Advances in Computational Design, *Vol. 7, No. 1 (2022) 37-56*
https://doi.org/10.12989/acd.2022.7.1.037

Fitting acyclic phase-type distributions by orthogonal distance

Reza Pulungan*[1] and Holger Hermanns[2a]

[1]*Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia*
[2]*Dependable Systems and Software, Saarland University, Saarbrücken, Germany*

(Received July 12, 2021, Revised November 09, 2021, Accepted November 12, 2021)

Abstract. Phase-type distributions are the distributions of the time to absorption in finite and absorbing Markov chains. They generalize, while at the same time, retain the tractability of the exponential distributions and their family. They are widely used as stochastic models from queuing theory, reliability, dependability, and forecasting, to computer networks, security, and computational design. The ability to fit phase-type distributions to intractable or empirical distributions is, therefore, highly desirable for many practical purposes. Many methods and tools currently exist for this fitting problem. In this paper, we present the results of our investigation on using orthogonal-distance fitting as a method for fitting phase-type distributions, together with a comparison to the currently existing fitting methods and tools.

Keywords: acyclic; fitting; orthogonal distance; phase-type distributions; traces

1. Introduction

Stochastic modeling has diverse applications in many areas, such as queuing theory (Neuts 1994), dependability (Böde *et al*. 2009), transportation (Pulungan and Hermanns 2009), building and steel structures (Alam *et al*. 2017, Chavan and Lal 2018), pre-clinical studies (Al-Khalidi and Schnell 1997), and even building information modeling (Sandoval *et al*. 2018). In many such models, phase-type distributions are widely used because they are versatile and tractable. Phase-type distributions generalize exponential distributions; hence, the tractability of exponential distributions is retained when phase-type distributions substitute in their stead. This enables analyses of richer stochastic models and properties.

The class of phase-type distributions has been proven to be topologically dense in the set of probability distributions with support on $[0,\infty)$ (Johnson and Taaffe 1988). Therefore, they can approximate other probability distributions or the traces of empirical distributions obtained from experimental observations. The approximations are usually carried out by fitting phase-type distributions to the empirical distributions.

Various fitting methods for phase-type distributions have existed for almost four decades. These

*Corresponding author, Dr.-Ing., E-mail: pulungan@ugm.ac.id
[a] Professor, E-mail: hermanns@cs.uni-saarland.de

Copyright © 2022 Techno-Press, Ltd.
http://www.techno-press.org/?journal=acd&subpage=7 ISSN: 2383-8477 (Print), 2466-0523 (Online)

methods can be roughly categorized into maximum likelihood, least-square, and moment matching methods. Several fitting tools have been produced based on these fitting methods over quite a long time, including (Bobbio and Cumani 1992), (Asmussen *et al*. 1996), (Horváth and Telek 2002), (Thümmler *et al*. 2006), (Reinecke *et al*. 2012), and most recently (Reinecke *et al*. 2013) and (Horváth and Okamura 2013).

Orthogonal distance fitting is a least-square-based fitting method. It is a parameter estimation problem, where a function of the sum of the squares of distances between points is minimized while determining the parameters. The main characteristic of orthogonal distance fitting is its notion of distance between points. The distance between points is defined in a geometrical way, as opposed to its counterpart, algebraic distance fitting.

In this paper, we report on our investigation on using orthogonal-distance fitting to fit phase-type distributions to empirical probability distributions. We further conduct some experiments with a prototypical implementation, comparing it to existing fitting tools in terms of their goodness-of-fit and computational performance.

This paper is organized as follows: Section 2 provides an overview of related work. Section 3 provides the general concepts of phase-type distributions, fitting, and orthogonal distance fitting. Section 4 presents our proposed method. Section 5 describes the prototypical implementation of the method, the experimental results, and the comparison of the resulting fitting tool with currently existing tools. Section 6 provides a case study showcasing the method's application in the field of transportation. Section 7 concludes the paper.

2. Related work

Phase-type distributions generalize exponential distributions: the class of phase-type distributions consists of all serial, parallel, and cyclical arrangements of exponential distributions. Furthermore, the class of phase-type distributions is a superset of all types of distributions formerly recognized as the generalizations of exponential distributions, such as hyperexponential, hypoexponential, Erlang distributions, and mixtures thereof. This means that the tractability and explicit solutions encountered when dealing with those distributions (exponential, Erlang, or the mixture of Erlang distributions, etc.) are retained when phase-type distributions are used instead. This is important, for then phase-type distributions, which possess richer stochastic properties than the exponential or Erlang distributions, can be used in modeling and analyzing stochastic behaviors.

Johnson and Taaffe (1988) showed that phase-type distributions are dense in the set of probability distributions with support on $[0,\infty)$. This means that for any probability distribution F with support on $[0,\infty)$, there exists a sequence of phase-type distributions that converges weakly to F. They first proved the denseness of infinite mixtures of Erlang distributions, which was then extended to finite mixtures of Erlang distributions and phase-type distributions with Coxian representations. Hence, given enough phases, any probability distribution with support on $[0,\infty)$ can be approximated arbitrarily closely by a phase-type distribution.

The denseness of the class of phase-type distributions has practical significances: first, phase-type distributions can be used as approximations for various distributions, whose inclusion in a model makes analyses difficult or impossible; and second, phase-type distributions can also be used as approximations for distributions that are obtained from the traces of experiments. In both cases, the use of phase-type distributions instead of the original distributions makes the analyses

tractable. In these circumstances, parameter estimation (fitting) algorithms, namely algorithms to determine the optimal values of the parameters of phase-type distributions given any distribution to approximate, play an important role.

In the following, several existing tools for fitting phase-type distributions using the maximum likelihood and least-square methods are described.

Bobbio *et al*. (1980) proposed an algorithm for fitting acyclic phase-type distributions to empirical distributions using the Kolmogorov-Smirnov goodness-of-fit test by minimizing the maximum absolute difference between the distributions' cumulative distribution functions. Acyclic phase-type distributions are a subset of phase-type distributions whose representations have matrices of triangular form. They then solved the estimation problem by reducing it to a non-linear constrained optimization problem.

Cumani (1982) presented three canonical forms of triangular phase-type distributions. These canonical forms reduce the number of parameters to be estimated in a given triangular phase-type distribution and moreover have unique representations. These canonical forms were used by Bobbio and Cumani (1992) to build an algorithm for maximum-likelihood-based phase-type distribution fitting. The estimation problem here was also reduced to a non-linear optimization problem. The algorithm was implemented in Fortran, and the resulting tool was named MLAPH.

In 1996, Asmussen *et al*. (1996) presented an expectation-maximization (EM) algorithm for maximum-likelihood-based phase-type distribution fitting. They imposed no restrictions on the structure of the phase-type distributions used in the fitting. The EM algorithm is an iterative procedure to compute the maximum-likelihood parameter estimates when data are incomplete. The algorithm consists of two steps: (1) E-step, in which the expected values of parameters are computed based on the available data; and (2) M-step, in which the log-likelihood of both distributions is calculated. The two steps are carried out repeatedly until some stopping criteria of convergence are met. The algorithm was implemented in C and is available as tool EMpht.

Horváth and Telek (2002) developed a new tool for fitting phase-type distributions called PhFit. The tool was based on a fitting method that separately approximates the main and tail parts of arbitrary probability distributions with phase-type distributions. The approximation was done by minimizing selected distance measures defined on the probability density function (PDF) or cumulative distribution function (CDF) of both approximated and approximating distributions. Three distance measures were defined: relative entropy, PDF area difference, and CDF area difference. They restricted to acyclic phase-type distributions as the approximating distributions for the main part. For this, they used the first canonical form (CF1) of acyclic phase-type distributions as defined in (Cumani 1982). The problem of minimizing the various distance measures and estimating the parameter were solved using non-linear optimization procedures. Following (Feldmann and Whitt 1998), they used hyperexponential distributions to approximate the tail part of the given distributions.

