A Benchmark for PH Eestimation Algorithms: Results for Acyclic-PH

Andrea BOBBIO

Department of Electronics
University of Brescia
25123 BRESCIA (Italy)

bobbio@bsvax.unibs.it

Miklós TELEK

Dept. of Telecommunications
Technical University of Budapest
1521 BUDAPEST (Hungary)

telek@plan.hit.bme.hu

Abstract Distribution functions that can be expressed as exponential polynomials have useful computational properties in applied stochastic modeling and have gained widespread acceptance in recent years. Nevertheless, the implementation of efficient numerical procedures for estimating the distribution parameters remains an open problem that limits the use of this class of distributions in applications. The difficulty of the fitting problem is largely related to the non-linearity of the model and to the number of the parameters to be estimated.

Many attempts have been presented in the literature. However, the lack of accepted and standardized test examples makes it difficult to establish a comparative merit among the various approaches. This paper proposes a benchmark based on the workshop on Fitting phase type distributions, organized by S. Asmussen in February 1991. It also presents the results obtained by applying the Maximum Likelihood (ML) estimation procedure to the canonical representation of Acyclic Phase Type (APH) distributions.

Key words: PH distributions, Acyclic PH distributions, Maximum likelihood estimation.

1. INTRODUCTION

Cox showed [10] that the class $\mathcal{R}_m$ of distribution functions, whose Laplace transforms are rational with m poles, can be formally considered as resulting from a series of m

exponential stages with complex valued transition rates γ_i. Each stage i is entered with probability $(1-q_i)$ while with probability q_i the process stops immediately. In the general case, q_i can be negative or complex valued. Fitting a coxian model of order m requires the estimation of $2m$ parameters (m transition rates γ_i, and m switching probabilities q_i), but no necessary and sufficient conditions are known to ascertain if the obtained result is a proper cdf. In order to overcome this problem, many authors have defined various restrictions of the coxian representation.

A simple and popular restriction consists of defining fixed configurations of the stages, often limited to mixtures of Erlangs. Bux and Herzog [8] have implemented a non-linear estimation approach based on the matching of the first two moments coupled with the minimization of a distance measure with respect to the empirical distribution. Their approach operates on mixtures of Erlangs. Singh and Billinton [33, 34] have considered a number of fixed configurations of series/parallel combinations of Erlang stages. The parameters are estimated by matching an equal number of moments by means of a Newton-Raphson numerical method. In a series of papers [18, 19, 16], Johnson and Taaffe have explored the problem of matching the first three moments to a mixture of two Erlangs. The shape of the density functions has been analyzed in [17]. The package MEDA, developed by Schmickler [32, 31], works on mixtures of two or more Erlang distributions and implements a criterion based on the matching of the first three moments combined with the minimization of a deviation measure (e.g. the area between the two cdf's). Malhotra and Reibman [23] investigate ad-hoc models to be inserted in the general purpose reliability and performance modeling tool *SHARPE* [30]. They also propose a hybrid method based on the matching of the first two moments and on the minimization of a suitable distance measure (least squares).

The class of Phase Type (PH) distributions, considered by Neuts in [24], has favorable computational properties. However, the PH representation is redundant and not unique [26] and does not appear as a good starting point for the fitting problem. Asmussen and Nerman [2] have recently presented a fitting procedure for the complete class of PH distributions, based on the EM algorithm of Dempster and al. [12] for incomplete data. Bobbio et al. [5] have elaborated an estimation procedure for the restricted class of Acyclic Phase Type (APH) distributions. More recently, they have proposed [4] a ML technique based on the canonical representation of APH [11], which is suited for the interpretation of censored data samples [3].

Harris et al. have introduced a subclass of $\mathcal{R}_m$ called GH (Generalized Hyperexponential) [7]. GH distributions are defined as mixtures of simple exponential terms

with non-necessarily positive mixing parameters. An ML estimation procedure for this class of functions has been described in [15]. The GH class contains functions that are not cdf's [14] and does not include the Erlang distribution (and, of course, mixtures of Erlangs) which, for a given order, is the PH distribution with the minimum coefficient of variation [1].

A representation of the cdf's in $\mathcal{R}_m$ called *Matrix Exponential (ME)*, has been proposed in [21]. The ME class includes the PH class while preserving most of its useful properties. Van de Liefvoort [35] discusses an algorithm to construct an ME distribution from its moments in a finite number of steps.

Up to now little effort has been directed to comparing the performance of different techniques; a noticeable exception is the work of Nielsen [25, 22]. In order to stimulate a common criterion for establishing a merit among the algorithms developed by various authors, the participants to the workshop on "Fitting phase type distributions", organized by S. Asmussen in Aalborg (DK), in February 1991, have formulated a benchmark. The present paper describes a modification of the benchmark in Section 3 and reports the results obtained by applying the ML method for APH distributions illustrated in [4]. The notation and an outline of the proposed technique are summarized in Section 2.

