An Efficient Brute Force Approach to Fit Finite Mixture Distributions

Falko Bause
LS Informatik IV
Department of Computer Science
TU Dortmund
44221 Dortmund
Germany
E-Mail: falko.bause@cs.tu-dortmund.de

Author prepared version of a paper published in Hermanns H. (eds) Measurement, Modelling and Evaluation of Computing Systems. MMB 2020. Lecture Notes in Computer Science, vol 12040. Springer, Cham

The final authenticated version is available online at `https://doi.org/10.1007/978-3-030-43024-5_13`

An Efficient Brute Force Approach to Fit Finite Mixture Distributions

Falko Bause

LS Informatik IV
Department of Computer Science
TU Dortmund
44221 Dortmund
Germany
falko.bause@cs.tu-dortmund.de

Abstract. This paper presents a brute force approach to fit finite mixtures of distributions considering the empirical probability density and cumulative distribution functions as well as the empirical moments. The fitting problem is solved using a non-negative least squares method determining a mixture from a larger set of distributions.
The approach is experimentally validated for finite mixtures of Erlang distributions. The results show that a feasible number of component distributions, which accurately fit to the empirical data, is obtained within a short CPU time.

Keywords: Mixture Distributions · Hyper-Erlang Distributions · Non-Negative Least Squares · Farey Sequences

1 Introduction

Mixture distributions are a well explored model type for the description of statistically varying events. In this paper, we focus on the fitting of continuous univariate finite mixture distributions and assume that all probability density functions and moments do exist. Mixture distributions are usually defined by a set of $G, G \in \mathbb{N}$, component distributions specified by their probability density functions (PDFs) $f_i(x|\boldsymbol{\theta}_i)$ with $\boldsymbol{\theta}_i \in \mathbb{R}^{m_i}, m_i \in \mathbb{N}$, denoting the component-specific parameters and mixing probabilities $\pi_i \in [0,1], i = 1, \ldots, G$ satisfying $\sum_{i=1}^{G} \pi_i = 1$ [14]. The PDF of the mixture distribution is defined by

$$f(x|(\boldsymbol{\pi}, \boldsymbol{\theta})) = \sum_{i=1}^{G} \pi_i f_i(x|\boldsymbol{\theta}_i) \tag{1}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ the vector containing all parameters and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. Additionally, the cumulative distribution function (CDF) F and the moments

$E[X^j]$ are given by a convex combination of the components' counterparts:

$$F(x|(\boldsymbol{\pi},\boldsymbol{\theta})) = \sum_{i=1}^{G} \pi_i F_i(x|\boldsymbol{\theta}_i) \tag{2}$$

$$E[X^j|(\boldsymbol{\pi},\boldsymbol{\theta})] = \sum_{i=1}^{G} \pi_i E[X_i^j|\boldsymbol{\theta}_i], \quad j \in \mathbb{N} \tag{3}$$

where X and X_i denote the random variables with CDFs F and F_i, respectively.

For performance modeling phase-type distributions (PHDs, [24,25]) are popular, since they allow analytical analysis approaches. Different from Eq. (1) general PHDs are usually specified in a more compact notation, since component parameters might be dependent. Common subclasses of PHDs are mixtures of exponential (hyper-exponential) and mixtures of Erlang (hyper-Erlang) distributions, especially since hyper-Erlang distributions can approximate any PDF of a nonnegative random variable [12]. In practice, applicability of mixture distributions depends on efficient fitting procedures which construct a mixture distribution approximating an empirical distribution given by trace data $T = (t_1, \ldots, t_n), t_i \in \mathbb{R}$. There exists a vast number of literature on fitting mixture distributions, see, e.g., [9,15,33] for an overview. In the following only a sketch is presented emphasizing those from a Markovian setting being relevant here.

Trace based fitting methods use T and try to determine $(\boldsymbol{\pi},\boldsymbol{\theta})$ which maximizes the likelihood or equivalently the log-likelihood $\sum_{i=1}^{n} \log\left(f(t_i|(\boldsymbol{\pi},\boldsymbol{\theta}))\right)$. Corresponding fitting procedures are commonly based on expectation maximization (EM) algorithms [2,27], some of them on the basis of sub-classes of PHDs, as, e.g., hyper-exponential [22] or hyper-Erlang distributions [32]. EM based methods often become inefficient for large traces, but there are attempts to overcome this problem, e.g., by aggregating the trace [28]. A different approach applicable to large traces is presented in [29,30], where the user identifies peaks of the empirical PDF being the basis for a cluster analysis of the trace. The cluster sizes determine parameter $\boldsymbol{\pi}$ and component Erlang distributions are fitted on the basis of the clustered data. Since being based on Eq. (1) trace based fitting methods usually fit the empirical PDF fairly precise, but have difficulties to approximate the empirical moments.

Moment matching methods are based on Eq. (3) and the empirical moments. Some approaches consider specific structures of PHDs to match a finite set of moments trying to cope with possible non-unique representations of the same distribution, but suffer from the problem that only a restricted set of values are feasible moments, which makes fitting difficult [5,19]. Other approaches use more flexible structures and some of them consider hyper-Erlang distributions (or variants), since they can match any set of moments of a distribution [18,21]. [8] considers acyclic PHDs with n states (being characterized by $(2n-1)$ feasible moments [31]) and iteratively specifies sequences $\boldsymbol{\pi}^{(i)}$ and $\boldsymbol{\theta}^{(i)}$, where $\boldsymbol{\pi}^{(i+1)}$ is determined solving a constrained non-negative least squares (NNLS) problem for given $\boldsymbol{\theta}^{(i)}$. $\boldsymbol{\theta}^{(i+1)}$ is computed by standard polynomial optimization techniques for given $\boldsymbol{\pi}^{(i+1)}$ to obtain the parameter setting for the next iteration. Hardly surprising, being based on Eq. (3) moment matching methods have difficulties to approximate the empirical PDF/CDF.

The approach presented in this paper heads towards fitting of mixture distributions approximating the empirical PDF/CDF as well as the empirical moments and profits from the existence of efficient algorithms for solving NNLS problems. The main idea is partly along the lines of [8], for given $\boldsymbol{\theta}$ all Eqs. (1)-(3) can be used to formulate the fitting problem as a constrained NNLS problem. The main problem is to find an appropriate setting for $\boldsymbol{\theta}$. In this paper, a method is proposed to construct a possibly large parameter vector $\tilde{\boldsymbol{\theta}}$ such that the solution $\tilde{\boldsymbol{\pi}}$ of the NNLS problem gives a distribution $f(x|(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\theta}}))$ approximating the empirical PDF/CDF and moments. For construction, values from Farey sequences are used, which are transformed to values possibly conforming with the empirical trace data. Experiment results for mixtures of Erlang distributions show that most of the entries of $\tilde{\boldsymbol{\pi}}$ are almost vanishing and can be neglected giving a mixture distribution with a moderate number of components.

