

NEW FRONTIERS IN APPLIED PROBABILITY

A Festschrift for SØREN ASMUSSEN
Edited by P. GLYNN, T. MIKOSCH and T. ROLSKI

Part 6. Statistics

FISHER INFORMATION AND STATISTICAL INFERENCE FOR PHASE-TYPE DISTRIBUTIONS

MOGENS BLADT, *Universidad Nacional Autónoma de México*

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, A.P. 20-726. 01000 México DF, Mexico.

Email address: bladt@sigma.iimas.unam.mx

LUZ JUDITH R. ESPARZA, *Technical University of Denmark*

Department of Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kgs. Lyngby, Denmark.

BO FRIIS NIELSEN, *Technical University of Denmark*

Department of Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kgs. Lyngby, Denmark.



APPLIED PROBABILITY TRUST
AUGUST 2011

FISHER INFORMATION AND STATISTICAL INFERENCE FOR PHASE-TYPE DISTRIBUTIONS

BY MOGENS BLADT, LUZ JUDITH R. ESPARZA AND BO FRIIS NIELSEN

Abstract

This paper is concerned with statistical inference for both continuous and discrete phase-type distributions. We consider maximum likelihood estimation, where traditionally the expectation-maximization (EM) algorithm has been employed. Certain numerical aspects of this method are 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 distribution; Fisher information; EM algorithm; Newton–Raphson

2010 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 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 their 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 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 overparameterization.

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 expectation-maximization (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 will 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 overparameterization in general persists, we will only consider distributions which have a unique representation. Confidence regions for parameters in models which are overparameterized or nonunique 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 $\pi_i = P(X_0 = i)$ 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 $\boldsymbol{T} = \{t_{ij}\}_{i,j=1,\dots,p}$ denote the transition matrix (intensity matrix) restricted to the transient states. Finally, let $\boldsymbol{t} = (t_1, \dots, t_p)^\top$ be the vector of exit probabilities (exit rates). With \boldsymbol{e} being a p -dimensional column vector of 1s, we have $\boldsymbol{t} = \boldsymbol{e} - \boldsymbol{T}\boldsymbol{e}$ ($\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$). We say that $\tau = \inf\{t \in I \mid X_t = p + 1\}$ has a phase-type distribution with representation $(\boldsymbol{\pi}, \boldsymbol{T})$, and write $\tau \sim \text{PH}_p(\boldsymbol{\pi}, \boldsymbol{T})$. Throughout this paper, the acronyms DPH and CPH are used for the discrete and continuous phase-type cases, respectively.

Sometimes it is convenient to allow for an atom at 0 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 $f(x) = \boldsymbol{\pi}\boldsymbol{T}^{x-1}\boldsymbol{t}$, ($\boldsymbol{\pi}\boldsymbol{e}^{\boldsymbol{T}x}\boldsymbol{t}$), $x > 0$. We will initially assume that the phase-type distributions under consideration do not have an atom at 0.

3. Maximum likelihood estimation of phase-type distributions

Consider M independent observations y_1, \dots, y_M from a $\text{PH}_p(\boldsymbol{\pi}, \boldsymbol{T})$ distribution, where throughout the paper the order of the distribution p is assumed to be known. We may use the Akaike information criterion [1] for estimating p , but this matter will not be pursued here. Let $\boldsymbol{y} = (y_1, \dots, y_M)$. We observe only 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 θ denote a vector containing the parameters (π, T, t) . The incomplete data likelihood functions are given by

$$L(\theta; y) = \prod_{k=1}^M \pi T^{y_k-1} t \quad \text{for the DPH}, \quad L(\theta; y) = \prod_{k=1}^M \pi e^{T y_k} t \quad \text{for the CPH}.$$

The log-likelihood function is defined as $\ell(\theta; y) = \log L(\theta; y)$.

3.1. EM algorithm

One approach to maximizing the incomplete likelihood function is via the EM algorithm [5] for which we will need the full data or complete likelihood function, L_f . Let $\mathbf{x} = (x_1, \dots, x_M)$ 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, i.e.

$$L_f(\theta; \mathbf{x}) = \begin{cases} \prod_{i=1}^p \pi_i^{B_i} \prod_{i,j=1}^p t_{ij}^{N_{ij}} \prod_{i=1}^p t_i^{N_i} & \text{for the DPH,} \\ \prod_{i=1}^p \pi_i^{B_i} \prod_{\substack{i,j=1 \\ i \neq j}}^p t_{ij}^{N_{ij}} e^{-t_{ij} Z_i} \prod_{i=1}^p t_i^{N_i} e^{-t_i Z_i} & \text{for the CPH,} \end{cases}$$

where B_i is the number of Markov chains (Markov jump processes) initiating in state i , N_{ij} is the number of transitions from state i to state j , N_i is the number of chains (processes) jumping from state i to the absorbing state, and Z_i is the total time the processes spent in state i .