2. NOTATION AND ML ESTIMATION PROCEDURE

Let $\{X(t),\ t \geq 0\}$ be a Continuous Time Markov Chain ($CTMC$) with state space $\{1,\ 2,\ \ldots, m+1\}$, initial distribution $(\underline{\alpha},\ \alpha_{m+1})$ and infinitesimal generator:

$$\widehat{\mathbf{A}} = \left| \begin{array}{cc} \mathbf{A} & \underline{A}^0 \\ \underline{0} & 0 \end{array} \right| \tag{1}$$

In (1), $\mathbf{A}$ is an $(m \times m)$ nonsingular matrix, $\underline{A}^0 = -\mathbf{A}\underline{u}$ ($\underline{u}$ being a column vector with all the entries equal to 1) and $\underline{0}$ a row vector with all the entries equal to 0. Let τ be the time to absorption of $X(t)$ in state $m+1$. By solving the Markov equation, subject to the initial condition $(\underline{\alpha}, \alpha_{m+1})$, the cdf $F(x)$ of the r.v. τ becomes [24]:

$$F(x) = Prob\{\tau \leq x\} = 1 - \underline{\alpha}\, exp(\mathbf{A}x)\, \underline{u} \quad . \tag{2}$$

A distribution with cdf of the type given in (2), is called a PH distribution with representation $(\mathbf{A},\ \underline{\alpha})$ of order m. In general, a PH distribution is characterized by $\omega_F = m^2 + m$ free parameters (m^2 pertinent to matrix $\mathbf{A}$ and m pertinent to vector $\underline{\alpha}$).

Let $F^{\sim}(s)$ denote the Laplace Stieltjes Transform (LST) of $F(t)$. From (2) we have:

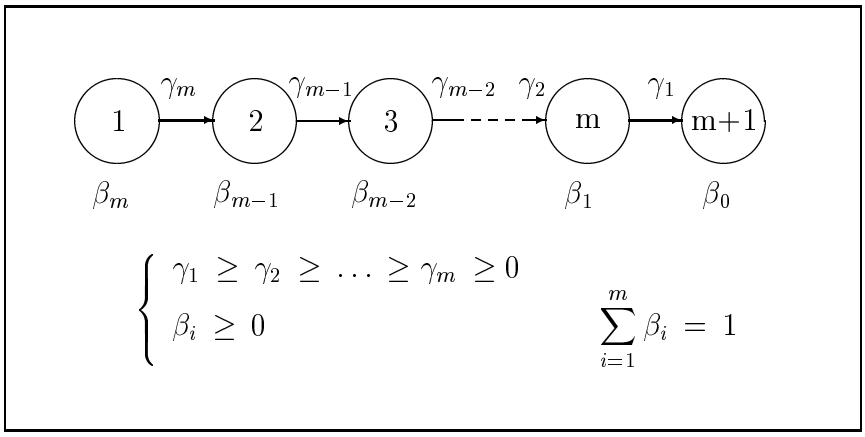


Figure 1 - The series Canonical Form *CF1*

$$F^{\sim}(s) = \alpha_{m+1} + \underline{\alpha}\,(\,s\mathbf{I} - \mathbf{A}\,)^{-1}\,\underline{A}^{0} = \alpha_{m+1} + \frac{N^{\sim}(s)}{D^{\sim}(s)} \tag{3}$$

where:

$$D^{\sim}(s) = \det\,(s\mathbf{I} - \mathbf{A}) = \prod_{i=1}^{m}(s - \gamma_i) \tag{4}$$

The γ_i's are the (generally complex-valued) eigenvalues of $\mathbf{A}$. By (4), $D^{\sim}(s)$ is a polynomial of order m, while $N^{\sim}(s)$ is, in general, of order $m-1$. Therefore, $F^{\sim}(s)$ in (3) has $\omega_G = 2m$ degrees of freedom. It is thus clear that the representation (1) is not minimal for a PH distribution (since $\omega_F > \omega_G$) and, moreover, it is not unique [26]. Thus, (1) is not a convenient representation for the parameter estimation problem.

When the Markov chain $X(t)$ is acyclic, the state space can always be ordered to have the matrix $\mathbf{A}$ in triangular form. PH distributions generated by acyclic $CTMC$'s form the subclass called Acyclic PH. From the above definition, an APH is characterized, in general, by $\nu_F = [\,m(m+1)/2 + m]$ free parameters. Cumani [11] has shown that any APH of order m admits a minimal representation in which the number of free parameters equals the number of degrees of freedom ($\omega_F = \omega_G = 2\,m$), and in which the transition rates are ordered in a non-increasing sequence. The minimal representation is referred to as the *Canonical form*, and conveys very useful properties in the estimation problem. From a computational point of view the more convenient canonical form is the one represented in Figure 1 and referred to as $CF1$.

The advantage of $CF1$ is that the LST can be written by inspection and has the following expression:

$$F^{\sim}(s,\underline{\beta},\underline{\gamma}) = \sum_{i=1}^{m} \beta_i F^{\sim(i)}(s,\underline{\beta},\underline{\gamma}) = \sum_{i=1}^{m} \beta_i \prod_{k=1}^{i} \frac{\gamma_k}{s+\gamma_k} \tag{5}$$

where $F^{\sim(i)}(s,\underline{\beta},\underline{\gamma})$ is the LST of the cdf of a series comprising the last i stages, $\underline{\beta}$ is the m-dimensional vector grouping the initial probabilities, and $\underline{\gamma}$ the m-dimensional vector grouping the transition rates. The definition of $CF1$ requires that the rates in $\underline{\gamma}$ are ordered in the non-increasing sequence $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_m \geq 0$. For the sake of simplicity, we assume $\beta_0 = 0$ so that only $(m-1)$ parameters are independent because of the normalization condition $\sum_1^m \beta_i = 1$. An algorithm to compute $F^{\sim}(s,\underline{\beta},\underline{\gamma})$ in (5) and its derivatives with respect to the model parameters has been described in [4].

