Acyclic Discrete Phase Type Distributions. Part 2: A Parameter Estimation Algorithm*

Andrea Bobbio[1], András Horváth[2], Miklós Telek[2†]

[1] Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale

[2] Department of Telecommunications, Technical University of Budapest

e-mail: bobbio@di.unito.it, {horvath, telek}@pyxis.hit.bme.hu

Abstract

In Part 1 of this paper, the properties of Discrete Phase Type (*DPH*) distributions have been investigated, which make the use of this class of distributions very convenient in the approximate analysis of discrete state stochastic systems. In particular, a minimal representation, called canonical form, has been derived for the subclass of the *DPH* distributions characterised by an acyclic transition graph. The present Part 2 provides an algorithm to estimate the parameters of an *ADPH* in canonical form given a continuous distribution function or a set of experimental samples. The algorithm is based on the maximum likelihood principle. In contrast with the previous fitting algorithms, that were based on expressions in the transform domain, the present paper proposes also a method based on time domain analysis. The implementation of the time domain algorithm proved to be simpler and more stable. The algorithm is then tested over a benchmark consisting of 10 different continuous distributions that have been already utilized in the past for a similar purpose.

Key words: Discrete Phase Type Distributions, Maximum Likelihood, Phase Type Fitting.

1 Introduction

In recent years, an increasing attention has been devoted to discrete-time models, due to several reasons: - *i)* the continuous time approximation of non-Markovian processes may provide poor results, particularly when the involved random variables have a low coefficient of variation; - *ii)* in many instances, there is a need of combining random times with constant durations in the same model (e.g. real time systems); - *iii)* discrete time models become more and more important in the modelling of high speed telecommunication networks, where packets are transmitted in slots; - *iv)* discrete models can be more closely related to physical observations [9, 10]; - *v)* greater emphasis has been put on discrete stochastic Petri Nets [5, 6, 11].

*This work has been performed under a cooperative visiting exchange program supported by the Italian Ministry of Foreign Affairs and the Hungarian Committee for Technologycal Development (OMFB).

†M. Telek was partially supported by Hungarian Scientific Research Fund (OTKA) under Grant No. F-23971.

Part 1 of this paper [3] was aimed at comparing *DPH* distributions with Continuous Phase Type (*CPH*) distributions and exploring the peculiar properties of the *DPH* family that could make convenient their use in the approximate analysis of discrete state non-Markovian stochastic systems. A minimal representation, called canonical form, for the subclass of Acyclic *DPH* (*ADPH*) was provided. An important theorem was presented, that relates the minimal coefficient of variation for the *ADPH* class to the order and the expected value. It was shown that below a given order (that is a function of the expected value) the minimal coefficient of variation of the *ADPH* family is always less than the minimal coefficient of variation of the *CPH* family. Moreover, since the unit step function is a member of the *ADPH* class, constant and random durations can be combined inside the proposed formalism.

The use of discrete-state discrete-time processes in applied stochastic modeling requires the possibility of estimating the parameters of a *DPH* distribution given a continuous distribution function or a set of experimental samples.

Starting from the results given in Part 1 [3], Part 2 presents an algorithm for the estimation of the *ADPH* model parameters based on the maximum likelihood (ML) principle. A z-transform version of the algorithm is derived from the continuous case [2], while a novel time domain version is provided. It is shown that the time domain algorithm is easier to implement and more stable.

The algorithm is then tested on a benchmark of 10 different continuous distributions that have been already utilized for a similar study in the continuous case [4]. However, since a continuous distribution needs to be discretized in order to feed the fitting algorithm, the role of the discretization interval on the performance of the algorithm and on the goodness of the fit is extensively discussed.

The structure of the paper is as follows. Section 2 introduces the basic definitions and notations mainly derived from [3]. Section 3 presents the ML estimation algorithm, both in z-transform domain and in time domain. Section 4 discusses the role of the discretization interval on the accuracy of the obtainable approximation, while Section 5 is devoted to present the results of the benchmark analysis. Finally, Section 6 concludes the paper.

2 Definition and Notation

The material of this Section is extracted from Part 1 [3]. A *DPH* distribution [7, 8] is the distribution of the time until absorption in a Discrete-State Discrete-Time Markov Chain *(DTMC)* with n transient states, and one absorbing state. If the transient states are numbered $1, 2, \ldots, n$ and the absorbing state is numbered $(n+1)$, the one-step transition probability matrix of the corresponding *DTMC* can be partitioned as:

$$\widehat{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{b} \\ \mathbf{0} & 1 \end{bmatrix}$$

where $\boldsymbol{B} = [b_{ij}]$ is the $(n \times n)$ matrix collecting the transition probabilities among the transient states, $\boldsymbol{b} = [b_{i,n+1}]^T$ is the $n-$dimensional column vector grouping the probabilities from any state to the absorbing one, and $\mathbf{0} = [0]$ is the zero vector. Since $\widehat{\boldsymbol{B}}$ is the transition probability matrix of a *DTMC*, the following relation holds: $\sum_{j=1}^{n} b_{ij} = 1 - b_{i,n+1}$.

The initial probability vector is an $(n+1)$ dimensional vector $\widehat{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}, \alpha_{n+1}]$, with $\sum_{j=1}^{n} \alpha_j = 1 - \alpha_{n+1}$. (In the present paper, we only consider the class of *DPH* distributions

with n finite and for which $\alpha_{n+1} = 0$). Let τ be the time till absorption of the *DTMC* into state $(n+1)$. We say that τ is a *DPH* r.v. of order n and representation $(\boldsymbol{\alpha}, \boldsymbol{B})$.

Let $f(k)$, $F(k)$ and $\mathcal{F}(z)$ be the probability mass function (*pmf*), the cumulative probability function and the probability generating function of τ, respectively. It follows:

$$f(k) = Pr\{\tau = k\} = \boldsymbol{\alpha}\, \boldsymbol{B}^{k-1}\, \boldsymbol{b} \qquad \text{for } k > 0 \tag{1}$$

$$F(k) = Pr\{\tau \leq k\} = \boldsymbol{\alpha} \sum_{i=1}^{k-1} \boldsymbol{B}^i\, \boldsymbol{b} = 1 - \boldsymbol{a}\boldsymbol{B}^k\, \boldsymbol{e} \tag{2}$$

$$\mathcal{F}(z) = E\{z^\tau\} = z\, \boldsymbol{\alpha}\, (\boldsymbol{I} - z\boldsymbol{B})^{-1}\, \boldsymbol{b} \tag{3}$$

where $\boldsymbol{e}$ is an n-dimensional column vector with all the entries equal to 1 and $\boldsymbol{I}$ is the $(n \times n)$ identity matrix.

