MESSAGE PASSING-BASED INFERENCE IN THE GAMMA MIXTURE MODEL

Albert Podusenko[⋆] *Bart van Erp*[⋆] *Dmitry Bagaev*[⋆] *İsmail Şenöz*[⋆] *Bert de Vries*[⋆†]

{a.podusenko, b.v.erp, d.v.bagaev, i.senoz, bert.de.vries}@tue.nl [⋆]TU Eindhoven, [†]GN Hearing

ABSTRACT

The Gamma mixture model is a flexible probability distribution for representing beliefs about scale variables such as precisions. Inference in the Gamma mixture model for all latent variables is non-trivial as it leads to intractable equations. This paper presents two variants of variational message passing-based inference in a Gamma mixture model. We use moment matching and alternatively expectation-maximization to approximate the posterior distributions. The proposed method supports automated inference in factor graphs for large probabilistic models that contain multiple Gamma mixture models as plug-in factors. The Gamma mixture model has been implemented in a factor graph package and we present experimental results for both synthetic and real-world data sets.

Index Terms— Expectation-Maximization, Factor Graphs, Gamma Mixture Model, Message Passing, Moment Matching, Probabilistic Inference

I. INTRODUCTION

Mixture models are commonly used in the literature to model probability density functions that are outside of the exponential family. Especially Gaussian mixture models are used often, for instance in the field of natural language processing [1]. However, this paper will focus on the less common Gamma mixture models (ΓMMs). The ΓMMs allow us to efficiently model skewed distributions with positive support [2]. For example, this model can be used as the conjugate prior for the precision parameter of a Gaussian distribution. In that case, the conjugate relationship supports modeling of processes with switching noise levels.

The ΓMM has been used in a variety of applications, such as in the detection of COVID-19 in medical images [3]. The literature describes a few approaches for performing inference in the ΓMM, or the generalized ΓMM, most notably a sampling approach [4] and a variational expectation-maximization method [2]. Unfortunately, these approaches are not modular by nature, which often leads to tedious and error-prone manual derivations when extending or applying the models in a different context. In this paper we propose a modular message passing-based probabilistic inference method for ΓMMs.

We represent the ΓMM as a composite factor (node) in a Forney-style Factor Graph (FFG) [5], [6]. A benefit of the FFG representation is that all (message passing) computations are local, and as a result the ΓMM factor can be used as a plug-in module in larger probabilistic models. More details on the FFGs will be provided in Section II, where we also specify the Gamma mixture (ΓM) model.

In Section III we specify the problem that we solve in this paper: how to perform message passing-based inference in the ΓMM. A solution proposal is presented in Section IV. Specifically, in Section IV-C we provide a local expectation-maximization extension to variational message passing, and in Section IV-D we propose a moment matching-based non-conjugate variational message passing method. These solutions are verified and validated in Section V. We discuss our findings and conclude the paper in Section VI.

II. MODEL SPECIFICATION

Let $\boldsymbol{x} \triangleq [x_1, \ldots, x_K]$, where $x_k \in \mathbb{R}_{>0}$ for every $k = 1, \ldots, K$, denote a vector of strictly positive independent and identically distributed (IID) observations. The likelihood for a ΓMM with M mixture components is given by

$$p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{k=1}^{K} \prod_{m=1}^{M} \Gamma(x_k|a_m, b_m)^{s_{km}}, \tag{1}$$

where $\Gamma(x_k|a_m, b_m)$ specifies the Gamma distribution for x_k with shape and rate parameters a_m and b_m, respectively. $\boldsymbol{a} \triangleq [a_1, \ldots, a_M]$ and $\boldsymbol{b} \triangleq [b_1, \ldots, b_M]$ are vectors of the parameters of the Gamma distributions such that $a_m, b_m \in \mathbb{R}_{>0}$ for every $m = 1, \ldots, M$. For each observation x_k we have a corresponding latent selector variable $\boldsymbol{s}_k$ comprising a 1-of-M binary vector with elements $s_{km} \in \{0, 1\}$, which are constrained by $\sum_m s_{km} = 1$. We denote the vector of selector variables by $\boldsymbol{s} \triangleq [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]$.