The full log-likelihood function ℓ_f is therefore given by

$$\ell_f(\theta; \mathbf{x}) = \begin{cases} \sum_{i=1}^p B_i \log(\pi_i) + \sum_{i,j=1}^p N_{ij} \log(t_{ij}) + \sum_{i=1}^p N_i \log(t_i) & \text{for the DPH,} \\ \sum_{i=1}^p B_i \log(\pi_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^p N_{ij} \log(t_{ij}) + \sum_{i=1}^p N_i \log(t_i) \\ - \sum_{\substack{i,j=1 \\ i \neq j}}^p t_{ij} Z_i - \sum_{i=1}^p t_i Z_i & \text{for the CPH.} \end{cases}$$

The full likelihood is easily maximized by applying, e.g. the method of Lagrange multipliers, attending the constraints. We find that the maximum likelihood estimators of π , T , and t are given by

$$\begin{aligned} \hat{\pi}_i &= \frac{B_i}{M}, & \hat{t}_{ij} &= \frac{N_{ij}}{J_i}, & \text{and } \hat{t}_i &= \frac{N_i}{J_i}, & \text{for the DPH,} \\ \hat{\pi}_i &= \frac{B_i}{M}, & \hat{t}_{ij} &= \frac{N_{ij}}{Z_i}, & \text{and } \hat{t}_i &= \frac{N_i}{Z_i}, & \text{for the CPH,} \end{aligned}$$

where J_i is the total number of jumps out of state i (DPH), and $\hat{t}_{ii} = 1 - \sum_{j \neq i} \hat{t}_{ij} - \hat{t}_i$ (DPH) and $\hat{t}_{ii} = - \sum_{j \neq i} \hat{t}_{ij} - \hat{t}_i$ (CPH).

The EM algorithm works as follows. Let $\theta_0 = (\pi_0, T_0, t_0)$ be (in principle) any choice of initial parameters.

1. Calculate $h: \theta \mapsto E_{\theta_0}(\ell_f(\theta; \mathbf{x}) \mid \mathbf{y})$.
2. Maximize h . Let $\hat{\theta} = (\hat{\pi}, \hat{T}, \hat{t})$ denote the point which maximizes h .
3. Set $\theta_0 = \hat{\theta}$ and go to step 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 point then follows by summing up the conditional expectations over all data points y_1, \dots, y_M .

First, we consider the discrete case (see also [3]). We note that $B_i = \mathbf{1}_{\{X_0=i\}}$ and, hence,

$$E(B_i \mid \tau = y) = P(X_0 = i \mid \tau = y) = \frac{P(\tau = y \mid X_0 = i) P(X_0 = i)}{P(\tau = y)} = \frac{\mathbf{e}_i^T \mathbf{T}^{y-1} \mathbf{t}}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}} \pi_i.$$

Here \mathbf{e}_i denotes a p -dimensional column vector with 1 in the i th entry and 0s elsewhere.

Concerning N_{ij} , if $\tau = y$ we have

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

Thus,

$$\begin{aligned} E(N_{ij} \mid \tau = y) &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} P(X_k = i, X_{k+1} = j \mid \tau = y) \\ &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \frac{P(\tau = y \mid X_{k+1} = j) P(X_{k+1} = j \mid X_k = i) P(X_k = i)}{P(\tau = y)} \\ &= \mathbf{1}_{\{y \geq 2\}} \sum_{k=0}^{y-2} \frac{\mathbf{e}_j^T \mathbf{T}^{y-(k+1)-1} \mathbf{t} \boldsymbol{\pi} \mathbf{T}^k \mathbf{e}_i}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}} t_{ij}. \end{aligned}$$

Similar calculations yield

$$E(N_i \mid \tau = y) = \frac{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{e}_i}{\boldsymbol{\pi} \mathbf{T}^{y-1} \mathbf{t}} t_i.$$

Finally, $E(J_i) = \sum_{j=1}^p E(N_{ij}) + E(N_i)$. The final EM algorithm for the discrete case then translates into the following.

0. Let $\theta_0 = (\pi_0, T_0, t_0)$.
1. Under θ_0 , calculate the three conditional expectations $E_{\theta_0}(B_i \mid \mathbf{y})$, $E_{\theta_0}(N_{ij} \mid \mathbf{y})$, and $E_{\theta_0}(N_i \mid \mathbf{y})$. Let $E(J_i \mid \mathbf{y}) = \sum_{j=1}^p E(N_{ij} \mid \mathbf{y}) + E(N_i \mid \mathbf{y})$.
2. Let $\hat{\pi}_i = E_{\theta_0}(B_i \mid \mathbf{y})/M$, $\hat{t}_{ij} = E_{\theta_0}(N_{ij} \mid \mathbf{y})/E_{\theta_0}(J_i \mid \mathbf{y})$, and $\hat{t}_i = E_{\theta_0}(N_i \mid \mathbf{y})/E_{\theta_0}(J_i \mid \mathbf{y})$.
3. Set $\theta_0 = (\pi_0, T_0, t_0) = (\hat{\pi}, \hat{T}, \hat{t})$ and go to step 1.

The EM algorithm for the CPH is similar, changing only 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, Asmussen *et al.* [2] used a discretization argument where they approximated the continuous process by the corresponding Markov chain formula derived above.