The next section presents the main idea behind the brute force approach and in Sect. 3 its adaption to the fitting of hyper-Erlang distributions is described. Sect. 4 shows results from experiments followed by an extension of the approach presented in Sect. 5.

2 A General Brute Force Approach

In the following, we assume that (empirical) PDF f_e, CDF F_e and (finite) moments m_e^j of order $j = 1, \ldots, K, K \in \mathbb{N}$, are given or can be derived from a given trace. For notational convenience all component distributions are assumed to belong to the same known family (thus $\boldsymbol{\theta}_i \in \mathbb{R}^m, \forall i$) although the approach can be easily extended to heterogeneous mixtures. The main idea is to define an appropriate set $\tilde{S}$ (of size $\tilde{G} \in \mathbb{N}$) of component distributions with parameter vector $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{\tilde{G}})$ and to determine mixing probabilities $\tilde{\boldsymbol{\pi}}$ by solving an appropriate NNLS problem such that the resultant mixture distribution approximates f_e, F_e and m_e^j.

Obviously, constructing such a set needs to be done systematically in order to promise better approximation results with increasing set size $\tilde{G}$. First a sequence of basic value sets $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$ with elements from $[0, 1]$ is defined which satisfy a denseness property within the unit interval. Then the elements of some set $\mathcal{F}_i$ are transformed by one or more transformation functions $TF_{(\cdot,j)}$ defined for the j-th component parameter. E.g., in Sect. 3, components with Erlang distributions are considered so that here two component parameters μ, k exist implying $j \in \{1, 2\}$. The transformed values are then taken to specify the component distributions of set $\tilde{S}$ by using all combinations of the component parameter values. In the following this construction process is described in more detail.

2.1 Farey Sequences as a Basic Value Set

As a basic value set Farey sequences, also known as Farey series [16] are utilized. A Farey series $\mathcal{F}_n$ is the increasing sequence of irreducible fractions in $[0, 1]$ with denominators not exceeding n. In the following, we will define Farey sequences as sets, since the order is irrelevant for our approach.

Definition 1 (Farey sequence (cf. [16])). *The Farey sequence* $\mathcal{F}_n, n \in \mathbb{N}$, *is defined as*

$$\mathcal{F}_n = \left\{ \frac{p}{q} \;\middle|\; p \in \mathbb{N}_0, q \in \mathbb{N} : 0 \leq p \leq q \leq n \text{ with } gcd(p,q) = 1 \right\}$$

where $gcd(p, q)$ *is the greatest common divisor of* p *and* q*.*

The elements of a Farey sequence are called Farey numbers and the first Farey sequences are

$$\mathcal{F}_1 = \left\{\frac{0}{1}, \frac{1}{1}\right\}, \mathcal{F}_2 = \left\{\frac{0}{1}, \frac{1}{2}, \frac{1}{1}\right\}, \mathcal{F}_3 = \left\{\frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}\right\}, \mathcal{F}_4 = \left\{\frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}\right\}$$

illustrating the following properties of Farey sequences.

Theorem 1 ((cf. [16])). *Let* $\frac{p}{q}, \frac{p'}{q'} \in \mathcal{F}_n$ *be two consecutive elements of* $\mathcal{F}_n$*, i.e.* $\forall x \in \mathbb{R} : \frac{p}{q} < x < \frac{p'}{q'} \Rightarrow x \notin \mathcal{F}_n$ *then*

(i) $qp' - pq' = 1$ *and* $q + q' > n$

(ii) $\frac{p'}{q'} - \frac{p}{q} = \frac{1}{qq'}$

(iii) $\mathcal{F}_n \subset \mathcal{F}_{n+1}, \quad \forall n \in \mathbb{N}$

(iv) Approximate cardinality of Farey sequences: $|\mathcal{F}_n| \approx \dfrac{3n^2}{\pi^2}$

where $\pi \approx 3.14$ is here the transcendental number.

Farey sequences have some favorable properties which support the design of a fitting procedure. Properties (i)+(ii) show that all elements of $[0, 1]$ can be approximated arbitrarily close selecting an appropriate $n \in \mathbb{N}$ [16]. Thus, theoretically Farey numbers can be used to approximate all elements of the unknown parameter vector $\boldsymbol{\theta}$ by means of appropriate transformations. Property (iii) is of decisive importance. It is the basis to ensure that an increasing effort, by using more Farey numbers ($''n \to n + 1''$) will not deteriorate fitting results. Finally, property (iv) gives hope that this increase might still lead to problem instances of manageable size.

2.2 Transformation of Basic Value Set

Since Farey numbers are inside the interval $[0, 1]$ they can not be used directly for an approximation of a single component parameter $\theta \in \mathbb{R}$ and have to be transformed accordingly. Generally, one can distinguish between the following two possibilities. A finite interval $[a, b], a, b \in \mathbb{R}, a < b$, might be assumed to contain parameter values of one or several components or due to the lack of information only an infinite interval can be supposed. In the first case, linear function $a + (b - a)x$ maps Farey numbers of $[0, 1]$ to $[a, b]$. For an infinite interval $[\alpha, \infty], \alpha \in \mathbb{R}, \alpha \geq 0$ we use a stereographic projection of the two-dimensional unit sphere mapping a point $(x_1, x_2), x_1^2 + x_2^2 = 1$, to $(x_1, x_1/(1 - x_2))$ giving a mapping of the interval $[0, 1]$ to $[\alpha, \infty]$ by

$$y = \left(\frac{x}{1 - \sqrt{1 - x^2}} - 1 + \alpha \right), \text{ if } x > 0 \qquad \text{and } y = \alpha, \text{ if } x = 0 \tag{4}$$

If useful, the values can be further transformed (stretched or compressed) by an exponential transformation $z = \exp(\beta y) - \exp(\beta\alpha) + \alpha$ with parameter $\beta \in \mathbb{R}, \beta > 0$ to ensure that also for small n the transformed values of basic value set $\mathcal{F}_n$ result in component definitions, such that Eqs. (1)-(3) might be fulfillable. Note that all equations specify convex combinations so that, e.g. considering Eq. (1), for all allowed $x \in \mathbb{R}$, components i, j have to exist in the mixture with $f(x|(\boldsymbol{\pi}, \boldsymbol{\theta})) \leq f_i(x|\boldsymbol{\theta}_i)$ and $f(x|(\boldsymbol{\pi}, \boldsymbol{\theta})) \geq f_j(x|\boldsymbol{\theta}_j)$. The same holds for Eqs. (2)-(3).

Similar transformations can be specified for co-domains $[-\infty, -\alpha]$ or $[-\infty, \infty]$ with $-y$ or by applying linear function $(2x - 1)$ and Eq. (4) in succession. If $\theta \in \mathbb{N}$, rounded values can be used. Generally, a transformation function $TF_{(I,j)}$ for interval I and the j-th parameter can be defined arbitrarily, but has to ensure that for all possible allowed values θ of the assumed finite or infinite interval I one has $\forall \epsilon > 0 : \exists n \in \mathbb{N}, x \in \mathcal{F}_n : |TF_{(I,j)}(x) - \theta| < \epsilon$, so that "denseness" of Farey sequences carries over to the range of transformed values.

