An Upper Bound on Normalized Maximum Likelihood Codes for Gaussian Mixture Models

So Hirai and Kenji Yamanishi *

Abstract

This paper shows that the normalized maximum likelihood (NML) code-length calculated in [1] is an upper bound on the NML code-length strictly calculated for the Gaussian Mixture Model. We call an upper bound on the NML code-length as uNML (upper bound on NML). When we use this uNML code-length, we have to change the scale of data sequence to satisfy the restricted domain. However, in the point of model selection, we find the fact that the model selection algorithm is essentially universal regardless of scale conversion of data in Gaussian Mixture Models, and the experimental results of paper [1] can be used as it is.

In addition to this, we correct the normalized maximum likelihood (NML) code-length for generalized logistic distribution calculated in [1].

1 Problem Setting

In this paper, we consider the problem of model selection in which we aims to calculate the number of clusters for Gaussian Mixture Model. Let us use the given sequence $\mathbf{x}^n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $\mathbf{x}_i = (x_{i1}, \cdots, x_{im})^\top$ $(i = 1, \cdots, n)$. Here, we use the Gaussian Model Class: $\mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m}$, and calculate NML code-length for Gaussian Model. The Gaussian distribution for data sequence $\mathbf{x}^n$ is defined as follows:

$$f(\mathbf{x}^n; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{mn}{2}} \cdot |\Sigma|^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \Big\}.$$

We define the NML distribution f_{NML} relative to a model class $\mathcal{M} = \{f(X^n; \theta) : \theta \in \Theta\}$ by

$$f_{\mathrm{NML}}(\mathbf{x}^n; \mathcal{M}) \stackrel{\mathrm{def}}{=} \frac{f(\mathbf{x}^n; \hat{\mu}(\mathbf{x}^n), \hat{\Sigma}(\mathbf{x}^n))}{\int_Y f(\mathbf{y}^n; \hat{\mu}(\mathbf{y}^n), \hat{\Sigma}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n}. \tag{1}$$

Here Y is the restricted domain for $\mathbf{x}^n$. By using this restrict, we can calculate the NML code-length without divergence.

The NML code-length for Gaussian Mixture Model is defined as below with latent variable z^n:

$$\begin{aligned}
L_{\mathrm{NML}}(\mathbf{x}^n, z^n; Y, \mathcal{M}(K)) \quad \stackrel{\mathrm{def}}{=} \quad & -\log f_{\mathrm{NML}}(\mathbf{x}^n, z^n) \\
= \quad & -\log f(\mathbf{x}^n, z^n; \mathcal{M}(K), \hat{\theta}(\mathbf{x}^n, z^n)) + \log C(\mathcal{M}(K), n), \tag{2} \\
C(\mathcal{M}(K), n) \quad = \quad & \sum_{w^n} \int_Y f(\mathbf{y}^n, w^n; \mathcal{M}(K), \hat{\theta}(\mathbf{y}^n, w^n)) \mathrm{d}\mathbf{y}^n. \tag{3}
\end{aligned}$$

Here $\theta = (\pi, \mu, \Sigma)$ is the set of parameters, We consider problem of model selection for Gaussian Mixture Model with using (2) as a criterion.

---

*S. Hirai and K. Yamanishi are with Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, JAPAN
E-mail: so_hirai@mist.i.u-tokyo.ac.jp
E-mail: yamanishi@mist.i.u-tokyo.ac.jp

2   Influence of scale conversion of data on model selection

When we use the NML code-length defined by (2), we have to change the scale of data sequence to safisfy the restricted domain Y (e.g. to multiply $1/\alpha$, etc.). In this section, we show that the model selection algorithm is essentially universal regardless of scale conversion of data.