It has been proved in Part-1 that any *DPH* with Acyclic graph *(ADPH)* admits of a minimal representation, called canonical form.

Definition 1 *Canonical Form (CF1)*
Any ADPH has a unique minimal representation called Canonical Form 1 (CF1). The DTMC associated to the CF1 is given in Figure 1, and its matrix representation $(\boldsymbol{a}, \boldsymbol{P})$ *takes the form:*

$$\boldsymbol{a} = [a_1, a_2, \ldots, a_n] \qquad , \qquad \boldsymbol{P} = \begin{bmatrix} q_1 & p_1 & 0 & 0 & \ldots & 0 \\ 0 & q_2 & p_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & q_n \end{bmatrix} \tag{4}$$

$$\text{with:} \qquad \sum_{1}^{n} a_i = 1 \quad \text{and:} \quad p_1 \leq p_2 \leq \ldots \leq p_n$$

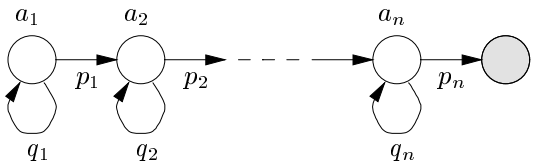


Figure 1: Canonical Form *CF1*.

Due to the particular structure of the *CF1*, we will use the notation $(\boldsymbol{a}, \boldsymbol{p})$ (instead of 4, where $\boldsymbol{a}$ is the n-dimensional initial probability vector and $\boldsymbol{p}$ is the n-dimensional vector containing the ordered exit probabilities $(p_1 \leq p_2 \leq \ldots \leq p_n \leq 1)$.

Given a canonical form *CF1* of order n and representation $(\boldsymbol{a}, \boldsymbol{p})$ (Figure 1), the mean, the second moment and the probability generating function are expressed as:

$$m = \sum_{i=1}^{n} a_i \sum_{j=i}^{n} \frac{1}{p_j} \tag{5}$$

$$d = \sum_{i=1}^{n} a_i \left[\sum_{j=i}^{n} \left(\frac{1}{p_j^2} - \frac{1}{p_j} \right) + \left(\sum_{j=i}^{n} \frac{1}{p_j} \right)^2 \right] \quad (6)$$

$$\mathcal{F}(z) = \sum_{i=1}^{n} a_i \prod_{j=i}^{n} \frac{p_j z}{1-(1-p_j)z} \quad (7)$$

The canonical form *CF1* has a number of important advantages that make it very suited as a basic structure for the development of a numerical estimation algorithm.

- *CF1* is a minimal representation for the *ADPH* class.
  A *CF1* of order n has $N_F = n(n-1)$ free parameters (n free parameters in vector $\boldsymbol{p}$ and $n-1$ in vector $\boldsymbol{a}$), and can represent any *ADPH* that in general may depend on $N_F = n^2(n-1)$ free parameters (n^2 in matrix $\boldsymbol{B}$ and $n-1$ in vector $\boldsymbol{\alpha}$). The *CF1* is minimal, because $N_F = n(n-1)$ is the minimal number of parameters that are necessary to describe a family of distributions whose z-transform (Equation 3) has a denominator of degree n and a numerator of degree at most $n-1$ [3].

- The *CF1* form is a unique representation for the *ADPH* class.

- The simplicity of the structure allows to derive the closed form expressions for the *pmf* and its derivatives with respect to the model parameter both in time domain and in z-transform domain.

3 A fitting algorithm for parameter estimation

We describe a fitting algorithm for estimating the parameters of an *ADPH* in *CF1* form, based on the Maximum Likelihood (ML) principle [2, 4]. We first derive the closed form expression for the *pmf* both in the z-transform domain and in the time domain, and for its derivatives with respect to the model parameters, then the implemented *ML* estimation algorithm is briefly sketched. The range of applicability of both techniques is finally discussed.

3.1 The probability mass function

The generating function of a *CF1-ADPH* with n phases and representation $(\boldsymbol{a}, \boldsymbol{p})$ is provided in (7) and may be written as

$$\mathcal{F}(z) = \sum_{i=1}^{n} a_i \mathcal{F}^{(i)}(z). \quad (8)$$

where (see Figure 1) $\mathcal{F}^{(i)}$ is the generating function of a path of length $(n-i+1)$ form phase i to $(n+1)$, and is given by:

$$\mathcal{F}^{(i)}(z) = \prod_{k=i}^{n} \frac{1-q_k}{z^{-1}-q_k}. \quad (9)$$

Let σ_i ($\sigma_i \leq n-i+1$) denote the number of distinct eigenvalues out of the set $\{q_i, q_{i+1}, ..., q_n\}$ and let us further denote by $(\hat{q}_1^{(i)}, \hat{q}_2^{(i)}, ..., \hat{q}_{\sigma_i}^{(i)})$ the σ_i-dimensional vector of the distinct eigenvalues and by $(\hat{m}_1^{(i)}, \hat{m}_2^{(i)}, ..., \hat{m}_{\sigma_i}^{(i)})$ the vector of their multiplicities. With

this notations, $\hat{m}_j^{(i)}$ is the multiplicity of $\hat{q}_j^{(i)}$ and $\sum_{j=1}^{\sigma_i} \hat{m}_j^{(i)} = n-1+i$. Equation (9) can be rewritten as:

$$\mathcal{F}^{(i)}(z) = \prod_{j=1}^{\sigma_i} \frac{(1-\hat{q}_j^{(i)})^{\hat{m}_j^{(i)}}}{(z^{-1}-\hat{q}_j^{(i)})^{\hat{m}_j^{(i)}}},$$

