Variational Bayesian Gaussian Mixture Models for Mapping

Shreyas Seshadri

March 29, 2017

1 Introduction

This document describes the use of Bayesian Gaussian mixture models (BGMM) for mapping. It has been used as the main mapping method in [1]. It is the Bayesian extension of the typically used GMM mapping learned with Expectation maximization (for example in voice conversion [2]).

Let us consider the problem of learning a mapping from source data, $\mathbf{x}_s$, to target data, $\mathbf{x}_t$. In this method we first learning a BGMM model for the concatenated training data as $\mathbf{x} = [\mathbf{x}_s, \mathbf{x}_t]^T$. Let training data $\mathbf{x}$ be D dimensional.

2 BGMM Modeling

Now $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_N]$ be modeled by a BGMM with K Gaussians with parameters $\{\theta_k\}_{i=1}^K$ and weights $\{\pi_k\}_{i=1}^K$. Hence the likelihood of $\mathbf{X}$ is defined as

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\theta}_k) \tag{1}$$

In the Bayesian setting we consider a prior on the model parameters and aim to infer their posterior distribution. The prior on the weights was chosen as the Dirichlet distribution i.e. $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0$ is a K-dimensional parameter. We consider full covariance Gaussians parameterized by the mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$, i.e. $\theta_k = \{\mu_k, \Lambda_k\}$. The conjugate prior is chosen for $\boldsymbol{\theta}$ as the Normal-Wishart distribution i.e. $\theta_k \sim \mathcal{NW}(\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0)$, where mean $\mathbf{m}_0$, scale matrix $\mathbf{W}_0$, real values $\beta_0 > 0$ and $\nu_0 > D - 1$ are parameters of the $\mathcal{NW}$ distribution [3]. Latent variables $\{z_i\}_{i=1}^N$ denote the Gaussian to which each of the N data points $\{\mathbf{x}_i\}_{i=1}^N$ are assigned.

Sections 2.1 and 2.2 details the generative process and inference of the BGMM respectively.

2.1 Generative process of BGMM

The Bayesian graph of the BGMM is shown in Figure 1. The generative process is as follows. Let $\boldsymbol{\Theta}$ be the space of all the GMM parameter values. Let H be a distribution on this space, from which the K GMM parameters $\{\boldsymbol{\theta}_k\}_{i=1}^K$ are sampled. As mentioned, we consider full covariance Gaussians and the

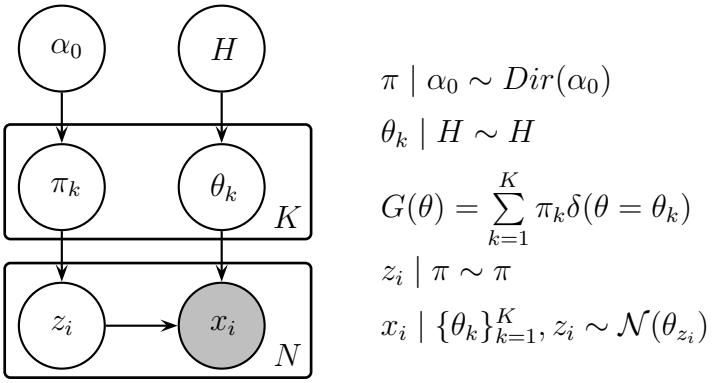

Figure 1: BGMM represented as a graph where nodes, arrows and plates indicate variables, dependencies and repetition respectively.

parameters are the mean, $\boldsymbol{\mu}$ and the precision, $\boldsymbol{\Lambda}$, i.e. $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}$. We consider the conjugate prior of the Gaussian to model H as the Normal-Wishart distribution, $\mathcal{NW}(\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0)$. The parameter vector associated with each observation $\{\mathbf{x}_i\}_{i=1}^N$ can be modeled as a random variable $\boldsymbol{\theta}$ with distribution

$$G = \textstyle\sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta} = \boldsymbol{\theta}_k), \tag{2}$$

where the weights $\{\pi_k\}_{k=1}^K$ are sampled from the prior Dirichlet distribution parameterized by a K dimensional vector $\boldsymbol{\alpha}_0$. The latent variables $\{z_i\}_{i=1}^N$ denote the cluster to which the N data points $\{\mathbf{x}_i\}_{i=1}^N$ are assigned, and are sampled from a multinomial distribution with component probabilities $\{\pi_k\}_{k=1}^K$. Finally, the observed variables $\mathbf{x}_i$ are then sampled from the Gaussian model $\mathcal{N}(\boldsymbol{\theta}_{z_i})$.

2.2 Variational Inference

There is no direct analytic solution for the posterior distribution of the BGMM parameters. This paper uses variational inference method [3] that approximates the analytically intractable posterior with a tractable distribution called variational distribution $q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. This is done by making the following independence assumption:

$$q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \approx q(\mathbf{z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{z}) q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{3}$$

Kullback–Leibler (KL) divergence to the true posterior is then minimized to find the variational distribution. Since we use conjugate priors, $q(\boldsymbol{\pi})$ is another Dirchlet distribution $Dir(\boldsymbol{\alpha})$, and $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ another Normal-Wishart distribution $\mathcal{NW}(\mathbf{m}_k, \beta_k, \mathbf{W}_k, \nu_k)$ (see [3] for details). In practice, the final update equations are similar to the expectation–maximisation (EM) algorithm that iterates between finding the probabilities $q(\mathbf{z})$ (called responsibilities) based on the current model $q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, and updating model parameters based on the current responsibilities.

