Estimation of the End-to-End Delay in 5G Networks Through Gaussian Mixture Models

Diyar Fadhil, Rodolfo Oliveira

Estimation of the End-to-End Delay in 5G Networks through Gaussian Mixture Models

Diyar Fadhil[1,2], Rodolfo Oliveira[1,2]

[1] Departamento de Engenharia Electrótecnica e de Computadores, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[2] Instituto de Telecomunicações, Portugal

Abstract. Network analytics provide a comprehensive picture of the network's Quality of Service (QoS), including the End-to-End (E2E) delay. In this paper, we characterize the E2E delay of heterogeneous networks when a single known probabilistic density function (PDF) is not adequate to model its distribution. To this end, multiple PDFs, denominated as components, are assumed in a Gaussian Mixture Model (GMM) to represent the distribution of the E2E delay. The accuracy and computation time of the GMM is evaluated for a different number of components. The results presented in the paper consider a dataset containing E2E delay traces sampled from a 5G network, showing that the GMM's accuracy allows addressing the rich diversity of probabilistic patterns found in 5G networks and its computation time is adequate for real-time applications.

Keywords: End-to-End delay, Quality of Service, Gaussian Mixture Model.

1 Introduction

Network analytics was introduced to analyze network problems caused by the enormous increase of connected devices [1]. The Quality-of-Service (QoS) metrics are performance indicators of the network status. One of the most tangible QoS metrics is the End-to-End (E2E) delay. Due to the different requirements of each service in terms of throughput, reliability, and time sensitivity, it is crucial to know the E2E delay probabilistic features. For instance, in 5G networks, Ultra-Reliable Low Latency Communication (URLLC) applications demand low latency networks [2], while other applications such as opportunistic sensing do not have such requirements.

The E2E delay is defined as the time needed to transfer a packet from one endpoint to another, i.e., the time between the instant the transmission starts at the source node and the instant the packet is completely received at the destination node. From a network management viewpoint, it is essential to identify the network E2E delay profile, so that its suitability to support different delay-constrained services can be assessed over time. Using probabilistic models to determine the E2E delay distribution is also crucial to support different delay management strategies [3].

In this section, we motivate the paper, its contributions, and the related work in the field. Section 2 introduces the scope of the paper. The proposed methodology and its performance evaluation are presented in Section 3 and Section 4, respectively. Finally, Section 5 concludes the paper.

Research Question and Motivation. The critical challenge in characterizing the E2E delay is determining which mixed distributions represent the set of experimental data collected over time. Due to the network's heterogeneous nature, especially due to the difference radio access and communication technologies available in 5G networks [4], the characterization of the E2E delay is a complex task due to different delay patterns imposed by the multiple technologies set up in the network. Consequently, the E2E delay often does not follow a single and known probability density function (PDF) but a mixture of them. This motivates a modeling approach based on probabilistic mixture models that combine two or more distributions to increase the model's accuracy. To this end, the hypothesis explored in this work is the evaluation of modeling the E2E through a Gaussian Mixture Model (GMM) [5]. The research questions addressed in this work are threefold:

(1) What is the impact of the number of GMM components on the model's accuracy error?
(2) How to estimate the parameters of each distribution adopted in the GMM?
(3) How much time does it take to compute the GMM model, and whether it is adequate for real-time applications?

Related Work. A GMM is defined as a parametric probability density function that consists of a linear combination of multiple Gaussian distributions [6]. Different approaches to estimate the distributions' parameters and weights values based on observed data include the Maximum Likelihood Estimation (MLE), Expectation-Maximization (EM), Minimum Message Length (MML), Moment Matching (MM), and Penalized Maximum Likelihood Expectation-Maximization (PML-EM) [5]. The MLE approach maximizes the likelihood function between a known distribution and the observed data. The MML method is an information measure for statistical comparison. The MM method finds the unknown parameters by obtaining the expected values of the random variable's powers of population distribution model equal to the sample moments. MM can be employed as an alternative approach for MLE in most complex problems due to its simple, easy, and quick computation. The PML-EM is an approach to estimate the parameters in cases when the likelihood is relatively flat, which makes the ML estimation determination difficult. The EM is an iterative algorithm that maximizes the likelihood expectation between data and a mixture of distributions. However, EM's convergence rate is influenced by the initialization random values, and it is hard to define the number of mixture model distributions and how they affect the accuracy of the approximation. EM's dependency on the initialized values is one of the main causes of slow convergence, as mentioned in [7].