After a partial fraction decomposition, we have:

$$\mathcal{F}^{(i)}(z) = \sum_{j=1}^{\sigma_i} \sum_{l=1}^{\hat{m}_j^{(i)}} \frac{c_{jl}^{(i)}}{(z^{-1}-\hat{q}_j^{(i)})^l} = \sum_{j=1}^{\sigma_i} \sum_{l=1}^{\hat{m}_j^{(i)}} \frac{c_{jl}^{(i)}\ z^l}{(1-\hat{q}_j^{(i)} z)^l}, \tag{10}$$

where $c_{jl}^{(i)}$ ($j \in \{1,2,\ldots,\sigma_i\}$ and $l \in \{1,2,\ldots,\hat{m}_j^{(i)}\}$) are coefficients determined by the partial fraction decomposition. In [2], a recursive algorithm is proposed for the computation of the coefficients $c_{jl}^{(i)}$.
Using the fact that the z-transform of a series like:

$$h(0) = 0, \text{ and } h(k) = c\ \frac{(k-1)(k-2)...(k-(l-1))}{(l-1)!} q^{k-l}, \text{ if } k \geq 1, l \geq 2$$

is:

$$\mathcal{H}(z) = \frac{c\ z^l}{(1-qz)^l},$$

the inverse of (10) is

$$f^{(i)}(k) = \sum_{j=1}^{\sigma_i} \left[c_{j1}^{(i)} (\hat{q}_j^{(i)})^{k-1} + \sum_{l=2}^{\hat{m}_j^{(i)}} c_{jl}^{(i)} \frac{(k-1)(k-2)...(k-(l-1))}{(l-1)!} (\hat{q}_j^{(i)})^{k-l} \right], \quad k \geq 1, \tag{11}$$

which is the conditional *pmf* of absorption in k steps in state $(n+1)$ assuming that the chain started from phase i. Applying (8), the unconditional *pmf* of absorption in k steps becomes:

$$f(k) = \sum_{i=1}^{n} a_i\ f^{(i)}(k) \tag{12}$$

In the time domain, the *pmf* of the time to absorption is obtained from (1):

$$f(k) = \boldsymbol{a}\ \boldsymbol{P}^{k-1}\ \boldsymbol{p}_n\ , \tag{13}$$

where $\boldsymbol{a}$ and $\boldsymbol{P}$ are given in (4), and $\boldsymbol{p}_n$ is a n-dimensional column vector whose first $n-1$ elements are equal to 0, and the n-th element is equal to p_n.

3.2 The derivatives of the probability mass function

In order to solve the non-linear constrained optimization problem that arises from the application of the ML principle (see Section 3.3), the derivatives of the *pmf* with respect to the parameters $(\boldsymbol{a}, \boldsymbol{p})$ are needed. Because the *pmf* depends linearly on the entries of $\boldsymbol{a}$, the derivatives with respect to these parameters are immediate. In order to express the derivative of the *pmf* with respect to p_j, we rearrange (7):

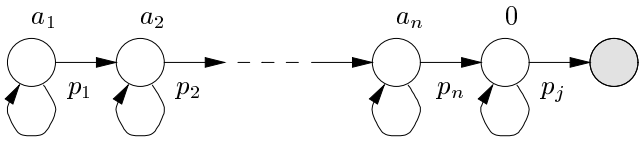


Figure 2: Structure used to determine the derivative with respect to p_j

$$\mathcal{F}(z) = \frac{p_j}{z^{-1} - 1 + p_j} \sum_{i=1}^{j} a_i \prod_{k=i,k\neq j}^{n} \frac{p_k}{z^{-1} - q_k} + \sum_{i=j+1}^{n} a_i \prod_{k=i}^{n} \frac{p_k}{z^{-1} - q_k}, \tag{14}$$

where the second term of the r.h.s. does not depend on p_j. The derivative of (14) with respect to p_j is:

$$\frac{\partial \mathcal{F}(z)}{\partial p_j} = \left[\frac{1}{z^{-1} - q_j} - \frac{p_j}{(z^{-1} - q_j)^2}\right] \sum_{i=1}^{j} a_i \prod_{k=i,k\neq j}^{n} \frac{p_k}{z^{-1} - q_k} =$$

$$\frac{1}{p_j}\left[\sum_{i=1}^{j} a_i \mathcal{F}^{(i)}(z) - \frac{p_j}{z^{-1} - q^j} \sum_{i=1}^{j} a_i \mathcal{F}^{(i)}(z)\right], \tag{15}$$

where $\mathcal{F}^{(i)}(z)$ is given in (9). The second term in the r.h.s. may be interpreted as the generating function of a *CF1*-like model that is obtained by adding one further phase, with exit probability p_j, and null initial probability (Figure 2) to the original *CF1* model. Hence, any algorithm that can be used to evaluate the *CF1* structure of Figure 1, can be utilized to evaluate the derivatives with respect to the p factors as in Figure 2.

Using the partial fraction decomposition method, the coefficients of the augmented model of Figure 2, may be calculated by iterating the same recursive algorithm described earlier [2], just one step more.

Using the matrix formulation (13), the time domain equivalent of (15) becomes:

$$\frac{\partial f(k)}{\partial p_j} = \frac{1}{p_j}\left[\hat{\boldsymbol{a}}\ \boldsymbol{P}^{k-1}\ \boldsymbol{p}_n\ -\ \hat{\boldsymbol{a}}^*\ (\boldsymbol{P}_j^*)^{k-1}\ \boldsymbol{p}_{n+1}^*\right] \tag{16}$$

where $\hat{\boldsymbol{a}}$ ($\hat{\boldsymbol{a}}^*$) is a row vector of length n ($n+1$) with elements $\hat{a}_i = \hat{a}_i^* = a_i$ if $1 \leq i \leq j$, and $\hat{a}_i = \hat{a}_i^* = 0$ otherwise, $\boldsymbol{p}_{n+1}^*$ is a column vector of length $n+1$ with elements $p_i^* = 0$ if $1 \leq i \leq n$, $p_{n+1}^* = p_j$, and

$$\boldsymbol{P}_j^* = \begin{bmatrix} 1-p_1 & p_1 & & \cdots & 0 \\ 0 & 1-p_2 & p_2 & \cdots & 0 \\ & \cdots & & \cdots & 0 \\ 0 & \cdots & \cdots & 1-p_n & p_n \\ 0 & \cdots & \cdots & 0 & 1-p_j \end{bmatrix}$$

is the transition probability matrix that is obtained by adding one more transient phase to the original *CF1* structure (as in Figure 2).

Since $\boldsymbol{P}$ and $\boldsymbol{P}_j^*$ ($j = 1, ..., n$) are upper triangular matrices whose only non-zero elements are located in the main diagonal or in the first (upper) subdiagonal the numerical solution of Equations (13) and (16) does not require the complexity of a general vector-matrix multiplication algorithm.