Let $Z = \{z_1, \ldots, z_\eta, z_{\eta+1}, \ldots, z_\nu\}$ be a sample of ν points, such that $\eta \leq \nu$ are non censored and $\nu - \eta$ are censored. Let us denote by $\widehat{\underline{\beta}}$ and by $\widehat{\underline{\gamma}}$ the ML estimators of $\underline{\beta}$ and $\underline{\gamma}$, respectively and by $f(\cdot,\underline{\beta},\underline{\gamma})$ and $\pounds(Z,\underline{\beta},\underline{\gamma})$ the density and the likelihood function, respectively. The estimation problem can be formulated in the following way:

find: $\widehat{\underline{\beta}}$ and $\widehat{\underline{\gamma}}$ such that

$$\pounds(Z,\underline{\beta},\underline{\gamma}) = \prod_{i=1}^{\eta} f(z_i,\underline{\beta},\underline{\gamma}) \cdot \prod_{j=\eta+1}^{\nu} \left[1 - F(z_j,\underline{\beta},\underline{\gamma})\right]$$

is maximum, under the canonical constraints:

$$\begin{cases} \gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_m \geq 0 \\ \beta_i \geq 0 \qquad\qquad \sum_{i=1}^{m} \beta_i = 1 \end{cases}$$

The ML procedure, described in [4], was originally intended to estimate the distribution parameters from experimental data, and in particular from censored samples originated from truncated life-time experiments [3]. The same procedure can also be adapted to fit continuous cdf's: the domain $(0-1)$ of the cdf is divided into ν equal intervals, and a discrete sample is generated by calculating the inverse function in the midpoint of each interval.

3. DESCRIPTION OF THE BENCHMARK

The continuous nonnegative distributions included in the benchmark are summarized in Table I. The nine cases were chosen with the following motivations:

- Weibull and Log-normal distributions (*W1, W2, L1, L2* and *L3*) are often encountered in the interpretation of experimental data in engineering (reliability, queuing, transmission, etc.);

- The uniform distribution (*U1* and *U2*) and the distribution *SE* (a mixture of an exponential and a shifted exponential) are difficult to closely approximate with a PH distribution;

- The distribution *ME* is in $\mathcal{R}_m$, thus has a matrix exponential representation, but is not PH.

Table I - *Test cases of the benchmark*

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

Let $F(\cdot)$ and $f(\cdot)$ denote the cdf and the pdf of the distribution to be fitted, and $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$ the cdf and the pdf of the approximating PH. Furthermore, $c_i(F)$ denotes

the ith cumulant of distribution F so that $c_1(F)$ is the mean, $c_2(F)$ is the variance and $c_3(F)$ is the centered third moment. The following five parameters [1] were chosen as a measure of the goodness of the fit:

1. Relative error in the 1st moment: $\hat{e}_1 = |\, c_1(\hat{F}) - c_1(F) \,| \,/\, c_1(F)$

2. Relative error in the 2nd moment: $\hat{e}_2 = |\, c_2(\hat{F}) - c_2(F) \,| \,/\, c_2(F)$

3. Relative error in the 3rd moment: $\hat{e}_3 = |\, c_3(\hat{F}) - c_3(F) \,| \,/\, c_3(F)$

4. Density absolute area difference: $\hat{D} = \int_0^\infty |\, \hat{f}(t) - f(t) \,| \, d\, t$

5. Minus the cross entropy: $-\hat{H} = \int_0^\infty \log \hat{f}(t) \, d\, F(t)$

For a discrete sample $-\hat{H}$ equals the log-likelihood [2] (normalized with respect to the number of sample points ν), and is, therefore, proportional, through the normalization constant ν, to the objective function used in the optimization procedure.

4. APPLICATION OF THE BENCHMARK

In order to test the accuracy of the approximation as a function of the increasing number of phases, each one of the nine distributions of the benchmark is fitted with three APH's, in canonical form $CF1$, of order $m = 2, 4$ and 8 respectively. Hence, 27 cases are examined. Each case is identified by the symbol *Ds_m* where *Ds* is one of the nine distributions (*W1, W2, ... , ME*) and m (= 2, 4, 8) is the order of the fitting APH. The cdf to be fitted is sampled by means of a discrete sample of 200 points.