Thümmler *et al*. (2006) developed a novel approach for fitting phase-type distributions to traces of probability distributions. Their method uses a specific class of phase-type distributions, namely the mixture of mutually independent Erlang distributions, which they referred to as hyper-Erlang distributions. The set of hyper-Erlang distributions is a subset of the triangular phase-type distributions. The fitting algorithm they used is the expectation maximization for the mixture of the hyper-Erlang densities. Because of the special structure of the hyper-Erlang distributions, the EM steps can be solved in a closed-form instead of in an iterative manner, which makes their implementation, called G-Fit, significantly faster than other existing tools.

The standard performance evaluation of parameter estimation algorithms for fitting phase-type

distributions is the *Aalborg benchmark* (Bobbio and Cumani 1992). The benchmark is the result of an international workshop on fitting phase-type distributions held in Aalborg, Denmark, in 1991. Bobbio and Telek (1994) extended the benchmark while evaluating the performance of their fitting algorithm.

3. Preliminaries

3.1 Phase-type distributions

We consider a continuous-time Markov chain with n+1 states, where the state (n+1) is absorbing (namely, having zero total outgoing rate), while the others are transient (namely, there is a non-zero probability that the state will never be revisited once left). The infinitesimal generator matrix of such a Markov chain is of the form:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A} & \vec{A} \\ \vec{0} & 0 \end{pmatrix}, \tag{1}$$

where $\mathbf{A}_{i,j} \geq 0$ ($1 \leq i,j \leq n$; $i \neq j$) represents the transition rate from state i to state j; $\mathbf{A}_{i,i} < 0$ ($1 \leq i \leq n$) is the negative sum of all transition rates outgoing from state i; and $\vec{A}$ is a column vector, where A_i ($1 \leq i \leq n$) represents the transition rate from state i to the absorbing state. Since $\mathbf{Q}$ is a Markov generator matrix, $\vec{A}=-\mathbf{A}\vec{1}$, where $\vec{1}$ is an n-dimensional column vector whose entries are all equal to 1. Moreover, because the first n states are transient, matrix $\mathbf{A}$ is non-singular. A probability distribution with support on $[0,\infty)$ is phase-type if and only if it is the distribution of the time until absorption in such an absorbing Markov chain (Neuts 1994).

The absorbing Markov chain is completely specified by the generator matrix $\mathbf{Q}$ and the initial probabilities vector $(\vec{\alpha},\alpha_{n+1})$, where $\vec{\alpha}$ is an n-dimensional row vector representing the initial probabilities of the transient states and $\vec{\alpha}\vec{1} + \alpha_{n+1} = 1$. A phase-type distribution associated with the Markov chain has representation $(\vec{\alpha},\mathbf{A})$. The cumulative distribution function of the distribution is given by $F(t) = 1 - \vec{\alpha}e^{\mathbf{A}t}\vec{1}$, while its probability density function is:

$$f(t) = \begin{cases} \alpha_{n+1}, & \text{for } t = 0, \\ \vec{\alpha}e^{\mathbf{A}t}\vec{A}, & \text{for } t > 0. \end{cases} \tag{2}$$

3.2 Fitting phase-type distributions

Fitting a phase-type distribution to a given trace of a probability distribution is a parameter estimation problem. In parameter estimation terminology, the phase-type distribution is the model to be estimated, while the trace is the observations of the system. In the following, we first provide the general concepts of parameter estimation, mainly based on (Bobbio and Cumani 1992).

Let $\vec{\Pi}$ be an n-dimensional vector of model's parameters to be estimated. The size and configuration of this vector are usually pre-determined and based on general assumptions made on the system. An observation of the system, on the other hand, is represented by Y, namely error-free information gathered from observing the system. The system itself is a function F relating the parameters and the observation such that:

$$F\left(\vec{\Pi}, Y\right) = 0.$$

However, in practice, observations are not free of errors but are affected by noises. Thus, what is actually observed from the system is:

$$Z = Y + \epsilon,$$

where ε is the noise in the measurement. To counteract the noise's effects, the observation is usually repeated, yielding Z_i ($1 \le i \le m$). These Z_i's are collected in an m-dimensional vector $\vec{Z}$, which corresponds to the given trace. Since the trace is noisy, the function valuations $F(\vec{\Pi},Z_i) = 0$ ($1 \le i \le m$) are no longer valid. Instead, we define a new function:

$$\mathcal{F}\left(\vec{\Pi}, \vec{Z}\right),$$

which is called the *objective function*. The parameters $\vec{\Pi}$ are estimated while optimizing (minimizing or maximizing) this function.

Many parameter estimation methods for probability distributions have been proposed in the last five decades. These methods can be roughly categorized into moment matching, maximum likelihood, and least-square methods. We are focusing on least-square methods in this work.

3.3 Orthogonal distance fitting

Orthogonal distance fitting (ODF) is a least-square fitting method that minimizes errors based on geometric distance. In this section, algebraic distance and geometric distance in fitting are compared. A general definition of orthogonal distance fitting follows. Each iteration of the ODF relies on determining minimum distance points between the approximating and approximated curves. An algorithm for determining these points is presented afterward. The last part of the section provides the algorithm for determining or updating the model's parameters. The primary source of this section is Ahn's monograph (Ahn 2004).

3.3.1 Algebraic and geometric fittings

Based on the definition of the error measure to minimize, least-square-based fitting is categorized into algebraic and geometric fittings. An algebraic curve in implicit representation can generally be described by the function:

$$F\left(\vec{\Pi}, \vec{X}\right) = 0, \tag{3}$$

where $\vec{\Pi}$ is a vector of the function's parameters (for instance, the coefficients if the function is a polynomial) and $\vec{X}$ is the definition of a point on the curve, namely a vector whose size is the dimension of the space in which the curve lies.

If $F(\vec{\Pi}, \vec{X}_i) \neq 0$ for some point $\vec{X}_i$, then the given point $\vec{X}_i$ is not on the curve. When the curve approximates a function describing a set of such points, there must be some error-of-fit. For point $\vec{X}_i$ particularly, the error measure is $F(\vec{\Pi},\vec{X}_i)$ itself, and it is called the algebraic distance. Using this error measure, least-square fitting of a set of given points $\{\vec{X}_i\}_{i=1}^m$ is carried out by determining the parameters $\vec{\Pi}$ that minimize the sum of the squared algebraic distance at each given point, namely:

$$\min_{\vec{\Pi}} \sum_{i=1}^{m} F^2(\vec{\Pi}, \vec{X}_i).$$

This least-square fitting is called algebraic fitting.

In the geometric fitting, on the other hand, the error measure is defined as the shortest distance between the given point and any point on the curve. This distance is called the geometric distance.

Hence, least-square fitting of a set of given points $\{\vec{X}_i\}_{i=1}^m$ is geometric fitting if the parameters $\vec{\Pi}$ are determined while minimizing the sum of the squared geometric distance at each given point, namely:

$$\min_{\vec{\Pi},\{\vec{X}_i'\}_{i=1}^m \subset F} \sum_{i=1}^{m} \left\|\vec{X}_i - \vec{X}_i'\right\|^2,$$

where $\{\vec{X}_i'\}_{i=1}^m$ are the closest points on the curve to the corresponding given points $\{\vec{X}_i\}_{i=1}^m$, and:

$$\left\|\vec{X}_i - \vec{X}_i'\right\| = \sqrt{(\vec{X}_i - \vec{X}_i')^{\mathrm{T}}(\vec{X}_i - \vec{X}_i')}.$$

In general, geometric fitting is non-linear to the model's parameters, which results in a higher complexity in the computation and minimization of the sum of the squared geometric distance compared to algebraic fitting, which is linear to the model's parameters and can often be solved in closed form. Nevertheless, algebraic fitting has numerous disadvantages, as listed by Ahn (2004), including the fact that the error definition of algebraic fitting does not comply with measurement guidelines. Specifically, in algebraic fitting, errors are always assumed not to occur in one of the variables defining the curves. Fig. 1 illustrates the difference between algebraic and geometric fittings in terms of their error definitions. Let $\vec{X}_i$ be a point to approximate. In algebraic fitting, the closest point in the curve is $\vec{X}_i''$, and the distance between them is d_i''. In geometric fitting, on the other hand, the closest point is $\vec{X}_i'$, and the distance between them is d_i'. It is evident from the figure that in the algebraic fitting of 2-dimensional curves, errors are assumed to occur only in the *y*-axis.