To complete the specification of the ΓMM we need to specify priors on $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{s}$. We choose the priors as

$$p(\boldsymbol{a}) = \prod_{m=1}^{M} \Gamma\left(a_m|\alpha_m^{(a)}, \beta_m^{(a)}\right) \quad \alpha_m^{(a)}, \beta_m^{(a)} \in \mathbb{R}_{>0} \tag{2}$$

$$p(\boldsymbol{b}) = \prod_{m=1}^{M} \Gamma\left(b_m|\alpha_m^{(b)}, \beta_m^{(b)}\right) \quad \alpha_m^{(b)}, \beta_m^{(b)} \in \mathbb{R}_{>0} \tag{3}$$

$$p(\boldsymbol{s}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \prod_{m=1}^{M} \pi_m^{s_{km}} \text{ such that } \sum_{m=1}^{M} \pi_m = 1 \tag{4}$$

978-1-7281-6338-3/21/$31.00 ©2021 IEEE

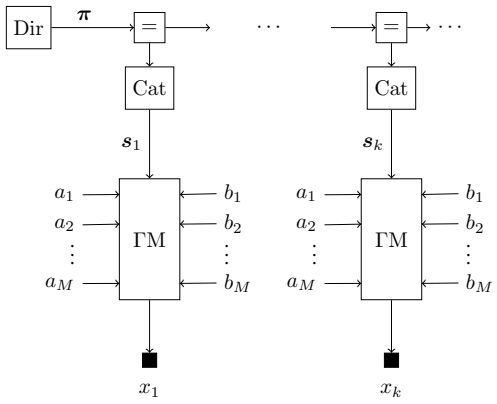


Fig. 1: An FFG representation of the ΓMM in (6). Dir and Cat denote Dirichlet and Categorical distributions respectively. The '=' nodes represent equality factors. Small black nodes denote observations. For brevity, we did not add the nodes corresponding to the distributions of shape a_m and rate b_m parameters. The inside of the ΓM node is further worked out in Table I.

and we choose a Dirichlet prior for the event probabilities $\boldsymbol{\pi} \triangleq [\pi_1, \ldots, \pi_M]$ of the categorical distribution $p(\boldsymbol{s}|\boldsymbol{\pi})$ as

$$p(\boldsymbol{\pi}) = \frac{1}{Z(\boldsymbol{\eta})} \prod_{m=1}^{M} \pi_m^{\eta_m - 1} \text{ with } Z(\boldsymbol{\eta}) = \frac{\prod \Gamma(\eta_m)}{\Gamma\left(\sum \eta_m\right)}, \quad (5)$$

where $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_M]$ are the concentration parameters with $\eta_m \in \mathbb{R}_{>0}$ for every $m = 1, \ldots, M$. The full ΓMM is then given by the joint distribution

$$p(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\pi}) = p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b}) p(\boldsymbol{a}) p(\boldsymbol{b}) p(\boldsymbol{s}|\boldsymbol{\pi}) p(\boldsymbol{\pi}). \quad (6)$$

An FFG is an undirected graph in which nodes represent factors of a global function and edges represent random variables [5]. In an FFG, each edge can be connected to a maximum of 2 factors, whereas a node can be connected to an arbitrary number of edges. Hence, FFGs usually contain multiple *equality nodes* with factors $\delta(x - x')\delta(x - x'')$ that constrain the beliefs over two "copy variables" x' and x'' to be equal to the belief over x [7]. As a matter of notational convention, in an FFG, factors are represented by square (unfilled) nodes and observations or fixed variables in these graphs are represented by small black squares, whose factors can be regarded as Dirac delta functions centered on the observed value. For a detailed explanation of the FFG formalism, we refer to [5], [6], [8]. FFGs corresponding to the ΓMM of (6) are presented in Table I and Fig. 1.

III. Problem Statement

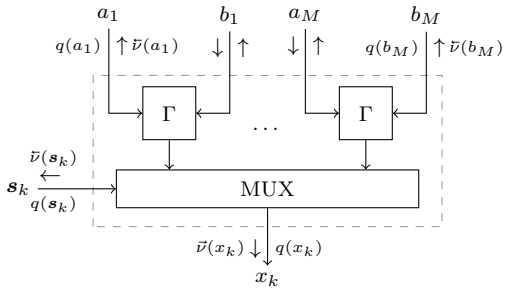
Given the ΓMM and a collection of observations $\boldsymbol{x}$ we are interested in obtaining the posterior distributions $p(\boldsymbol{a}|\boldsymbol{x}), p(\boldsymbol{b}|\boldsymbol{x}), p(\boldsymbol{s}|\boldsymbol{x})$ and $p(\boldsymbol{\pi}|\boldsymbol{x})$. Computation of the posteriors requires the integration and summation of the model (6) with respect to all remaining model variables:

$$p(\boldsymbol{a}|\boldsymbol{x}) = \frac{\sum_{\boldsymbol{s}} \int p(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{b} \, \mathrm{d}\boldsymbol{\pi}}{\sum_{\boldsymbol{s}} \int p(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{a} \, \mathrm{d}\boldsymbol{b} \, \mathrm{d}\boldsymbol{\pi}}. \quad (7)$$

Even though (7) is the exact solution to one of the inference tasks, it is intractable because the integrals involving $\boldsymbol{a}$ and $\boldsymbol{b}$ do not yield known analytical solutions. In this paper, the problem we address is how to compute approximate posteriors for the ΓMM.