3.3 ML estimation

Let $\boldsymbol{\phi} = \phi_1, ..., \phi_\nu$ be a set of ν integer data samples. Values in $\boldsymbol{\phi}$ may derive from experimental observations or from the discretization of a continuous *cdf*. Let us denote $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{p}}$ the maximum likelihood estimators of $\boldsymbol{a}$ and $\boldsymbol{p}$, respectively. The likelihood function has the form

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{p}) = \prod_{i=1}^{\nu} f(\phi_i, \boldsymbol{a}, \boldsymbol{p}).$$

The estimation problem consists of finding the parameters ($\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{p}}$) such that the likelihood function $\mathcal{L}(\boldsymbol{\phi}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{p}})$ is maximal, under the constraints of the *CF1* form:

- $0 \leq p_1 \leq p_2 \leq \cdots \leq p_n \leq 1$,
- $a_i \geq 0, \quad \sum_{i=1}^{n} a_i = 1.$

The estimation problem is then formulated in terms of a non-linear constrained optimization problem, that is solved by resorting to an iterative application of a linear programming algorithm. The logarithm of the likelihood function is linearized around the current point by means of a first order series expansion:

$$\begin{aligned} &\log \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a} + \Delta, \boldsymbol{p} + \Delta) = \\ &\quad \log \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{p}) + \frac{\partial \log \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{p})}{\partial \boldsymbol{a}} \; \Delta \; \boldsymbol{a}^T + \frac{\partial \log \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{p})}{\partial \boldsymbol{p}} \; \Delta \; \boldsymbol{p}^T \end{aligned} \tag{17}$$

Given an initial guess $\boldsymbol{a}_0, \boldsymbol{p}_0$, $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{p})$ is linearized according to (17) and the maximum of the linearized function is found inside a small box around $\boldsymbol{a}_0, \boldsymbol{p}_0$. The solution of this step is used as the initial guess in the subsequent step of the iterative procedure, and the procedure is iterated until a preassigned tolerance level is reached or the maximum number of iterations is exceeded.

3.4 Comparison of the algorithms

The z-transform algorithm is based on a partial fraction decomposition method applied to Equations (9) and (10) for the computation of the *pmf*, and to Equation (15) for the computation of the derivatives. The most time consuming and unstable part of the algorithm is the evaluation of the coefficients $c_{jl}^{(i)}$ in (10). The instability comes from the fact that when two eigenvalues tend to be equal, the associated coefficients grow unboundedly, and when the eigenvalues are coincident the expression of the partial fraction expansion changes.

During the iterative estimation procedure the p parameters may become closer, and a criterion should be set to decide whether two close eigenvalues are "coincident" and to modify the partial fraction expansion accordingly. Practically, a small quantity ϵ is assigned, and when the difference between two eigenvalues becomes less than ϵ, they are considered to be coincident. However, this procedure introduces numerical instabilities and inaccuracies.

Once the coefficients $c_{jl}^{(i)}$ are determined, the evaluation of the *pmf* and of its derivatives even for a large k can be done recursively at a very low computational cost.

On the other hand, using the time domain analysis, the *pmf* is evaluated through Equation (13) while the derivatives are evaluated through Equation (16). The solution of

both equations requires a vector matrix multiplication that must be replicated k times. Due, however, to the very special and sparse structure of the involved matrices and vectors a very specialized algorithm can be used. Moreover, since all the entries in the vectors and matrices are non-negative numbers less than 1, the vector matrix multiplication remains very stable for any value of k. Of course, in this case, the complexity of the algorithm increases with k. Hence, the time domain algorithm is much simpler to be implemented, more stable and faster. Only when the time span k over which the solution is required becomes very high, the use of the z-transform algorithm may be justified.

We have implemented and experimented both algorithms, but the results we show in the next section are all obtained by means of the time-domain algorithm.

4 Approximating continuous distributions

When using *ADPH* distributions to approximate random variables arising in practical problems, there are cases in which a discrete sample of data points is directly derived from the application. But there are also cases in which the distributions to be approximated are not discrete. For example, *ADPH* distributions can be utilized to approximate continuous distributions.

The *ADPH* approximation of a continuous distribution requires two steps:

1. The distribution is discretized according to a given discretization step. Indeed, discrete samples and associated mass probability values are generated.

2. The *ADPH* estimation algorithm is run over the discrete sample provided in the previous step.

The discretization of a continuous distribution is a delicate step that introduces errors, and the amplitude of the introduced errors is mainly related to the size of the discretization interval. Therefore, the role of the discretization interval and its impact on the goodness of fit of *DPH* estimation algorithms is investigated in the following sections.

4.1 The role of the discretization interval

There are several ways to "discretize" a general distribution, i.e., to assign a probability mass to the elements of a discrete, finite (ordered) set $\mathcal{S} = \{x_1, x_2, x_3, \ldots\}$ (where $x_1 < x_2 < x_3 < \ldots$). The most common case of discretization is when the elements of the discrete set are integer multiples of a discretization interval (δ), i.e., $x_i = i\delta$.

Given a r.v. X whose *cdf* is $F_X(x)$, a simple rule for discretizing $F_X(x)$ over the discrete set $\mathcal{S} = \{x_1, x_2, x_3, \ldots\}$ is to use the following:

$$p_i = F_X\left(\frac{x_i + x_{i+1}}{2}\right) - F_X\left(\frac{x_{i-1} + x_i}{2}\right), \quad i > 1, \quad \text{and} \quad p_1 = F_X\left(\frac{x_1 + x_2}{2}\right) \tag{18}$$

where p_i is the probability associated with x_i. This discretization does not preserve the moments of the distribution.

Since there are various potential ways to discretize a given distribution function, here we try to collect some general guidelines that are valid in the majority of cases.

In general, the smaller is the discretization interval the closer is the discretized distribution to the original one. Hence, on the one hand, the discretization error decreases by

decreasing the discretization interval: this remark suggests the use of a small δ. On the other hand, the discretization interval changes the scale of the representation. Indeed, let X be the original (non-negative) random variable expressed in a natural time unit (e.g., seconds); its discretized counterpart X_d is expressed in δ unit. For any reasonable discretization procedure, we must have:

$$E(X^i) \sim \delta^i \; E(X_d^i), \qquad i \geq 1 \; , \tag{19}$$

being $E(X^i)$ and $E(X_d^i)$ the i-th moment of X and X_d^i, respectively.