3.3.2 Definition

The purpose of orthogonal distance fitting of a model to a set of given points is to determine the model's parameters that minimize the sum of the squared minimum distances of each given point to any point on the model. There are two ways to define the sum of the squared distance (also called the *cost function*), namely:

$$\sigma_0^2 = \left\|\mathbf{W}\vec{d}\right\|^2 = \vec{d}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\vec{d} \text{ and} \tag{4}$$

$$\sigma_0^2 = \left\|\mathbf{W}(\vec{X} - \vec{X}')\right\|^2 = (\vec{X} - \vec{X}')^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\vec{X} - \vec{X}'), \tag{5}$$

where $\vec{X}$ and $\vec{X}'$ are the coordinate column vectors of the given ($\{\vec{X}_i\}_{i=1}^m$) and the model's corresponding points ($\{\vec{X}_i'\}_{i=1}^m$), respectively; $\vec{d}$ is the distance column vector, namely $\vec{d}^{\mathrm{T}} = (d_1, d_2, \cdots, d_m)$ and $d_i = \left\|\vec{X}_i - \vec{X}_i'\right\|$; and $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ is the weighting or error covariance matrix.

An ODF algorithm that uses Eq. (4) as the cost function is a distance-based algorithm, while that which uses Eq. (5) is a coordinate-based algorithm. This paper selects the distance-based algorithm because the resulting algorithm for phase-type distributions is simpler and easier to implement.

ODF determines the model's parameters and minimizes the sum of the squared minimum distance of each given point to any point on the model. If determining the parameters and minimizing the sum are carried out simultaneously, the ODF is said to use the total method (Ahn 2004), in which case the ODF is described by:

$$\min_{\vec{\Pi},\{\vec{X}_i'\}_{i=1}^m \subset F} \sigma_0^2.$$

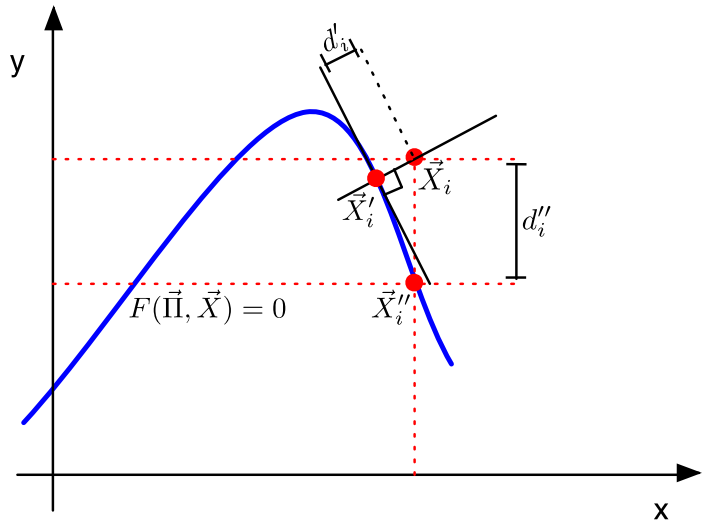

Fig. 1 Algebraic and geometric distances

On the other hand, if they are carried out separately, the ODF uses the variable-separation method, and the ODF is described by:

$$\min_{\vec{\Pi}} \min_{\{\vec{X}_i'\}_{i=1}^m \subset F} \sigma_0^2. \tag{6}$$

We use the variable-separation method in this paper for the reason of simplicity of the resulting ODF algorithm.

3.3.3 Determining the minimum distance points

The inner minimization of Eq. (6) involves searching for the minimum distance points $\{\vec{X}_i'\}_{i=1}^m$ on the model from the given points $\{\vec{X}_i\}_{i=1}^m$. This paper uses the generalized Newton method to determine the minimum distance points.

For a point $\vec{X}_i'$ on a model with implicit representation (cf. Eq. (3)) to have the minimum distance from a given point $\vec{X}_i$, it is necessary that the line connecting $\vec{X}_i'$ and $\vec{X}_i$ be parallel to $\nabla F = \left(\frac{\partial F}{\partial \vec{X}_i'}\right)^{\mathrm{T}}$ at $\vec{X}_i'$, hence:

$$\nabla F \times \left(\vec{X}_i - \vec{X}_i'\right) = \vec{0}.$$

This equation combined with Eq. (3) allows us to define the function:

$$\mathbb{F}\left(\vec{\Pi}, \vec{X}_i, \vec{X}_i'\right) = \begin{pmatrix} F \\ \nabla F \times \left(\vec{X}_i - \vec{X}_i'\right) \end{pmatrix} = \vec{0},$$

which can be solved by using the generalized Newton method starting from the initial point $\vec{X}_0 = \vec{X}_i$ as follows:

$$\left.\frac{\partial \mathbb{F}}{\partial \vec{X}_i'}\right|_k \Delta\vec{X}_i' = -\mathbb{F}\left(\vec{X}_i'\right)\big|_k; \qquad \vec{X}_i'\big|_{k+1} = \vec{X}_i'\big|_k + \alpha\Delta\vec{X}_i', \tag{7}$$

where $\vec{X}_i'\big|_k$ is the value of $\vec{X}_i'$ at iteration k.

3.3.4 Updating the model's parameters

This part presents the ODF distance-based algorithm using the variable-separation method to determine the model's parameters. This essentially boils down to the outer minimization of Eq. (6). The necessary condition for the cost function (cf. Eq. (4)), as a function of the model's parameters $\vec{\Pi}$, to be minimum is given by:

$$\left(\frac{\partial}{\partial\vec{\Pi}}\sigma_0^2\right)^{\mathrm{T}} = 2\mathbf{J}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\vec{d} = \vec{0}, \tag{8}$$

where $\mathbf{J} = \frac{\partial\vec{d}}{\partial\vec{\Pi}}$. Eq. (8) can be iteratively solved using the Gauss-Newton method as follows:

$$\mathbf{W}\mathbf{J}|_k\Delta\vec{\Pi} = -\mathbf{W}\vec{d}\big|_k; \qquad \vec{\Pi}\big|_{k+1} = \vec{\Pi}\big|_k + \alpha\Delta\vec{\Pi}, \tag{9}$$

with termination conditions:

$$\left\|\mathbf{J}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\vec{d}\right\| \approx 0 \quad \text{or} \quad \left\|\Delta\vec{\Pi}\right\| \approx 0 \quad \text{or} \quad \sigma_0^2\big|_{k+1} - \sigma_0^2\big|_k \approx 0.$$

The second part of Eq. (9) updates the model's parameters, and the updates are weighted by a constant of proper choice α. The first of the termination conditions is the original minimum requirement, namely Eq. (8). Eq. (9) itself is a system of linear equations:

$$\mathbf{W}\underbrace{\begin{pmatrix}\mathbf{J}_{d_1,\vec{\Pi}}\\ \vdots \\ \mathbf{J}_{d_m,\vec{\Pi}}\end{pmatrix}}_{m\times n}\Delta\vec{\Pi} = -\mathbf{W}\underbrace{\begin{pmatrix}d_1\\ \vdots\\ d_m\end{pmatrix}}_{m\times\mathbf{1}}, \quad \text{where } \mathbf{J}_{d_i,\vec{\Pi}} = \frac{\partial d_i}{\partial\vec{\Pi}}. \tag{10}$$

