Available online at www.sciencedirect.com

SciVerse ScienceDirect

Systems Engineering Procedia 2 (2011) 344 – 351

Procedia
Systems Engineering

SS&IAGA-EM-based Algorithm for Fitting a Continuous PH Distribution

Lu Hu[a,b], Yangsheng Jiang [a,b], Luxi Zhang[a]

[a]*College of Transportation and logistics, Southwest Jiaotong University, Chengdu 610031,Sichuan,China*
[b]*Key Laboratory of Comprehensive Transportation of Sichuan Province*

Abstract

It is an important and difficult task in model analysis of traffic engineering to fit with the general distribution or test data whether which fit the distribution or not via PH distribution. Although there are lots of methods to fit continuous PH distribution, which are all lack of efficiency and numerical stability. For giving consideration to both fitting effect and efficiency of a continuous PH distribution, we apply the maximum likelihood method to the dense subset HErD of PH distribution and design a SS & IAGA-EM algorithm for study. In the data fitting test for long-tailed distribution function, partial peak distribution function and heavy-tailed distribution function with a sample size of 104, the maximum error of the algorithm is 7.32%. When operated in a standard PC with 2.5 GHz Pentium CPU running under the operating system of Windows XP, the maximum operating time of the algorithm is 100s, which meets the demand for effectiveness and efficiency.

Keywords：PH; SS&IAGA-EM algorithm; Maximum likelihood method; HErD; Traffic engineering;

1.Introduction

Phase-Type (PH) distribution has good versatility, computability, which is also analytical. Therefore it has become an important tool for stochastic analysis in many research areas which includes queuing system analysis, reliability modeling and analysis, performance analysis and optimization of communications systems and so on. PH distribution has replaced the special status of exponential distribution in the stochastic model analytic processing, and quickly became a powerful tool for contemporary analysis of stochastic models. We can say that the analysis of stochastic model has developed to a new stage where PH distribution plays a significant role [1]. The research on theory and application of PH distribution has become a hot spot in many disciplines.

We called the parameter estimate of PH distribution which is basing on sampling data and the process of fitting other known distribution PH fitting. The data fitting method (or call parameter estimation method) of PH distribution is an important basis for solving practical problems via the application of PH distribution, which is the first step to use the PH distribution.

The data fitting problem of PH distribution has the following characteristics [2]: (1) it is highly nonlinear; (2) it usually needs a lot of estimated parameters; (3) the relationship between the parameters of PH distribution and the

* Lu Hu. Tel.: +86-15884529679.
E-mail address: hulu361@126.com.

doi:10.1016/j.sepro.2011.10.048

shape of its probability distribution function is very complex; (4) the representation of PH distribution is not unique.

These characteristics make the research on data fitting problem for the PH distribution a big difficulty and a big challenge for researchers [3].Nowadays, there are two main methods for the study of PH distribution: moment matching method and maximum likelihood estimators.

Moment matching method for PH distribution generally use a PH distribution to match a general distribution or the first orders of origin moments of observation data. This method has the advantage of little computation and high efficiency, while the disadvantages are as following [4]: (1)The fitting of the first three moments can not describe the shape of the experience probability density function very well. Fitting more moments can improve fitting results, but meanwhile it will greatly enhance the complexity of estimation algorithm. (2) When fitting the observation data of unknown distributions, the results from moment estimation method are worse than those from maximum likelihood estimation method. (3) Moment estimation method can not be easily used for fitting, which is because the moments of censored data can not know for sure. Therefore, moment estimation only apply to somewhat more restrictive models, ie, models that consists of a few exponential phases only.