For the definition of appropriate finite intervals one can exploit characteristics of the trace. E.g., the minimum and maximum value of a trace might give a very rough interval for an estimation of the components expected values. Narrower intervals might be obtained from empirical quantile values, which are, e.g., used in Sects. 3 and 4.

Parameter values of several mixture components $\tilde{\boldsymbol{\theta}}_i$ can now be obtained vector componentwise from Farey sequences possibly of different sizes, the assumed finite or infinite intervals and the corresponding transformation functions. For simplicity, we assume that a single Farey sequence $\mathcal{F}_{n_j}, n_j \in \mathbb{N}$, is used for the j-th component parameter. Let $I_{i,j}, i = 1, \ldots, k_j, j = 1, \ldots, m$ denote the i-th interval for the j-th parameter, where $k_j \in \mathbb{N}$ denotes the number of intervals defined for the j-th parameter. Note that all component distributions are assumed to belong to the same family. Then, a set of transformed values for the j-th parameter is given by $V_j = \{TF_{(I_{i,j},j)}(x) | i = 1, \ldots, k_j, x \in \mathcal{F}_{n_j}\}$ and set $\tilde{S}$ is composed from component definitions where all combinations are taken into account $V_{\tilde{S}} = \{(\tilde{\theta}_1, \dots, \tilde{\theta}_m) | \tilde{\theta}_j \in V_j, j = 1, \dots, m\}$. Since the order of mixture components is irrelevant for the approach presented here, an arbitrary vectorization of all elements of $V_{\tilde{S}}$ can be used to define $\tilde{\boldsymbol{\theta}}$.

2.3 Non-Negative Least Squares Problem Definition

With given components a NNLS problem can be specified as follows. Assume that empirical PDF f_e, CDF F_e and a finite number of moments $m_e^j, j \in \mathbb{N}$, are given or can be derived from a given trace. With $c, p \in \mathbb{N}$, let $P_{PDF} = \{x_1, \dots, x_p\}, x_i \in \mathbb{R}$, be a finite set such that for all $x_i \in P_{PDF}$ $f_e(x_i), \tilde{f}(x_i|\cdot)$[§] are defined and let $P_{CDF} = \{x_1, \dots, x_c\}, x_i \in \mathbb{R}$, be a finite set such that for all $x_i \in P_{CDF}$ $F_e(x_i), \tilde{F}(x_i|\cdot)$ are defined. Considering a finite set of K moments and Eqs. (1)–(2) at $x \in P_{PDF}$ and $x \in P_{CDF}$ respectively, results in a finite set of equations, such that the fitting problem can be formulated as a constrained NNLS problem for which very efficient algorithms exist being able to solve large problem instances [23]. Since numerical values, especially of the moments, might differ orders of magnitude, it is common to introduce appropriate weights, here $\gamma_{PDF}, \gamma_{CDF}, \gamma_j, j = 1, \dots, K$, with $\gamma_* \in \mathbb{R}$. Defining with $i = 1, \dots, \tilde{G}$

$$
\begin{aligned}
&\boldsymbol{A} = \gamma_{PDF} \left(\tilde{f}_i(x_j|\boldsymbol{\theta}_i) \right), && \boldsymbol{a} = \gamma_{PDF} \left(f_e(x_j) \right), && x_j \in P_{PDF} \\
&\boldsymbol{B} = \gamma_{CDF} \left(\tilde{F}_i(x_j|\boldsymbol{\theta}_i) \right), && \boldsymbol{b} = \gamma_{CDF} \left(F_e(x_j) \right), && x_j \in P_{CDF} \\
&\boldsymbol{C} = \left(\gamma_j E[\tilde{X}_i^j|\boldsymbol{\theta}_i] \right), && \boldsymbol{c} = \left(\gamma_j m_e^j \right), && j = 1, \dots, K. \\
&\boldsymbol{D} = \left(\boldsymbol{A}|\boldsymbol{B}|\boldsymbol{C} \right), && \boldsymbol{d} = \left(\boldsymbol{a}|\boldsymbol{b}|\boldsymbol{c} \right) && \qquad (5)
\end{aligned}
$$

the NNLS problem is $\quad \min_{\boldsymbol{\pi}} \|\boldsymbol{d} - \boldsymbol{\pi}\boldsymbol{D}\|_2^2$ subject to $\sum_{i=1}^{\tilde{G}} \pi_i = 1, \pi_i \geq 0.$

The weights γ_* can be used to control the impact of PDF, CDF and moments within the fitting process. Defining e.g., as in Sect. 4, weights γ_* such that $\sum_j a_j = \sum_j b_j = \sum_j c_j$ results in a similar contribution of the PDF, CDF and moments within the NNLS problem definition.

3 Fitting Finite Mixtures of Erlang Distributions

In the following, the approach of Sect. 2 is applied to mixtures of Erlang distributions with PDF, CDF and moments of an Erlang distribution given by

$$
\begin{aligned}
f(x|(\mu,k)) &= \left(\frac{k}{\mu}\right)^k \frac{x^{k-1}}{(k-1)!} \exp\left(-\frac{k}{\mu}x\right) \\
F(x|(\mu,k)) &= 1 - \exp\left(-\frac{k}{\mu}x\right) \sum_{r=0}^{k-1} \frac{(kx)^r}{\mu^r r!} \\
E[X^j|(\mu,k)] &= \frac{(k+j-1)!}{(k-1)!} \left(\frac{\mu}{k}\right)^j
\end{aligned}
$$

[§] $g(x|\cdot)$ denotes $g(x|(\boldsymbol{\pi}, \boldsymbol{\theta}))$ for arbitrary parameters $(\boldsymbol{\pi}, \boldsymbol{\theta})$.

where exp denotes the exponential function, $\mu \in \mathbb{R}^+$ is the expected value and $k \geq 1, k \in \mathbb{N}$, denotes the number of phases. Common definitions of an Erlang distribution use a parameter $\lambda = k/\mu$, here the expected value is used to make the fitting approach directly applicable.

Obviously, the first parameter is a candidate for the definition of transformation functions based on finite intervals, since it seems reasonable to assume, e.g., that the expected values of all component distributions might be bounded by the minimum and maximum values T_{min}, T_{max} of the trace, although this is of course theoretically not guaranteed. As mentioned, empirical quantiles $q_e(r)$ of order r can be used here for the definition of a set of contiguous intervals, assuming that the expected values of some components might be covered by an interval. For all later experiments 10 quantile values $q_e(i/11), i = 1, \ldots, 10$ and T_{min}, T_{max} have been used to define a set of intervals. If the minimum/maximum of the quantile values and the minimum/maximum of the trace differ orders of magnitude, i.e., if $(q_e(1/11)/T_{min}) > 10$ or $(T_{max}/q_e(10/11)) > 10$ larger finite intervals might occur. Additional quantile values of orders $1/10^z$ and $(10^z - 1)/10^z, z \in \mathbb{N}$, give narrower intervals. In experiments these additional quantile values have been used, increasing $z \in \mathbb{N}$ until the mentioned quantities do not differ orders of magnitude. Large differences might occur if the trace contains outliers or if the empirical skewness is significant.