Matrix $\mathbf{J}$ in Eq. (9) is called the Jacobian matrix, whose components, represented by $\mathbf{J}_{d_i,\vec{\Pi}}$, is solved as follows:

$$\mathbf{J}_{d_i,\vec{\Pi}} = \frac{\partial d_i}{\partial\vec{\Pi}} = -\frac{\left(\vec{X}_i - \vec{X}_i'\right)^{\mathrm{T}}}{\left\|\vec{X}_i - \vec{X}_i'\right\|}\frac{\partial\vec{X}}{\partial\vec{\Pi}}\bigg|_{\vec{X}=\vec{X}_i'}.$$

From the derivative of Eq. (3):

$$\frac{\partial F}{\partial\vec{X}}\frac{\partial\vec{X}}{\partial\vec{\Pi}} + \frac{\partial F}{\partial\vec{\Pi}} = 0; \quad \frac{\partial F}{\partial\vec{X}}\frac{\partial\vec{X}}{\partial\vec{\Pi}} = \nabla^{\mathrm{T}}F\frac{\partial\vec{X}}{\partial\vec{\Pi}} = -\frac{\partial F}{\partial\vec{\Pi}}.$$

and since $\left(\vec{X}_i - \vec{X}_i'\right)$ is parallel to $\nabla F|_{\vec{X}=\vec{X}_i'}$, we have:

$$\mathbf{J}_{d_i,\vec{\Pi}} = \left(\frac{\mathrm{sgn}\left(\left(\vec{X}_i - \vec{X}_i'\right)^{\mathrm{T}}\nabla F\right)}{\|\nabla F\|}\underbrace{\frac{\partial F}{\partial\vec{\Pi}}\bigg|_{\vec{X}=\vec{X}_i'}}_{1\times n}\right).$$

4. Proposed method

This section describes the application of ODF for fitting phase-type distributions to given traces

of probability distributions. The ODF algorithm described in Section 3 fits a particular curve of a certain parameter structure of unknown parameter values to a set of given points in space. The curve itself is represented in implicit form. Algorithm 1 depicts the proposed orthogonal-distance-based fitting algorithm especially tailored for a specific type of phase-type distributions.

Phase-type distributions used in the proposed ODF algorithm are restricted to the class of acyclic phase-type (APH) distributions. These are phase-type distributions whose generator matrices are upper triangular or contain no cycles when viewed as graphs. This class of phase-type distributions has several canonical forms. This paper uses the ordered bidiagonal form, first defined by Cumani (1982). The n-state ordered bidiagonal representation $(\vec{\alpha},\mathbf{A})$ has the following structure:

$$\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n) \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} -A_1 & A_1 & 0 & \cdots & 0 \\ 0 & -A_2 & A_2 & \cdots & 0 \\ 0 & 0 & -A_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -A_n \end{pmatrix},$$

where $A_n \geq A_{n-1} \geq \cdots \geq A_2 \geq A_1$. This particular structure is selected to ease the computation of both its probability density and cumulative distribution functions. Furthermore, the number of free parameters to estimate is reduced to $2n$ compared to $2n^2$ for the general phase-type distributions. Although this class is a proper subset of the whole class of phase-type distributions, the use of the restricted class does not seem to limit the flexibility of the estimation, as was also observed by Thümmler *et al*. (2006).

At the beginning of Algorithm 1, acyclic phase-type distributions and the corresponding parameters to estimate are initialized. Function Initialize-APH(n) performs the initialization of an ordered bidiagonal representation $(\vec{\alpha},\mathbf{A})$ of size n. Two ways are considered for this, namely:

1. Initializing the representation by an Erlang distribution (Stewart 2009) of n phases (states), whose rate is fixed such that its mean is equal to the given trace's mean.

2. Initializing the representation by the best (in terms of the cost function in Eq. (4)) of k randomly generated ordered bidiagonal representations.

To fix the structure of the parameters to estimate $\vec{\Pi}$ from the representation $(\vec{\alpha},\mathbf{A})$ is basically to select a particular structure for the non-zero components of vector $\vec{\alpha}$ and matrix $\mathbf{A}$. Function Initialize-Parameters($\vec{\alpha},\mathbf{A}$) accomplishes this initialization of the vector of parameters $\vec{\Pi}$ by setting:

$$\Pi_i = \begin{cases} \alpha_i, & \text{for } 1 \leq i \leq n, \\ A_{i-n}, & \text{for } n+1 \leq i \leq 2n. \end{cases}$$

The fitting of the phase-type distributions is carried out in the probability density function domain. This means that the phase-type distributions are fitted to the probability density function of the given traces. The probability density function of a phase-type distribution with representation $(\vec{\alpha}, \mathbf{A})$ is Eq. (2), whose implicit representation is given by:

$$F(\vec{\Pi}, \vec{X}) = \begin{cases} f(t) - \alpha_{n+1} = 0, & \text{for } t = 0, \\ f(t) - \vec{\alpha} e^{\mathbf{A}t} \vec{A} = 0, & \text{for } t > 0. \end{cases}$$

where $\vec{X} = (f(t),t)^{\mathrm{T}}$.

The probability density function of the phase-type distribution is computed by using the uniformization (randomization) method (Jensen 1953, Gross and Miller 1984). This method arises in the transient analysis of Markov chains. Given a Markov chain as in Eq. (1), the transient analysis is the computation of the transient state-probabilities vector:

Algorithm 1 The proposed orthogonal distance fitting algorithm

1: function Orthogonal-APH-Fit($n, \{\vec{X}_i\}_{i=1}^m, \epsilon, maxit$): $(\vec{\alpha}, \mathbf{A})$
2: $(\vec{\alpha}, \mathbf{A}) \leftarrow$ Initialize-APH(n)
3: $\vec{\Pi} \leftarrow$ Initialize-Parameters($\vec{\alpha}, \mathbf{A}$)
4: $k \leftarrow 0$
5: repeat
6: $k \leftarrow k + 1$
7: for $i \leftarrow 1$ to m do
8: $\vec{X}_i' \leftarrow$ Minimum-Distance($\vec{X}_i, \vec{\alpha}, \mathbf{A}$) ▷Eq. (7)
9: end for
10: $\vec{\Pi} \leftarrow$ Update-Parameters($\{\vec{X}_i'\}_{i=1}^m, \vec{\alpha}, \mathbf{A}$) ▷Eq. (9)
11: $(\vec{\alpha}, \mathbf{A}) \leftarrow$ Update-APH($\vec{\Pi}$)
12: until ($\left\|\mathbf{J}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\vec{d}\right\| < \epsilon$ or $\left\|\Delta\vec{\Pi}\right\| < \epsilon$ or $|\sigma_0^2|_{k+1} - \sigma_0^2|_k| < \epsilon$ or $k > maxit$)
13: return $(\vec{\alpha}, \mathbf{A})$
14: end function

$$\vec{\pi}(t) = (\vec{\alpha}, \alpha_{n+1})e^{\mathbf{Q}t} = (\vec{\alpha}, \alpha_{n+1}) \sum_{i=0}^{\infty} \frac{(\mathbf{Q}t)^i}{i!}, \tag{11}$$

where $\vec{\pi}(t)$ is the state-probabilities vector at time-instant t. For $\Lambda \geq \max_i\{-\mathbf{Q}_{i,i}\}$ and $\mathbf{P} = \mathbf{I} + \mathbf{Q}/\Lambda$, where $\mathbf{I}$ is the corresponding identity matrix, the uniformization method solves Eq. (11) as follows:

$$\vec{\pi}(t) = (\vec{\alpha}, \alpha_{n+1})e^{\Lambda t(\mathbf{P}-\mathbf{I})} = (\vec{\alpha}, \alpha_{n+1}) \sum_{i=0}^{\infty} \frac{e^{\Lambda t}(\Lambda t)^i}{i!} \mathbf{P}^i.$$

Hence, in this case the cumulative distribution function of the phase-type distribution is $F(t) = \pi_n(t)$ and its derivative (namely, the probability density function), for the first canonical form of acyclic phase-type distributions, is simply given by $f(t) = F'(t) = \pi_n(t)A_n$.