As to the research on maximum likelihood estimation methods for PH distribution, the general idea is: first select an appropriate set of PH distributions (a general PH distribution or a subset of PH distribution, such as APH, HED, HErD and CHED, etc. The relationship between these distributions is shown in Figure 1), and then use a variety of algorithms for selected sets to carry on maximum likelihood estimation. All techniques based on maximum likelihood estimation can be divided into the following four kinds according to the classification of algorithm:

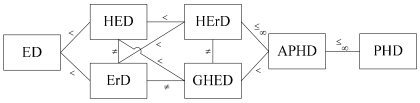


Fig. 1. Relationship of subclasses of phase-type distributions

- EM algorithm

Asmussent [3] proposed EM algorithm to fit the data of general PH distribution, and used the C language to implement the general PH distribution fitting tool Empht [5]; The method has the advantage of using the general PH distribution, which can take advantage of the flexibility of PH distribution to achieve the required fitting results with minimum order. The disadvantage is that the fitting effect of EM algorithms depends heavily on the initial parameters, and the convergence speed is very slow; Meanwhile, the number of parameters a general N-order PH distribution needing to be estimated is up to $N^2 + N$, so the computation of algorithm is very large.

- Nonlinear programming algorithm

Khoshgoftaar [4] applied gradient method directly to fit data to mixed exponential distribution with two branches. The advantage is that the fitting efficiency is extremely high, while the disadvantage is that the fitting ability is very limited. In [6],Harris gives maximum likelihood estimation algorithm of GHED using the method of nonlinear programming. The method has the advantage of high fitting efficiency and unique optimal solution, but the disadvantage is that the obtained distribution don't necessarily have Markovian nature.

- Heuristic algorithm

Feidmann, etc. [7] fitted HED which is a subset of PH distribution to heavy-tailed distribution, using the method of recursive heuristic algorithm. It can fit heavy-tailed Pareto and Weibull distribution very well whose probability density function is monotone decreasing. For non-monotonic heavy-tailed distribution, this method can not attain good fitting effects; Moreover, the algorithm can only fit a known distribution, which can not fit the observed data directly. So the actual use of this algorithm should be divided into two steps: first fit data to heavy-tailed distribution such as Pareto or Weibull, etc. , then fit HED to the obtained distribution, which makes the algorithm inconvenient to be applied in practice.

- Improved EM algorithm

For the defects of EM algorithm, Riska [8] proposed a new method which is called D&C-EM (divide-and-conquer-EM). This method has the advantage of fitting the heavy-tailed data directly, and can fit the data in despite of whose empirical probability density function is monotone or not; The disadvantage is that we must determine whether the data is monotonous in advance, and it's very difficult to analyse data online as well as divide the main part of non-monotonic data and the tail part reasonably. Huang Zhuo, etc. [9]proposed accelerate numerical EM algorithm, which can not only guarantee the convergence of the algorithm, but also improve the convergence speed when fitting the data of PH distribution to EM algorithm effectively. However, there is still an issue of the sensitivity to initial value.

It can be discovered from the analysis of the status that there is still no data fitting method which can strike a balance between fitting effect and efficiency currently.

From the view of estimating methods, moment estimate only applies to somewhat more restrictive models. When fitting "the observed data whose distribution is unknown", the result attained from the method of moment estimation is even worse than that from the method of maximum likelihood estimation. This paper applies maximum likelihood estimation.

From the view of the distribution we takes, using general PH distribution can take full advantage of the fitting ability of each state. But because of too many parameters and too complex solution space, it makes the fitting to general PH distribution difficult and inefficient; While the number of parameters of APH distribution is significantly less than that of general PH distribution, the efficiency of data fitting remains low, and the calculation of correlation function for APH distribution exists serious problem of numerical instability; The special structure of HErD makes its value calculation very stable, which also makes HErD as strong as APH and general PH distribution which has no order-limits in theory [3]. Although it can strike a balance between fitting effect and efficiency, determining the number of order and branches is still a problem, so this paper carries on further research on HErD distribution.

From the view of algorithm, optimization methods which are aimed at solving the data fitting problem of PH distribution mostly base on grads, such as nonlinear programming, EM algorithm, etc. The grads-based optimization methods can only guarantee the local optimality of solutions. So how to avoid the data fitting method for PH distribution into a local optimal solution is one of the problems which should be solved. Some scholars tried heuristic algorithm and the improved EM algorithm, making some progress but also bringing some problems such as premature convergence. As the improved adaptive genetic algorithm (IAGA)is significantly effective for global optimization of complex problem [10], scatter searching algorithm (SS) can search in a wide range and converge fast, which enhances the ability of global optimization of IAGA algorithms [11].In addition, EM algorithm has an efficient local search capability, we syncretize this three algorithms and present SS & IAGA-EM algorithm which is used to fit HErD distribution. The presented algorithm can strike a balance between fitting effect and efficiency.