The corresponding formulae for the CPH (see [2]) are given by

$$\begin{aligned}
 E(B_i \mid \tau = y) &= \frac{e_i^\top e^{T y} \mathbf{t}}{\boldsymbol{\pi} e^{T y} \mathbf{t}} \pi_i, \\
 E(N_{ij} \mid \tau = y) &= \frac{\int_0^y \boldsymbol{\pi} e^{T u} e_i e_j^\top e^{T(y-u)} \mathbf{t} \, du}{\boldsymbol{\pi} e^{T y} \mathbf{t}} t_{ij}, \\
 E(N_i \mid \tau = y) &= \frac{\boldsymbol{\pi} e^{T y} e_i}{\boldsymbol{\pi} e^{T y} \mathbf{t}} t_i, \\
 E(Z_i \mid \tau = y) &= \frac{\int_0^y \boldsymbol{\pi} e^{T u} e_i e_i^\top e^{T(y-u)} \mathbf{t} \, du}{\boldsymbol{\pi} e^{T y} \mathbf{t}}.
 \end{aligned}$$

If 0 is contained in the data, we also need to include an atom of a certain size at 0 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. (i) Let $\hat{\pi}_{p+1}$ denote the proportion of 0s in the data set. (ii) Eliminate the 0s from the data. (iii) Fit a phase-type distribution $\text{PH}_p(\hat{\boldsymbol{\pi}}, \hat{\mathbf{T}})$ to the remaining data. This procedure, indeed, produces a maximum likelihood estimator for the full model which contains an atom at 0.

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 the condition that all elements of the transition matrix are equal. If some parameter t_{ij} is set to 0 initially then all subsequent values of t_{ij} through the iterations will remain 0. This makes it possible to estimate subclasses of general discrete phase-type distributions by adequately specifying 0s of certain transition probabilities from the beginning. If other subclasses or reparameterizations, such as, e.g. letting all remaining t_{ij} depend only on i , are to be considered then we need to intervene directly in 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 proposed using 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 $E(N_{ij} \mid \tau = y)$. The advantage of uniformization is the higher numerical precision. In most cases we found uniformization to be superior in terms of efficiency too; although, for a very high number of observations, our implementation was outperformed with respect to speed by the Runge–Kutta implementation of [2].

We now describe the EM algorithm for the CPH using uniformization. In standard uniformization (see [6]) we let $\mathbf{K} = (1/c)\mathbf{T} + \mathbf{I}$, where $c = \max\{-t_{ii} : 1 \leq i \leq p\}$ and \mathbf{I} is the

identity matrix of appropriate dimension $p \times p$. We have

$$e^{T y} = \sum_{r=0}^{\infty} e^{-c y} \frac{(c y)^r}{r!} \mathbf{K}^r.$$

Also, for $y \in \{y_1, \dots, y_M\}$, we have to evaluate the integral $\mathbf{J}(y) = \int_0^y e^{T(y-u)} \mathbf{t} \boldsymbol{\pi} e^{T u} du$ for which we will use uniformization. Here

$$\begin{aligned} \mathbf{J}(y) &= \int_0^y \left(e^{-c(y-u)} \sum_{k=0}^{\infty} \frac{(c \mathbf{K}(y-u))^k}{k!} \right) \mathbf{t} \boldsymbol{\pi} \left(e^{-c u} \sum_{j=0}^{\infty} \frac{(c \mathbf{K} u)^j}{j!} \right) du \\ &= e^{-c y} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left(\int_0^y \frac{(c u)^j}{j!} \frac{(c(y-u))^k}{k!} du \right) \mathbf{K}^j \mathbf{t} \boldsymbol{\pi} \mathbf{K}^k \\ &= e^{-c y} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{(c y)^{j+k+1}}{j! k!} \frac{j! k!}{(j+k+1)!} \mathbf{K}^j \frac{1}{c} \mathbf{t} \boldsymbol{\pi} \mathbf{K}^k \\ &= e^{-c y} \sum_{s=0}^{\infty} \frac{(c y)^{s+1}}{(s+1)!} \mathbf{D}_{\mathbf{J}}(s), \end{aligned} \tag{1}$$

where $\mathbf{D}_{\mathbf{J}}(s) = \sum_{j=0}^s \mathbf{K}^j \mathbf{t} \boldsymbol{\pi} \mathbf{K}^{s-j} / c$, 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 interarrival distribution $\text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$ (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:

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

With this formula we can 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, owing 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

$$q_i = \frac{i \lambda^i}{\lambda i!} e^{-\lambda}, \quad i = 0, 1, 2, \dots,$$

or

$$q_i = \frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda}, \quad i = 1, 2, \dots,$$

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 $\mathbf{D}_{\mathbf{J}}(s)$ in the expression for $\mathbf{J}(y)$, we see that all row sums of $\mathbf{D}_{\mathbf{J}}(s)$ 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 will 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 Nielsen and Beyer [9], we want to work with an unconstrained system, and use a package for unconstrained optimization written by Madsen *et al.* [7]. Their program, as well as many other standard routines available for unconstrained optimization, finds 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 will refer to this method as the direct method (DM) since it does not use the underlying probabilistic structure.