The non-linear constrained optimization problem has been solved by an iterative linearization method [4]. At each step, the objective function is linearized around the current solution by a first order expansion. The accuracy of the linear approximation is preserved by constraining the decision variables to be confined into a small box centered around the current solution. Then, a modified simplex method is applied [28], coupled with a line-search along the direction indicated by the optimal linear solution. This procedure is computationally very intensive but is feasible since the use of the $CF1$

[1] The definition of the test indicators differs, in some cases, from the one suggested at the Aalborg workshop. The Appendix motivates this choice.

makes possible the evaluation in closed form of the derivatives of the likelihood function. The same package compiled and executed on different machines (say IBM-RISC 6000 and DEC-VAXstation 2000) may provide appreciably different results. Some of the difficulties and instabilities in convergence are inherent in the problem itself and are shared by other methods [25]. Indeed, the same order of complexity has been documented in the companion problem of fitting functions (or data) by sums of exponentials [20, 29, 36]. The difficulties are mainly caused by the non-linear dependence of the objective function on the exponential parameters, the number of parameters to be simultaneously optimized, the presence of local minima and the flatness of the objective function.

In order to reduce the variability of the results from the computational conditions, the final goodness of fit measures are calculated by averaging the outcomes of 10 different runs, obtained by starting with different initial guesses. The numerical procedure assumes as initial guess the $CF1$ with the largest value of the likelihood function, among a defined number (200 in the examples) of $CF1$'s generated at random with the expected value equal to that of the distribution to be fitted.

The results obtained over the 10 runs are reported in detail in [6]. The analysis reveals that the variability in the test measures, and in particular in the cross entropy related to the objective function, is lower than the variability in the estimated APH parameters $\underline{\widehat{\beta}}$ and $\underline{\widehat{\gamma}}$. This observation supports the conjecture that the dominant factor in the low rate of convergence of the numerical optimization procedure is the flatness of the objective function around the maximum. At each iteration, the improvement of the objective function is very small with respect to the variation of the model parameters. The low convergence rate, coupled with the numerical instabilities, has forced us to bound at $m = 10$ the maximum order for the approximating APH.

Two criteria have been adopted for stopping the iterative estimation algorithm. The first criterion tests whether the Euclidean norm between an iteration and the preceding one is less than a given value (actually $\varepsilon = 1.0 \cdot 10^{-6}$); the second one stops the procedure when the total number of iterations exceeds a threshold (actually $niter = 200$). In general, as the order of the APH increases, the algorithm tends to stop because the maximum allowable number of iterations is exceeded even without reaching the prescribed accuracy on the Euclidean norm.

The integrals from 0 to ∞, implied in the test measures $\hat{D}$ and $-\hat{H}$, are evaluated by resorting to a Gaussian quadrature method over an interval from 0 to $50 \cdot c_1(F)$.

The program is composed of three main sections: *a) - initialization (sample and initial guess generation)*; *b) - iterative optimization* and *c) - computation of the final*

measures. The execution times of the three sections are approximately in the ratio *1 : 7 : 4*, and are increasing with the order of the PH distribution. The approximate execution times, on an IBM RISC 6000-320H computer, are as follows: less than 20 sec for PH_2, less than 1 min for PH_4 and less than 4 min for PH_8.

5. RESULTS

The test indicators are presented and discussed in the following.

Density Plots - Figure 2) plots the densities of the original distributions of the benchmark and of the approximating PH of order $m = 2$, 4 and 8, respectively. A visual inspection shows that for the cdf's *W1, W2, L1, L2* the approximation is satisfactory already with order $m = 4$. Thus, "well behaved" functions can be closely approximated by APH models of low order. For the other cases (namely: *L3, U1, U2, SE, ME)*, a sharp disagreement in the density curves is evident (and expected) even though some of the five test indicators still have satisfactory values. A rigid limitation to the use of an m-order PH is that the minimal squared coefficient of variation is equal to $1/m$ and is reached by the Erlang of order m [1]. Thus, when the distribution to be fitted is less variable than this lower bound, the APH attains the corresponding Erlang (see the cases *W1_2, L3_8* (with some numerical instabilities), *L3_4, L3_2, U2_4, U2_2, SE_2*). This is the primary reason for preventing the distribution *L3* (and in some sense also *U2*) to be adequately fitted with PH of order less than 8. However, this limitation does not appear to be too restrictive in many engineering applications, where it is unlikely to encounter data samples having very low variability.

In addition, PH distributions of low order are not suited to model abrupt variations in the density shape or minima in the density close to zero [27]. However, it is interesting to observe how the approximation evolves in the *U1* and in the *SE* cases going from order 2 to order 8, since increasing the order increases the number of points where the derivative of the density can be equal to zero. For the *U1*, the approximating APH tends to ripple over the plateau and for the *SE* the bimodality clearly appears with $m = 4$ and 8 (a APH of order 2 cannot be bimodal).