Let us consider the NML code-length for Gaussian Mixture Model as $L_{\mathrm{NML}}(\mathbf{x}^n, z^n; Y, \mathcal{M}(K))$, and we can get the definition of code-length as (2). The term influenced by the scale of the data is the first term of the formula (2). Here, in order to evaluate the influence of the first term from the scale conversion of the data, we use the dataset $\mathbf{x}_\alpha^n$ ($= \mathbf{x}^n/\alpha$) which we calculate by multiplting $\mathbf{x}^n$ by $1/\alpha$. We consider model selection when each data $\mathbf{x}^n$ or $\mathbf{x}_\alpha^n$ is used, and evaluate the difference between them. Since it is important for model selection to evaluate the difference between $\mathcal{M}(K_1)$ and $\mathcal{M}(K_2)$, we focus on the difference in the first term of equation (2).

$$-\log f(\mathbf{x}_\alpha^n, z^n; \mathcal{M}(K_1), \hat{\theta}(\mathbf{x}_\alpha^n, z^n)) - (-\log f(\mathbf{x}_\alpha^n, z^n; \mathcal{M}(K_2), \hat{\theta}(\mathbf{x}_\alpha^n, z^n)))$$

$$= C + \sum_{k=1}^{K_1} \sum_{j=1}^{m} \frac{h_k}{2} \log \hat{\lambda}_j(\mathbf{x}_\alpha^n) - \sum_{k=1}^{K_2} \sum_{j=1}^{m} \frac{h_k'}{2} \log \hat{\lambda}_j'(\mathbf{x}_\alpha^n)$$

$$= C + \sum_{k=1}^{K_1} \sum_{j=1}^{m} \frac{h_k}{2} \left\{ \log \hat{\lambda}_j(\mathbf{x}^n) - 2\log\alpha \right\} - \sum_{k=1}^{K_2} \sum_{j=1}^{m} \frac{h_k'}{2} \left\{ \log \hat{\lambda}_j'(\mathbf{x}^n) - 2\log\alpha \right\} \tag{4}$$

$$= C + \sum_{k=1}^{K_1} \sum_{j=1}^{m} \frac{h_k}{2} \log \hat{\lambda}_j(\mathbf{x}^n) - \sum_{k=1}^{K_2} \sum_{j=1}^{m} \frac{h_k'}{2} \log \hat{\lambda}_j'(\mathbf{x}^n)$$

$$= -\log f(\mathbf{x}^n, z^n; \mathcal{M}(K_1), \hat{\theta}(\mathbf{x}^n, z^n)) - (-\log f(\mathbf{x}^n, z^n; \mathcal{M}(K_2), \hat{\theta}(\mathbf{x}^n, z^n))) \tag{5}$$

where we define

$$C \overset{\mathrm{def}}{=} -\sum_{k=1}^{K_1} h_k \log \hat{\pi}_k + \sum_{k=1}^{K_2} h_k' \log \hat{\pi}_k' + \sum_{k=1}^{K_1} \frac{mh_k}{2} \log 2\pi\mathrm{e} - \sum_{k=1}^{K_2} \frac{mh_k'}{2} \log 2\pi'\mathrm{e}$$

and each h_k, h_k' represents the number of data which belong to k under the model class $\mathcal{M}(K_1), \mathcal{M}(K_2)$. This shows us that the difference of code-length is not effected by scale conversion of data, and it is possible for us to process the data so as to satisfy the restriction and use it for model selection without changing the result.

Here, in this paper, we define the restricted domain with maximum likelihood estimator (MLE) of parameter $\hat{\mu}, \hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_m)$, where each $\hat{\lambda}_j$ is j-th eigen values of $\hat{\Sigma}$. These MLEs are changed by scale conversion of data as follows:

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^{n} x_{i\alpha} = \frac{1}{n\alpha} \sum_{i=1}^{n} x_i = \frac{1}{\alpha} \hat{\mu}, \tag{6}$$

$$\hat{\Sigma}_\alpha = \frac{1}{n} \sum_{i=1}^{n} (x_{i\alpha} - \hat{\mu}_\alpha)(x_{i\alpha} - \hat{\mu}_\alpha)^\top = \frac{1}{n\alpha^2} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top = \frac{1}{\alpha^2} \hat{\Sigma} = \frac{1}{\alpha^2} U\Lambda U^\top$$

$$\Leftrightarrow \hat{\lambda}_{j\alpha} = \frac{1}{\alpha^2} \hat{\lambda}_j. \tag{7}$$

Here the data sequence $\mathbf{x}^n$ denotes data assigned to a cluster. This shows that the MLEs of parameters can be converted into arbitory size by scale conversion of data.

3   An upper bound on NML code-length

By the discussion of section 2, the model selection algorithm is essentially universal regardless of scale conversion of data. Using this feature, we restrict the domain in order to calculate NML code-length. We show that the code-length calculated by [1] is an upper bound on the the NML code-length strictly calculated for the Gaussian Mixture Model, and we can use this code-length called uNML for model selection.

3.1 1st Change Point of [1]

We use the Shtarkov's minmax regret with restricted domain as follows:

$$\min_{Q} \max_{\mathbf{x}^n \in Y(R,\epsilon_1,\epsilon_2)} \left\{ -\log Q(\mathbf{x}^n) - \min_{\theta}(-\log P(\mathbf{x}^n|\theta)) \right\} \tag{8}$$

$$Y(R,\epsilon_1,\epsilon_2) \overset{\text{def}}{=} \{\mathbf{y}^n | \,\|\hat{\mu}(\mathbf{y}^n)\|^2 \le R, \epsilon_{1j} \le \hat{\lambda}_j(\mathbf{y}^n) \le \epsilon_{2j} \le \epsilon_2 < 1 \ (j=1,\cdots,m),$$

$$\frac{\mathrm{Vol}(O(m))}{2^m} \cdot \epsilon_2^{\frac{m(m-1)}{2}} \le 1, \ \mathbf{y}^n \in \mathcal{X}^n \}. \tag{9}$$

3.2 2nd Change Point for [1]

Using the restricted domain (9), we can calculate an upper bound on NML code-length as follows:

$$
\begin{aligned}
C(\mathcal{M},n) &= \int_{Y(R,\epsilon_1,\epsilon_2)} f(\mathbf{y}^n; \hat{\mu}(\mathbf{y}^n), \hat{\Sigma}(\mathbf{y}^n)) \, \mathrm{d}\mathbf{y}^n \\
&= \int \delta(\hat{\mu}(\mathbf{x}^n) = \hat{\mu}, \hat{\Sigma}(\mathbf{x}^n) = \hat{\Sigma}) \, \mathrm{d}\mathbf{y}^n \int_{\|\hat{\mu}\|^2 \le R} \mathrm{d}\mu \int g(\hat{\lambda}) \, \mathrm{d}\hat{\Sigma} \\
&= \int_{\|\hat{\mu}\|^2 \le R} \mathrm{d}\mu \int \mathrm{d}U \int_{\epsilon_{1j} \le \hat{\lambda}_j \le \epsilon_{2j}} g(\hat{\lambda}) \prod_{1 \le i < j \le m} |\hat{\lambda}_i - \hat{\lambda}_j| \, \mathrm{d}\hat{\lambda} \quad (10) \\
&< \int \mathrm{d}U \cdot \epsilon_2^{\frac{m(m-1)}{2}} \int_{\|\hat{\mu}\|^2 \le R} \mathrm{d}\mu \int_{\epsilon_{1j} \le \hat{\lambda}_j \le \epsilon_{2j}} g(\hat{\lambda}) \, \mathrm{d}\hat{\lambda} \\
&= \frac{\mathrm{Vol}(O(m))}{2^m} \cdot \epsilon_2^{\frac{m(m-1)}{2}} \cdot \frac{2^{m+1} R^{\frac{m}{2}} (\prod_{j=1}^m \epsilon_{1j}^{-\frac{m}{2}} - \prod_{j=1}^m \epsilon_{2j}^{-\frac{m}{2}})}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot \left(\frac{n}{2\mathrm{e}}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})} \\
&< \frac{2^{m+1} R^{\frac{m}{2}} \prod_{j=1}^m \epsilon_{1j}^{-\frac{m}{2}}}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)} \times \left(\frac{n}{2\mathrm{e}}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})} \\
&= B(m,R,\epsilon) \times \left(\frac{n}{2\mathrm{e}}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})}. \quad (11)
\end{aligned}
$$