The DM we employ assumes that the parameters are unbounded. This is obviously not the case for the phase-type intensities, so we consider a reparameterization τ of the parameters. We also need to provide the gradient at a given point of the transformed parameters,

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

Let $\theta = (\theta_1, \dots, \theta_{p^2+(p-1)})$. By the chain rule, this vector can be obtained as

$$\frac{\partial \ell(\theta; y)}{\partial \tau} = \frac{\partial \ell(\theta; y)}{\partial \theta} \frac{\partial \theta}{\partial \tau}, \tag{2}$$

where $\partial \ell(\theta; y)/\partial \theta$ is a $(p^2 + (p - 1))$ -dimensional row vector and $\partial \theta/\partial \tau$ is the Jacobian matrix. Taking the derivative of the log-likelihood function with respect to θ yields

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

where f is the density of the phase-type distribution parameterized by θ . Thus, the problem reduces to finding the derivative of f with respect to the original parameters. To do this, we introduce

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

By substituting $\pi = \sum_{j=1}^{p-1} \pi_j e_j^\top + (1 - \sum_{j=1}^{p-1} \pi_j) e_p^\top$, the density of the phase-type distribution evaluated at y is given by

$$f(y) = \sum_{j=1}^{p-1} \pi_j e_j^\top \Psi(y) t + \left(1 - \sum_{j=1}^{p-1} \pi_j \right) e_p^\top \Psi(y) t,$$

and its partial derivatives with respect to the original parameters are given by

$$\begin{aligned} \frac{\partial f(y)}{\partial \pi_m} &= \mathbf{e}_m^\top \Psi(y) \mathbf{t} - \mathbf{e}_p^\top \Psi(y) \mathbf{t}, \\ \frac{\partial f(y)}{\partial t_{mn}} &= \pi \frac{\partial \Psi(y)}{\partial t_{mn}} \mathbf{t}, \quad m \neq n, \\ \frac{\partial f(y)}{\partial t_m} &= \pi \Psi(y) \mathbf{e}_m + \pi \frac{\partial \Psi(y)}{\partial t_m} \mathbf{t}. \end{aligned}$$

In order to compute the partial derivatives of Ψ with respect to θ_m for $m \in \{1, \dots, p^2 + (p - 1)\}$, we will need the derivatives of T^r for $r \geq 1$, and the derivative of e^{T^y} . In general, we have

$$\frac{\partial T^r}{\partial \theta_m} = \sum_{k=0}^{r-1} T^k \frac{\partial T}{\partial \theta_m} T^{r-1-k}, \quad r \geq 1, \tag{3}$$

where $[\partial T / \partial t_{ij}]_{ij} = 1$, $[\partial T / \partial t_{ij}]_{ii} = -1$, and $[\partial T / \partial t_i]_{ii} = -1$.

Concerning the derivative of e^{T^y} , we will use a uniformization argument similar to (1). We obtain

$$\frac{\partial e^{T^y}}{\partial \theta_m} = e^{-cy} \sum_{s=0}^{\infty} \frac{(cy)^{s+1}}{(s+1)!} \mathbf{D}_m(s) + \frac{\partial c}{\partial \theta_m} y e^{T^y} (\mathbf{K} - \mathbf{I}), \tag{4}$$

where $\mathbf{D}_m(s) = \partial \mathbf{K}^{s+1} / \partial \theta_m$, which is calculated as in (3). Since $e^{T(x+y)} = e^{T^x} e^{T^y}$, we can obtain a recursive version of (4) given by

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

In order to deal with unconstrained parameters in the optimization, we propose the following transformation. For $m = 1, \dots, p^2 + (p - 1)$, let $-\infty < \tau_m < \infty$ be such that

$$\pi_i = \frac{\exp(\tau_i)}{1 + \sum_{s=1}^{p-1} \exp(\tau_s)}, \quad i = 1, \dots, p - 1, \quad \pi_p = \frac{1}{1 + \sum_{i=1}^{p-1} \exp(\tau_i)},$$

and, for $i, j = 1, \dots, p, i \neq j$,

$$\begin{aligned} t_{ij} &= \frac{\exp(\tau_{ip+(j-1)})}{1 + \sum_{s=1}^p \exp(\tau_{ip+(s-1)})} \quad \text{and} \quad t_i = \frac{\exp(\tau_{ip+(i-1)})}{1 + \sum_{s=1}^p \exp(\tau_{ip+(s-1)})} \quad \text{for the DPH,} \\ t_{ij} &= \exp(\tau_{ip+(j-1)}) \quad \text{and} \quad t_i = \exp(\tau_{ip+(i-1)}) \quad \text{for the CPH.} \end{aligned}$$