Moment Relative Errors - The results obtained for $\hat{e}_1$, are summarized in Table II. Even if a pure ML criterion is adopted, the fit to the first moment is usually satisfactory; nevertheless, a strict relationship between increasing accuracy and increasing order is not always respected. The error is often lower for distributions for which the visual

Table II - *Relative error in the 1st moment*

Distr.	$c_1(F)$	Relative Error $\hat{e}_1$		
		order 2	*order* 4	*order* 8
W1	0.9027	$5.8168E-02$	$6.0447E-04$	$7.9532E-04$
W2	2.0000	$1.2966E-02$	$1.4206E-02$	$3.6991E-02$
L1	1.0000	$6.6181E-02$	$9.0828E-02$	$1.2890E-01$
L2	1.0000	$3.8239E-03$	$4.6680E-03$	$3.6624E-03$
L3	1.0000	$4.2886E-02$	$4.7874E-03$	$1.7026E-03$
U1	0.5000	$1.7410E-02$	$1.3146E-01$	$5.1761E-02$
U2	1.5000	$1.2904E-03$	$5.7054E-02$	$6.0070E-04$
SE	1.5000	$3.2515E-02$	$1.4458E-03$	$1.5287E-03$
ME	1.0494	$2.0964E-03$	$1.6879E-02$	$2.3281E-03$

agreement is the poorest (see e.g. *L3, U2, SE*) than for distributions with a long tail (see e.g. *W2* and *L1*).
The $2nd$ moment relative error $\hat{e}_2$, reported in Table III, strictly decreases with increasing order, except for the distributions with a long tail (e.g. *W2* and *L1*). This test indicator makes apparent that the poor approximation in the *L3* and *U2* cases is due to the inability of the APH to attain a sufficiently low coefficient of variation.

Many results in queuing theory depend purely on the moments of the arrival and service times. This fact has induced many authors to develop hybrid techniques, in which the first two (or three) moments are equated separately. In the light of the present results, it should be interesting to investigate to what extent a hybrid estimation affects the approximation, and whether it could simplify or speed up computations since the variability is reduced by adding further constraints. A remarkable feature of this technique is that the first two moments are closely approximated, when possible.

The $3rd$ moment relative error $\hat{e}_3$ is reported in Table IV. Since $\hat{e}_3$ is undefined for the distributions for which $c_3(F) = 0$ (like *U1* and *U2*), Table IV reports in these two cases (marked with *) the corresponding absolute measure $c_3(\hat{F})$. The parameter decreases with increasing order, except in cases *W2_8* and *L1_8*.

Table III - *Relative error in the 2nd moment*

Distr.	Original Distr.		Relative Error $\hat{e}_2$		
	$c_2(F)$	$(c.v.)^2$	*order* 2	*order* 4	*order* 8
W1	$3.7569E-01$	$4.6103E-01$	$1.6817E-01$	$1.1899E-02$	$4.8555E-03$
W2	$2.0000E+01$	$5.0000E-00$	$4.5348E-01$	$2.0092E-01$	$2.9776E-01$
L1	$2.4534E+01$	$2.4534E+01$	$8.3709E-01$	$7.6972E-01$	$8.1851E-01$
L2	$8.9648E-01$	$8.9648E-01$	$2.6153E-01$	$8.7586E-02$	$7.0375E-02$
L3	$4.0811E-02$	$4.0811E-02$	$1.0385E+01$	$5.5880E-00$	$2.6277E-00$
U1	$8.3333E-02$	$3.3333E-01$	$6.3528E-01$	$5.1237E-01$	$2.9476E-01$
U2	$8.3333E-02$	$3.7037E-02$	$1.2534E+01$	$7.7728E-00$	$2.9260E-00$
SE	$1.2500E-00$	$5.5555E-01$	$2.3837E-01$	$3.9395E-02$	$1.2140E-02$
ME	$9.5308E-01$	$8.6539E-01$	$1.5199E-01$	$6.9499E-02$	$2.3009E-02$

Entropy and Area Difference - Let us define, in the usual way, the intrinsic entropy of the original distribution $H = -\int_0^\infty \log f(t)\, d\,F(t)$.

The cross entropy $\hat{H}$ converges from above to the intrinsic entropy H as the approximating distribution converges to the original one [9]. Therefore, their difference, called relative entropy H_r, provides a measure of the goodness of the approximation:

$$H_r = \hat{H} - H = \int_0^\infty \log \frac{f(t)}{\hat{f}(t)}\, d\,F(t) \ .$$

The three quantities are compared in Table V. The better the visual approximation, the lower the relative entropy. Furthermore, the relative entropy generally decreases as a function of the number of phases, except in cases *W2* and *L1*. This anomaly, shared also by other test indicators, requires further investigation. In any case, it is a symptom that the asymptotic exponential decay of the *APH* may be inappropriate to fit long tail distributions.

The difference between the two densities $\hat{D}$ is known as the $L1$ distance of two random variables and is related to the relative entropy by the following inequality [13]:

$$H_r \geq (1/2)\ \hat{D}^2 \ .$$

The results obtained for $\hat{D}$ are reported in Table VI.