The nucleus of the algorithm is the repeated computation of Eq. (6) in the form of determining the minimum distance points and then updating the model's parameters until one of the termination conditions is satisfied (lines 5-12).

The first part is accomplished by applying the function Minimum-Distance$(\vec{X}_i, \vec{\alpha}, \mathbf{A})$ to compute the orthogonally closest point in the curve of the current phase-type representation $(\vec{\alpha}, \mathbf{A})$ to each point $\vec{X}_i$ in the given trace. This function represents the computation of Eq. (7), namely solving the minimum orthogonal-distance points by using the generalized Newton method.

The second part is carried out by the function Update-Parameters($\{\vec{X}_i'\}_{i=1}^m, \vec{\alpha}, \mathbf{A}$), which represents the computation of Eq. (9) by using the singular value decomposition (SVD) method (Press *et al*. 2007). The components of the Jacobian matrix in Eq. (10), which are essentially the partial derivative of the distance measure relative to each parameter, are computed by a numerical derivation method. The distance measure between points is not weighted; hence, we use $\mathbf{W} = \mathbf{I}$ in Eq. (4) and in subsequent equations that use the error covariance matrix. In each iteration of the

evaluation of Eq. (9), the parameters are updated by $\Delta\vec{\Pi}$. However, updating the parameter by $\Delta\vec{\Pi}$ simultaneously can result in worse and diverging approximants. To avoid this, a similar method as described by Horváth and Telek (2002) is used to search for the optimal direction inside the linearized area of changes as specified by $\Delta\vec{\Pi}$.

As the last step of the iteration, the curve (namely, the PDF of the acyclic phase-type distribution) represented by ($\vec{\alpha}$,**A**) is updated as directed by changes to the vector of parameters $\vec{\Pi}$ through the function Update-APH($\vec{\Pi}$). This function ensures that the structure of the ordered bidiagonal canonical form is maintained, namely that:

1. $\alpha_1 > 0$ and $\alpha_i \geq 0$ for $2 \leq i \leq n$,
2. $\sum_{i=1}^{n} \alpha_i = 1 - \alpha_{n+1}$,
3. $A_i > 0$ for $1 \leq i \leq n$, and
4. $A_n \geq A_{n-1} \geq \cdots \geq A_2 \geq A_1$.

5. Results and discussions

A prototype of the proposed orthogonal-distance fitting method for acyclic phase-type distributions has been implemented in C++. The prototype requires a special implementation of sparse matrix operations tailored to the first canonical form of acyclic phase-type distributions.

Six goodness-of-fit measures, as specified in (Bobbio and Telek 1994), are used to evaluate the performance of the proposed method. The measures are presented in Table 1 In the table, $F(t)$ and $f(t)$ are the cumulative distribution function (CDF) and the probability density function (PDF) of the phase-type distribution (curve), respectively, $\tilde{F}(t)$ and $\tilde{f}(t)$ are the CDF and the PDF of the approximated distribution (trace), respectively, $c_1(F)$ is the mean, $c_2(F)$ is the variance, and $c_3(F)$ is the third central moment of $F(t)$.

For the test cases of distributions to be fitted, we use the seven distributions whose traces are provided by the tool G-Fit (Thümmler *et al*. 2006). Each of the seven distributions is represented by 10,000 points in the PDF domain. The test cases are quite similar to the test cases in (Bobbio

Table 1 Goodness-of-fit measures (Bobbio and Telek 1994)

Measure	Definition
Relative error in the 1st moment	$e_1 = \dfrac{\lvert c_1(F) - c_1(\tilde{F})\rvert}{c_1(\tilde{F})}$
Relative error in the 2nd moment	$e_2 = \dfrac{\lvert c_2(F) - c_2(\tilde{F})\rvert}{c_2(\tilde{F})}$
Relative error in the 3rd moment	$e_3 = \dfrac{\lvert c_3(F) - c_3(\tilde{F})\rvert}{c_3(\tilde{F})}$
PDF absolute area difference	$D_{pdf} = \int_0^\infty \lvert f(t) - \tilde{f}(t)\rvert dt$
CDF absolute area difference	$D_{cdf} = \int_0^\infty \lvert F(t) - \tilde{F}(t)\rvert dt$
Minus the cross entropy	$-H = \int_0^\infty \log(f(t)) d\tilde{F}(t)$

Table 2 Benchmark probability distributions used to generate the traces

Distribution	Probability density function
Weibull$(a,b), a>0, b>0$	$f(t;a,b)=\frac{b}{a}\left(\frac{t}{a}\right)^{b-1}e^{-\left(\frac{t}{a}\right)^{b}}$
ParetoII$(a,b), a>0, b>0$	$f(t;a,b)=\frac{b^{a}e^{-b/t}}{\Gamma(a)}t^{-a-1}$
Uniform$(a,b), -\infty<a<b<\infty$	$f(t;a,b)=\begin{cases}\frac{1}{b-a}, & a\le t\le b\\ 0, & \text{otherwise}\end{cases}$
Lognormal$(a,b), -\infty<a<\infty, b>0$	$f(t;a,b)=\frac{1}{tb\sqrt{2\pi}}e^{-\frac{(\ln(t)-a)^{2}}{2b^{2}}}$
Shifted exponential	$f(t;a,b)=\begin{cases}\frac{1}{2}e^{-t}, & 0\le t<1\\ \frac{1}{2}e^{-t}+\frac{1}{2}e^{-(t-1)}, & t\ge 1\end{cases}$
Matrix exponential	$f(t)=\left(1+\frac{1}{(2\pi)^{2}}\right)(1-\cos(2\pi t))e^{-t}$

and Telek 1994), but for some changes in the parameters of the distributions. Specifically, we use Weibull (1.0,0.5), Weibull (1.0,5.0), ParetoII (1.5,2.0), Uniform (0.5,1.5), Lognormal (1.0,1.5), Shifted exponential, and Matrix exponential, as specified in Table 2.

To evaluate the goodness-of-fit and the performance of the prototype, we compare it with G-Fit according to the benchmark goodness-of-fit measures by fitting phase-type distributions to the traces of the test cases. We decide to compare the proposed method with G-Fit because, according to Thümmler *et al*. (2006), its goodness-of-fit and performance were better than the other tools. Furthermore, even with the latest advancements in phase-type fitting tools (Esparza 2011, Reinecke *et al*. 2012, 2013, Horváth and Okamura 2013, Buchholz *et al*. 2014, Okamura and Dohi 2016), G-Fit remains the best in both measures.

In the comparison, each test case distribution is fitted by both G-Fit (GF) and our prototypical tool (OD) to phase-type distributions of size 4, 8 and 16 phases. The resulting goodness-of-fit of both tools can be observed in Tables 3, 4 and 5, where the approximating phase-type distributions are of size 4, 8 and 16 phases, respectively. Table 6 shows the performance of both tools in terms of the number of required iterations to convergence and the required computation time.