The second parameter of the Erlang distributions is a candidate for the definition of an infinite interval and corresponding transformation function, since it seems difficult to set up one or several reasonable finite intervals. Some information can be obtained to support fitting. For $k \to \infty$ an Erlang distribution tends toward a normal distribution with expected value μ and variance $\sigma^2 = \mu^2/k$. Since the maximum of the PDF of a normal distribution is at $x = \mu$ with $f(\mu) = 1/\sqrt{2\pi\sigma^2}$, $(\pi \approx 3.14)$, the number of phases of one of the component distributions has to be at least $k^* = 2\pi \cdot \max_x\{(xf_e(x))^2\}$. This fact can be used to define an appropriate transformation onto the infinite interval $[1, \infty]$ followed by rounding the resultant values. For experiments $\beta > 0$ has been determined iteratively, such that the maximum of the set of transformed values of a fixed set size ($|V_2| = 30$ in all experiments) exceeds k^* significantly ($100 \cdot k^*$ in all experiments) and β has been kept fixed for all experiments with the same trace, so that property (iii) of Th. 1 in essence also holds for the transformed values. Since rounded values are used, a strict $\subset$ relation might not always hold and even large Farey sequences might result in relatively small value sets after transformation and rounding.

4 Experimental Results

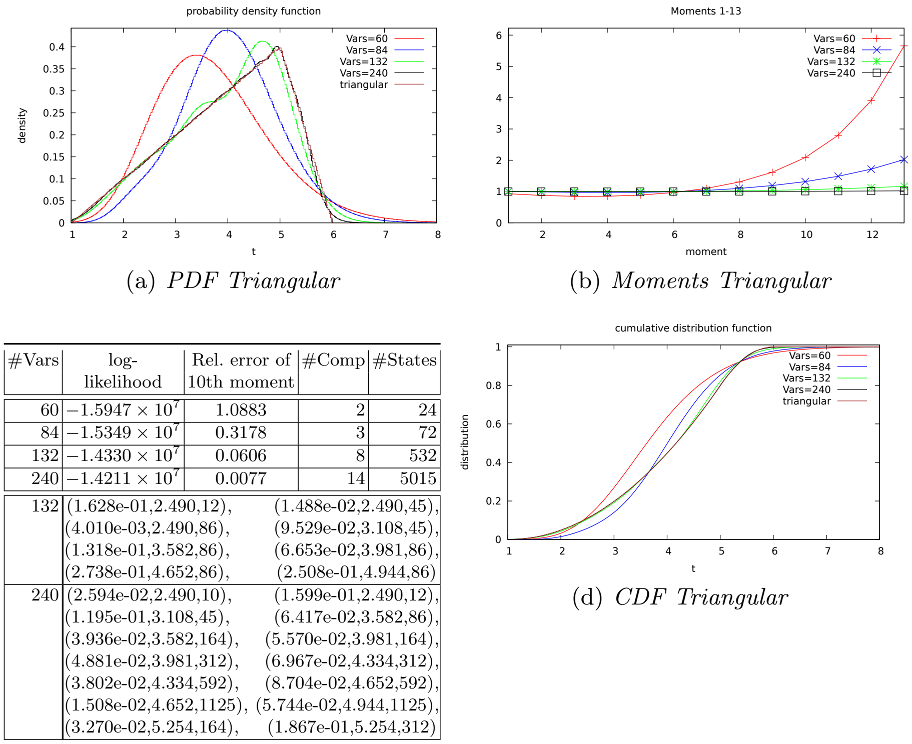
The brute force approach has been implemented in MATLAB (release R2017b). For experiments several synthetically generated traces and two real traces have been used and reported CPU times are from runs on a computer using a single core of an E5-2699 processor with 64GB RAM. Results are compared with those from the tools G-FIT [32] and MomFit [8]. G-FIT uses an EM algorithm to find the maximum likelihood estimates and MomFit is designed to fit an acyclic PHD to the empirical moments. G-FIT's parameters have been set to a maximum of 20 states and to aggregate the trace [28]. The brute force approach of Sects. 2 and 3 is named *shotgun* in corresponding figures.

The Freedman–Diaconis rule [13] has been used for the representation of the empirical PDF defining the width of the histogram bins by $2(Q_3 - Q_1)/\sqrt[3]{n}$ with Q_i denoting the i-th quartile and n being the number of trace elements. The set P_{PDF} is formed by the midpoints of the histogram bins. The size of set P_{CDF} is a user input giving equidistant points $x_j \in [T_{min}, T_{max}]$. In all experiments $|P_{CDF}|$ has been set such that the resultant NNLS problems are not underdetermined. $K = 10$ moments have been used for fitting and weights γ_* have been set such that $\sum_{x_j \in P_{PDF}} \gamma_{PDF} f_e(x_j) = \sum_{x_j \in P_{CDF}} \gamma_{CDF} F_e(x_j) = 1$. Likewise, the weights for the moments have been defined by $\gamma_j = \frac{2(K+1-j)}{K(K+1)m_e^j}$, $j = 1, \ldots, K$, emphasizing lower order moments, however giving $\sum_{j=1}^{K} \gamma_j m_e^j = 1$ as well. Irrelevant components ($\pi_i \leq 10^{-12}$) have been deleted from the results. The ProFiDo toolset [3] has been used for plots, generating traces from the fitted distributions, if necessary.

First, fitting of a triangular distribution is discussed. Fig. 1 shows several results for different sets of components with set size $\tilde{G}$ giving the number of variables for the NNLS problem (#Vars). Fig. 1(b) depicts the first moments of the fitted distribution relative to the empirical moments of the trace. Note that a few more higher order moments are shown than having been utilized for fitting. The table of Fig. 1(c) presents corresponding numerical values, amongst the log-likelihood, the relative error of the 10-th moment is given, since in all experiments this moment shows the largest relative error compared to lower order moments. Column #Comp gives the number of resultant components and column #States shows the sum of the phases of all Erlang branches. The last part presents concrete fitting results for $\tilde{G} = 132$ and 240. Not surprisingly, tiny NNLS problem instances give bad results, but a bit unexpectedly, even relatively small instances lead to satisfactory fitting results. The results show that PDF, CDF and the requested moments are fitted accurately, but at the price of Erlang distributions with a large number of phases. An additional interesting effect, which is also exhibited in other experiments, is the significantly reduced number of resultant components compared to the initial size of set $\tilde{S}$.