Table IV - *Relative (absolute*) error in the 3rd moment*

Distr.	$c_3(F)$	Relative (Absolute*) Error $\hat{e}_3$		
		order 2	*order* 4	*order* 8
W1	$2.4685E-01$	$5.2179E-01$	$1.0678E-01$	$2.2066E-02$
W2	$5.9200E+02$	$8.7081E-01$	$5.3194E-01$	$6.5082E-01$
L1	$1.6573E+04$	$9.9803E-01$	$9.9508E-01$	$9.9689E-01$
L2	$3.1315E-00$	$6.8388E-01$	$3.2754E-01$	$2.9681E-01$
L3	$5.0643E-03$	$8.7844E+01$	$2.7941E+01$	$7.6730E-00$
*U1**	$0.0000E-00$	$7.1969E-02$	$4.0905E-02$	$1.9239E-02$
*U2**	$0.0000E-00$	$1.6941E+00$	$6.9410E-01$	$1.4334E-01$
SE	$2.0000E-00$	$4.4231E-01$	$2.3433E-01$	$7.2139E-02$
ME	$1.9929E-00$	$3.0223E-01$	$2.1078E-01$	$1.2695E-01$

6. CONCLUSION

In order to establish a homogeneous criterion for evaluating the performance of different estimation algorithms, the paper has proposed and applied a benchmark developed in conjunction with other researchers working in the field. Besides the achievable values of the goodness of fit measures, the choice of a PH model depends on the following issues [23]: *i) the number of states in the approximation; ii) the sparsity of the resulting Markovian model: iii) the time and memory requirements of the numerical implementation.*

The paper demonstrates that the estimation procedure based on APH in canonical forms responds primarily to points *i)* and *ii)*, even if its implementation is more intensive than some other approaches [25]. The canonical representation of APH distributions is very effective for fitting "well behaved" cdf's with phase models of low order. For "non-PH-like behaved" functions the approximation could be inappropriate, even if some test indicators still maintain satisfactory values, and some features of the density (e.g. moments, multimodality, etc.) are preserved. The ability of keeping the model of low order and sparse is of paramount importance when the estimation package is used as the first step in a complete tool for the modeling and analysis of non-exponential systems.

Table V - *Intrinsic H, cross $\hat{H}$, and relative entropy H_r*

Distr.	Intrinsic Entropy H	Cross Entropy $\hat{H}$			Relative Entropy H_r		
		order 2	*order* 4	*order* 8	*order* 2	*order* 4	*order* 8
W1	0.7869	0.7996	0.7875	0.7872	0.0127	0.0006	0.0003
W2	1.1546	1.2648	1.1772	1.1778	0.1102	0.0226	0.0232
L1	0.3745	0.4265	0.3826	0.3848	0.0520	0.0081	0.0103
L2	0.8757	0.9046	0.8836	0.8781	0.0289	0.0079	0.0024
L3	−0.2104	0.6381	0.3409	0.0916	0.8485	0.5513	0.3020
U1	0.0000	0.2061	0.1922	0.1601	0.2061	0.1922	0.1601
U2	0.0000	1.0356	0.7963	0.4829	1.0356	0.7963	0.4829
SE	1.2950	1.3559	1.3263	1.3194	0.0609	0.0313	0.0244
ME	0.7277	0.9655	0.9399	0.8808	0.2378	0.2122	0.1531

APPENDIX - MEASURES OF THE GOODNESS OF FIT

Some of the measures discussed in this paper differ from the ones originally suggested at the Aalborg workshop. This appendix motivates the reasons for these modifications.

a) - The standardized $2nd$ moment error measure $e_2^\star$ was originally defined as:

$$e_2^\star \;=\; |\, c_2(\hat{F}) \,-\, c_2(F) \,|\, / \, c_1(F)^2 \;=\; \hat{e}_2 \; (c.v.)^2 \quad .$$

$e_2^\star$ was intended to provide a meaningful measure to compare the outcomes of different approximation methodologies applied to a given distribution of the benchmark. However, in comparing how a given approximation procedure performs for various distributions, $e_2^\star$ does not appear an appropriate measure of goodness of fit, since it does not reflect the relative error when the variance is small with respect to the mean (i.e. for distributions

Table VI
Density absolute area difference $\hat{D}$

Distr.	*order* 2	*order* 4	*order* 8
W1	$1.1297E-01$	$1.7697E-02$	$1.1376E-02$
W2	$3.3503E-01$	$1.3219E-01$	$1.2586E-01$
L1	$2.4927E-01$	$6.0071E-02$	$5.3945E-02$
L2	$1.4607E-01$	$6.4253E-02$	$2.6830E-02$
L3	$1.1159E+00$	$8.7772E-01$	$6.2398E-01$
U1	$4.3162E-01$	$3.5707E-01$	$3.5026E-01$
U2	$1.2787E+00$	$1.0845E+00$	$7.4539E-01$
SE	$2.8260E-01$	$1.8746E-01$	$1.4968E-01$
ME	$5.4445E-01$	$4.9624E-01$	$3.7720E-01$

with a low coefficient of variation). Indeed, in the cases *L3* and *U2*, $e_2^\star$ has a small value in contrast to the poor visual agreement and with the fact that the misfit is primarily caused by a mismatch in the second moment. Also $e_2^\star$ usually decreases with the increasing order of the APH distribution.

b) - The standardized $3rd$ moment error measure $e_3^\star$ was originally defined as:

$$e_3^\star \;=\; |\, c_3(\hat{F}) \,-\, c_3(F) \,|\,/\, c_1(F)^3 \;=\; \hat{e}_3\,[\, c_3(F)\,/c_1(F)^3\,]\;\;.$$

The definition of $e_3^\star$ is exposed to the same criticism given for $e_2^\star$. Indeed $e_3^\star$ penalizes the cdf's with high $c_3(F)$ with respect to the mean. The proposed parameter $\hat{e}_3$ has the disadvantage of leading to an undefined measure when $c_3(F) = 0$ as for the variates *U1* and *U2* in the benchmark. We propose, in these cases, an absolute measure given by the *3rd* moment itself $c_3(\hat{F})$.

c) - The area difference between the original and the approximating distribution was

proposed in terms of the cdf's:

$$\hat{A} \;=\; \frac{1}{c_1(F)} \int_0^\infty \,|\, \hat{F}(t) \;-\; F(t) \,|\, d\, t$$

The density area difference $\hat{D}$ provides the distance between two normalized areas, and is, thus, a measure normalized between 0 and 2. $\hat{D}$ is also related to the relative entropy [13].