2.Maximum likelihood estimation model of HErD

Let X to be any non-negative random variables, then denote a set of simple random sample of observations by x_1, x_2, ..., x_n. The problem of HErD density estimation is to establish the mixture density model basing on the sample of observations such as the type shown in *(1)*, which includes determining the number of branches m and the values of model parameters. In most cases, maximum likelihood estimation has many excellent features, such as strong consistency, consistent asymptotic normality, optimal asymptotic normality, etc. Therefore, people usually first consider the maximum likelihood method to estimate the parameters of probability distribution both in theory and practice. So is parameters estimation for HErD density.

Assume that different samples have statistical independence, then

$$F(X|\theta) = F(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

This is a function with respect to θ, which can also be called a likelihood function of θ respect to x. Because of the monotonicity of the logarithmic function and the particularity of " product can be simplified into sum ", we define the log-likelihood function

$$L(\theta) = \ln \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

Estimating θ via maximum likelihood is to evaluate the maximum of log-likelihood function. Thus, the maximum likelihood estimation model for HErD can be written as

$$L(\hat{\theta}) = max\, L(\theta)$$

$$s.t.\begin{cases}\theta = [\alpha_1\ \alpha_2\ \dots\ \alpha_m\ r_1\ r_2\ \dots\ \alpha_m\ \lambda_1\ \lambda_2\ \dots\ \lambda_m] \\ L(\theta) = \sum_{i=1}^{n} \ln \sum_{j=1}^{m} \alpha_j \frac{(\lambda_j x_i)^{r_j-1}}{(r_j-1)!}\lambda_j e^{-\lambda_j x_i} \\ \sum_{i=1}^{m} \alpha_i = 1,\ \ \alpha_i \ge 0, i = 1,2,\dots,m \\ 1 \le r_1 \le r_2\ \dots \le r_m; r_i \in N^*, i = 1,2,\dots,m \\ \lambda_i > 0, i = 1,2,\dots,m \end{cases} \tag{1}$$

Where, selecting the values of *m* and r_i which are discrete parameter of HErD is always an important and complex problem which should be solved. Since the total number of branches m determines the number of parameters, the process of parameter optimization will seriously affect efficiency. Therefore this article still asks user to determine the branch number *m*. In addition, if automatically optimizing the order number r_i of each branch via traditional algorithm without restriction, there will be kind of serious numerical instability problems. So this paper adopt the strategy of "M is given by user to limit the maximum branch order, that is $r_m \le M$" to narrow the search range (the value of M can't be too small, which should meet Theorem 3), and designed SS & IAGA-EM algorithm to automatically optimize the order number of each branch and the other parameters to avoid numerical instability and enhance fitting efficiency and effect as a whole as well.

3.SS&IAGA-EM Algorithm Design

In the field of evolutionary algorithms, although SS & IAGA algorithm is efficient and robust, evolutionary algorithms are not the most successful methods for any particular areas generally. They are usually not a patch on the algorithm which is specialized to deal with the problems in this area. We call the latter original algorithm. So how can we apply SS & IAGA algorithm into practice? One of the most effective way is to adopt a mixed strategy, that is, integrating the original algorithm with SS & IAGA algorithm effectively to design a new hybrid algorithm, the performance of which is better than both that of original algorithm and SS&IAGA algorithm.

As described in 1, EM algorithm is basing on gradient ascent, which can guarantee the likelihood of model for training data increases after each iteration. However, as a gradient-ascent mountain-climbing algorithm it has a big flaw, that is, usually EM algorithm can only get local optimal solution. SS & IAGA algorithm can solve the local minima problem efficiently by combining the strong points of IAGA and SS algorithm. Therefore, in this paper we propose a hybrid algorithm (that is, SS & IAGA-EM algorithm)which mostly bases on SS & IAGA algorithm and combines EM algorithm to take both the fitting effect and efficiency of continuous PH distribution into account. Fig. 2. shows the flow chart of this algorithm, whose main idea is: in order to avoid a large number of operations of summation and logarithmic, we select the minimum sample from the total sample randomly, globally optimize log-likelihood estimate of minimum sample via SS & IAGA algorithm, then sequentially determine the basic outline and parameters of HErD fitting curves, and finally further optimize parameters for the total sample via EM algorithm until it reaches the desired accuracy.