Table I: Table containing (top) the Forney-style factor graph representation of the Gamma mixture node. The node indicated by MUX represents a multiplexer node, which selects the mixture component. (middle) An overview of the chosen approximate posterior distributions. Here the $\hat{\cdot}$ accent refers to the parameters of these distributions. The choice of functional form for $q(a_m)$ depends on the approximation method (Section IV). (bottom) The derived messages for the Gamma mixture node. The definitions of ζ_{km} and ρ_{km} are presented in the supplementary material at http://github.com/mlsp2021-gmm.

Factor graph



Marginals	**Functional form**
$q(a_m)$	$\delta(a_m - \hat{a}_m)$ or $\Gamma\left(a_m \mid \hat{\alpha}_m^{(a)}, \hat{\beta}_m^{(a)}\right)$ $\hat{\alpha}_m^{(a)}, \hat{\beta}_m^{(a)} \in \mathbb{R}_{>0}$
$q(b_m)$	$\Gamma\left(b_m \mid \hat{\alpha}_m^{(b)}, \hat{\beta}_m^{(b)}\right)$ $\hat{\alpha}_m^{(b)}, \hat{\beta}_m^{(b)} \in \mathbb{R}_{>0}$
$q(\boldsymbol{s}_k)$	$\prod_{m=1}^{M} \hat{\pi}_m^{s_{km}}$ such that $\sum_{m=1}^{M} \hat{\pi}_m = 1$
$q(x_k)$	$\Gamma\left(x_k \mid \hat{\alpha}_k^{(x)}, \hat{\beta}_k^{(x)}\right)$ $\hat{\alpha}_k^{(x)}, \hat{\beta}_k^{(x)} \in \mathbb{R}_{>0}$
Messages	**Functional form**
$\overrightarrow{\nu}(a_m)$	$\exp\left(\hat{\pi}_k \left(a_m \zeta_{km} - \log\Gamma(a_m)\right)\right)$
$\overrightarrow{\nu}(b_m)$	$\Gamma\left(b_m \middle\vert 1 + \hat{\pi}_m \mathrm{E}\left[a_m\right], \hat{\pi}_m \frac{\hat{\alpha}_k^{(x)}}{\hat{\beta}_k^{(x)}}\right)$
$\overleftarrow{\nu}(\boldsymbol{s}_k)$	$\prod_{m=1}^{M} \rho_{km}^{s_{km}}$ such that $\sum_{m=1}^{M} \rho_{km} = 1$
$\overrightarrow{\nu}(x_k)$	$\Gamma\left(x_k \middle\vert \sum_{m=1}^{M} \hat{\pi}_m \mathrm{E}\left[a_m\right], \sum_{m=1}^{M} \hat{\pi}_m \frac{\hat{\alpha}_m^{(b)}}{\hat{\beta}_m^{(b)}}\right)$

IV. Approximate Message Passing-Based Inference

In this section we first introduce message passing in an FFG as a probabilistic inference methodology. Next, we will derive messages for the Gamma mixture (ΓM) node using variational message passing (VMP) [9], [10], which allows us to perform probabilistic inference in the ΓMM. However, one of the VMP messages leads to an approximate posterior distribution, whose closed-form solution is the result of a non-conjugate multiplication that cannot be normalized analytically. We propose two approaches to resolve this problem. First, we propose to use expectation-maximization for bypassing the need of calculating the normalization constant. Secondly, we apply moment matching to approximate the moments of the approximate posterior distribution through importance sampling [11, Ch.7].

A. Variational message passing

Because of the conditional independencies in the generative model we can perform execution of (7) through a dis-

tributed set of smaller local computations called messages. Unfortunately, the intractability in these computations limits us in performing exact message passing-based inference, also known as the sum-product algorithm [12] or belief propagation [13]. To resolve this, we will resort to VMP [9], [10]. Consider the generative model $p(\boldsymbol{x}, \boldsymbol{z})$ for the ΓMM, where $\boldsymbol{z} \triangleq [\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\pi}]$, with intractable posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$, in which $\boldsymbol{x}$ and $\boldsymbol{z}$ are the observed and latent variables, respectively. Variational inference approximates the exact posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by a tractable approximate posterior distribution $q(\boldsymbol{z})$ through minimization of the variational free energy (VFE) functional

$$F[q] = D_{\mathrm{KL}}[q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})] - \log p(\boldsymbol{x})\,. \tag{8}$$

where D_{KL} is the Kullback-Leibler divergence. In VMP the variational free energy is optimized by iteratively updating the approximate posterior distributions. In order to facilitate optimization of VFE, $q(\boldsymbol{z})$ is often constrained by a mean-field factorization

$$q(\boldsymbol{z}) = \prod_i q(z_i)\,. \tag{9}$$

For a generic node $f(z_1, z_2, \ldots, z_M)$ the outgoing variational message $\vec{\nu}(z_j)$, under the mean-field assumption, can be evaluated as [10]

$$\vec{\nu}(z_j) \propto \exp \int \prod_{i \neq j} q(z_i) \log f(z_1, z_2, \ldots, z_M) \mathrm{d}\boldsymbol{z}_{\setminus j}. \tag{10}$$