References

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

[2] S. Asmussen and O. Nerman. Fitting PH distributions via the EM algorithm. In *Proceedings: "Symposium i Advent Statistik"*, 335–346, Copenhagen, 1991.

[3] A. Bobbio and A. Cumani. Markov models: a new class of distributions for the analysis of lifetime data samples. In *Proceedings of the 4-th EUREDATA Conference*, Paper: 10.3/1-13, Venice, 1983.

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

[5] A. Bobbio, A. Cumani, A. Premoli, and O. Saracco. Modelling and identification of non-exponential distributions by homogeneous Markov processes. In *Proceedings of the 6-th Advances in Reliability Technology Symposium*, pages 373–392, Bradford (UK), 1980.

[6] A. Bobbio and M. Telek. *Parameter estimation of Phase type distributions.* Technical Report, R.T. 423, Istituto Elettrotecnico Nazionale G. Ferraris, August 1992.

[7] R.F. Botta, C.M. Harris, and W.G. Marchal. Characterizations of generalized hyperexponential distribution functions. *Stochastic Models*, 3:115–148, 1987.

[8] W. Bux and U. Herzog. The phase concept: approximation of measured data and performance analysis. In K.M. Chandy and M. Reiser, editors, *Computer Performance*, pages 23–38, North-Holland, Amsterdam, 1977.

[9] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[10] D.R. Cox. A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Phylosophical Society*, 51:313–319, 1955.

[11] A. Cumani. On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Reliability*, 22:583–602, 1982.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[13] L. Devroye and L. Györfi. *Non Parametric Density Estimation. The L_1 View*. Wiley, New York, 1985.

[14] C.M. Harris, W.G. Marchal, and R.F. Botta. A note on generalized hyperexponential distributions. *Stochastic Models*, 8:179–191, 1992.

[15] C.M. Harris and E.A. Sykes. Likelihood estimation for generalized mixed exponential distributions. *Naval Research Logistic Quarterly*, 34:251–279, 1987.

[16] M.A. Johnson. Selecting parameters of phase type distributions: combining nonlinear programming, heuristics and Erlang distributions. *ORSA Journal on Computing*, 69–83, 1993.

[17] M.A. Johnson and M.R. Taaffe. Matching moments to Phase distributions: density function shapes. *Stochastic Models*, 6:283–306, 1990.

[18] M.A. Johnson and M.R. Taaffe. Matching moments to Phase distributions: mixtures of Erlang distribution of common order. *Stochastic Models*, 5:711–743, 1989.

[19] M.A. Johnson and M.R. Taaffe. Matching moments to Phase distributions: nonlinear programming approaches. *Stochastic Models*, 6:259–281, 1990.

[20] D.W. Kammler. Least squares approximation of completely monotonic functions by sums of exponentials. *SIAM Journal on Numerical Analysis*, 16:801–818, 1979.

[21] L. Lipsky. *Queueing theory - A linear algebraic approach.* Mcmillan, New York, 1992.

[22] H. Madsen and B.F. Nielsen. The use of Phase-type distributions for modelling packet-switched traffic. In A. Jensen and V.B. Iversen, editors, *Teletraffic and Data-traffic (Proceedings 13-th International Teletraffic Congress, ITC-13)*, pages 593–599, Elsevier Science Publisher, 1991.

[23] M. Malhotra and A. Reibman. Selecting and implementing phase approximations for semi-Markov models. *Stochastic Models*, 9, 1993.

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

[25] B.F. Nielsen. *Modelling of multiple access systems with phase type distributions.* Technical Report, Phd Thesis, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, 1988.

[26] C.A. O'Cinneide. On non-uniqueness of representations of Phase-type distributions. *Stochastic Models*, 5:247–259, 1989.

[27] C.A. O'Cinneide. *Order and triangular order of phase-type distributions.* Technical Report, Purdue University, 1992.

[28] A. Premoli. Piecewise-linear programming: the compact cplp algorithm. *Mathematical Programming*, 36:210–227, 1986.

[29] A. Ruhe. Fitting empirical data by positive sums of exponentials. *SIAM Journal on Scientific and Statistical Computing*, 1:481–498, 1980.

[30] R. Sahner and K.S. Trivedi. Reliability modeling using SHARPE. *IEEE Transactions on Reliability*, R-36:186–193, 1987.

[31] L. Schmickler. MEDA: mixed Erlang distributions as Phase-type representations of empirical distribution functions. *Stochastic Models*, 8:131–156, 1992.