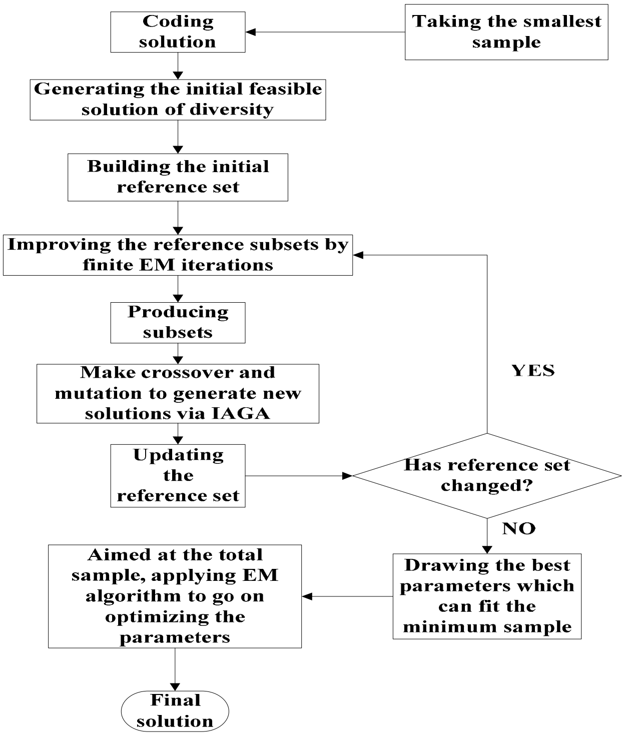
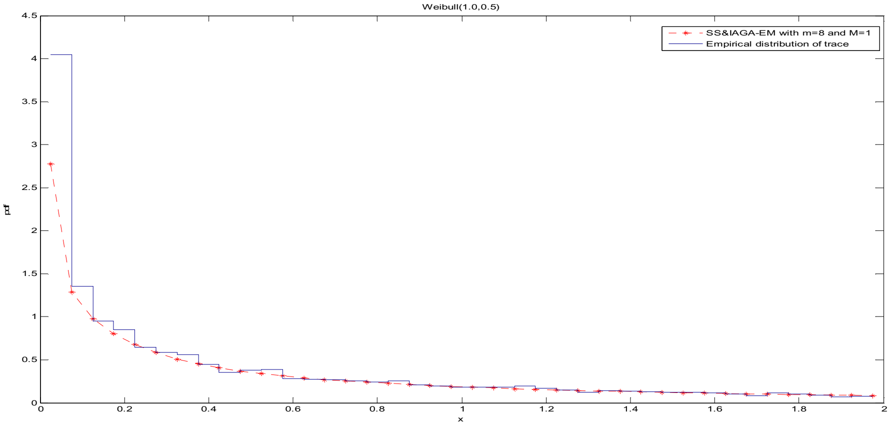

Coding solution
Taking the smallest sample
Generating the initial feasible solution of diversity
Building the initial reference set
Improving the reference subsets by finite EM iterations
Producing subsets
Make crossover and mutation to generate new solutions via IAGA
Updating the reference set
Has reference set changed?
YES
NO
Drawing the best parameters which can fit the minimum sample
Aimed at the total sample, applying EM algorithm to go on optimizing the parameters
Final solution

Fig. 2. Algorithmic process of SS&IAGA-EM

4.Case Study

In this section, in order to test and verity the effectiveness in striking a balance between effect and efficiency of HErD fitting SS & IAGA-EM algorithm (has been programmed by MATLAB software) presented in this paper, we adopt 10^4 data samples which are generated from long-tailed distribution function Weibull (1.0,0.5), partial peak function Weibull (1.0,5) and heavy-tailed distribution function Pareto-II (1.5,2.0) to operate on a standard PC with 2.5 GHz Pentium CPU running the operating system Windows XP.