The approximate posterior can then be updated by the normalized multiplication of the messages on the corresponding edge as

$$q(z_j) = \frac{\vec{\nu}(z_j)\overleftarrow{\nu}(z_j)}{\int \vec{\nu}(z_j)\overleftarrow{\nu}(z_j)\mathrm{d}z_j}\,. \tag{11}$$

In the VMP algorithm, (10) and (11) are iteratively repeated for all variables until convergence [10].

B. Variational message passing in the Gamma mixture node

The ΓM node of (1) has been visualized in Table I. We will assume a mean-field factorization over the joint approximate posterior distribution as

$$q(x_k, \boldsymbol{s}_k, \boldsymbol{a}, \boldsymbol{b}) = q(x_k)q(\boldsymbol{s}_k) \prod_{m=1}^{M} q(a_m)q(b_m), \tag{12}$$

where the distributions of the individual factors are presented in Table I. To support modular usage of the ΓM node, the variable x_k is not assumed to be observed for the derivations of the messages. The variational messages of Table I have been derived by the substitution of the approximate posterior distributions into (10).[1]

All messages, except for $\overleftarrow{\nu}(a_m)$, are of the same functional form as the corresponding approximate posterior distribution. Since the Gamma and categorical distributions are closed under multiplication, the resulting updated approximate posterior distributions remain in the same family of distributions. However, the message $\overleftarrow{\nu}(a_m)$ has a functional form that makes a closed-form result for the approximate posterior distribution infeasible. Therefore, to make calculations tractable we will approximate $q(a_m)$ by a parametric distribution, see Table I. In the remainder of this section we will propose two solutions: (1) expectation-maximization and (2) moment matching.

[1]The derived messages are available in the supplementary material at https://github.com/mlsp2021-gmm/gmm-experiments.

C. Solution 1: Expectation-maximization (VMP-EM)

The first proposed solution uses VMP in conjunction with expectation-maximization (VMP-EM) to approximate the resulting posterior distribution of a_m using message passing, inspired by [14]. Here the posterior distribution $q(a_m)$ is fixed to a Dirac delta function

$$q(a_m) = \delta(a_m - \hat{a}_m) \tag{13}$$

instead of the Gamma distribution from Table I. This distribution is located at $\hat{a}_m$, whose value is obtained through expectation-maximization using message passing according to [14]. The location $\hat{a}_m$ is determined by

$$\hat{a}_m = \arg\max_{a_m} \; \log \vec{\nu}(a_m) + \log \overleftarrow{\nu}(a_m), \text{ s.t. } a_m > 0\,, \tag{14}$$

where the message $\overleftarrow{\nu}(a_m)$ represents the variational message from Table I.

Theorem 1. *The solution of the constrained maximization problem given by* (14) *exists and is unique.*

Proof. From Table I we know the functional form $\log \overleftarrow{\nu}(a_m) = \hat{\pi}_k \left(a_m \zeta_{km} - \log \Gamma(a_m)\right)$. Since the logarithm of the Gamma function is strictly convex when restricted to positive real numbers (Bohr-Mollerup theorem) [15], $\log \overleftarrow{\nu}(a_m)$ is strictly concave as it is a summation of an affine and a strictly concave term [16, Ch. 2.3]. Because the prior message $\vec{\nu}(a_m)$ is proportional to a Gamma distribution, $\log \vec{\nu}(a_m)$ is either affine or concave depending on the shape parameter. Hence, $\log \vec{\nu}(a_m)\overleftarrow{\nu}(a_m)$ is always strictly concave. Because it is concave the maximum exists by strong duality [16, Ch. 5.3.2] and is unique because concavity is strict. □

D. Solution 2: Moment matching (VMP-MM)

Expectation-maximization provides us with a single estimate of the parameter a_m. If instead we would like to retain uncertainty about this parameter, we could approximate the resulting marginal distribution by a Gamma distribution using VMP with moment matching (VMP-MM), realized by importance sampling (IS) [11, Ch.7]. The IS procedure approximates the target distribution $q(a_m)$ by drawing L samples $a_m^{(l)}$ from an *importance distribution* $\tilde{q}(a_m)$ as

$$a_m^{(l)} \sim \tilde{q}(a_m) = \frac{\vec{\nu}(a_m)}{\int_{\mathbb{R}_{>0}} \vec{\nu}(a_m)\mathrm{d}a_m}, l = 1, \ldots, L. \tag{15}$$

We choose the normalized forward message $\tilde{q}(a_m)$ as the importance distribution. We can make this choice, because the support of the importance distribution is $\mathbb{R}_{>0}$, which coincides with the support of the multiplication $\vec{\nu}(a_m)\overleftarrow{\nu}(a_m)$. The mean and variance of a_m can then be approximated by

Table II: Shape and rate parameters of the ΓMMs used for data generation.