Another synthetically generated trace is from a mixture of three Beta distributions (*"Beta3"*) with PDF $\frac{1}{5} f_B(x|(1,30)) + \frac{3}{5} f_B(x|(10,10) + \frac{1}{5} f_B(x|(25,1))$ where $f_B(x|(\alpha,\beta))$ is the PDF of a Beta distribution with shape parameters $\alpha, \beta \in \mathbb{R}^+$. Figs. 2(a) and 2(b) show corresponding fitting results. Illustration of the CDF is omitted here and in the following experiment results, since corresponding curves are quite close. Fig. 2(a) shows that the right part of the distribution is not fitted exactly. Experiments with larger sets of components showed slightly better fittings. Since the PDFs of PHDs and thus hyper-Erlang distributions exhibit an exponential decay [26] exact fits cannot be expected in practice.

In addition, three traces have been selected which were also used for fitting with G-FIT [32]: traces from a uniform distribution on interval $[0.5, 1.5]$ (*"Uniform"*)



(a) *PDF Triangular*

(b) *Moments Triangular*

#Vars	log-likelihood	Rel. error of 10th moment	#Comp	#States
60	-1.5947×10^7	1.0883	2	24
84	-1.5349×10^7	0.3178	3	72
132	-1.4330×10^7	0.0606	8	532
240	-1.4211×10^7	0.0077	14	5015
132	(1.628e-01,2.490,12),	(1.488e-02,2.490,45),		
	(4.010e-03,2.490,86),	(9.529e-02,3.108,45),		
	(1.318e-01,3.582,86),	(6.653e-02,3.981,86),		
	(2.738e-01,4.652,86),	(2.508e-01,4.944,86)		
240	(2.594e-02,2.490,10),	(1.599e-01,2.490,12),		
	(1.195e-01,3.108,45),	(6.417e-02,3.582,86),		
	(3.936e-02,3.582,164),	(5.570e-02,3.981,164),		
	(4.881e-02,3.981,312),	(6.967e-02,4.334,312),		
	(3.802e-02,4.334,592),	(8.704e-02,4.652,592),		
	(1.508e-02,4.652,1125),	(5.744e-02,4.944,1125),		
	(3.270e-02,5.254,164),	(1.867e-01,5.254,312)		

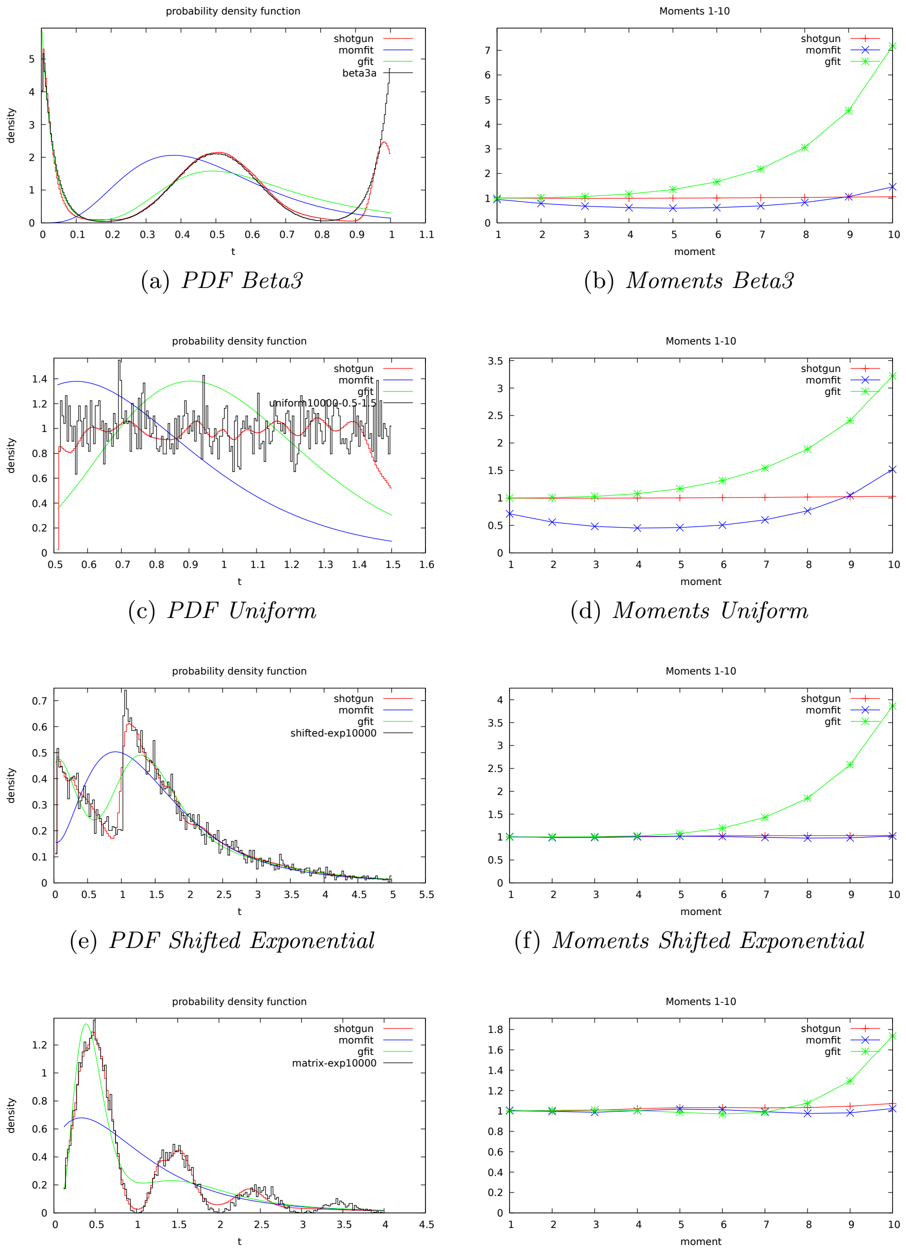
(c) *Results* (*Mixture Spec.:* (π_i, μ_i, k_i))

(d) *CDF Triangular*

Fig. 1: Fitting results for a trace from a triangular distribution

and traces from a Shifted Exponential and a Matrix Exponential Distribution [6]. Corresponding results are shown in Fig. 2 and numerical results are presented in Tab. 1, which extends the table of Fig. 1(c) by a few columns: The first two of them denote the name as well as the length of the trace and the tool used, the last column gives the CPU time in seconds needed for fitting. Since the fitting time of MomFit was more than 130s in all cases, we give figures depicting the first moments of the fitted distributions relative to the empirical moments of the trace and present numerical results for G-FIT only. For the fitting of all three traces accurate fitting results have been received in short CPU time. Tab. 1 shows that CPU times increase with larger traces. Most of the fitting time is used for the determination of the empirical PDF and CDF and thus for the specification of the NNLS problem. Solving the NNLS problem usually takes only a few seconds down to less than a second.