Equation (19) shows that a discretization procedure modifies the mean of the distribution ($E(X) \sim \delta \; E(X_d)$) (since the mean of the discretized distribution is δ times lower than the mean of the original distribution), but leaves (almost) unchanged its coefficient of variation ($cv(X) \sim cv(X_d)$). Since the minimal cv of a *DPH* distribution is a function of its mean the chosen discretization interval may play a significant role in the variability of the *DPH* and, hence, in the goodness of the fit.

4.2 Bounds of the discretization interval

The following considerations provide practical upper and lower bounds to guide in the choice of a suitable discretization interval δ, and are mainly based on the dependence of the minimal coefficient of variation of a *ADPH* on the order n and on the mean m.

Since we only consider *DPH* distributions with no mass at zero, the mean of any *DPH* distribution is greater than 1, which means that, δ should be less than $E(X)$. However, given the order n of the *DPH*, in order to completely exploit the flexibility associated with the n phases, a better upper bound is:

$$\delta \leq \frac{E(X)}{n-1} \; . \tag{20}$$

If the squared coefficient of variation of the distribution to be approximated ($cv^2(X_d)$) is greater than $1/n$ (i.e., $cv^2(X) \sim cv^2(X_d) > 1/n$), any small value of δ provides a suitable discretization interval. Instead, if $cv^2(X) \sim cv^2(X_d) \leq 1/n$, in order to allow the *ADPH* to reach this low coefficient of variation (lower than the bound of any *CPH* as established in the well known theorem of Aldous and Shepp [1]), δ should satisfy the following relation

$$\delta > \left(\frac{1}{n} - cv^2(X)\right) E(X) \tag{21}$$

based on the theorem about the minimal cv of *ADPH* distributions [3].

The effect on the goodness of the attainable approximation of different discretization intervals (δ) is illustrated utilizing a Lognormal distribution with parameters $(1, 0.2)$, whose mean is 1 and cv^2 is 0.0408 (this Lognormal distribution is a test case of the benchmark considered in Section 5). Figure 3 reports the discretized Lognormal distribution together with the best fit *ADPH*s of order $n = 2$, $n = 4$ $n = 8$ and $n = 12$ (obtained applying the ML algorithm), for two different values of the discretization interval $\delta = 0.05$ and $\delta = 0.025$. The lower and upper bounds of δ, computed from Equations (21) and (20), are reported in Table 1 as a function of the order of the *ADPH* (the same orders $n = 2, 4, 8, 12$ as in Figure 3 are used).

From Figure 3, it can be seen that when δ is less than its lower bound the required low cv cannot be attained; while when δ is in the proper range (e.g. $n = 12; \delta = 0.05$ and $n = 16; \delta = 0.025$) a reasonably good fit is obtained.

n	lower bound of δ equation (21)	upper bound of δ equation (20)
4	0.2092	0.333
8	0.0792	0.1428
12	0.0425	0.0909
16	0.0217	0.0666

Table 1: Upper and lower bound for δ as a function of the order

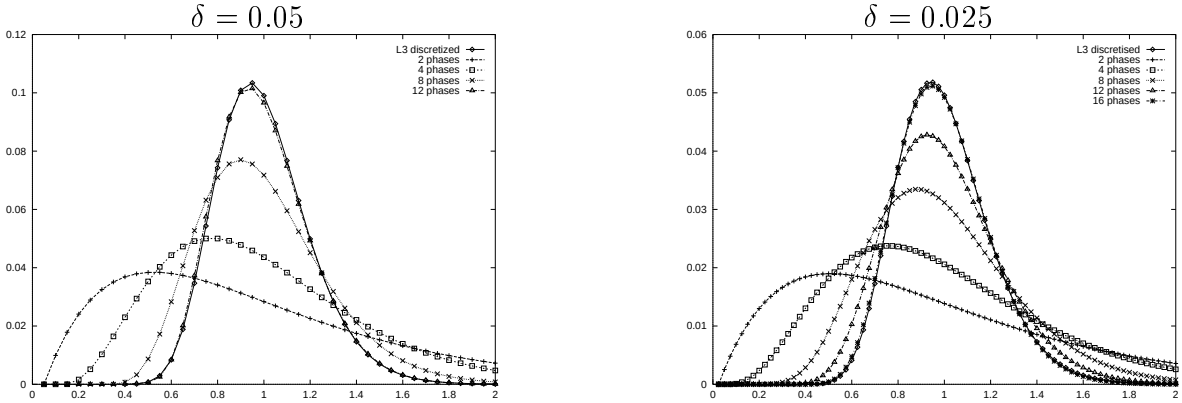


Figure 3: Effect of changing the discretization step

4.3 The required number of phases with a given δ

In Figure 3, it is also visible that the lower δ we use (the higher the mean of the discretized distribution with respect to the discretization interval) the more phases are needed in order to achieve the same goodness of fit. In fact, according to the theorem derived in Part 1 [3] about the minimal cv of the $ADPH$ family, more phases are needed to attain a given coefficient of variation. The minimal number of phases (n) that are needed to reach a given cv^2 when the mean is $E(X_d)$ is given by the next expression

$$n \geq \frac{E(X_d)}{cv^2(X_d)\ E(X_d) + 1}\ , \qquad \text{if } cv^2 > \frac{R(E(X_d))(1 - R(E(X_d)))}{E(X_d)^2}\ ,$$

where $I(\cdot)$ is the integer and $R(\cdot)$ is the fractional part of a real number. Table 2 reports, for the lognormal distribution of Figure 3, the mean $E(X_d)$ and the coefficient of variation $cv^2(X_d)$ of the discretized distribution together with the minimal number of phases needed to reach the coefficient of variation of the original distribution ($cv^2 = 0.0408$), as a function of different discretization steps.