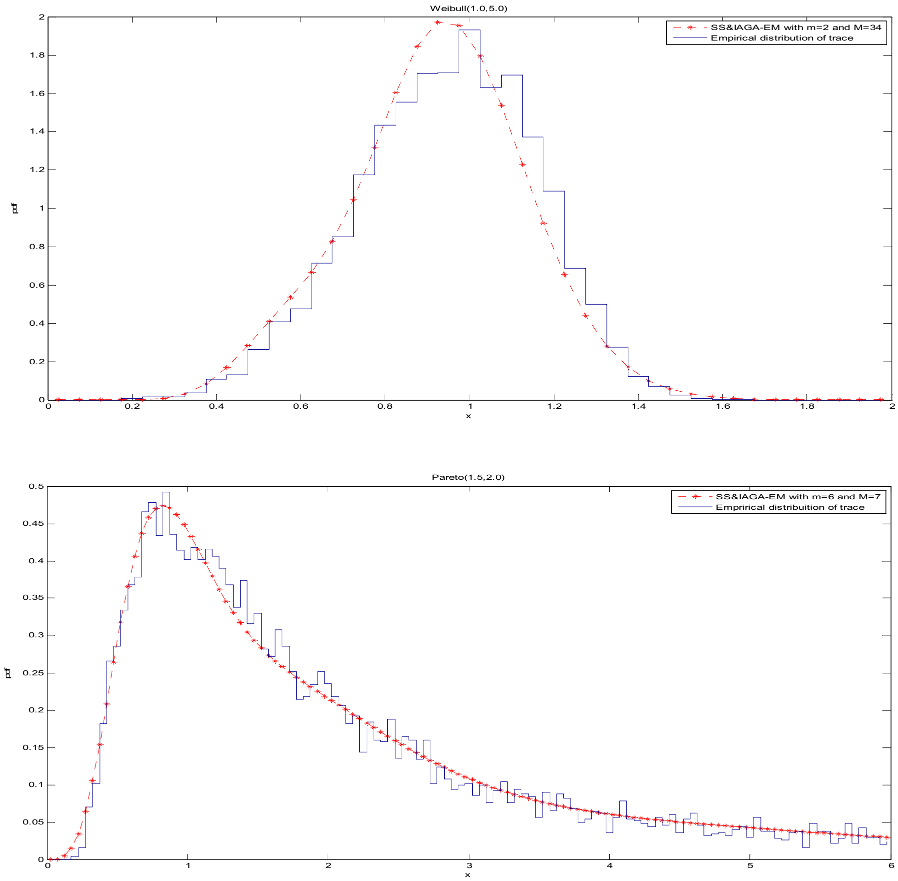


Fig. 3. The fitting effect diagram of SS&IAGA-EM algorithm

Table 1. The fitting parameters and indexes of SS&IAGA-EM algorithm

Weibull(1.0,0.5)	Input parameters	number of branches	8
		Order of the largest branch	1
	Fitting indexes	Log-likelihood	-10999.03
		Average relative error [%]	3.41
		Number of iteration	391
		CPU time[sec]	12
	Output parameters	States	8
		Number of phases	1,1,1,1,1,1,1,1
		weight	0.07,0.28,0.21,0.15,0.15,0.09,0.04,0.01
		lambda	0.09,0.32,1.04,2.14,7.06,51.80,625.85,28993.04
Weibull(1.0,5)	Input parameters	number of branches	2
		Order of the largest branch	34
	Fitting indexes	Log-likelihood	1402.30
		Average relative error [%]	7.32
		Number of iteration	820
		CPU time[sec]	100
	Output parameters	States	50
		Number of phases	16,34
		weight	0.23,0.77
		lambda	23.13,34.31
Pareto-II(1.5,2.0)	Input parameters	number of branches	6
		Order of the largest branch	7
	Fitting indexes	Log-likelihood	-87643.01
		Average relative error [%]	6.47
		Number of iteration	613

