Computers and Mathematics with Applications 64 (2012) 3840–3851

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Cluster-based fitting of phase-type distributions to empirical data

Philipp Reinecke *, Tilman Krauß, Katinka Wolter

Institute of Computer Science, Freie Universität Berlin, Berlin, Germany

ARTICLE INFO

Keywords:
Phase-type distribution
PH fitting
Data modelling

ABSTRACT

We present a clustering-based fitting approach for phase-type distributions that is particularly suited to capture common characteristics of empirical data sets. The distributions fitted by this approach are especially useful in efficient simulation approaches. We describe the Hyper-* tool, which implements the algorithm and offers a user-friendly interface to efficient phase-type fitting. We provide a comparison of cluster-based fitting with segmentation-based approaches and other algorithms and show that clustering provides good results for typical empirical data sets.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Accurate data modelling is required in a wide range of scientific disciplines. Phase-type (PH) distributions are a very useful tool in modelling measurement data. They combine high flexibility in fitting data with nice properties in application. In particular, in performance and dependability evaluation, PH distributions enable efficient evaluation methods, as they allow fast generation of random variates as well as analytical approaches.

Approaches for fitting PH distributions to data include moment-matching [1], non-linear optimisation [2] and expectation–maximisation (EM) algorithms [3,4]. Partitioning of the data set can improve fitting quality. In particular, the family of segmentation-based approaches (e.g. [4]) aims to capture oscillations in the density as well as heavy-tailed behaviour by splitting the data set into segments with low variability, which can be fitted by simple distributions using the EM algorithm.

In this paper, we propose the use of clustering to detect important features of the density and to partition the data set such that clusters contain samples belonging to the same feature. Clusters are fitted by distributions with a simple structure, which then form the branches of the overall PH model. This approach yields results that fit the density well even with data sets that are difficult to fit by a PH distribution. It produces mixtures of distributions, which enable efficient algorithms for random-variate generation [5,6]. Furthermore, with clustering the user can control the quality of the fit by setting initial cluster centres on a plot of the histogram, which is a very intuitive way of adjusting parameters of the fitting algorithm.

This paper is structured as follows. We first introduce some basic concepts of phase-type distributions. In Section 3, we describe our cluster-based approach to PH fitting. Section 4 places our algorithm in the context of other methods, with a particular focus on approaches that partition the data prior to fitting. Section 5 gives an overview of our implementation of cluster-based fitting in the extensible Hyper-* tool. We evaluate the method in Section 6 with three typical data sets, employing standard quality measures for PH distributions, such as log-likelihood values and moment errors to assess fitting quality. We provide a comparison to previous partitioning approaches as well as to two standard tools for PH fitting. Section 7 concludes the paper with an outlook on future work.

* Corresponding author.
E-mail addresses: philipp.reinecke@fu-berlin.de (P. Reinecke), tilman.kraussfg@fu-berlin.de (T. Krauß), katinka.wolter@fu-berlin.de (K. Wolter).

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.03.016

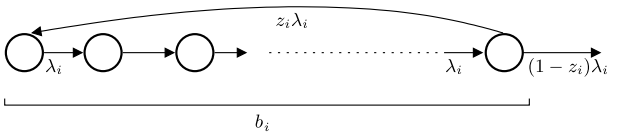

Fig. 1. Structure of a feedback-erlang block.

2. Mathematical background

Continuous phase-type (PH) distributions are defined as the distribution of time to absorption in a Continuous-Time Markov Chain (CTMC) with one absorbing state [7]. They are commonly represented by a vector-matrix tuple $(\boldsymbol{\alpha}, \mathbf{Q})$, where

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} -\lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & -\lambda_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n} \tag{1}$$

with $\lambda_{ij} \geq 0$, $\lambda_{ii} > 0$, $\mathbf{Q1I} \leq \mathbf{0}$, $\mathbf{Q}$ is non-singular, and $\boldsymbol{\alpha}\mathbf{1I} = 1$, where $\mathbf{1I}$ is the column vector of ones of the appropriate size.