δ	$E(X_d)$	$cv^2(X_d)$	min. phases needed
0.1	9.9649	0.04164	8
0.05	19.9136	0.04101	12
0.025	39.8079	0.04086	16

Table 2: Minimal number of phases as a function of δ

Density	Symbol	Numerical Cases
Weibull $f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(\frac{t}{\eta})^{\beta}}$	*W1*	$\eta = 1 \quad \beta = 1.5$
	W2	$\eta = 1 \quad \beta = 0.5$
Lognormal $f(t) = \frac{1}{\sigma\, t\sqrt{2\,\pi}} \exp\left[-\frac{(\log(t/\phi)+\sigma^2/2)^2}{2\,\sigma^2}\right]$	*L1*	$\phi = 1 \quad \sigma = 1.8$
	L2	$\phi = 1 \quad \sigma = 0.8$
	L3	$\phi = 1 \quad \sigma = 0.2$
Uniform on (a, b)	*U1*	$a = 0 \quad b = 1$
	U2	$a = 1 \quad b = 2$
Shifted Exponential $f(t) = \frac{1}{2}\, e^{-t} + \frac{1}{2}\, e^{-(t-1)}\, I(t \geq 1)$	*SE*	
Matrix Exponential $f(t) = \left(1 + \frac{1}{(2\,\pi)^2}\right)(1 - \cos(2\,\pi\, t))\, e^{-t}$	*ME*	
Exponential $f(t) = \lambda e^{-\lambda t}$	*EX*	$\lambda = 1$

Figure 4: Test cases of the benchmark

Table 2 also shows how the discretization modifies the mean and the cv as a function of the discretization step.

5 Examples for the estimation process

This section reports the results of the numerical experiments that have been carried out to test the goodness of fit of the proposed ML fitting algorithm. The experiments are based on a benchmark (composed of continuous distributions only) already proposed in [4] to test the goodness of fit of algorithms for *CPH* distributions (the origin and the motivations behind the proposed benchmark are discussed in [4]). Hence, the present results allows us to compare the features of the discrete and the continuous phase type fitting.

Table 5 summarises the distributions that compose the benchmark. In Table 5, the continuous exponential distribution has been added, which was not present in the original benchmark in [4], since the continuous exponential is not a *DPH* distribution.

Since in our experiments we have to approximate continuous distributions, we have to discretize them before approximation. In the present experiments, we have used the following discretization method. We conventionally assume that the largest sample in the discretized distribution corresponds to the discrete point closest to $\hat{x}$ where $F(\hat{x}) = 0.995$, and we assign a probability mass to all points from 1 to $\hat{x}$ based on the rule in (18). As

1.	Relative error in the 1st moment:	$\hat{e}_1 = \lvert c_1(F) - c_1(\bar{F})\rvert / c_1(F)$
2.	Relative error in the 2nd moment:	$\hat{e}_2 = \lvert c_2(F) - c_2(\bar{F})\rvert / c_2(F)$
3.	Relative error in the 3rd moment:	$\hat{e}_1 = \lvert c_1(F) - c_1(\bar{F})\rvert / c_1(F)$
4.	*pmf* absolute area difference:	$\hat{D} = \sum_{i=1}^{\infty} \lvert f_d(i) - \bar{f}(i)\rvert$
5.	Minus cross entropy:	$-\hat{H} = \sum_{i=1}^{\infty} f_d(i)\ log(\bar{f}(i))$

Table 3: Measures for evaluating the goodness of fit

mentioned in the previous section, this discretization rule does not preserve the moments, so that the moments of the discretized distribution (including the expected value) are not coincident with the ones of the original continuous distribution.

For further reference let us denote $F(\cdot)$ $f(\cdot), F_d(\cdot)$ $f_d(\cdot), \bar{F}(\cdot)$ $\bar{f}(\cdot)$ the *cdf* and *pmf* of the original distribution, the discretized distribution, and the one resulting from the *ML* estimation algorithm, respectively.

According to [4], five different measures have been chosen to evaluate the goodness of the fit. The five measures are defined on the left of Table 3, where $c_1(F)$, $c_2(F)$ and $c_3(F)$ represent the first three centered moments of $F(\cdot)$.

While in [4], measures 4. and 5. were defined over continuous functions (as integrals over the support of the distribution), in Table 3 the discretized version has been reported. Hence, the first three measures in Table 3 are computed between the original and the *ML*-estimation, the last two measures are computed between the discretized and the *ML*-estimation.

5.1 Results

Figure 5 plots the results obtained for the 10 distributions of the benchmark in term of their *pmf*'s. For each distribution of Table 5, Figure 5 reports the discretized distribution $f_d(\cdot)$ (in solid line) and the *ML*-estimations $\bar{f}(\cdot)$, computed for *CF1* with 2, 4 and 8 phases, respectively. The discretization step is assumed $\delta = 0.1$ in all the plots.

A detailed description of the measures obtained for all the distributions in the benchmark is reported from Table 4 to Table 8. In each table, the measures are reported for *CF1* of order 2, 4 and 8 respectively, and for two discretization intervals, namely: $\delta = 0.1$ (as in Figure 5) and $\delta = 0.05$.

For most of the cases the results of the discrete approximation are comparable with the results obtained from the continuous approximation [4]. However, for the cases where the distribution has a low coefficient of variation the *DPH* approximation shows a better fit, which is in line with the result on the minimal *cv* of the *ADPH* class, discussed in Part 1 [3]. As the relation between the order n and the discretization interval δ fits the bounds established in Section 4, the *ADPH* approximation can attain lower coefficients of variations with respect to the *CPH*s of the same order. This can be seen for the test

Dist.	$c_1(F)$	Relative error $\hat{e}_1$					
		2 phases		4 phases		8 phases	
		$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$
W1	0.9027	0.0144	0.0352	0.0018	0.0142	0.0031	0.0043
W2	2.0000	0.5148	0.6417	0.5203	0.6705	0.5297	0.6711
L1	1.0000	0.2195	0.3703	0.3215	0.4721	0.3224	0.4945
L2	1.0000	0.0918	0.0339	0.0751	0.0794	0.0744	0.0758
L3	1.0000	0.0194	0.0226	0.0107	0.0337	0.0036	0.0072
U1	0.5000	0.0904	0.0667	0.0996	0.0375	0.1000	0.0492
U2	1.5000	0.0308	0.0024	0.0215	0.0387	0.0014	0.0147
SE	1.5000	0.0148	0.0503	0.0238	0.0200	0.0355	0.0623
ME	1.0494	0.0706	0.0831	0.0709	0.1334	0.0709	0.0755
EX	1.0000	0.0181	0.0560	0.0170	0.0467	0.0176	0.0353

Table 4: Relative error in the 1st moment

case L3, when $n = 8$ and $\delta = 0.1$.