The relative errors in the first, second, and third moments are presented in the first six rows of Tables 3, 4, and 5. Both G-Fit and the proposed method produce consistent relative errors of moments, in the sense that the relative errors decrease as the number of phases increases. Aside from this, a few patterns can be further observed. In all experiments, G-Fit shows better results in the relative errors of the first moment than the proposed method. G-Fit produces fitting with almost no first moment error and approximates the mean of the empirical distribution well. In

Table 3 Goodness-of-fit comparison of G-Fit and the proposed method (4 phases)

Measure	Wei. 1	Wei. 2	Pareto	Uniform	Lognorm.	Sh. exp.	Mat. exp.
1st moment (GF)	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1st moment (OD)	0.00000	0.00000	0.00056	0.00241	0.00000	0.00492	0.00301
2nd moment (GF)	0.21083	3.79841	0.63891	1.94629	0.16833	0.00424	0.02412
2nd moment (OD)	0.20021	2.90156	1.47393	2.82377	0.16720	0.62119	0.41782
3rd moment (GF)	0.64104	-45.857	0.90748	346.243	0.64772	0.12340	0.00615
3rd moment (OD)	0.57091	-43.218	1.95679	403.451	0.63892	0.79129	0.58931
PDF area diff. (GF)	1.64613	2.20822	1.33205	2.10464	1.06309	1.27079	1.44163
PDF area diff. (OD)	1.58321	2.19320	1.06481	3.23901	1.41832	1.10547	1.28494
CDF area diff. (GF)	0.13909	0.19692	0.49026	0.15374	0.04399	0.06527	0.08439
CDF area diff. (OD)	0.57321	0.21029	0.53561	0.12531	0.11782	0.11024	0.20612
Cross entropy (GF)	0.29979	-1.84399	-1.62523	0.00799	-0.40339	-1.10627	-0.90572
Cross entropy (OD)	0.16121	-2.19054	-2.05631	0.00319	-0.61980	-1.57199	-1.14198

Table 4 Goodness-of-fit comparison of G-Fit and the proposed method (8 phases)

Measure	Wei. 1	Wei. 2	Pareto	Uniform	Lognorm.	Sh. exp.	Mat. exp.
1st moment (GF)	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1st moment (OD)	0.00000	0.00000	0.00000	0.00161	0.00000	0.00251	0.00135
2nd moment (GF)	0.08566	1.39920	0.38259	0.47315	0.07857	0.03636	0.02665
2nd moment (OD)	0.07189	1.29430	1.12761	1.16731	0.07729	0.44091	0.33931
3rd moment (GF)	0.40724	-12.214	0.64704	85.8107	0.34642	0.16784	0.12418
3rd moment (OD)	0.36831	-11.529	1.39811	97.9540	0.32981	1.38561	1.17658
PDF area diff. (GF)	8.02902	2.30474	1.34247	2.05356	1.07779	1.32691	1.51665
PDF area diff. (OD)	5.87231	2.10723	1.10542	2.87490	2.18965	1.20551	1.10744
CDF area diff. (GF)	0.04143	0.09376	0.27928	0.05416	0.01537	0.04099	0.06438
CDF area diff. (OD)	0.09710	0.32712	0.89261	0.05098	0.20951	0.08497	0.10617
Cross entropy (GF)	9.19468	-1.75304	-1.59501	0.01207	-0.39229	-1.12367	-0.88617
Cross entropy (OD)	7.18732	-1.56390	-1.70319	0.00569	-0.59120	-1.73109	-2.12018

the second and third moments, the proposed method produces smaller relative errors for Weibull 1, Weibull 2, and Lognormal test cases, while G-Fit performs better for the other four test cases. This is also consistent with when the approximating acyclic phase-type distributions are set to have 8 or 16 phases. However, it must be noted that in the second and third moments, the relative errors of the proposed method can be orders of magnitude larger than those of G-Fit for the cases of Shifted exponential and Matrix exponential.

The PDF absolute area difference measure is especially important since the proposed method, in essence, performs approximation by minimizing the orthogonal distance between every point in the trace to the closest point in the curve of the PDF. This minimization is also supposed to proportionally minimize the area below the PDF curve. However, even for this measure, the proposed method cannot become the absolute winner. The proposed method performs relatively

Table 5 Goodness-of-fit comparison of G-Fit and the proposed method (16 phases)

Measure	Wei. 1	Wei. 2	Pareto	Uniform	Lognorm.	Sh. exp.	Mat. exp.
1st moment (GF)	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1st moment (OD)	0.00000	0.00000	0.00000	0.00082	0.00000	0.00076	0.00041
2nd moment (GF)	0.06712	0.19960	0.12971	0.07138	0.05158	0.00610	0.03769
2nd moment (OD)	0.05679	0.17514	0.30287	0.22901	0.05076	0.08910	0.16230
3rd moment (GF)	0.34897	-3.80358	0.19095	44.9169	0.23726	0.01043	0.14837
3rd moment (OD)	0.31543	-3.17534	0.34082	50.5929	0.22876	0.10871	0.20459
PDF area diff. (GF)	9.72018	2.43744	1.35283	2.08429	1.08734	1.36746	1.57195
PDF area diff. (OD)	8.12785	2.08356	1.20844	3.70891	1.48703	1.10875	1.39753
CDF area diff. (GF)	0.03549	0.02953	0.08281	0.05009	0.01070	0.02104	0.05460
CDF area diff. (OD)	0.09276	0.12289	0.09719	0.10851	0.06719	0.14813	0.06571
Cross entropy (GF)	11.5548	-1.61246	-1.61012	0.01373	-0.39081	-1.14263	-0.89862
Cross entropy (OD)	10.2398	-1.81921	-2.01876	0.00498	-0.71093	-1.91034	-1.02984

Table 6 Performance comparison of G-Fit and the proposed method

Measure	Distribution	4 phases		8 phases		16 phases	
		GF	OD	GF	OD	GF	OD
Number of iterations	Weibull (1.0,0.5)	393	25	3813	25	79700	30
	Weibull (1.0,5.0)	78	18	808	20	16565	21
	ParetoII (1.5,2.0)	116	22	1661	20	39989	27
	Uniform (0.5,1.5)	81	26	849	23	17972	29
	Lognormal (1.0,1.5)	302	26	3480	32	61775	30
	Shifted exponential	360	36	3039	36	52463	32
	Matrix exponential	150	27	2031	25	35261	29
Computation times (seconds)	Weibull (1.0,0.5)	3	899	35	1862	765	4064
	Weibull (1.0,5.0)	1	719	5	1612	124	2905
	ParetoII (1.5,2.0)	1	928	11	1406	472	3739
	Uniform (0.5,1.5)	1	677	6	1365	143	3746
	Lognormal (1.0,1.5)	2	765	22	1934	590	3343
	Shifted exponential	1	1242	17	2582	461	4890
	Matrix exponential	1	620	13	1570	302	3319

better than G-Fit for test cases Weibull 1, Weibull 2, Pareto II, Shifted exponential and Matrix exponential. On the other hand, G-Fit produces smaller PDF absolute area differences for Uniform and Lognormal test cases. This result is consistent for the cases of 4, 8 or 16 phases.

For the last two measures, the CDF absolute area difference and negative cross entropy, the G-Fit performs better in almost all test cases. The proposed method produces smaller CDF absolute area differences for the Uniform distribution test case, but only when the approximating distribution is of phase 4 or 8. When the phase is 16, the proposed method again produces larger CDF absolute area differences.

Table 6 provides the comparison of G-Fit and the proposed method in terms of computational performance. The table immediately reveals that the proposed method requires far more computation time in each iteration than G-Fit. On average, each iteration consumes around 33 seconds when the proposed method is fitting a 4-phase phase-type distribution to the traces. Most of the computation effort is spent in determining the closest orthogonal distance between the points of the approximating and the approximated distributions. The proposed method requires 16 to 32 iterations to perform the fitting. In more than half of the cases, however, the iterations are terminated because optimal directions cannot be found anymore in the linearized area of the parameters' updates. These cases usually arise when the minimization procedure results in local minima.

G-Fit, on the other hand, consumes much less amount of time per iteration compared with the proposed method. This is because the parameters' updates are not computed iteratively but in closed form. G-Fit fits hyper-Erlang distributions to empirical distributions. Hyper-Erlang distributions are basically mixtures of Erlang distributions. However, for a given number of states, there are many ways of mixing Erlang distributions. For 4, 8 and 16 phases, the numbers of possible mixtures of Erlang distributions are 5, 22 and 231, respectively. This explains the high number of iterations in the G-Fit. On average, for the 16-phase cases, G-Fit requires around 187 iterations to fit a particular mixture of Erlang distributions to the traces.