The elements in the diagonal of T are defined as $t_{ii} = 1 - \sum_{j=1, j \neq i}^p t_{ij} - t_i$ in the DPH, and as $t_{ii} = -\sum_{j=1, j \neq i}^p t_{ij} - t_i$ in the CPH. Note that 0s for π_i and t_{ij} are not a possibility in this reparameterization. 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 $i, j = 1, \dots, p - 1$, 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 $i, j = 1, \dots, p$ and $m = p, \dots, p^2 + (p - 1)$, the $(ip + (j - 1), m)$ th element of the matrix is given by $\partial t_{ij} / \partial \tau_m$ if $i \neq j$ and $\partial t_i / \partial \tau_m$ if $i = j$, where

$$\frac{\partial t_{ij}}{\partial \tau_m} = \begin{cases} t_{ij} \mathbf{1}_{\{m=ip+(j-1)\}} - t_{ij} \sum_{r=1}^p (t_i \mathbf{1}_{\{i=r\}} + t_{ir} \mathbf{1}_{\{i \neq r\}}) \mathbf{1}_{\{m=ip+(r-1)\}} & \text{for the DPH,} \\ t_{ij} \mathbf{1}_{\{m=ip+(j-1)\}} & \text{for the CPH,} \end{cases}$$

$$\frac{\partial t_i}{\partial \tau_m} = \begin{cases} t_i \mathbf{1}_{\{m=ip+(i-1)\}} - t_i \sum_{r=1}^p (t_i \mathbf{1}_{\{i=r\}} + t_{ir} \mathbf{1}_{\{i \neq r\}}) \mathbf{1}_{\{m=ip+(r-1)\}} & \text{for the DPH,} \\ t_i \mathbf{1}_{\{m=ip+(i-1)\}} & \text{for the CPH.} \end{cases}$$

4. Fisher information

Fisher information is a key concept in the theory of statistical inference and essentially describes the amount of information data provided 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 subclasses, such as generalized Erlang or hyperexponential distributions, where several intensities are assumed to be 0. The corresponding Fisher information is then calculated with the same formulae, but summing over indices where the parameters are different from 0. We present methods for calculating the Fisher information matrix for both the EM algorithm and the Newton–Raphson method.

Throughout, we will 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 in [10]. Considering L , the incomplete data likelihood which is maximized by the EM algorithm, the Fisher information matrix is given by

$$\frac{\partial^2 L(\theta; \mathbf{y})}{\partial \theta^2} = \left\{ \frac{\partial^2 Q(\hat{\theta} | \theta)}{\partial \hat{\theta}^2} + \frac{\partial^2 Q(\hat{\theta} | \theta)}{\partial \theta \partial \hat{\theta}} \right\}_{\hat{\theta}=\theta}, \tag{5}$$

where

$$Q(\hat{\theta} | \theta) = E_{\theta}(\ell_f(\hat{\theta}; \mathbf{x}) | \mathbf{y}). \tag{6}$$

Define

$$U_i = \sum_{l=1}^M \frac{e_i^{\top} \Psi(y_l) t}{f(y_l)}, \tag{7}$$

$$W_i = \sum_{l=1}^M \frac{\pi \Psi(y_l) e_i}{f(y_l)}, \tag{8}$$

$$V_{ij} = \begin{cases} \sum_{l=1}^M \mathbf{1}_{\{y_l \geq 2\}} \frac{1}{f(y_l)} \sum_{k=0}^{y_l-2} e_j^{\top} T^{y_l-k-2} t \pi T^k e_i & \text{for the DPH,} \\ \sum_{l=1}^M \frac{1}{f(y_l)} \int_0^{y_l} e_j^{\top} e^{T(y_l-u)} t \pi e^{Tu} e_i du & \text{for the CPH.} \end{cases} \tag{9}$$

Then (6) becomes

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

where

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

The elements of the Fisher information matrix (5) are given as follows. For $i, j = 1, \dots, p - 1$, the (i, j) th element is given by

$$\frac{\partial U_i}{\partial \pi_j} - \frac{\partial U_p}{\partial \pi_j};$$

for $m = 1, \dots, p - 1$ and $i, j = 1, \dots, p$, the $(ip - 1 + j, m)$ th element is given by

$$\frac{\partial U_m}{\partial t_{ij}} - \frac{\partial U_p}{\partial t_{ij}} \quad \text{if } i \neq j, \quad \frac{\partial U_m}{\partial t_i} - \frac{\partial U_p}{\partial t_i} \quad \text{if } i = j;$$

the $(m, ip - 1 + j)$ th element is given by

$$\frac{\partial V_{ij}}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m} \quad \text{if } i \neq j, \quad \frac{\partial W_i}{\partial \pi_m} - \frac{\partial V_{ii}}{\partial \pi_m} \quad \text{if } i = j;$$

and, finally, for $i, j, m, n = 1, \dots, p$, the $(ip - 1 + j, mp - 1 + n)$ th element is given by

$$\frac{\partial V_{ij}}{\partial t_{mn}} - \frac{\partial V_{ii}}{\partial t_{mn}} \quad \text{if } i \neq j, m \neq n, \quad \frac{\partial V_{ij}}{\partial t_m} - \frac{\partial V_{ii}}{\partial t_m} \quad \text{if } i \neq j, m = n, \\ \frac{\partial W_i}{\partial t_{mn}} - \frac{\partial V_{ii}}{\partial t_{mn}} \quad \text{if } i = j, m \neq n, \quad \frac{\partial W_i}{\partial t_m} - \frac{\partial V_{ii}}{\partial t_m} \quad \text{if } i = j, m = n.$$

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 DM, we take the second derivative of (2), which at the optimum gives

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

where $\partial^2 \ell(\theta; \mathbf{y}) / \partial \bar{\theta} \partial \theta$ is a square matrix of second-order partial derivatives. For this, we need the second derivatives of the density f with respect to the original parameters (see Appendix B).