Here equation (11) equals to code-length calculated in [1], and represents an upper bound on NML code-length. From this fact, we can use the uNML (11) for model selection with the data which satisfy the restricted domain (9).

Here, we can define an upper bound on NML (uNML) code-length as follows:

$$
\begin{aligned}
L_{\mathrm{uNML}}(\mathbf{x}^n; Y, \mathcal{M}) &\overset{\text{def}}{=} -\log f_{\mathrm{uNML}}(\mathbf{x}^n; \mathcal{M}) \\
&= -\log f(\mathbf{x}^n; \mathcal{M}, \hat{\theta}(\mathbf{x}^n)) + \log C_{\mathrm{u}}(\mathcal{M},n), \\
C_{\mathrm{u}}(\mathcal{M},n) &= B(m,R,\epsilon) \times \left(\frac{n}{2\mathrm{e}}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})}.
\end{aligned}
$$

In calculating uNML code-length for Gaussian Mixture Models, we can use this definition of normalization term in the same way as [1].

3.3 Points to be aware of when handling this code length

Using this uNML code-length, we have to change the scale of data sequence to safisfy the restricted domain (9). From the discussion in section 2, in computation, we can use it for model selection without scale conversion of data. In the experiments of paper [1], as we can scale artificial dataset to safisfy the restricted domain (9), the experimental results of paper [1] can be used as it is.

4 Summary of modified [1]

Here, the following text quotes the sentence of the paper [1] and it is modified to describe the calculation of an upper bound on NML code-length.

Let an observed data sequence be $\mathbf{x}^n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ where $\mathbf{x}_i = (x_{i1}, \cdots, x_{im})^\top$ $(i = 1, \cdots, n)$. We use a class of Gaussian distributions: $\mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^m$ is a mean vector, $\Sigma \in \mathbb{R}^{m \times m}$ is a covariance matrix, and m is the dimension of $\mathbf{x}_i$. A probability density function of $\mathbf{x}^n$ for the Gaussian distribution is given by

$$f(\mathbf{x}^n; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{mn}{2}} \cdot |\Sigma|^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \right\},$$

and the NML distribution based on the Gaussian distribution is defined as follows:

$$f_{\mathrm{NML}}(\mathbf{x}^n) \stackrel{\text{def}}{=} \frac{f(\mathbf{x}^n; \hat{\mu}(\mathbf{x}^n), \hat{\Sigma}(\mathbf{x}^n))}{\int_{Y(R, \epsilon_1, \epsilon_2)} f(\mathbf{y}^n; \hat{\mu}(\mathbf{y}^n), \hat{\Sigma}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n} \tag{12}$$

where $\hat{\mu}(\mathbf{x}^n)$ and $\hat{\Sigma}(\mathbf{x}^n)$ are the MLEs of μ and Σ respectively:

$$\hat{\mu}(\mathbf{x}^n) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

$$\hat{\Sigma}(\mathbf{x}^n) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mu}(\mathbf{x}^n))(\mathbf{x}_i - \hat{\mu}(\mathbf{x}^n))^\top.$$