	$\boldsymbol{a}$	$\boldsymbol{b}$
$M = 2$	[9, 90]	[27, 270]
$M = 3$	[40, 6, 200]	[20, 1, 20]
$M = 4$	[200, 400, 600, 800]	[100, 100, 100, 100]

$$\mathbb{E}[a_m] \approx \sum_{l=1}^{L} a_m^{(l)} \vec{\nu}(a_m^{(l)})/Z \tag{16a}$$

$$\mathrm{Var}[a_m] \approx \sum_{l=1}^{L} (a_m^{(l)} - \mathbb{E}[a_m])^2 \vec{\nu}(a_m^{(l)})/Z\,, \tag{16b}$$

where $Z = \sum_{l=1}^{L} \vec{\nu}(a_m^{(l)})$ is the normalization constant. In our implementation, we employ adaptive resampling [11, Ch.7] to avoid the degeneracy problem for the estimates obtained by (16a) and (16b).

Theorem 2. *For $L \to \infty$ the summations given by* (16) *converge to the true mean and variance of $q(a_m)$.*

Proof. The numerator of (16a) $\sum_{l=1}^{L} a_m^{(l)} \vec{\nu}(a_m^{(l)})$ is the average of $a_m q(a_m^{(l)})/\tilde{q}(a_m^{(l)})$ under sampling from $\tilde{q}(a_m^{(l)})$. These numerators for different L are independent and identically distributed random variables with mean $\mathbb{E}[a_m]$ [17]. The strong law of large numbers gives

$$\mathbb{P}\left\{ \lim_{L\to\infty} \sum_{l=1}^{L} a_m^{(l)} \vec{\nu}(a_m^{(l)})/Z = \mathbb{E}[a_m] \right\} = 1. \tag{17}$$

The denominator of (16a) Z converges to 1. □

With the mean and the variance the parameters of the Gamma distribution $q(a_m)$ from Table I can be determined as

$$\hat{\alpha}_m^{(a)} = \frac{\mathbb{E}[a_m]^2}{\mathrm{Var}[a_m]}, \qquad \hat{\beta}_m^{(a)} = \frac{\mathbb{E}[a_m]}{\mathrm{Var}[a_m]}\,. \tag{18}$$

Note that unlike VMP-EM that yields a point estimate by determining (14), VMP-MM results in a proper posterior distribution for a_m.

V. Experiments

All experiments were implemented in the Julia programming language [18].[2] We used the following computer configuration: *Operating system*: macOS Big Sur, *Processor*: 2,7 GHz Quad-Core Intel Core i7, *RAM*: 16GB.

A. Verification

For the verification stage, we followed the setup from [4], where Markov chain Monte Carlo was used for inference in a ΓMM. We generated data using three distinct ΓMMs, each specified by likelihood (1) with a different number of mixture components $M = \{2, 3, 4\}$. We fixed the shape and rate parameters a_m and b_m to the values from Table II.

Each of these models exhibits a different behavior as illustrated in Fig. 2. For $M = 2$, the mixture components have equal means, but different variances. For $M = 3$, two mixture components are well separated and have low variances. The third mixture has a large variance and overlaps with the other two mixtures. Finally, for $M = 4$ we have four well separated mixtures. For each of the models, we generated 10 distinct data sets with different mixing coefficients. These mixing coefficients were sampled from a standard uniform distribution and were normalized by dividing by the sum of the coefficients. Each data set contains $K = 2500$ observations (in total $3 \times 10 \times 2500$ data points). To verify the proposed inference method, we selected three generative models of which we assumed the number of components to be known. We then performed probabilistic inference through message passing for two situations. The first situation (known shape-rate) uses informative priors for $\boldsymbol{a}$ and $\boldsymbol{b}$ and a vague prior for $\boldsymbol{\pi}$. The second setup (known mixing) employs an informative prior for the mixing coefficient $\boldsymbol{\pi}$, but uninformative priors for $\boldsymbol{a}$ and $\boldsymbol{b}$. With informative priors, we imply that the distributions are centered at an ϵ-area ($\epsilon > 0, \epsilon^2 \approx 0$) of the values that were used for data generation. The priors were chosen such that they do not violate the properties of the corresponding distributions. We motivate the usage of informative priors for either mixing coefficients or parameters of gamma distribution by two reasons. First, based on a Bayesian analysis of the Gamma distribution [19], the choice of uninformative priors for small data sets generally leads to low accuracy. We should choose the priors of the ΓMM carefully as its parameter space is significantly larger than that of a single Gamma distribution. Secondly, due to the non-convexity of the mean-field assumption, we have multiple solutions for our inference problem [20, Ch.5]. Thus, the initialization of vague priors for all parameters of ΓMM may lead to undesirable local minima. The inference task, as specified in Section III, computes the quantities $q(\boldsymbol{a}|\boldsymbol{x}_{1:K})$, $q(\boldsymbol{b}|\boldsymbol{x}_{1:K})$, $q(\boldsymbol{s}|\boldsymbol{x}_{1:K})$ and $q(\boldsymbol{\pi}|\boldsymbol{x}_{1:K})$. The notation $q(\cdot|\boldsymbol{x}_{1:K})$ refers to the marginals after observing the data. In this experiment, we first want to ensure that the proposed algorithm recovers the unknown parameters of the mixture components. Additionally, we want to verify the convergence of the proposed methodology by monitoring the VFE $F[q(\cdot)]$.