Also a larger trace from a heavy tailed Pareto-II distribution (cf. [17, 32]) has been selected for fitting. [32] used a smaller trace with 10^4 elements. Fig. 3 shows results for different sets of components initially used for fitting. Since the first 10 quantile values significantly differ from T_{min}, T_{max} fitting of this trace



(a) *PDF Beta3*

(b) *Moments Beta3*

(c) *PDF Uniform*

(d) *Moments Uniform*

(e) *PDF Shifted Exponential*

(f) *Moments Shifted Exponential*

(g) *PDF Matrix Exponential*

(h) *Moments Matrix Exponential*

Fig. 2: Synthetic Traces (Part I)

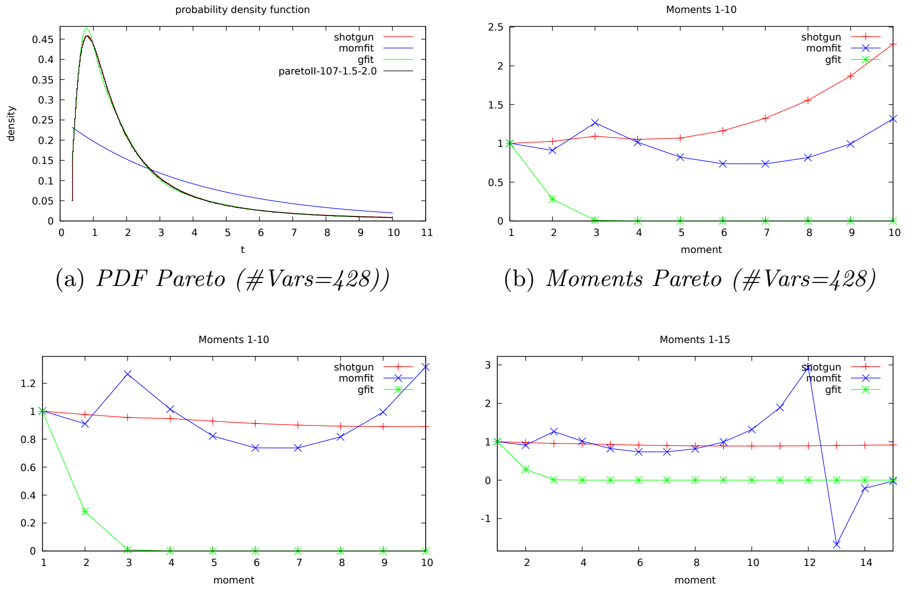
greatly benefits from additional quantile values used for the definition of finite

Table 1: Fitting results

Trace	Tool	#Vars	log-	Rel. error of	#Comp	#States	Fitting
(Length)			likelihood	10th moment			time (s)
Beta3	shotgun	400	1.9807×10^6	5.8253×10^{-2}	14	6392	15.32
(5.0×10^6)	G-FIT		6.1510×10^5		3	20	8.34
Uniform	shotgun	240	-5.2868×10^2	3.0065×10^{-2}	14	10 030	0.70
(1.0×10^4)	G-FIT		-1.8276×10^3		9	20	4.24
Shifted Exp.	shotgun	352	-1.3047×10^4	7.7144×10^{-3}	21	1825	0.79
(1.0×10^4)	G-FIT		-1.3179×10^4		5	20	10.44
Matrix Exp.	shotgun	333	-7.7801×10^3	4.1290×10^{-2}	28	3153	0.75
(1.0×10^4)	G-FIT		-8.7481×10^3		4	20	7.34
Pareto	shotgun	428	-2.0084×10^7	4.7756×10^{-1}	24	691	19.50
(1.0×10^7)	shotgun	667	-1.9966×10^7	1.1212×10^{-1}	28	1648	22.48
	shotgun	864	-1.9872×10^7	3.9117×10^{-2}	28	2045	27.95
	G-FIT		-1.9831×10^7		6	20	11.12
LBL3	shotgun	275	-1.5995×10^6	3.5725×10^{-2}	15	283	3.35
$(\approx 1.79 \times 10^6)$	G-FIT		-1.634×10^6		5	20	17.14
pAug	shotgun	314	-7.9409×10^5	4.2109×10^{-2}	27	12 845	2.53
(1.0×10^6)	G-FIT		-8.0743×10^5		4	20	15.68

intervals. Tab. 1 shows that G-FIT obtained a higher log-likelihood value, but did not match the moments accurately, see Fig. 3. Fig. 3(d) shows results for the first 15 moments. Note that only 10 moments have been used for fitting. Plots have been generated from traces of the fitted mixture distributions and are here not completely conforming with the numerical results of Tab. 1. The "Pareto" trace is from a heavy tailed distribution (e.g. kurtosis is approx. 10^6) and fitting has given one component with an expected value of T_{max} and a larger number of phases (55 for the case of #Vars=428 and 512 for #Vars=864), but with mixing probability less than 10^{-7}, so that large traces might be needed to be in accordance with the numerical values of Tab. 1.

As real traces two well-known traces from the Internet Traffic Archive [20] have been selected, which also have been used in several other publications: the LBL-TCP-3 trace (*"LBL3"*) and the BC-pAug89 trace (*"pAug"*), with all values normalized to a mean value of one. Corresponding results are depicted in Fig. 4 and also exhibit an accurate fitting. The large number of states (and components) for trace pAug (see Tab. 1) reveal the immanent danger of the brute force approach to overfit the model.



(a) *PDF Pareto (#Vars=428))*

(b) *Moments Pareto (#Vars=428)*

(c) *Moments Pareto (#Vars=864)*

(d) *Additional Moments Pareto (#Vars=864)*

Fig. 3: Synthetic Traces (Part II)

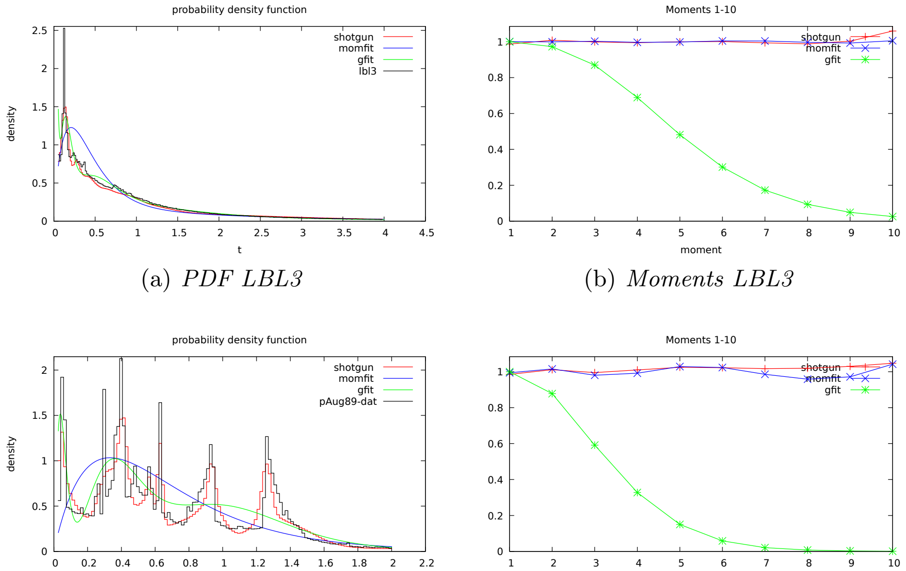
5 Extending the Approach

The presented approach fits mixture distributions to data taking account of the PDF/CDF and moments. Also other criteria can be considered as long as they can be encoded into the NNLS problem definition. Typical criteria might be those which help to reduce the risk of overfitting, as, e.g., the Akaike information criterion [1]. In the following, the model size is considered.