The GMM adopts Gaussian distributions and every local population optimum for the MLE problem is globally optimal. The GMM has received significant attention in the literature, particularly to support the estimation of QoS network parameters. The work

in [8] investigates how to estimate the link-delay distributions based on end-to-end multicast measurements and adopts an MLE-based GMM model. In [9], it is proposed a known conditional distribution and an unknown finite Gaussian mixture to approximate the weighting of the GMM components. The work in [10] proposed an improved EM algorithm to select the number of components of GMM and simultaneously estimate the weight of these components and unknown parameters. The work in [11] suggests a new clutter elimination method based on GMM and EM estimation, which attempts to estimate and perform fast clutter with a small amount of data. The work in [12] provides a comprehensive analysis of actual latency values collected among various data center locations and a GMM approximation is proposed to simulate and emulate the deployment of applications and services in the cloud.

Contributions. The main contribution of this paper is the characterization of the influence of the number of GMM components and the number of data samples on the accuracy of the E2E delay model of a 5G network. By doing so, the GMM's error and its computing time can be drawn as a function of the number of samples and GMM components, which is of high importance to assess the accuracy of the model and its applicability in terms of computing time to be used in practical 5G networks scenarios.

2 Technological Innovation for Digitalization and Virtualization

Digitalization and virtualization are being adopted in non-real-time scenarios, such as offline data analytics adopted in several trading platforms, but also demanding real-time applications, such as autonomous driving. A common need in all these scenarios is the support of efficient communication systems and networks, capable of exchanging huge amounts of data in a short period of time. In critical real-time applications very low latency and E2E delay are required. The work reported in the paper advances the knowledge about the E2E delay of complex networks, by characterizing the E2E delay of 5G networks through mixture distributions. This knowledge is crucial to select the networks according to the requirements of the specific virtualization systems, being a piece of high importance in the implementation of virtualization systems.

3 Methodology

Although Gaussian distributions can model an impressive number of probabilistic scenarios, certain phenomena follow unknown distributions demanding more complex modeling approaches. The mixture models can be a possible solution for these cases. Given that any natural process may depend on several independent factors that form several subpopulations, the GMM models the subpopulations that can then be mixed to describe the whole distribution of the population.

A GMM is defined as a parametric PDF consisting of a linear combination of multiple Gaussian PDFs. The mixture models are usually used for multimodal or multi-peak PDF data. Fitting the multimodal data with a single distribution is not proper and

accurate. The mixture models are used to combine different distributions that better match the data density. The Gaussian distribution is adopted in the mixture models due to its theoretical and computational benefits to represent massive datasets [6]. Each Gaussian component is used to represent the subpopulations within an overall population. Three primary parameters define each component of a GMM: the mean, the standard deviation, and a weight. The Gaussian mixture model can be represented as follows

$$p(x) = \sum_{i=1}^{K} w_i \, \mathcal{N}(x|\mu_i \sigma_i), \tag{1}$$

where x is the data measurement, w_i, is the mixture weight for the i-th component, $i = 1, \ldots, K$, and K is the number of mixed Gaussian distributions, aka GMM components. Each component of the GMM is defined as

$$\mathcal{N}(\mathrm{x}|\mu_\mathrm{i}, \sigma_\mathrm{i}) = \frac{1}{\sigma_\mathrm{i}\sqrt{2\pi}} \exp\left(-\frac{(\mathrm{x}-\mu_\mathrm{i})^2}{2\sigma_\mathrm{i}^2}\right), \tag{2}$$

with mean μ_i and standard deviation σ_i. The mixture weights satisfy the constraint

$$\sum_{i=1}^{K} w_i = 1. \tag{3}$$

Therefore, the GMM parameters vector is represented by λ as follows

$$\lambda = \{w_i, \mu_i, \sigma_i\} \quad i = 1, \ldots, K. \tag{4}$$

The parameters λ are estimated based on training or measurement data. The maximum likelihood (ML) estimation is a well-known method that aims to find the mixture model parameters and maximize the GMM likelihood function with given training data. We assume that the vector of the training data samples is represented by $X = x_1, \ldots, x_T$, where the samples $x_1, \ldots, x_T$, are independent. Therefore, the likelihood function is written as