We now highlight the results of the verification stage in Fig. 2. For the VMP-MM approach we used $L = 5000$. Both algorithms recover the parameters of the ΓMM in the aforementioned situations. Both algorithms converge, which is reflected by the evolution of the VFE in Fig. 3. The VMP-EM approach converges more slowly than the VMP-MM approach as a function of iteration count, but for this experimental setup VMP-MM is on average approximately 30 times slower in evaluation time than VMP-EM due to the relatively expensive sampling procedure.

B. Validation

For the validation of our model, we used the country data set from Kaggle.[3] This data set contains socio-economic

[2]Experiments are available at https://github.com/mlsp2021-gmm/gmm-experiments.

[3]https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

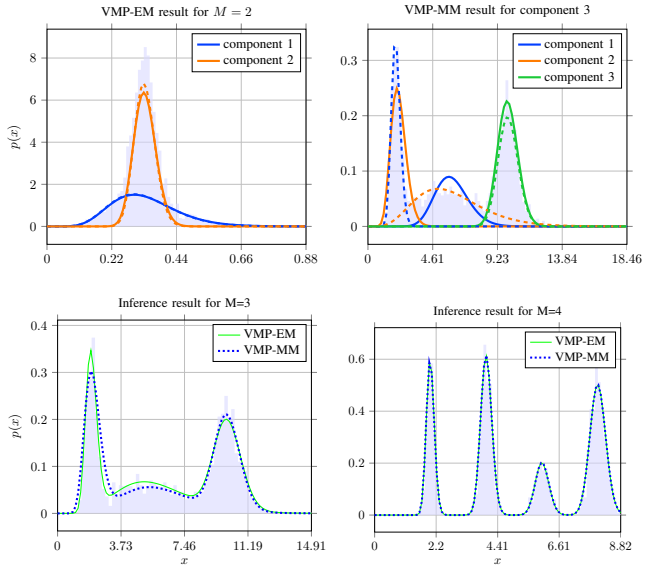


Fig. 2: Verification results. The shaded light-blue bar plots in the background denote the normalized histograms of the generated data. (Top) The dashed and solid lines denote the actual and estimated density functions, respectively. (Top-Left) Inference results for the VMP-EM approach for two components with informative shape and rate parameters. The estimated and actual densities match meaning that the mixing coefficient are inferred properly. (Top-Right) Inference results of the VMP-MM approach for three components with known mixing coefficients. The estimated mixture components 1 and 2 were swapped. The variance of the estimated component 1 is lower than the corresponding actual component 2. In contrast, the estimated component 2 has a larger variance than actual component 1. The estimated component 3 features shape and rate parameters that are close to the parameters of the corresponding generated mixture. (Bottom) The dashed and solid lines denote the density functions estimated by VMP-EM and VMP-MM, respectively. (Bottom-Left) Comparison of both algorithms for three components with informative mixing coefficients. Both algorithms provide reasonable estimates of the shape and rate parameters for each mixture. (Bottom-Right) Comparison of two algorithms for a mixture of four components with informative shape and rate parameters. Both algorithms lead to correct mixing posteriors.