For given constants $R, \epsilon_1, \epsilon_2$, we set a restricted domain as follows:

$$Y(R, \epsilon_1, \epsilon_2) \stackrel{\text{def}}{=} \{\mathbf{y}^n | \; ||\hat{\mu}(\mathbf{y}^n)||^2 \le R, \epsilon_{1j} \le \hat{\lambda}_j(\mathbf{y}^n) \le \epsilon_{2j} \le \epsilon_2 < 1 \; (j = 1, \cdots, m),$$
$$\frac{\mathrm{Vol}(O(m))}{2^m} \cdot \epsilon_2^{\frac{m(m-1)}{2}} \le 1, \; \mathbf{y}^n \in \mathcal{X}^n\}, \tag{13}$$

where $\hat{\lambda}_j(\mathbf{y}^n)$ $(j = 1, \cdots, m)$ are eigen values of $\hat{\Sigma}(\mathbf{y}^n)$. This restriction makes the calculation of an upper bound on normalization term $C(\mathcal{M}, n)$ easier, as shown below.

First, by substituting MLE $\hat{\mu}(\mathbf{x}^n)$, $\hat{\Sigma}(\mathbf{x}^n)$ into the formula (12), the numerator of Eq. (12) can be expressed as follows:

$$f(\mathbf{x}^n; \hat{\mu}(\mathbf{x}^n), \hat{\Sigma}(\mathbf{x}^n)) = \prod_{i=1}^n \frac{1}{(2\pi)^{\frac{m}{2}} |\hat{\Sigma}(\mathbf{x}^n)|^{\frac{1}{2}}} \times \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \hat{\mu}(\mathbf{x}^n))^\top \hat{\Sigma}(\mathbf{x}^n)^{-1} (\mathbf{x}_i - \hat{\mu}(\mathbf{x}^n)) \right\} \tag{14}$$

$$= (2\pi e)^{-\frac{mn}{2}} \prod_{j=1}^m \hat{\lambda}_j(\mathbf{x}^n)^{-\frac{n}{2}}. \tag{15}$$

Next, we calculate the denominator of formula (12). Using the fact that $\hat{\mu}(\mathbf{x}^n)$ and $\hat{\Sigma}(\mathbf{x}^n)$ are sufficient statistics, we can calculate the normalization term as an integral with respect to $\hat{\mu}, \hat{\Sigma}$. As MLEs are sufficient statistics, $f(\mathbf{x}^n; \mu, \Sigma)$ is decomposed as follows:

$$f(\mathbf{x}^n; \mu, \Sigma) = f(\mathbf{x}^n | \hat{\mu}(\mathbf{x}^n), \hat{\Sigma}(\mathbf{x}^n)) \cdot g_1(\hat{\mu}(\mathbf{x}^n); \mu, \Sigma) \cdot g_2(\hat{\Sigma}(\mathbf{x}^n); \Sigma).$$

where

$$g_1(\hat{\mu}(\mathbf{x}^n); \mu, \Sigma) \stackrel{\text{def}}{=} \frac{1}{(2\pi/n)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2/n} (\hat{\mu}(\mathbf{x}^n) - \mu)^\top \Sigma^{-1} (\hat{\mu}(\mathbf{x}^n) - \mu) \right\},$$

$$g_2(\hat{\Sigma}(\mathbf{x}^n); \Sigma) \stackrel{\text{def}}{=} \frac{|\hat{\Sigma}(\mathbf{x}^n)|^{\frac{n-m-2}{2}}}{2^{\frac{m(n-1)}{2}} |\frac{1}{n}\Sigma|^{\frac{n-1}{2}} \Gamma_m(\frac{n-1}{2})} \times \exp\left\{ -\frac{1}{2} \mathrm{Tr}(n\Sigma^{-1}\hat{\Sigma}(\mathbf{x}^n)) \right\}.$$