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathrm{x}_t|\lambda), \tag{5}$$

where T represents the number of samples. It is impracticable to maximize the non-linear function of parameters λ that is expressed on the ML function [11]. A practical solution for the ML estimation of the parameters λ can be determined by a numerical approach such as iterative Expectation-Maximization (EM) or similar ones [9]. In EM, initial random values are assigned to λ, which are used to determine the subsequent estimation of parameters λ. In the first step, the initial values of all parameters are determined and employed as an input of the iterative Expectation and Maximization algorithm to reach the convergent values of the different parameters. The initial values of the parameters, λ_0, are computed as follows

$$\mu_i = \frac{T * rand(i)}{2} \sum_{t=1}^{T} x_t, \tag{6}$$

$$\sigma_i = \left(\frac{1}{T}\sum_{t=1}^{T} (x_t - \mu_i)^2\right)^{\frac{1}{2}}, \tag{7}$$

$$w_i = \frac{1}{K}, \tag{8}$$

where $rand\ (i)$ represents a random number sampled from a uniform distribution between zero and one. An iterative cycle is then started until the estimated parameters λ reach a specific convergence threshold. The EM algorithm relies on an iterative approach divided in two steps:

E-step: In the Expectation step, the expectation of the likelihood function is calculated based on observed data X and the current model parameters λ_m.

M-step: In the Maximization step, the expectation of the likelihood function is used to compute new model parameters λ_{m+1} that maximize the conditional distribution given by the parameters λ_m and the observed data X. The symbols m, and $m+1$ indicate consecutive iterations. In the E-step, λ_m is used to indicate the current model parameters. In the M-step, λ_m is used to determine the subsequent model parameter λ_{m+1}. Expanding the E-step and taking separate derivatives concerning the different parameters (M-step), we get the equations as follows

$$\widehat{w}_i = \frac{1}{T}\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_m), \tag{9}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_m)\mathbf{x}_t}{\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_m)}, \tag{10}$$

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_m)x_t^2}{\sum_{t=1}^{T} P(i|\mathbf{x}_t, \lambda_m)} - \hat{\mu}_i{}^2, \tag{11}$$

and the subsequent GMM estimation parameters vector is given by

$$\lambda_{m+1} = \{\widehat{w}_i, \hat{\mu}_i, \hat{\sigma}_i\} \quad i = 1, \dots, K. \tag{12}$$

The GMM parameters can be estimated by employing the iterative Expectation-Maximization (EM) algorithm and Maximum a Posteriori (MAP) estimation [12]. Although different approaches to parametrize the GMM parameters may provide closed-form expressions [9], its computation complexity is severely increased by the number of distributions used in the mixture model. For instance, if we assume that the GMM consists of $K = 7$ components, the MM estimation needs to calculate and solve 21 equations of the moments to determine the estimations of the parameters. Therefore, the EM algorithm is a well-known approach often adopted to estimate the GMM parameters due to its iterative behavior and improved computational performance.

4 Performance Evaluation

In this section we present the simulation results and evaluate the performance of the method described in Section 3. In the EM algorithm a policy was adopted to avoid exceeding the iteration cycle when the level of convergence described by a difference parameter D_m is lower than the convergence threshold γ. D_m is given by

$$D_m = \sum_{k=1}^{K} \left(\left|\mu_{m+1_k} - \mu_{m_k}\right| + \left|\sigma_{m+1_k} - \sigma_{m_k}\right| + \left|w_{m+1_k} - w_{m_k}\right|\right). \tag{13}$$