For mixtures of Erlang distributions the number of states might be an appropriate indicator for the overall size of the mixture distribution. Consequently, penalizing those Erlang distributions with several phases might reduce the model size. Extending Eqs. (5) by an additional penalty factor $p_f \in \mathbb{R}_0^+$ and column vector $\boldsymbol{p} = \left(p_f \left(1 + k_i / \left(\sum_{j=1}^{\tilde{G}} k_j\right)\right)\right)$ gives $\boldsymbol{D} = (\boldsymbol{A}|\boldsymbol{B}|\boldsymbol{C}|\boldsymbol{p}), \boldsymbol{d} = (\boldsymbol{a}|\boldsymbol{b}|\boldsymbol{c}|p_f)$.

Like the weights γ_*, factor p_f controls the impact state penalization has within the fitting process. E.g., setting $p_f = 1$ gives the same weight to this criterion as given to other parts of vector $\boldsymbol{d}$.

For additional experiments a trace from a hyper-Erlang distribution with 2 components and 3 states has been selected using an initial set of 215 Erlang distributions. Again 10 moments have been used for fitting and other parameters have been set as described in Sect. 4. Fitting with no penalty factor results in a mixture of 20 components and 222 states. G-FIT determines a mixture with

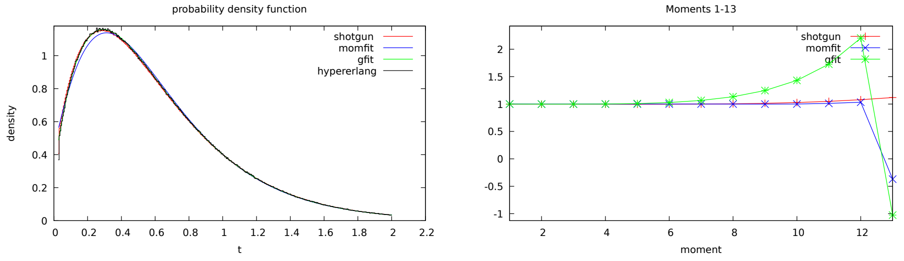


(a) *PDF LBL3* (b) *Moments LBL3*

(c) *PDF pAug* (d) *Moments pAug*

Fig. 4: Real Traces

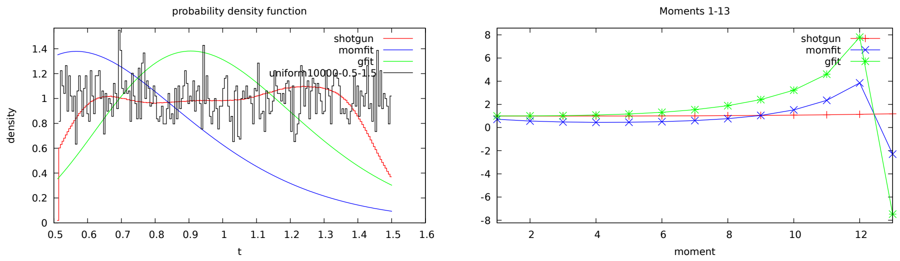
15 components and 20 states, whereas MomFit fits an acyclic PHD of order 5. Introducing a penalty factor $p_f = 10$ reduces the number of components as well as the number of states without worsening the fitting accuracy significantly (see. Fig. 5†) giving a hyper-Erlang distribution with 4 components and 8 states.



(a) *PDF Hyper-Erlang* (b) *Moments Hyper-Erlang*

Fig. 5: Hyper-Erlang distribution $\big(\boldsymbol{\pi} = (1/\pi^2, 1 - 1/\pi^2), \boldsymbol{\theta} = ((1/\mathrm{e}, 1), (2/\pi, 2))\big)$ fitting result: $(\boldsymbol{\pi'} = (1.19\text{e-}01, 7.12\text{e-}01, 9.9\text{e-}02, 7.0\text{e-}02)$, $\boldsymbol{\theta'} = ((3.51\text{e-}01, 1), (6.09\text{e-}01, 2), (7.01\text{e-}01, 2), (8.80\text{e-}01, 3)))$

$^\dagger\pi, \mathrm{e}$ the transcendental numbers



(a) *PDF Uniform*

(b) *Moments Uniform*

Fig. 6: Fitting of uniform distribution with penalty factor $p_f = 10$. Fitting result: $(\boldsymbol{\pi}' = (1.47\text{e-}01, 4.88\text{e-}04, 3.10\text{e-}01, 2.47\text{e-}01, 2.32\text{e-}01, 1.26\text{e-}02, 5.14\text{e-}02)$, $\boldsymbol{\theta}' = ((6.09\text{e-}01, 33), (6.09\text{e-}01, 45), (8.24\text{e-}01, 17), (1.05, 33), (1.28, 86), (1.39, 164), (1.39, 312)))$

Fig. 6 shows the results from fitting the trace of the uniform distribution presented in Sect. 4 with penalty factor $p_f = 10$. In this case, the size of the fitted distribution has been reduced significantly from 10030 states to 690 states still giving good fitting results.

6 Conclusions

This paper has presented a general brute force approach to fit finite mixture distributions taking the PDF/CDF as well as the moments into account. Since "dense" Farey sequences are used, the fitting approach is designed to approximate unknown parameter values arbitrarily close, provided transformation functions and intervals are suitably defined. Experiments have shown that only a small number of values is necessary and that the number of components resulting from solving the NNLS problem is also relatively small. An additional advantage of the approach is that no assumption on the number of resultant components is needed for fitting.

Experiments with hyper-Erlang distributions have given very accurate results in reasonable CPU times, but also indicate an immanent danger of overfitting as the resultant hyper-Erlang distributions tend to increase in size. In case of fitting mixture distributions to empirical data accurate fits might be problematic. E.g., estimating higher order moments is unreliable even for large traces [8], also the specification of the empirical PDF (and set P_{PDF}) is crucial. Apart from this aspect, hyper-Erlang distributions with a large number of phases might not be avoidable, since it is known that mixtures of fixed delays and Erlang distributions are necessary to approximate general positive distributions arbitrarily close [11]. As shown, the approach can be extended, so that the number of phases might be reduced still giving accurate results. However, concise representations are not always desirable. E.g., if also correlation structures need to be considered, some methods apply (similarity) transformations [10, 31] to obtain a suitable, often enlarged representation of the distribution for further fitting steps (cf. [4, 7]).