In the benchmark, there are test cases whose discretized version is a *DPH* distribution. For example, the two uniform distributions (U1 and U2), using a discretization interval $\delta = 0.1$, can be represented as, respectively:

$$\boldsymbol{a}_1 = [0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1], \boldsymbol{p}_1 = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1],$$

and

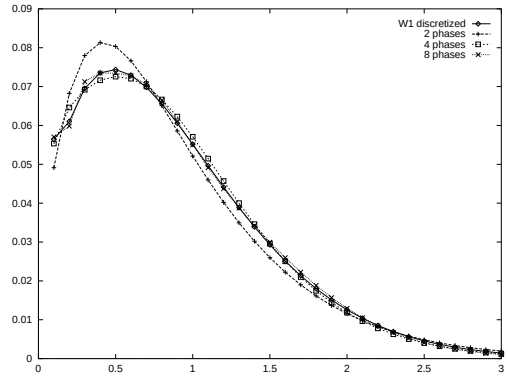
$$\boldsymbol{a}_2 = [0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0],$$

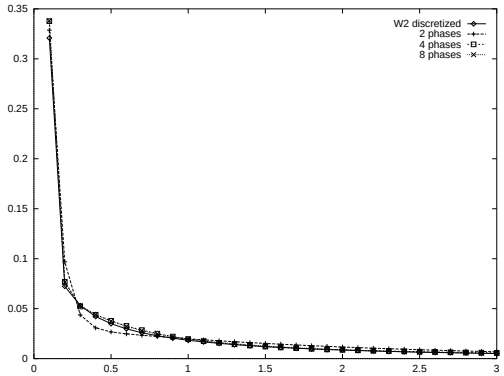
$$\boldsymbol{p}_2 = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1],$$

The proposed estimation algorithm is able the find exactly these forms. For example, with $\delta = 0.1$, $n = 10$ phases are needed to represent exactly the discretized version of U1; however, it is interesting to observe by a visual inspection of Figure 5 how the approximating *ADPH* improves the fit passing from $n = 2$ to $n = 8$.

6 Conclusion

A maximum likelihood estimation procedure for the evaluation of the parameters of a *ADPH* distribution in canonical form *CF1* has been presented. While previous estimation





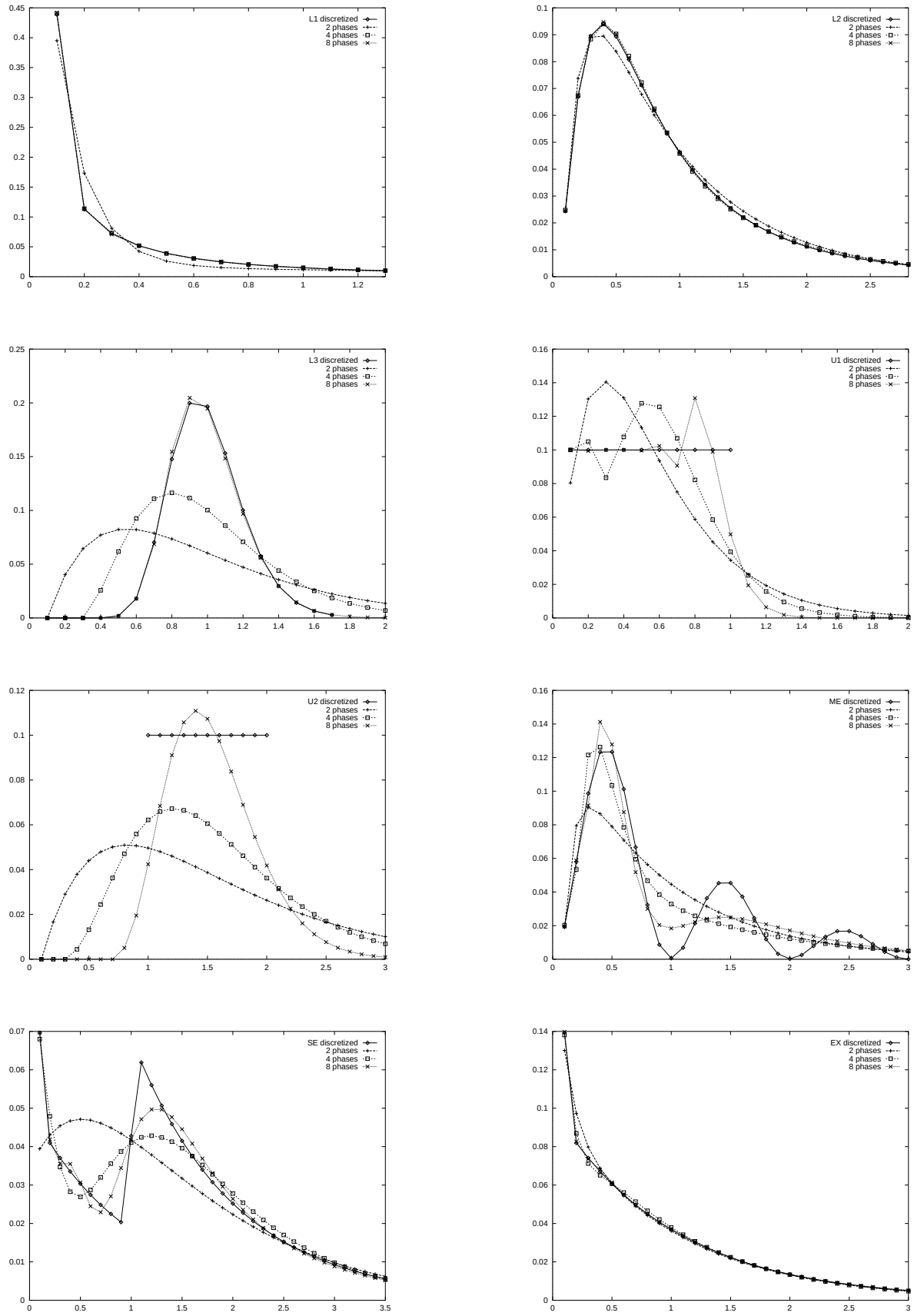


Figure 5: Probability mass functions of the discretized an the approximating *ADPH* distributions

Dist.	Original Dist.		Relative error $\hat{e}_2$					
	$c_2(F)$	$(c.v.)^2$	2 phases		4 phases		8 phases	
			$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$
W1	0.3756	0.4610	0.1704	0.0402	0.0136	0.0497	0.0288	0.0262
W2	20.000	5.0000	0.9161	0.9600	0.9057	0.9618	0.9128	0.9629
L1	24.534	24.534	0.9282	0.9604	0.9347	0.9714	0.9350	0.9736
L2	0.8964	0.8964	0.4661	0.3977	0.4194	0.4510	0.4286	0.4392
L3	0.0408	0.0408	0.9922	10.405	0.9926	4.2794	0.9928	0.8861
U1	0.0833	0.3333	0.6109	0.7544	0.1611	0.1915	0.0070	0.0522
U2	0.0833	0.0370	11.479	10.308	4.2027	4.3402	0.5621	1.5680
SE	1.2500	0.5555	0.0890	0.1224	0.1732	0.2044	0.1873	0.2765
ME	0.9530	0.8653	0.3267	0.3765	0.1358	0.3561	0.3080	0.3190
EX	1.0000	1.0000	0.0396	0.1953	0.0773	0.2075	0.1246	0.1884