6. Case study

To demonstrate the applicability of the proposed method, we revisit and extend a case study we previously presented in (Pulungan and Hermanns 2009), which was based on the models developed by Meester and Muns (2007). Fig. 2 depicts a part of the Netherlands' intercity railway network connecting 10 cities and Table 7 provides relevant information on the lines and segments in the network. We are interested in studying the propagation of delays experienced on each line's segments at the line's end in both directions. To this end, we make assumptions on the delays, model them by Erlang, Weibull, and Uniform distributions, and then fit them to APH distributions using the proposed method.

Especially important to note from the figure is that some line segments become feeders to connecting trains on other line segments. There are seven such synchronizations, namely 1-01 to 3-02, 1-52 to 3-53, 2-01 to 3-03, 3-01 to 1-02, 3-51 to 2-54, 3-51 to 2-02, and 3-52 to 1-53. Aside from the delays accrued up the line, the departure of the connecting trains is further constrained by the arrival of the feeder trains because passengers may need to transfer. The complication of the model arises from these synchronizations.

Let D be a random variable governing the departure delay at the beginning of a segment, and T be a random variable representing the segment's actual travel time subtracted by its minimum travel time t (in minutes). Note that when $T - r > 0$, the maximum buffer time r (in minutes) is exceeded, and a delay occurs. In our experiments, each of the 10 distinct values of r (cf. Fig. 2) will be modeled by a distinct random variable T.

Let $[x]^+ = \max\{x, 0\}$. The departure delay of the next segment in a line is $[D + T - r]^+$. For synchronization cases, the departure delay of the next segment after the feeder f is $[D_f + T_f - r]^+$, while the departure delay of the connecting segment c is given by $\max\{[D_f + T_f - r]^+, [\mathrm{D_c} + \mathrm{T_c} - r]^+\}$. All those random variables are approximated by (fitted to) APH distributions. Addition and maximum operations on the random variables can be accomplished by convolution and maximum

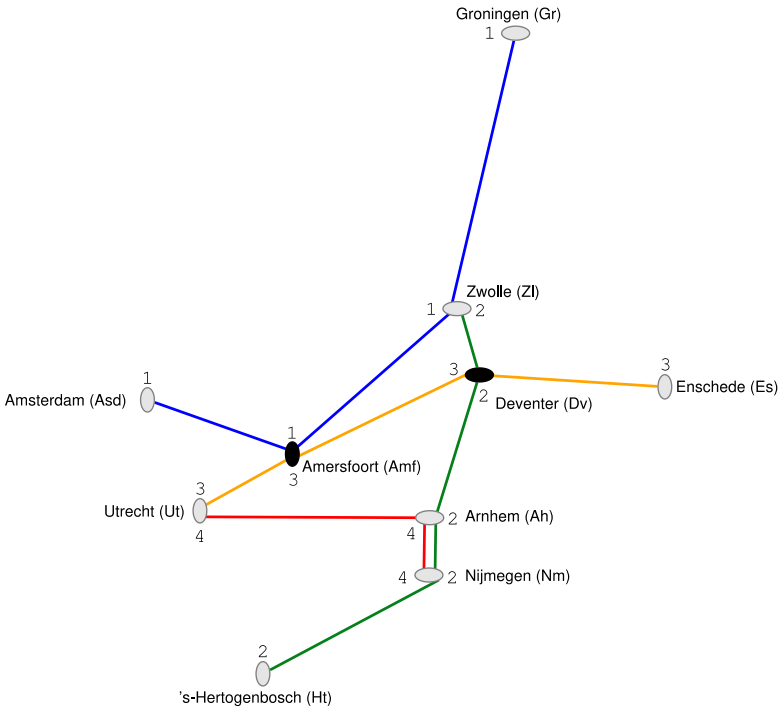

Fig. 2 The intercity railway network connecting Amersfoort, Amsterdam, Arnhem, Deventer, Enschede, Groningen, Neijmegen, 's-Hertogenbosch, Utrecht, and Zwolle

operations on APH distributions, respectively. Moreover, if $(\vec{\alpha},\mathbf{A})$ is the APH distribution governing random variable X, the APH distribution of $[X-r]^+$ is given by $(\vec{\alpha}e^{\mathbf{A}r},\mathbf{A})$.

In this case study, we experiment on three types of delay models governing the random variables T, using Erlang, Weibull, and Uniform distributions. In the Erlang experiment, for a specific number of phases, the Erlang rate is adjusted such that r is around the 80-th percentile of the Erlang distribution. In the Weibull experiment, the shape parameter of the Weibull distribution is set to 2.4 when r is 1, 2, or 3; to 1.8 when r is 4, 5, or 6; and to 1.4 when r is 7, 8, 9, or 11. The scale parameter, on the other hand, is adjusted in a similar fashion, namely such that r is around the 80-th percentile of the Weibull distribution. In the Uniform experiment, the random variables T are governed by Uniform (0, r). Since Erlang distributions are a subset of APH distributions, they require no fitting. The Weibull and Uniform distributions, however, are first fitted to APH distributions by using the proposed method before the transient analysis of the resulting delay models.

Table 8 presents the results of our experiments on varying the delay distribution models and the number of phases of APH distributions approximating those delay distributions. The table presents the resulting probability that the overall delay when a train arrives at the end of the line (in both directions) is no more than 3 minutes ($\Pr(A \leq 3)$) for the six different lines. The results indicate that the Weibull delay model produces the largest probabilities almost in all cases, followed by Erlang and then Uniform models. Furthermore, varying the number of phases does not readily result in a discernible impact on the probability. Nevertheless, all delay models agree that the probability that the overall delay is no more than 3 minutes is quite large, more than 85% in all

Table 7 Lines and segments joining the 10 cities in the railway network, and each segment's minimum travel time *t* and maximum buffer time *r*

Line	Segment	Origin	Destination	t	r
1	01	Amsterdam	Amersfoort	29	5
	02	Amersfoort	Zwolle	32	5
	03	Zwolle	Groningen	59	11
	51	Groningen	Zwolle	59	9
	52	Zwolle	Amersfoort	32	6
	53	Amersfoort	Amsterdam	28	8
2	01	Zwolle	Deventer	17	2
	02	Deventer	Arnhem	32	3
	03	Arnhem	Nijmegen	14	1
	04	Nijmegen	's-Hertogenbosch	26	8
	51	's-Hertogenbosch	Nijmegen	25	3
	52	Nijmegen	Arnhem	13	1
	53	Arnhem	Deventer	32	3
	54	Deventer	Zwolle	18	7
3	01	Utrecht	Amersfoort	13	3
	02	Amersfoort	Deventer	33	4
	03	Deventer	Enschede	39	9
	51	Enschede	Deventer	39	5
	52	Deventer	Amersfoort	36	5
	53	Amersfoort	Utrecht	14	6

Table 8 Probability that a line's overall delay is no more than 3 minutes ($\Pr(A \leq 3)$) for different delay models (Erlang, Weibull, and Uniform) and different phases of the approximating APH

APH phases	Line segment	$\Pr(A \leq 3)$		
		Erlang	Weibull	Uniform
10	1-03	0.94460584	0.98750249	0.85562310
	1-53	0.94866633	0.98473806	0.85758391
	2-04	0.97077302	0.98481936	0.86661845
	2-54	0.97382524	0.98495819	0.86674911
	3-03	0.95994336	0.98787838	0.86539341
	3-53	0.96052168	0.96486199	0.86434618
20	1-03	0.97260921	0.97636953	0.86558931
	1-53	0.97888762	0.97582951	0.86712758
	2-04	0.99010732	0.98919711	0.87830552
	2-54	0.99239673	0.98239760	0.87887365
	3-03	0.98340104	0.97321550	0.87346501
	3-53	0.98628849	0.94047354	0.87475398

cases, even when delays associated with transferring passengers to connecting trains are considered. This bespeaks the efficiency of the Netherlands' trains and railway network.