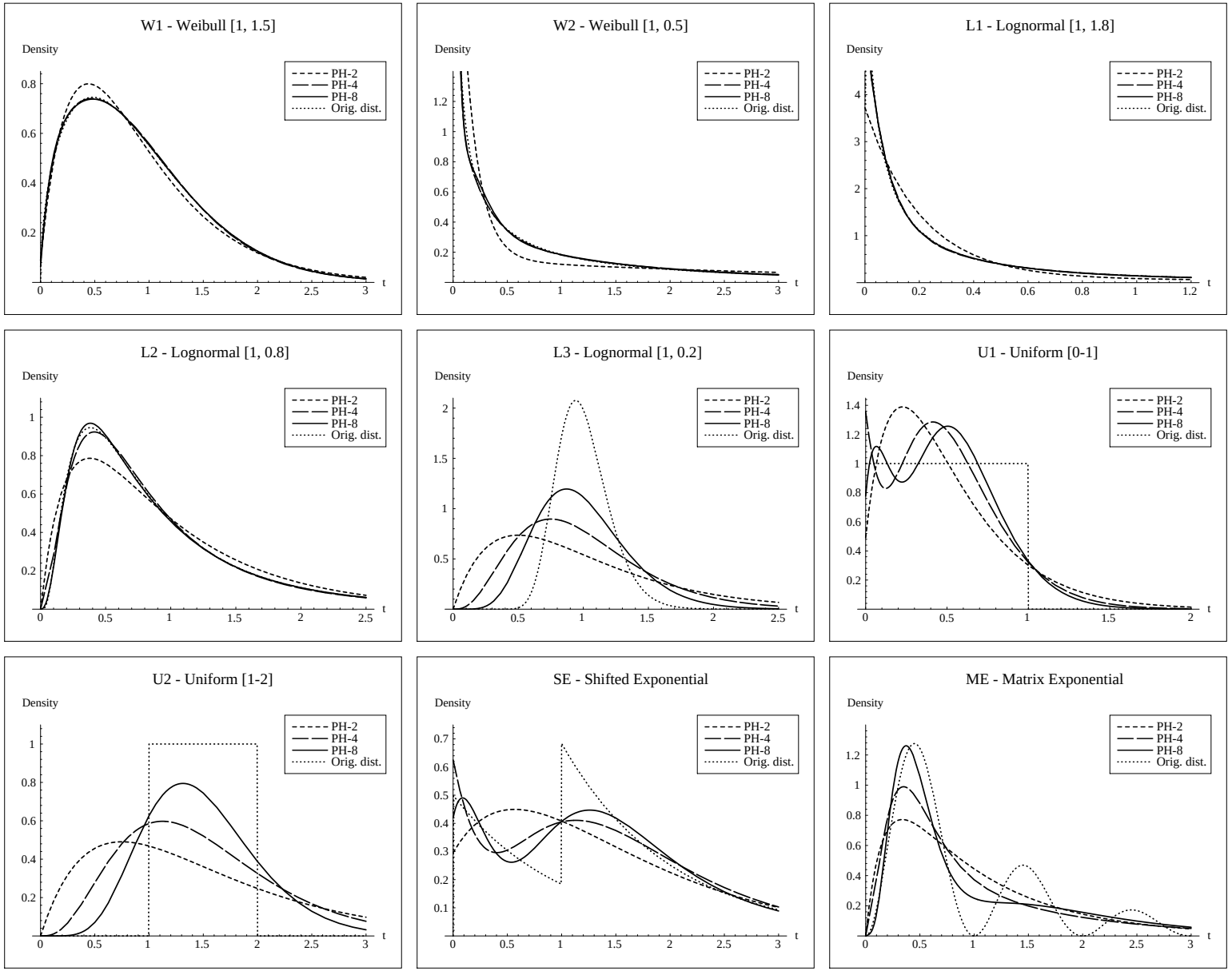
[32] L. Schmickler and H.D. Südhofen. Representation of data network measurement results by the mixed Erlang approximation method MEDA. *Archiv für Elektronik und Übertragungstechnik*, 43:8–12, 1989.

[33] C. Singh and R. Billinton. *System Reliability Modelling and Evaluation.* Hutchinson, London, 1977.

[34] C. Singh, R. Billinton, and S. Lee. The method of stages for non-markovian models. *IEEE Transactions on Reliability*, R-26:135–137, 1977.

[35] A. van de Liefvoort. *The moment problem for continuous distributions.* Technical Report, University of Missouri, WP-CM-1990-02, Kansas City, 1990.

[36] J.M. Varah. On fitting exponentials by nonlinear least squares. *SIAM Journal on Scientific and Statistical Computing*, 6:30–44, 1985.


W1 - Weibull [1, 1.5]
W2 - Weibull [1, 0.5]
L1 - Lognormal [1, 1.8]
L2 - Lognormal [1, 0.8]
L3 - Lognormal [1, 0.2]
U1 - Uniform [0-1]
U2 - Uniform [1-2]
SE - Shifted Exponential
ME - Matrix Exponential
Density
t
PH-2
PH-4
PH-8
Orig. dist.