Here we define the function $f(\mathbf{x}^n | \hat{\mu}(\mathbf{x}^n), \hat{\Sigma}(\mathbf{x}^n)) = \delta(\hat{\mu}(\mathbf{x}^n) = \hat{\mu}, \hat{\Sigma}(\mathbf{x}^n) = \hat{\Sigma})$. We fix values $\hat{\mu}(\mathbf{x}^n) = \hat{\mu}, \hat{\Sigma}(\mathbf{x}^n) = \hat{\Sigma}$, and let

$$g(\hat{\lambda}) \quad \overset{\text{def}}{=} \quad g_1(\hat{\mu}; \hat{\mu}, \hat{\Sigma}) \cdot g_2(\hat{\Sigma}; \hat{\Sigma}) \tag{16}$$

$$= \quad \frac{n^{\frac{mn}{2}}}{2^{\frac{mn}{2}} \pi^{\frac{m}{2}} e^{\frac{mn}{2}} \Gamma_m(\frac{n-1}{2})} \cdot \prod_{j=1}^{m} \hat{\lambda}_j^{-\frac{m}{2}-1}. \tag{17}$$

Letting the normalization term of formula (12) be $C(\mathcal{M}, n)$, we can calculate an upper bound on it by integrating $g(\hat{\lambda})$ with respect to $\hat{\mu}, \hat{\Sigma}$ over the restricted domain as follows:

$$
\begin{aligned}
C(\mathcal{M}, n) &= \int_{Y(R,\epsilon_1,\epsilon_2)} f(\mathbf{y}^n; \hat{\mu}(\mathbf{y}^n), \hat{\Sigma}(\mathbf{y}^n)) \, \mathrm{d}\mathbf{y}^n \\
&= \int \delta(\hat{\mu}(\mathbf{x}^n) = \hat{\mu}, \hat{\Sigma}(\mathbf{x}^n) = \hat{\Sigma}) \, \mathrm{d}\mathbf{y}^n \int_{||\hat{\mu}||^2 \le R} \mathrm{d}\mu \int g(\hat{\lambda}) \, \mathrm{d}\hat{\Sigma} \\
&= \int_{||\hat{\mu}||^2 \le R} \mathrm{d}\mu \int \mathrm{d}U \int_{\epsilon_{1j} \le \hat{\lambda}_j \le \epsilon_{2j}} g(\hat{\lambda}) \prod_{1 \le i < j \le m} |\hat{\lambda}_i - \hat{\lambda}_j| \, \mathrm{d}\hat{\lambda} \tag{18} \\
&< \int \mathrm{d}U \cdot \epsilon_2^{\frac{m(m-1)}{2}} \int_{||\hat{\mu}||^2 \le R} \mathrm{d}\mu \int_{\epsilon_{1j} \le \hat{\lambda}_j \le \epsilon_{2j}} g(\hat{\lambda}) \, \mathrm{d}\hat{\lambda} \\
&= \frac{\text{Vol}(O(m))}{2^m} \cdot \epsilon_2^{\frac{m(m-1)}{2}} \cdot \frac{2^{m+1} R^{\frac{m}{2}} (\prod_{j=1}^m \epsilon_{1j}^{-\frac{m}{2}} - \prod_{j=1}^m \epsilon_{2j}^{-\frac{m}{2}})}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot \left(\frac{n}{2e}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})} \\
&< \frac{2^{m+1} R^{\frac{m}{2}} \prod_{j=1}^m \epsilon_{1j}^{-\frac{m}{2}}}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)} \times \left(\frac{n}{2e}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})} \\
&= B(m, R, \epsilon) \times \left(\frac{n}{2e}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})}. \tag{19}
\end{aligned}
$$

where the formula (18) owes to [2], and we define $B(m, R, \epsilon)$ by

$$B(m, R, \epsilon) \overset{\text{def}}{=} \frac{2^{m+1} R^{\frac{m}{2}} \prod_{j=1}^m \epsilon_{1j}^{-\frac{m}{2}}}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)}.$$

$B(m, R, \epsilon)$ does not depend on a number of data n. Since (19) is finite, an upper bound on normalization term $C(\mathcal{M}, n)$ does not diverge.