7. Conclusions

In this paper, we have reported on our investigation in applying expectation-minimization-based orthogonal distance fitting to acyclic phase-type distributions. The resulting method minimizes the orthogonal distances of points in the curve of the approximating phase-type distribution and points in the empirical distribution. The minimization problem is turned into a non-linear optimization problem, which is then solved by an iterative linearization method. In every expectation-minimization iteration, the proposed method must search for the (orthogonally) closest point in the approximating curve to each point in the trace. This search turns out to be computationally substantial.

We have also compared the resulting fitting tool with the currently best fitting tool. Initially, we had expected that, although the computational requirement of orthogonal distance fitting is relatively heavier than other existing methods for the same number of phases, the orthogonal distance fitting would converge faster. The comparison produces no clear-cut judgment. It reveals that the orthogonal distance for fitting phase-type distributions to traces of probability distributions is not superior and only beneficial in certain cases.

Nevertheless, based on our experience in implementing the proposed method, we are still convinced of its potential and see several possible avenues for improving it further:

1. Investigating the best initialization of the parameters. Our experience indicates that initializing the parameters randomly often results in being trapped in local optima while initializing them by Erlang distributions results in sluggish convergence. The latter is still a viable method since it avoids randomness. However, it must be augmented by strategies to quicken convergence.

2. Improving the method for the simultaneous updating of the parameters. Since ordered bidiagonal representations impose a particular structure on the parameters, parameter changes specified by $\Delta\vec{\Pi}$ cannot always be readily applied. Devising an optimal updating strategy will greatly improve the overall fitting method.

3. Optimizing the computation of the minimum distance points.

These are for future endeavors. Since fitting is usually a pre-processing step, in which an approximating model is computed, heavy and time-consuming computation is still well invested in this step if it results in better and closer models. Further tuning in the proposed method is worth investigating.

Acknowledgments

This work was supported by Hibah Penelitian Dosen Departemen Ilmu Komputer dan Elektronika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada. This work was also partially supported by the ERC Advanced Investigators Grant 695614 (POWVER), by the German Research Foundation (DFG) under grant No. 389792660, as part of TRR 248, see https://perspicuous-computing.science, and by the Key-Area Research and Development Program of Guangdong Province (Grant 2018B010107004).

References

Ahn, S.J. (2004), *Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space*, volume 3151 of *Lecture Notes in Computer Science*, Springer, Heidelberg, Germany.

Al-Khalidi, H.R. and Schnell, D.J. (1997), "Application of a continuous-time Markov chain to a preclinical study", *Drug Inf. J.*, **31**(2), 607-613. https://doi.org/10.1177/009286159703100228.

Alam, J., Kim, D. and Choi, B. (2017), "Seismic probabilistic risk assessment of weir structures considering the earthquake hazard in the Korean Peninsula", *Earthq. Struct.*, **13**(4), 421-427. https://doi.org/10.12989/EAS.2017.13.4.421.

Asmussen, S., Nerman, O. and Olsson, M. (1996), "Fitting phase-type distributions via the EM algorithm", *Scand. J. Stat.*, **23**(4), 419-441. https://www.jstor.org/stable/4616418.

Bobbio, A. and Cumani, A. (1992), "ML estimation of the parameters of a PH distribution in triangular canonical form", *Proceedings of the 5th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS)*, Torino, Italy, February.

Bobbio, A., Cumani, A., Premoli, A. and Saracco, O. (1980), "Modeling and identification of non-exponential distributions by homogeneous Markov processes", *Proceedings of the 6th Advances in Reliability Technology Symposium*, West Yorkshire, United Kingdom, April.

Bobbio, A. and Telek, M. (1994), "A benchmark for PH estimation algorithms: results for acyclic-PH", *Commun. Stat. Stochastic Models*, **10**(3), 661-677. https://doi.org/10.1080/15326349408807315.

Böde, E., Herbstritt, M., Hermanns, H., Johr, S., Peikenkamp, T., Pulungan, R., Rakow, J., Wimmer, R. and Becker, B. (2009), "Compositional dependability evaluation for STATEMATE", *IEEE T. Software Eng.*, **35**(2), 274-292. https://doi.org/10.1109/TSE.2008.102.

Buchholz, P., Kriege, J. and Felko, I. (2014), *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*, Springer, Heidelberg, Germany.

Chavan, S.G. and Lal, A. (2018), "Stochastic bending characteristics of finite element modeled nano-composite plates", *Steel Compos. Struct.*, **26**(1), 1-15. https://doi.org/10.12989/SCS.2018.26.1.001.

Cumani, A. (1982), "On the canonical representation of homogeneous Markov processes modelling failure-time distributions", *Microelectron. Reliab.*, **22**(3), 583-602. https://doi.org/10.1016/0026-2714(82)90033-6.

Esparza, L.J.R. (2011), "Maximum likelihood estimation of phase-type distributions", Ph.D. Dissertation; Technical University of Denmark, Kongens Lyngby, Denmark.

Feldmann, A. and Whitt, W. (1998), "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models", *Perform. Evaluation*, **31**(3), 245-279. https://doi.org/10.1016/S0166-5316(97)00003-5.

Gross, D. and Miller, D.R. (1984), "The randomization technique as a modeling tool and solution procedure for transient Markov processes", *Oper. Res.*, **32**(2), 343-361. https://doi.org/10.1287/opre.32.2.343.

Horváth, A. and Telek, M. (2002), "PhFit: a general phase-type fitting tool", *Proceedings of the 12th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS)*, London, United Kingdom, April.

Horváth, G. and Okamura, H. (2013), "A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure", *Proceedings of the 10th European Workshop on Performance Engineering (EPEW)*, Venice, Italy, September

Jensen, A. (1953), "Markoff chains as an aid in the study of Markoff processes", *Scand. Actuar. J.*, **1953**(sup1), 87-91. https://doi.org/10.1080/03461238.1953.10419459.

Johnson, M.A. and Taaffe, M.R. (1988), "The denseness of phase distributions", Research Memorandum No. 88-20; School of Industrial Engineering, Purdue University, IN, USA.

Meester, L.E. and Muns, S. (2007), "Stochastic delay propagation in railway networks and phase-type distributions", *Transport. Res. B-Meth.*, **41**(2), 218-230. https://doi.org/10.1016/j.trb.2006.02.007.

Neuts, M.F. (1994), *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Dover Publications, New York, NY, USA.

Okamura, H. and Dohi, T. (2016), "Fitting phase-type distributions and Markovian arrival processes: algorithms and tools", *Principles of Performance and Reliability Modeling and Evaluation: Essays in Honor of Kishor Trivedi on his 70th Birthday*, Springer, Heidelberg, Germany, April.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007), *Numerical Recipes: The Art of Scientific Computing, (3rd Edition)*, Cambridge University Press, New York, NY, USA.

Pulungan, R. and Hermanns, H. (2009), "Acyclic minimality by construction-almost", *2009 Proceedings of the 6th International Conference on the Quantitative Evaluation of Systems (QEST)*, Budapest, Hungary, September.

Reinecke, P., Krauß, T. and Wolter, K. (2012), "Cluster-based fitting of phase-type distributions to empirical data", *Comput. Math. Appl.*, **64**(12), 3840-3851. https://doi.org/10.1016/j.camwa.2012.03.016.

Reinecke, P., Krauß, T. and Wolter, K. (2013), "Phase-type fitting using HyperStar", *Proceedings of the 10th European Workshop on Performance Engineering (EPEW)*, Venice, Italy, September.

Sandoval, C.A.O., Tizani, W. and Koch, C. (2018), "A method for discrete event simulation and building information modelling integration using a game engine", *Adv. Comput. Des.*, **3**(4), 405-418. https://doi.org/10.12989/ACD.2018.3.4.405.

Stewart, W.J. (2009), *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling*, Princeton University Press, Princeton, NJ, USA.

Thümmler, A., Buchholz, P. and Telek, M. (2006), "A novel approach for phase-type fitting with the EM algorithm", *IEEE T. Depend. Secure*, **3**(3), 245-258. https://doi.org/10.1016/10.1109/TDSC.2006.27.

WT