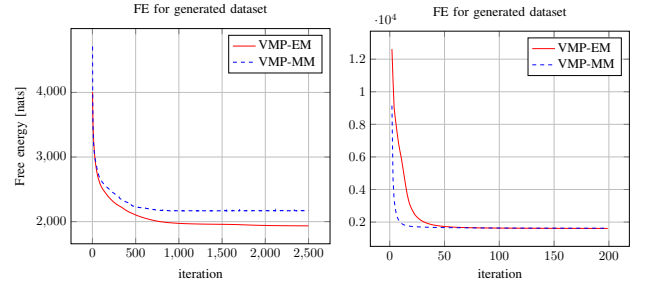


Fig. 3: Verification results. Evolution of the variational free energy for the VMP-EM and VMP-MM algorithms, averaged over their corresponding data sets. (Left) Situation with informative mixing coefficients. (Right) Situation with informative shape and rate parameters

and health data for all countries in the world. The task is to categorize the countries based on a set of data features. Most of the individual features represent positive real values, therefore the ΓMM appears as a possible approach to modelling. For the brevity of the experiment, we transformed the "inflation" feature (4.8% entries are negative) to a positive real range. Unlike the experiments on the generated data sets, we now have to deal with multivariate observations. Each observation x_k is now represented by a vector of $N = 7$ features as $x_k = [x_k^{(1)}, ..., x_k^{(N)}]$, where the superscript de-

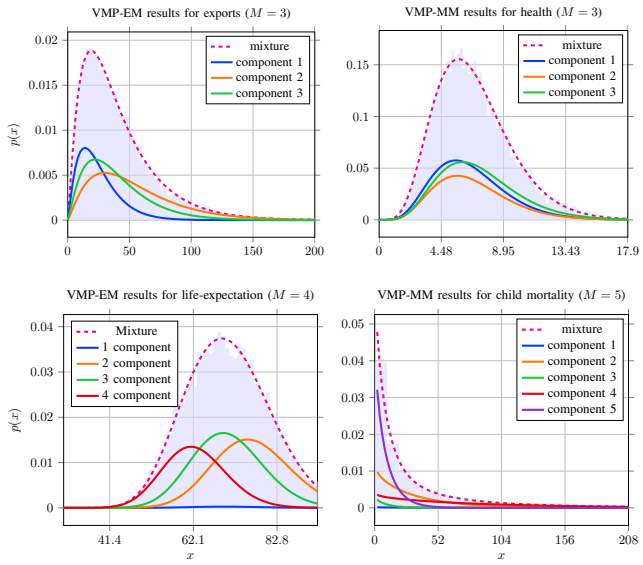


Fig. 4: Validation results for separated features. Dashed line denotes estimated ΓM. Solid lines correspond to individual mixture components. (Top-left) Inference result of VMP-EM algorithm for "exports" feature when $M = 3$. (Top-right) Inference result of VMP-MM algorithm for health feature when $M = 3$. (Bottom-left) Inference result of VMP-EM algorithm for "life expectation" feature when $M = 4$. (Bottom-right) Inference result of VMP-MM algorithm for "child mortality" feature when $M = 5$.

notes the feature, indexed by n, and where $x_k^{(n)} \in \mathbb{R}_{>0}$ for all $n = 1, \ldots, N$. For modelling the multivariate observations, we model each feature independently using a separate ΓMM, where the features are modelled by the same selector variable $\boldsymbol{s}_k$. The likelihood model (1) then changes to

$$p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{n=1}^{N} \Gamma\left(x_k^{(n)}|a_m^{(n)}, b_m^{(n)}\right)^{s_{km}} \tag{19}$$

and we change (2) and (3) to contain $M \times N$ independent mixture components, such that each feature is modelled by its own set of mixture components.

In this setup, we do not have any prior information about the mixing coefficients. To obtain informative priors for the shape and rate parameters of the mixture components, we extracted the empirical means and variances of each feature and converted those to the shape and rate parameters of a Gamma distribution using (18). To disentangle the priors of shape and rate parameters, we added a positive random jitter term to the each shape and rate parameters of the prior distributions. To determine the optimal number of mixture components, we tracked the values of mixing coefficients $\boldsymbol{\pi}$ for different numbers of mixture components $M = \{2, \ldots, 10\}$. Mixing coefficients that converge to 0 indicate the absence of the corresponding cluster [21, Ch. 10]. We highlight the inference results of the proposed algorithms in Figure 4. Based on this approach, both VMP-MM and VMP-EM experiments show that $M = 3$ is the optimal number of components.

To visualize the inferred components, we used the t-distributed stochastic neighbor embedding (tSNE) [22].

tSNE provides an intuition of how the high-dimensional data is arranged by mapping the data onto a lower dimensional space. Figure 5 shows the result of the tSNE projection for the countries data set. We colored the data points according to $\arg\max_{\boldsymbol{s}_k} q(\boldsymbol{s}_k)$, i.e., the most likely mixture component of the corresponding marginal. In this way, the labels provided by VMP-EM and VMP-MM are identical.

ACKNOWLEDGMENTS

This work was partly financed by research program ZERO with project number P15-06, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

VI. DISCUSSION AND CONCLUSIONS

The proposed inference methods, VMP-EM and VMP-MM, converge and correctly identify the parameters of ΓMM. Although VMP-MM yields a "full" posterior distribution, it suffers from a slower evaluation time. In contrast, while VMP-EM enjoys a relatively fast evaluation time, it provides only point estimates for the shape parameters of the mixture components. This makes VMP-EM difficult to employ in an online learning scenario when new observations become available in sequential order.