		CPU time[sec]	61
	Output parameters	States	31
		Number of phases	3,5,5,5,6,7
		weight	0.01,0.23,0.05,0.08,0.38,0.25
		lambda	0.03,1.07,0.31,6.32,3.06,7.67

Figure 3 shows the fitting effect of three samples under the accuracy of the relative likelihood function of 10^{-6} , while in Table 1 there are detailed records of the fit indexes. It can be seen from the figure that SS & IAGA-EM algorithm can make HErD fit to the distribution of sample in a high degree. From the fit indexes (average relative error and CPU time) in Table 1 we can find: the maximum average relative error of SS & IAGA-EM algorithm exists in the fitting to partial peak function Weibull (1.0,5), that is 7.32%, while the minimum exists in the fitting to long-tailed distribution function Weibull (1.0,0.5), that is 3.41%, which shows that accuracy is satisfactory; the maximum and minimum CPU time are 100s and 12s, which is amazing that a fitting of 10^4 data samples can have such speed, so efficiency is satisfactory as well. In a word, SS & IAGA-EM algorithm of HErD data fitting proposed in this paper can meet the needs of effectiveness and efficiency.

5.Conclusion

Aimed at the shortcomings of falling into the local optimal solution easily in continuous PH distribution fitting via EM algorithm, nonlinear programming algorithms, heuristics, improved EM algorithm or being inefficient, we apply the maximum likelihood method to select the dense subset HErD of PH distribution and design a SS & IAGA-EM algorithm in this paper in order to strike a balance between effect and efficiency. Then use Matlab to test long-tailed distribution function, partial peak function and heavy-tailed distribution function of 10^4 sample data, which shows a good fitting result and efficiency.

In order to illustrate the efficiency and stability of the algorithm better, the fitting test for more types of actual sample data and comparing with other existing algorithms remains to be done, thess will be the author's next work.

6.Copyright

All authors must sign the Transfer of Copyright agreement before the article can be published. This transfer agreement enables Elsevier to protect the copyrighted material for the authors, but does not relinquish the authors' proprietary rights. The copyright transfer covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microfilm or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce any figures for which copyright exists.

References

1.Tian N,Li Q.The PH Distribution and its Applications in Various stochastic Models [J].Communication On Applied Mathematics and Computation,1995,9(2):1-15.

2.Thtummler A,Buchholz P,Telek M.A novel approach for phase-type fitting with the EM algorithm [J].IEEE Trans.On Dependable and Secure Computing,2006,3(3):245-258.

3.Asmussen S,Nerman O,Olsson M.Fitting phase-type distributions via the EM algorithm[J].Scandinavian Journal of Statistics,1996,23(4):419-441.

4.Khoshgohaar T,Perros H G.A comparison of three methods of estimation for approximating general distributions by a COXian distribution[C]//Proceedings 3rd International Workshop on Modelling Techniques and Performance Evaluation,Paris,1987:169-77.

5.Haggstrom O,Asmussen S,Nerman O.EMPHT-a program for fitting phase-type distributions[R].Department of Mathematics,Chalmers University of Technology,Goteborg,Sweden,1992.

6.Harris C M,Sykes E A.Likelihood estimation for generalized mixed exponential distributions[J].Naval Research Logistic Quarterly,1987,34(2):251-279.

7.Feldmann A,Whitt W.Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models[J].Performance Evaluation,1998,31(3-4):245-279.

8.Risha A,Diev V,Smirni E.An EM-based technique for approximating long-tailed data sets with PH distributions [J].P erformance

Evaluation,2004,55(2):147-164.

9.Huang Z,Pan X,GUO B.Numerical Acceleration EM Algorithm for PH Distribution Data Fitting[J].Computer Engineering,2008,34(14):1-6.

10.Liu T K, Tsai J T, &Chou J H. Improved genetic algorithm for job-shop scheduling problem.International Journal of Advanced Manufacturing Technology[J].2006,27:1021–1029.

11.Sagarna R, Lozano J A. Scatter Search in software testing comparison and collaboration with Estimation of Distribution Algorithms [J]. European Journal of Operational Research(S0377-2217),2005,169(2): 392-412.