Definition 1. Let $(\boldsymbol{\alpha}, \mathbf{Q})$ be a representation of a phase-type distribution. The *size* of the representation is n. The *probability density function (PDF)*, *cumulative distribution function (CDF)*, and kth *moment*, respectively, are defined as follows [7,2,1]:

$$f(t) = \boldsymbol{\alpha} e^{\mathbf{Q}t}(-\mathbf{Q1I}), \tag{2}$$

$$F(t) = 1 - \boldsymbol{\alpha} e^{\mathbf{Q}t}\mathbf{1I}, \tag{3}$$

$$E\left[X^k\right] = k!\boldsymbol{\alpha}(-\mathbf{Q})^{-k}\mathbf{1I}.. \tag{4}$$

Fitting algorithms for PH distributions often benefit from special structures of the tuple $(\boldsymbol{\alpha}, \mathbf{Q})$, and special structures may also enable more efficient simulation [5,8,6].

Definition 2. Let $(\boldsymbol{\alpha}, \mathbf{Q}_c)$ and $(\boldsymbol{\alpha}, \mathbf{Q}_b)$ with

$$\mathbf{Q}_c = \begin{pmatrix} \mathbf{Q}_1 & -\mathbf{Q}_1\mathbf{1}\mathbf{e}_1 & & \\ & \ddots & \ddots & \\ & & & -\mathbf{Q}_{m-1}\mathbf{1}\mathbf{e}_1 \\ & & & \mathbf{Q}_m \end{pmatrix} \quad \text{and} \quad \mathbf{Q}_b = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{0} & & \\ & \ddots & \ddots & \\ & & & \mathbf{0} \\ & & & \mathbf{Q}_m \end{pmatrix}, \tag{5}$$

be representations of size n of a phase-type distribution and let $\mathbf{Q}_i \in \mathbb{R}^{b_i\times b_i}$, $i = 1, \ldots, m$ be square matrices such that $\sum_{i=1}^{m} b_i = n$. Then, $\mathbf{Q}_c$ is in *chain structure*, while $\mathbf{Q}_b$ is in *branch structure*.

Typically, all of the block matrices $\mathbf{Q}_i$ have the same structure. Two important chain structures are the CF-1 form for Acyclic Phase-type distributions [9] and the Monocyclic form for general PH distributions [10]. In the CF-1 form, all block matrices $\mathbf{Q}_i$ are of size 1, while in the Monocyclic form the block matrices are given by Feedback-Erlang blocks, defined as follows:

Definition 3 (*[10]*)**.** A *Feedback-Erlang (FE) block* (Fig. 1) is defined by the tuple (b, λ, z), where $\lambda > 0$, $b \geq 1$, and $z \in [0, 1)$. The associated matrix $\mathbf{F} \in \mathbb{R}^{b\times b}$ has the following structure:

$$\mathbf{F} = \begin{pmatrix} -\lambda & \lambda & & \\ & \ddots & \ddots & \\ & & & \lambda \\ z\lambda & & & -\lambda \end{pmatrix}. \tag{6}$$

In the CF-1 and Monocyclic forms the block matrices $\mathbf{Q}_i$ are ordered by increasing absolute value of the dominant eigenvalue. The CF-1 form is canonical for the Acyclic Phase-type (APH) class, which is a true sub-class of general PH distributions. The Monocyclic form is canonical for the PH class. Thus every APH distribution has a representation in CF-1 form [9], and every PH distribution has a representation in Monocyclic form [10] (the CF-1 form is a special case of the Monocyclic form). Canonical forms are attractive both for fitting and for simulation since they require a low number of parameters and have a simple structure.

Download English Version:

https://daneshyari.com/en/article/468327

Download Persian Version:

https://daneshyari.com/article/468327

Daneshyari.com