The maximum number of iterations was limited to 25000. The MSE is used to find out how the estimated PDF is close to the observed one, and is defined as follows

$$MSE = \frac{1}{T}\sum_{t=1}^{T} (x_t - \hat{x}_t)^2. \tag{14}$$

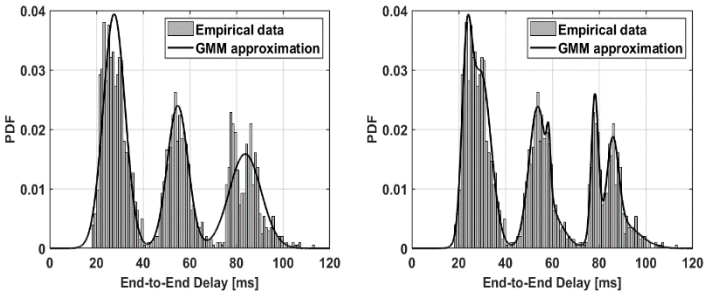
The 5Gophers dataset [13] was used as observation data. The dataset contains experimental data simultaneously obtained from three different 5G carriers. Two carriers run mmWave networks while the third one runs a mid-band network. The E2E delay dataset was gathered in three cities in the US with different urban environments. In the further analysis, we have used 2054 data records to determine the accuracy of the model as a function of the number of GMM components $K = 2, 3, 5, 8, 10$, and 12, and the number of samples T= 10, 25, 50, 100, 200, 500, and 1000. The threshold γ was set to 0.00001.

Table I summarizes the number of iterations of the EM algorithm and the MSE achieved for the different number of components. The EM algorithm was computed using the 2054 data records. Based on the numerical results, when the number of components increases, the number of iterations required to reach the convergence threshold increases because more parameters need to be estimated. On the other hand, the MSE decreases with the number of components because a higher number of components leads to a higher number of degrees of freedom to model the data.

TABLE I. Number of iterations and MSE to compute the GMM.

Components	Number of iterations	MSE
2	24	1.37E-04
3	33	1.00E-04
5	332	6.06E-05
8	2076	3.79E-05
10	2858	2.94E-05
12	4042	2.29E-05

The PDFs obtained with the GMM for 3 and 8 components are represented in Fig. 1. As can be seen in both Table I and Fig. 1, the adoption of 8 components increases the model's accuracy when compared to 3 components.



(a) 3 GMM component (b) 8 GMM components

Fig. 1. GMM approximation for different number of components.

Next, we characterize the impact of the number of E2E delay samples on the GMM's accuracy. The MSE values for a different number of GMM components and samples are summarized in Table 2. For a specific number of GMM components, the MSE decreases as the number of samples increases, which means that the increase of the number of samples leads to a lower error due to the best approximation of the likelihood expectation. In addition, for a fixed number of samples, the MSE decreases with the number of components because the GMM uses more PDFs to represent the experimental data. The computation time of the GMM is presented in Table 3. As a

general trend, the computation time increases with the number of samples for a specific number of components. This is due to the longer vector of data samples. In addition, for each number of samples the computation time increases with the number of components due to the increase of GMM complexity.

TABLE 2. MSE as a function of the number of GMM components and samples.

	T= 10	T= 25	T= 50	T= 100	T= 200	T= 500	T= 1000
K= 2	3.78E-02	1.25E-02	5.91E-03	2.86E-03	1.41E-03	5.62E-04	2.80E-04
K=3	1.85E-02	9.51E-03	4.47E-03	2.14E-03	1.05E-03	4.14E-04	2.06E-04
K=5	2.83E-03	1.72E-03	1.45E-03	9.25E-04	5.21E-04	2.52E-04	1.25E-04
K=8	5.64E-04	3.80E-04	3.13E-04	2.66E-04	2.20E-04	1.25E-04	6.92E-05
K=10	1.32E-04	1.21E-04	1.07E-04	9.32E-05	8.52E-05	7.69E-05	5.15E-05
K=12	1.13E-04	9.97E-05	8.91E-05	7.09E-05	6.03E-05	4.70E-05	3.50E-05

TABLE 3. GMM computation time [*ms*] varying the number of GMM components and samples.