For the validation experiments, we transformed the "inflation" feature to a positive real range, although this approach is undesirable as it breaks the natural support of the corresponding random variable. Alternatively, we could have substituted the ΓM node that models "inflation" by a Gaussian Mixture (GM) node [23], leading to a hybrid model that connects ΓM and GM nodes through selector variables.

We presented a variational message-passing approach for inferring the parameters in Gamma mixture models. The required variational messages are summarized in Table I. We proposed two approaches for computing the marginal distribution of the shape parameters of the Gamma mixture model. Furthermore, we demonstrated the convergence of the inference procedure through the minimization of variational free energy. The correctness of the message-passing scheme was verified on a synthetic data set. The Gamma mixture node can now be used as a plug-in node in any graphical model that supports message passing-based inference. Owing to the locality and modularity of the FFG framework, we showed how the Gamma mixture model can be easily extended to tackle multi-dimensional problems such as clustering of countries. In future work, we plan to use the Gamma mixture node for probabilistic modeling of time-series that exhibit switching behavior.

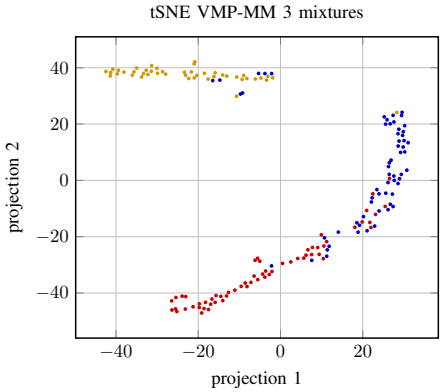


Fig. 5: tSNE visualization of validation experiments. The data points are colored according to $\arg\max_{\boldsymbol{s}_k} q(\boldsymbol{s}_k)$, i.e., the most likely mixture component of the corresponding marginal.

VII. REFERENCES

[1] S.J. Rennie, J.R. Hershey, and P.A. Olsen, "Single-Channel Multitalker Speech Recognition," *IEEE Signal Processing Magazine*, 2010.

[2] Chi Liu, Heng-Chao Li, Kun Fu, Fan Zhang, Mihai Datcu, and William J. Emery, "Bayesian estimation of generalized Gamma mixture model based on variational EM algorithm," *Pattern Recognition*, 2019.

[3] Hassen Sallay, Sami Bourouis, and Nizar Bouguila, "Online learning of finite and infinite gamma mixture models for COVID-19 detection in medical images," *Computers*, 2021.

[4] Michael Wiper, David Rios Insua, and Fabrizio Ruggeri, "Mixtures of Gamma Distributions with Applications," *Journal of Computational and Graphical Statistics*, 2001.

[5] G.David Forney, "Codes on graphs: normal realizations," *IEEE Transactions on Information Theory*, 2001.

[6] Hans-Andrea. Loeliger, "An introduction to factor graphs," *Signal Processing Magazine, IEEE*, 2004.

[7] Sascha Korl, *A factor graph approach to signal modelling, system identification and filtering*, 2005.

[8] Hans-Andrea Loeliger, Justin Dauwels, Junli Hu, Sascha Korl, Li Ping, and Frank R. Kschischang, "The Factor Graph Approach to Model-Based Signal Processing," *Proceedings of the IEEE*, 2007.

[9] John Winn and Christopher M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, 2005.

[10] Justin Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE International Symposium on Information Theory*, 2007.

[11] Simo Särkkä, *Bayesian Filtering and Smoothing*, 2013.

[12] Frank R. Kschischang, Brendan J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *Information Theory, IEEE Transactions on*, 2001.

[13] Judea Pearl, "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach," in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, 1982.

[14] J. Dauwels, S. Korl, and H.-A. Loeliger, "Expectation maximization as message passing," in *International Symposium on Information Theory, 2005. ISIT 2005. Proceedings*, 2005.

[15] Milan Merkle, "Logarithmic convexity and inequalities for the gamma function," *Journal of Mathematical Analysis and Applications*, 1996.

[16] Stephen P. Boyd and Lieven Vandenberghe, *Convex optimization*, 2004.

[17] Art B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[18] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, 2017.

[19] Fernando Antonio Moala, Pedro Luiz Ramos, and Jorge Alberto Achcar, "Bayesian Inference for Two-Parameter Gamma Distribution Assuming Different Noninformative Priors," *Revista Colombiana de Estadística*, 2013.

[20] Martin J. Wainwright and Michael I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, 2008.

[21] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 2006.

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 2008.

[23] Marco Cox, Thijs van de Laar, and Bert de Vries, "A factor graph approach to automated design of Bayesian signal processing algorithms," *International Journal of Approximate Reasoning*, 2019.