Table 5: Relative error in the 2nd moment

Dist.	$c_3(F)$	Relative error $\hat{e}_3$					
		2 phases		4 phases		8 phases	
		$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$
W1	0.2468	0.7011	0.3739	0.0263	0.1875	0.0899	0.0723
W2	592.00	0.9917	0.9977	0.9889	0.9976	0.9902	0.9977
L1	16573	0.9995	0.9998	0.9995	0.9999	0.9995	0.9999
L2	3.1315	0.8061	0.7974	0.7833	0.8149	0.8053	0.8152
L3	0.0050	74.82	83.019	11.362	18.847	0.1713	2.1075
U1	0.0000	0.3949	0.4090	0.2826	0.2987	0.2482	0.2625
U2	0.0000	8.5742	7.4103	4.8749	5.3990	3.3664	3.8605
SE	2.0000	0.1740	0.4032	0.4917	0.6468	0.4752	0.1260
ME	1.9929	0.4958	0.6240	0.1261	0.5418	0.5922	0.6278
EX	2.0000	0.0490	0.4366	0.1249	0.4579	0.3114	0.4689

Table 6: Relative error in the 3rd moment

Dist.	H	Cross Entropy $\hat{H}$					
		2 phases		4 phases		8 phases	
		$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$
W1	0.7869	0.7740	0.7868	0.7688	0.7810	0.7686	0.7803
W2	1.1546	0.8707	0.5096	0.8478	0.5061	0.8577	0.5138
L1	0.3745	0.1620	0.0866	0.1315	0.0420	0.1314	0.0457
L2	0.8756	0.8940	0.8284	0.8910	0.8256	0.9034	0.8257
L3	-0.2104	0.5300	0.5845	0.0982	0.2141	-0.1978	-0.1252
U1	0.0000	0.1528	0.1822	0.0722	0.1016	0.0172	0.0562
U2	0.0000	1.0041	1.0071	0.5801	0.6525	0.2201	0.3280
SE	1.2950	1.3327	1.2820	1.2983	1.2382	1.2883	1.2287
ME	0.7277	0.9439	0.9049	0.9013	0.8712	0.8609	0.8201
EX	1.0000	0.9455	0.9317	0.9442	0.9312	0.9460	0.9286

Table 7: Cross entropy

Dist.	Area Difference					
	2 phases		4 phases		8 phases	
	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$
W1	0.0710	0.0948	0.0248	0.0208	0.0088	0.0103
W2	0.1879	0.2095	0.1445	0.1976	0.1512	0.1972
L1	0.2259	0.2782	0.0099	0.0191	0.0063	0.0054
L2	0.0613	0.0951	0.0396	0.0358	0.0371	0.0306
L3	1.0233	1.0733	0.6274	0.7574	0.0248	0.3099
U1	0.4270	0.4162	0.2773	0.2886	0.1227	0.1892
U2	1.2994	1.2107	0.7871	0.8975	0.3091	0.4832
SE	0.2788	0.2770	0.1592	0.1824	0.0913	0.1260
ME	0.4698	0.5006	0.3845	0.4184	0.2668	0.3151
EX	0.0458	0.0275	0.0284	0.0341	0.0147	0.0254

Table 8: pmf absolute area difference

algorithms for the *CPH* family were based on a transform domain analysis, we have shown that the time domain analysis is also possible, and the estimation algorithm based on time domain expressions is easier to implement and numerically simpler and more stable.

The goodness of fit of this new algorithm has been tested with respect to a benchmark composed of 10 different continuous distributions. Here, the same benchmark, already proposed for testing the performances of *CPH* estimation algorithms, has be considered. However, in order to apply the proposed procedure to a continuous distribution, the continuous function must be discretized according to a given discretization interval. The role of the discretization interval has been discussed, and the way to chose a suitable discretization interval as a function of the mean and of the coefficient of variation has been indicated.

As it could have been expected from the properties of the *ADPH* family discussed in Part 1, the fitting algorithm performs better than the *CPH* one in the cases in which the coefficient of variation is low, and in the cases of distributions with finite support (like the uniform). Moreover, due to the trivial observation the the unit step function is a member of the *ADPH* family, the use of discrete *PH* distributions in applied modeling allows the user to consider, more accurately than in any previous methodology, distributions with finite and infinite support together with constant durations.

References

[1] D. Aldous and L. Shepp. The least variable phase type distribution is Erlang. *Stochastic Models*, 3:467–473, 1987.

[2] A. Bobbio and A. Cumani. ML estimation of the parameters of a PH distribution in triangular canonical form. In G. Balbo and G. Serazzi, editors, *Computer Performance Evaluation*, pages 33–46. Elsevier Science Publishers, 1992.

[3] A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions. Part 1: Properties and canonical forms. Technical report, Technical University of Budapest, 1999.

[4] A. Bobbio and M. Telek. A benchmark for PH estimation algorithms: results for Acyclic-PH. *Stochastic Models*, 10:661–677, 1994.

[5] G. Ciardo. Discrete-time markovian stochastic Petri nets. In *Proceedings of the 2-nd International Workshop on Numerical Solution of Markov Chains*, pages 339–358, 1995.

[6] G. Ciardo and R. Zijal. Discrete deterministic and stochastic Petri nets. In *ICASE Technical Report No. 96-72*, pages 1–25. NASA, 1996.

[7] M.F. Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. University of Louvain, 1975.

[8] M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore, 1981.

[9] W.J. Padgett and J.D. Spurrier. On discrete failure models. *IEEE Transactions on Reliability*, 34:253–256, 1985.

[10] A.A. Salvia. Some results on discrete mean residual life. *IEEE Transactions on Reliability*, 45:359–361, 1996.

[11] M. Scarpa and A. Bobbio. Kronecker representation of stochastic Petri nets with discrete PH distributions. In *International Computer Performance and Dependability Symposium - IPDS98*, pages 52–61. IEEE CS Press, 1998.