	T=10	T=25	T=50	T=100	T=200	T=500	T=1000
K=2	0.68	0.78	0.87	0.91	1.18	1.84	2.53
K=3	0.70	1.23	1.60	2.09	2.94	5.09	9.96
K=5	0.75	1.42	3.47	7.68	19.98	67.1	153.4
K=8	0.77	1.46	4.85	12.04	33.28	190.9	963.3
K=10	0.83	1.49	5.42	12.65	39.87	253.9	1699.3
K=12	0.89	1.53	5.72	16.65	46.76	541.8	1772.2

Fig. 2 illustrates the logarithmic plot of the MSE and computation time as a function of the number of components. Each curve represents a specific number of samples. As can be seen in the results, to obtain MSE errors below 1E-04 it is advantageous to decrease the number of samples (T) and increase the number of GMM components (K), as the computational time is more affected by the number of samples than by the number of GMM components. This can be observed for T=25 and K=12, where the MSE is below 1E-04 and the computation time is 1.53 ms, which compares for instance with T=50 and K=10 where the MSE increases (slightly above 1E-04) and the computation time also increases to 5.42 ms.

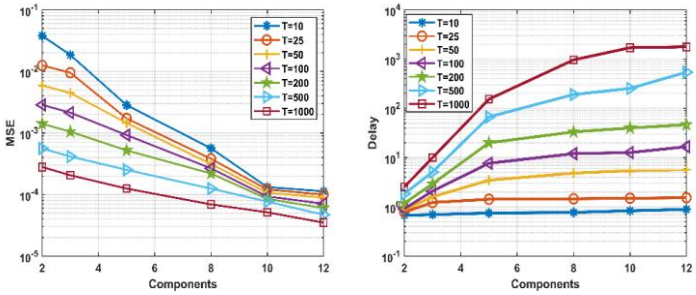


Fig. 2. MSE and computation time (Delay in milliseconds) for different number of GMM components and samples.

The tradeoff between the MSE and the computation time is of great importance because the results show that lower errors can be achieved with the cost of increased computation time. Selecting the best GMM configuration for each network depends on the network operator's requirement, their policies, and the kind of required application.

However, the number of GMM components can be varied in such a way that coarser or more accurate E2E delay distributions can be computed.

5 Conclusions

This paper characterized the accuracy and the computation time of the GMM to approximate 5G networks' E2E delay. The results indicate that higher accuracy is achieved as the number of samples and the number of components increases. However, the computation time also increases with the number of samples and components, as identified by the trade-off between the model's error and its computational time presented in Section 4. Future work includes the adoption of machine learning approaches to identify the GMM model parameters through unsupervised deep learning neural networks.

References

[1] M. Hung. Leading the iot, gartner insights on how to lead in a connected world. *Gartner Research, pages 1–29, 2017*

[2] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck. Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys & Tutorials, 20(3):2429–2453, 2018*

[3] B. G. Banavalikar. Quality of service (qos) for multi-tenant aware overlay virtual networks, *January 2019. US Patent 10,177,936*

[4] Q. Ye, W. Zhuang, X. Li and J. Rao, "End-to-End Delay Modeling for Embedded VNF Chains in 5G Core Networks," *in IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 692-704, Feb. 2019

[5] G. J. M., S. X. Lee, and S. I, Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019

[6] D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009.

[7] M. Yang, C. Lai, and C. Lin. A robust EM clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012

[8] E. Lawrence, G. Michailidis, and V, Nair. Maximum likelihood estimation of internal network link delay distributions using multicast measurements. In *Proceedings of the 37th Conference on Information Sciences and Systems*. Citeseer, 2003

[9] R. Orellana, R. Carvajal, and J. C. Aguero. Maximum likelihood infinite mixture distribution estimation utilizing finite Gaussian mixtures. *IFAC-PapersOnLine (Elsevier)*, 51(15):706–711, 2018

[10] T. Huang, H. Peng, and K. Zhang. Model selection for Gaussian mixture models. *Statistica Sinica*, pages 147–169, 2017

[11] L. Rahman, J. A. Zhang, X. Huang, Y. Jay Guo, and Z. Lu. Gaussian-mixture-model based clutter suppression in perceptive mobile networks. *IEEE Communications Letters*, 25(1):152– 156, 2021

[12] W. Cerroni, L. Foschini, G. Y. Grabarnik, L. Shwartz, and M. Tortonesi. Estimating delay times between cloud datacenters: A pragmatic modeling approach. *IEEE Communications Letters*,22(3):526–529, 2018

[13] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z. Zhang. A first look at commercial 5g performance on smartphones. In *Proceedings of The Web Conference 2020*, pages 894–905, 2020