The brute force approach has been implemented for fitting mixtures of Erlang distributions and the method is easily adaptable to families of distributions with similar parameter definitions where component specifications are independent of each other, as, e.g., mixtures of normal distributions. For other types of mixtures an adaption needs more research. E.g., phase-type distributions (PHDs) can be represented as mixture distributions, but the parameters of the individual components are usually dependent, making the definition of set $\tilde{S}$ not that easy. Even when considering sub-classes of PHDs, as e.g. acyclic PHDs, the definition of set $\tilde{S}$ might still offer multiple options, so that it is not directly evident how information on trace data can be used for the definition of appropriate finite intervals for component parameters. Similar to some other fitting methods for PHDs, a possible approach might be to consider also here specific finite PHD structures.

References

1. H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
2. S. Asmussen, O. Nerman, and M. Olsson. Fitting Phase-Type Distributions via the EM Algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.
3. F. Bause, P. Buchholz, and J. Kriege. ProFiDo - The Processes Fitting Toolkit Dortmund. In *Proc. of the 7th International Conference on Quantitative Evaluation of SysTems (QEST 2010)*, pages 87–96. IEEE Computer Society, 2010.
4. F. Bause and G. Horvath. Fitting Markovian Arrival Processes by Incorporating Correlation into Phase Type Renewal Processes. In *Proceedings of the 2010 Seventh International Conference on the Quantitative Evaluation of Systems*, QEST '10, pages 97–106, Washington, DC, USA, 2010. IEEE Computer Society.
5. A. Bobbio, A. Horváth, and M. Telek. Matching Three Moments with Minimal Acyclic Phase Type Distributions. *Stochastic Models*, 21(2-3):303–326, 2005.
6. A. Bobbio and M. Telek. A Benchmark for PH Estimation Algorithms: Results for Acyclic-PH. *Communications in Statistics. Stochastic Models*, 10(3):661–677, 1994.
7. P. Buchholz, I. Felko, and J. Kriege. Transformation of Acyclic Phase Type Distributions for Correlation Fitting. In A. Dudin and K. De Turck, editors, *Analytical and Stochastic Modeling Techniques and Applications*, pages 96–111, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
8. P. Buchholz and J. Kriege. A Heuristic Approach for Fitting MAPs to Moments and Joint Moments. In *Proc. of the 6th International Conference on Quantitative Evaluation of SysTems (QEST 2009)*, pages 53–62, Los Alamitos, CA, USA, 2009. IEEE Computer Society.
9. P. Buchholz, J. Kriege, and I. Felko. *Input Modeling with Phase-Type Distributions and Markov Models - Theory and Applications.* SpringerBriefs in Mathematics. Springer, 2014.
10. P. Buchholz and M. Telek. Stochastic Petri Nets with Matrix Exponentially Distributed Firing Times. *Perform. Eval.*, 67(12):1373–1385, Dec. 2010.
11. A. David and S. Larry. The Least Variable Phase Type Distribution is Erlang. *Communications in Statistics. Stochastic Models*, 3(3):467–473, 1987.
12. Y. Fang. Hyper-Erlang Distribution Model and its Applications in Wireless Mobile Networks. *Wireless Network*, 7:211–219, 2001.

13. D. Freedman and P. Diaconis. On the Histogram as a Density Estimator: L_2 Theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981.
14. S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, 2006.
15. S. Frühwirth-Schnatter, G. Celeux, and C. Robert. *Handbook of Mixture Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Taylor & Francis Group, 2019.
16. G. Hardy and E. Wright. *An Introduction to the Theory of Numbers*. Oxford Science Publications. Clarendon Press, Oxford, 5th edition, 1979.
17. A. Horváth and M. Telek. Markovian Modeling of Real Data Traffic: Heuristic Phase Type and MAP Fitting of Heavy Tailed and Fractal Like Samples. In M. C. Calzarossa and S. Tucci, editors, *Performance Evaluation of Complex Systems: Techniques and Tools*, pages 405–434, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.
18. G. Horváth. Moment Matching-Based Distribution Fitting with Generalized Hyper-Erlang Distributions. In A. Dudin and K. De Turck, editors, *Analytical and Stochastic Modeling Techniques and Applications*, pages 232–246, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
19. G. Horváth and M. Telek. On the Canonical Representation of Phase Type Distributions. *Performance Evaluation*, 66(8):396–409, 2009. Selected papers of the Fourth European Performance Engineering Workshop (EPEW) 2007 in Berlin.
20. Internet Traffic Archive. `ftp://ita.ee.lbl.gov/html/traces.html`. Accessed: 2019-11-13.
21. M. A. Johnson and M. R. Taaffe. Matching Moments to Phase Distributions: Mixtures of Erlang Distributions of Common Order. *Communications in Statistics. Stochastic Models*, 5(4):711–743, 1989.
22. R. E. A. Khayari, R. Sadre, and B. R. Haverkort. Fitting World-wide Web Request Traces with the EM-algorithm. *Perform. Eval.*, 52(2-3):175–191, Apr. 2003.
23. C. Lawson and R. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, 1995.
24. M. Neuts. A Versatile Markovian Point Process. *Journal of Applied Probability*, 16(4):764—-779, 1979.
25. M. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, 1981.
26. C. A. O'Cinneide. Phase-Type Distributions: Open Problems and a Few Properties. *Communications in Statistics. Stochastic Models*, 15(4):731–757, 1999.
27. H. Okamura, T. Dohi, and K. S. Trivedi. A refined EM algorithm for PH distributions. *Performance Evaluation*, 68(10):938 – 954, 2011.
28. A. Panchenko and A. Thümmler. Efficient Phase-Type Fitting with Aggregated Traffic Traces. *Perform. Eval.*, 64(7-8):629–645, Aug. 2007.
29. P. Reinecke, T. Krauß, and K. Wolter. Cluster-based fitting of phase-type distributions to empirical data. *Computers & Mathematics with Applications, Theory and Practice of Stochastic Modeling*, 64(12):3840–3851, December 2012.
30. P. Reinecke, T. Krauß, and K. Wolter. Phase-Type Fitting Using HyperStar. In M. S. Balsamo, W. J. Knottenbelt, and A. Marin, editors, *Computer Performance Engineering*, pages 164–175, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
31. M. Telek and G. Horvath. A minimal representation of Markov arrival processes and a moments matching method. *Performance Evaluation*, 64(9–12):1153 – 1168, 2007.

32. A. Thümmler, P. Buchholz, and M. Telek. A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM Algorithm. In *Dependable Systems and Networks, DSN 2005. Proceedings. International Conference on*, pages 712–721, June 2005.
33. D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.