For $m, n \in \{1, \dots, p^2 + (p - 1)\}$, and taking the second derivative of (3), we obtain

$$\frac{\partial^2 \mathbf{T}^r}{\partial \theta_n \partial \theta_m} = \sum_{k=0}^{r-1} \mathbf{T}^k \frac{\partial \mathbf{T}}{\partial \theta_m} \frac{\partial \mathbf{T}^{r-1-k}}{\partial \theta_n} + \frac{\partial \mathbf{T}^k}{\partial \theta_n} \frac{\partial \mathbf{T}}{\partial \theta_m} \mathbf{T}^{r-1-k}. \tag{10}$$

In the same way, from (4) we have

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

where $\partial^2 \mathbf{K}^r / \partial \theta_n \partial \theta_m$ can be calculated as in (10).

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

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

5. Simulation results

The phase-type representation of a given distribution is, in general, nonunique and nonminimal. Hence, we explore a subclass of phase-type distributions for which the representation is an acyclic graph (APH). 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

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

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 phase-type 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 hyperexponential, and has a phase-type 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

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

$$\mathbf{t} = \begin{pmatrix} \pi_1\lambda_1 + \pi_2\lambda_2 + \pi_3\lambda_3 \\ \frac{\pi_2\lambda_2(\lambda_1 - \lambda_2) + \pi_3\lambda_3(\lambda_1 - \lambda_3)}{\pi_2(\lambda_1 - \lambda_2) + \pi_3(\lambda_1 - \lambda_3)} \\ \lambda_3 \end{pmatrix}.$$

The method to obtain the canonical form is given in [4]. All estimation is performed using the canonical form.

After finding the maximum likelihood estimator, the Fisher information (FI) matrix was obtained considering only the nonzero parameters. As the inverse of the FI is the empirical variance–covariance matrix, we could obtain the standard deviation of the parameters (see Tables 1 and 2). The corresponding correlations are given in Tables 3 and 4.

TABLE 1: Maximum likelihood estimators (MLEs) and standard deviations (SDs) of the shifted negative binomial(3, 0.2), considering 10 000 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 (MLEs) and standard deviations (SDs) of the hyperexponential, considering 20 000 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).

Parameter	\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

TABLE 4: Correlations of the hyperexponential.

Parameter	\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

6. Concluding remarks

The paper by Asmussen *et al.* [2] provided the statistical framework for obtaining maximum likelihood estimates of continuous phase-type distributions using the EM algorithm. In this paper we have demonstrated how one can obtain uncertainty estimates of the parameters in cases where the phase-type distribution is not overparameterized. The development is done for discrete as well as for continuous phase-type 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 method 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 demonstrated 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 phase-type 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 using the EM algorithm

Let $R_i(u) = \pi \Psi(u)e_i$ and $Q_i(u) = e_i^\top \Psi(u)t$. Then, their derivatives are given by

$$\begin{aligned} \frac{\partial R_i(u)}{\partial \pi_m} &= e_m^\top \Psi(u)e_i - e_p^\top \Psi(u)e_i, & \frac{\partial Q_i(u)}{\partial \pi_m} &= 0, \\ \frac{\partial R_i(u)}{\partial t_{mn}} &= \pi \frac{\partial \Psi(u)}{\partial t_{mn}} e_i, \quad m \neq n, & \frac{\partial Q_i(u)}{\partial t_{mn}} &= e_i^\top \frac{\partial \Psi(u)}{\partial t_{mn}} t, \quad m \neq n, \\ \frac{\partial R_i(u)}{\partial t_m} &= \pi \frac{\partial \Psi(u)}{\partial t_m} e_i, & \frac{\partial Q_i(u)}{\partial t_m} &= e_i^\top \Psi(u)e_m + e_i^\top \frac{\partial \Psi(u)}{\partial t_m} t. \end{aligned}$$

