis of transformations. *Royal Statistical Society*, 26:211–252, 1964.

[24] J. Callut and P. Dupont. Sequence Discrimination using Phase-type Distributions. *Springer-Verlag Berlin Heidelberg*, pages 78–89, 2006.

[25] G. Casella and E. I. George. Explaining the Gibbs Sampler. *The American Stat.*, 46:167–174, 1992.

[26] D. R. Cox and H. D. Miller. *The theory of Stochastic Processes.* Chapman & Hall, paperback edition, 1977.

[27] A. Cumani. On the canonical representation of homogeneous Markov Process modelling failure-time distributions. *Microelectronics and Reliability*, 22:583–602, 1982.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[29] A. Erlang. Sandsynlighedsregning og telefonsamtaler. *Nyt tidsskrift for Matematik*, 20:33–39, 1909.

[30] P. Fearnhead and C. Sherlock. An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:767–784, 2006.

[31] A. E. Gelfand and A. F. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *J. American Stat. Assoc.*, 85:398–409, 1990.

[32] S. German and D. German. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[33] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–109, 1970.

[34] Q. He and H. Zhang. On matrix-exponential distributions. *Advances in Applied Probability*, 39:271–292, 2007.

[35] A. Hovarth and M. Telek. PhFit: A general phase-type fitting tool. *Proceedings International Conference on Dependable Systems and Networks*, page 543, 2002.

[36] A. Jensen. *A distribution model applicable to economics.* Munksgaard, Copenhagen, 1953.

[37] S. Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions.* Jhon Weley & Sons, 2000.

[38] A. Krishnamoorthy, S. Babu, and C. Narayanan. The MAP/(PH/PH)/1 queue with self-generation of priorities and non-preemtive service. *European Journal of Operational Research*, 195:174–185, 2009.

[39] V. G. Kulkarni. A new class of multivariate phase-type distributions. *Operations Research*, 37:151–158, 1989.

[40] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA, SIAM, 1999.

[41] K. Madsen, H. Nielsen, and J. Sondergaard. Robust subroutines for non-linear optimization. Technical Report IMM-REP-2002-02, Technical University of Denmark, 2002.

[42] R. S. Maier and C. A. O'Cinneide. A closure characterisation of phase-type distributions. *J. App. Probab.*, 29:92–103, 1992.

[43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[44] M. F. Neuts. Probability distributions of phase-type. *In Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206, 1975.

[45] M. F. Neuts. *Matrix Geometric solutions in stochastic models*, volume 2. Jhons Hopkins University Press, Baltimore, Md., 1981.

[46] M. F. Neuts. *Structured stochastic matrices of the M/G/1 type and their applications*, volume 5 of *Probability: pure and applied*. Marcel Dekker Inc, New York, 1989.

[47] M. F. Neuts. *Algorithmic probability*. Stochastic modeling series. Chapman & Hall, London, 1995.

[48] B. F. Nielsen and J. E. Beyer. Estimation of Interrupted Poisson Process Parameters from Counts. Technical report, Technical University of Denmark, 2005.

[49] B. F. Nielsen, F. Nielson, and H. R. Nielson. Model checking multivariate state rewards. *IEEE Computer Society, QEST*, 17:7–16, 2010.

[50] J. R. Norris. *Markov chains*. Cambridge University press, Cambridge, 1997.

[51] C. A. O' Cinneide. On non-uniqueness of representations of Phase-type distributions. *Stochastic models*, 5:247–259, 1989.

[52] D. Oakes. Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society*, 61:479–482, 1999.

[53] C. A. O'Cinneide. Characterization of phase-type distributions. *Comm. Statist. Stochastic Models*, 6:1–57, 1990.

[54] M. Olsson. Estimation of phase-type distributions from censored data. *Scand. J. Statist.*, 23:443–460, 1996.

[55] O. Perron. *Die lehre von den Kettenbruchen*. B. G. Teubner Verlagsgesellschaft, Stuttgart, 1957.

[56] V. Ramaswami. A duality theorem for the matrix paradigms in queueing theory. *Stochastic Models*, 6:151–161, 1990.

[57] S. M. Ross. *Stochastic processes*. John Wiley & Sons, Inc, University of California, Berkeley, second edition edition, 1996.

[58] E. Seneta. *Non-negative matrices and Markov chains*. Springer, New York, 2006.

[59] J. G. Shanthikumar. Bilateral phase type distributions. *Naval Res. Log. Quarterly*, 32:119–136, 1985.

[60] A. Thummler, P. Buchholz, and M. Telek. A novel approach for phase-type fitting with the EM algorithm. *IEEE Transactions on dependable and secure computing*, 3(3), 2006.

[61] A. Van de Liefvoort. The moment problem for continuous distributions. Technical report, University of Missouri, 1990.