3 BGMM Mapping

During application, the new source data, $\mathbf{y}_s$, needs to be mapped to the target, $\mathbf{y}_t$. Let us first calculate the probability of data $\mathbf{y} = [\mathbf{y}_s, \mathbf{y}_t]^T$ given data $\mathbf{X}$ (modeled by the BGMM), $p(\mathbf{y}|\mathbf{X})$, called as the ***posterior predictive*** [3], as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \frac{1}{\hat{\alpha}} \sum_{k=1}^{K} \alpha_k S_t(\mathbf{y}|\mathbf{m}_k, \boldsymbol{\Sigma}_k, \nu_k + 1 - D) \\ \text{where, } \boldsymbol{\Sigma}_k &= \frac{1+\beta_k}{(\nu_k + 1 - D)\beta_k} \mathbf{W}_k^{-1} \end{aligned} \tag{4}$$

That is, a mixture of multivariate Student's t-distributions S_t with kth component having means $\mathbf{m}_k$, covariance $\boldsymbol{\Sigma}_k$ and $\nu_k + 1 - D$ degrees of freedom; and α_k is the kth term in $\boldsymbol{\alpha}$ and $\hat{\alpha} = \sum_k \alpha_k$ [3].

Sections 3.1 details the marginal and conditional distributions of a multivariate Student-t distribution and Section 3.2 shows how these can be used for the MMSE of the target data in BGMM mapping.

3.1 Marginal and Conditional of Multivariate Student-t

Let $\mathbf{a}$ be a d dimensional variable that is distributed as a multivariate student-t as

$$\mathbf{a} \sim S_t(\boldsymbol{\zeta}, \boldsymbol{\Upsilon}, \rho) \tag{5}$$

with means $\boldsymbol{\zeta}$, covariance $\boldsymbol{\Upsilon}$ and ρ degrees of freedom. Let $\mathbf{a}$ be split into two blocks $\mathbf{a}_1$ and $\mathbf{a}_2$ of dimensions d_1 and d_2; means $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$; covariances $\boldsymbol{\Upsilon}_{11}$, $\boldsymbol{\Upsilon}_{22}$ and cross covariances $\boldsymbol{\Upsilon}_{12}$ and $\boldsymbol{\Upsilon}_{21}$ respectively.

Then the marginal distribution is simply [4]

$$\begin{aligned} \mathbf{a}_1 &\sim S_t(\boldsymbol{\zeta}_1, \boldsymbol{\Upsilon}_{11}, \rho) \\ \mathbf{a}_2 &\sim S_t(\boldsymbol{\zeta}_2, \boldsymbol{\Upsilon}_{22}, \rho) \end{aligned} \tag{6}$$

The conditional $\mathbf{a}_1|\mathbf{a}_2$ [4] can be calculated as

$$\begin{aligned} \mathbf{a}_1|\mathbf{a}_2 &\sim S_t(\boldsymbol{\zeta}_{1|2}, \boldsymbol{\Upsilon}_{1|2}, \rho + d_1) \\ \boldsymbol{\zeta}_{1|2} &= \boldsymbol{\zeta}_1 + \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}(\mathbf{a}_2 - \boldsymbol{\zeta}_2) \\ \boldsymbol{\Upsilon}_{1|2} &= h_{1|2}(\boldsymbol{\Upsilon}_{11} - \boldsymbol{\Upsilon}_{12}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}_{12}^T) \\ h_{1|2} &= \frac{1}{\rho + d_2}[\rho + (\mathbf{a}_2 - \boldsymbol{\zeta}_2)^T \boldsymbol{\Upsilon}_{22}^{-1}(\mathbf{a}_2 - \boldsymbol{\zeta}_2)] \end{aligned} \tag{7}$$

3.2 MMSE Estimate

Now let us consider the parameters of the kth multiavariate Student's t in Eq. (4) as block matrices $\mathbf{m}_k = [\mathbf{m}_s, \mathbf{m}_t]^T$ and $\boldsymbol{\Sigma}_k = \left[\begin{smallmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{st} \\ \boldsymbol{\Sigma}_{ts} & \boldsymbol{\Sigma}_{tt} \end{smallmatrix}\right]$. Now the MMSE estimate of $\boldsymbol{x}_t$ can be calculated, similar to a typical GMM mapping [2], as

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p(k|\mathbf{y}_s, \mathbf{X})[\mathbf{m}_t + \boldsymbol{\Sigma}_{ts}\boldsymbol{\Sigma}_{ss}^{-1}(\mathbf{y}_s - \mathbf{m}_s)] \tag{8}$$

where $p(k|\mathbf{y}_s, \mathbf{X})$ is the marginal probability of the kth component in Eq. (4), and the other term is the mean of the kth component in the conditional over

the posterior predictive in Eq. (4) (see Section 10.7 of [4]). MATLAB codes for the BGMM mapping are available under an open source license[1].

References

[1] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, *submitted*, 2017.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. of ICASSP*, Seattle, USA, 1998, pp. 285–288.

[3] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[4] K. P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," Tech. Rep., 2007.

[1] https://github.com/shreyas253/BGMM_Mapping