Then U_i , W_i , and V_{ij} (see (7), (8), and (9)) become

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

Hence, for $n \in \{1, \dots, p^2 + (p - 1)\}$, the derivatives with respect to θ_n are given by

$$\begin{aligned} \frac{\partial U_i}{\partial \theta_n} &= \sum_{l=1}^M \frac{1}{f(y_l)^2} \left(f(y_l) \frac{\partial Q_i(y_l)}{\partial \theta_n} - Q_i(y_l) \frac{\partial f(y_l)}{\partial \theta_n} \right), \\ \frac{\partial W_i}{\partial \theta_n} &= \sum_{l=1}^M \frac{1}{f(y_l)^2} \left(f(y_l) \frac{\partial R_i(y_l)}{\partial \theta_n} - R_i(y_l) \frac{\partial f(y_l)}{\partial \theta_n} \right), \\ \frac{\partial V_{ij}}{\partial \theta_n} &= \begin{cases} \sum_{l=1}^M \mathbf{1}_{\{y_l \geq 2\}} \sum_{k=0}^{y_l-2} \frac{1}{f(y_l)^2} \left[f(y_l) \left(Q_j(y_l - k - 1) \frac{\partial R_i(k + 1)}{\partial \theta_n} \right. \right. \\ \qquad \qquad \qquad \left. \left. + \frac{\partial Q_j(y_l - k - 1)}{\partial \theta_n} R_i(k + 1) \right) \right. \\ \qquad \qquad \qquad \left. - \frac{\partial f(y_l)}{\partial \theta_n} Q_j(y_l - k - 1) R_i(k + 1) \right] & \text{for the DPH,} \\ \sum_{l=1}^M \frac{1}{f(y_l)^2} \left[f(y_l) \int_0^{y_l} \left(Q_j(y_l - u) \frac{\partial R_i(u)}{\partial \theta_n} \right. \right. \\ \qquad \qquad \qquad \left. \left. + \frac{\partial Q_j(y_l - u)}{\partial \theta_n} R_i(u) \right) du \right. \\ \qquad \qquad \qquad \left. - \frac{\partial f(y_l)}{\partial \theta_n} \int_0^{y_l} Q_j(y_l - u) R_i(u) du \right] & \text{for the CPH.} \end{cases} \end{aligned}$$

Concerning the computation of $\partial V_{ij}/\partial \theta_n$ for the CPH, we define the integrals

$$J_1(y; \mathbf{M}) = \int_0^y e^{T(y-u)} \mathbf{M} e^{Tu} du = e^{-cy} \sum_{s=0}^{\infty} \frac{(cy)^{s+1}}{(s+1)!} \mathbf{D}_{J_1}(s),$$

$$J_2(y; \theta_n, \mathbf{M}) = \int_0^y e^{T(y-u)} \mathbf{M} \frac{\partial e^{Tu}}{\partial \theta_n} du = e^{-cy} \sum_{s=0}^{\infty} \frac{(cy)^{s+2}}{(s+2)!} (\mathbf{D}_{J_2,1}(s, \theta_n) + \mathbf{D}_{J_2,2}(s, \theta_n)),$$

$$J_3(y; \theta_n, \mathbf{M}) = \int_0^y \frac{\partial e^{T(y-u)}}{\partial \theta_n} \mathbf{M} e^{Tu} du = e^{-cy} \sum_{s=0}^{\infty} \frac{(cy)^{s+2}}{(s+2)!} (\mathbf{D}_{J_3,1}(s, \theta_n) + \mathbf{D}_{J_3,2}(s, \theta_n)),$$

where \mathbf{M} is a $p \times p$ matrix and

$$\mathbf{D}_{J_1}(s) = \sum_{j=0}^s \mathbf{K}^j \frac{1}{c} \mathbf{M} \mathbf{K}^{s-j},$$

$$\mathbf{D}_{J_2,1}(s, \theta_n) = \sum_{j=0}^s \mathbf{K}^j \frac{1}{c} \mathbf{M} \frac{\partial \mathbf{K}^{s-j+1}}{\partial \theta_n},$$

$$\mathbf{D}_{J_2,2}(s, \theta_n) = \sum_{j=0}^s \mathbf{K}^j (s+1-j) \frac{1}{c^2} \frac{\partial c}{\partial \theta_n} \mathbf{M} (\mathbf{K} - \mathbf{I}) \mathbf{K}^{s-j},$$

$$\mathbf{D}_{J_3,1}(s, \theta_n) = \sum_{j=0}^s \frac{\partial \mathbf{K}^{s-j+1}}{\partial \theta_n} \frac{1}{c} \mathbf{M} \mathbf{K}^j,$$

$$\mathbf{D}_{J_3,2}(s, \theta_n) = \sum_{j=0}^s \mathbf{K}^j (j+1) \frac{1}{c^2} \frac{\partial c}{\partial \theta_n} (\mathbf{K} - \mathbf{I}) \mathbf{M} \mathbf{K}^{s-j}.$$

Then

$$\frac{\partial V_{ij}}{\partial \pi_m} = \sum_{k=1}^M \frac{1}{f(y_k)^2} \left[f(y_k) (\mathbf{e}_j^\top \mathbf{J}_1(y_k; \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i - \mathbf{e}_j^\top \mathbf{J}_1(y_k; \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i) - \frac{\partial f(y_k)}{\partial \pi_m} \mathbf{e}_j^\top \mathbf{J}_1(y_k; \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i \right],$$

$$\frac{\partial V_{ij}}{\partial t_{mn}} = \sum_{k=1}^M \frac{1}{f(y_k)^2} \left[f(y_k) (\mathbf{e}_j^\top \mathbf{J}_2(y_k; t_{mn}, \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j^\top \mathbf{J}_3(y_k; t_{mn}, \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i) - \frac{\partial f(y_k)}{\partial t_{mn}} \mathbf{e}_j^\top \mathbf{J}_1(y_k; \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i \right],$$

$$\frac{\partial V_{ij}}{\partial t_m} = \sum_{k=1}^M \frac{1}{f(y_k)^2} \left[f(y_k) (\mathbf{e}_j^\top \mathbf{J}_2(y_k; t_m, \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j^\top \mathbf{J}_1(y_k; e_m \boldsymbol{\pi}) \mathbf{e}_i + \mathbf{e}_j^\top \mathbf{J}_3(y_k; t_m, \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i) - \frac{\partial f(y_k)}{\partial t_m} \mathbf{e}_j^\top \mathbf{J}_1(y_k; \mathbf{t} \boldsymbol{\pi}) \mathbf{e}_i \right].$$

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 at the end of Section 3.1. The row sums of the matrix $\mathbf{D}_{J_1}(s)$

are like those for $D_J(s)$ bounded by $s + 1$, while the row sums of $D_{J_2,1}(s, \cdot)$, $D_{J_2,2}(s, \cdot)$, $D_{J_3,1}(s, \cdot)$, and $D_{J_3,2}(s, \cdot)$ 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

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

which represents the summation of the first- and second-order moment distributions 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 yields

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

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 T , the second derivatives are given by

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

and, with respect to the exit probabilities, they are given by

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

Finally,

$$\begin{aligned} \frac{\partial^2 f(y)}{\partial \pi_m \partial t_{ij}} &= \frac{\partial^2 f(y)}{\partial t_{ij} \partial \pi_m} = \mathbf{e}_m^\top \frac{\partial \Psi(y)}{\partial t_{ij}} \mathbf{t} - \mathbf{e}_p^\top \frac{\partial \Psi(y)}{\partial t_{ij}} \mathbf{t}, & i \neq j, \\ \frac{\partial^2 f(y)}{\partial \pi_m \partial t_i} &= \frac{\partial^2 f(y)}{\partial t_i \partial \pi_m} = \mathbf{e}_m^\top \Psi(y) \mathbf{e}_i - \mathbf{e}_p^\top \Psi(y) \mathbf{e}_i + \mathbf{e}_m^\top \frac{\partial \Psi(y)}{\partial t_i} \mathbf{t} - \mathbf{e}_p^\top \frac{\partial \Psi(y)}{\partial t_i} \mathbf{t}, \\ \frac{\partial^2 f(y)}{\partial t_{mn} \partial t_i} &= \pi \frac{\partial \Psi(y)}{\partial t_{mn}} \mathbf{e}_i + \pi \frac{\partial^2 \Psi(y)}{\partial t_{mn} \partial t_i} \mathbf{t}, & m \neq n, \\ \frac{\partial^2 f(y)}{\partial t_i \partial t_{mn}} &= \pi \frac{\partial \Psi(y)}{\partial t_{mn}} \mathbf{e}_i + \pi \frac{\partial^2 \Psi(y)}{\partial t_i \partial t_{mn}} \mathbf{t}, & m \neq n. \end{aligned}$$

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. *IEEE Trans. Automatic Control* **19**, 716–723.
- [2] ASMUSSEN, S., NERMAN, O. AND OLSSON, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Statist.* **23**, 419–441.
- [3] CALLUT, J. AND DUPONT, P. (2006). Sequence discrimination using phase-type distributions. In *Machine Learning: ECML 2006 (Lecture Notes Artificial Intelligence 4212)*, Springer, Berlin, pp. 78–89.
- [4] CUMANI, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectron. Reliab.* **22**, 583–602.
- [5] DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **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, Philadelphia, PA.
- [7] MADSEN, K., NIELSEN, H. AND SONDERGAARD, J. (2002). Robust subroutines for non-linear optimization. Tech. Rep. IMM-REP-2002-02, Technical University of Denmark.
- [8] NEUTS, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models* (Johns Hopkins Ser. Math. Sci. **2**). Johns Hopkins University Press, Baltimore, MD.
- [9] NIELSEN, B. F. AND BEYER, J. E. (2005). Estimation of interrupted Poisson process parameters from counts. Tech. Rep. IML-R- -21-04/05- -SE+fall, Institute Mittag–Leffler.
- [10] OAKES, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Statist. Soc. B* **61**, 479–482.

MOGENS BLADT, *Universidad Nacional Autónoma de México*

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, A.P. 20-726. 01000 México DF, Mexico. Email address: bladt@sigma.iimas.unam.mx

LUZ JUDITH R. ESPARZA, *Technical University of Denmark*

Department of Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kgs. Lyngby, Denmark.

BO FRIIS NIELSEN, *Technical University of Denmark*

Department of Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kgs. Lyngby, Denmark.