Here, we can define an upper bound on NML (uNML) code-length as follows:

$$
\begin{aligned}
L_{\text{uNML}}(\mathbf{x}^n; Y, \mathcal{M}) &\overset{\text{def}}{=} -\log f_{\text{uNML}}(\mathbf{x}^n; \mathcal{M}) \\
&= -\log f(\mathbf{x}^n; \mathcal{M}, \hat{\theta}(\mathbf{x}^n)) + \log C_{\text{u}}(\mathcal{M}, n), \\
C_{\text{u}}(\mathcal{M}, n) &= B(m, R, \epsilon) \times \left(\frac{n}{2e}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})}.
\end{aligned}
$$

In calculating uNML code-length for Gaussian Mixture Models, we can use this definition of normalization term in the same way as [1].

5 Correction of NML for Generalized logistic distribution

Here, the following text quotes the sentence of the paper [1] and it is modified to describe the correction of NML for generalized logistic distribution.

In the paper [1], we use the generalized logistic distribution as an example of exponential family. The density function of x^n for a generalized logistic distribution with a parameter θ is defined as

$$f(x^n; \theta) = \prod_{i=1}^{n} \frac{\theta e^{-x_i}}{(1 + e^{-x_i})^{\theta+1}}.$$

The MLE of θ is analytically obtained as $\hat{\theta}(x^n) = n/(\sum_{i=1}^{n} \log(1 + e^{-x_i}))$. Thus the joint density of x^n is written as

$$
\begin{aligned}
f(x^n; \theta) &= \theta^n \cdot \exp\left\{ -\sum_{i=1}^{n} x_i - \frac{n(\theta + 1)}{\hat{\theta}(x^n)} \right\} \\
&= H(x^n | \hat{\theta}(x^n)) \cdot g(\hat{\theta}(x^n); \theta),
\end{aligned}
$$

where $n/\hat{\theta}(x^n)$ is distributed according to the Gamma distribution with a shape parameter n and a scale parameter $1/\theta$.

Here, we correct the function $g(\hat{\theta}(x^n); \theta)$ and the result of NML code-length changes. First, the function $g(\hat{\theta}(x^n); \theta)$ is written as:

$$g(\hat{\theta}(x^n); \theta) = \frac{\theta^n}{(n-1)!} \cdot \frac{n^n}{\hat{\theta}(x^n)^{n+1}} \cdot \exp\left\{ -\theta \cdot \frac{n}{\hat{\theta}(x^n)} \right\}.$$

Fix $\hat{\theta}(x^n) = \hat{\theta}$. Then we have

$$g(\hat{\theta}; \hat{\theta}) = \frac{n^n}{e^n (n-1)!} \cdot \frac{1}{\hat{\theta}}.$$

Then the normalization term: $C(\mathcal{M})$ is calculated by taking an integral of $g(\hat{\theta}; \hat{\theta})$ with respect to $\hat{\theta}$. Here we use hyper-parameters $\theta_{\min}, \theta_{\max}$ to restrict the domain for the integral to be taken as follows:

$$Y(\theta_{\min}, \theta_{\max}) = \left\{ y^n | \theta_{\min} \le \hat{\theta}(y^n) \le \theta_{\max} \right\}.$$

Then we have

$$
\begin{aligned}
C(\mathcal{M}) &= \int_{Y(\theta_{\min}, \theta_{\max})} g(\hat{\theta}; \hat{\theta}) \mathrm{d}\hat{\theta} \\
&= \frac{n^n}{e^n (n-1)!} \int_{\theta_{\min}}^{\theta_{\max}} \frac{1}{\hat{\theta}} \, \mathrm{d}\hat{\theta} \\
&= \frac{n^n}{e^n (n-1)!} \cdot \log \frac{\theta_{\max}}{\theta_{\min}}.
\end{aligned}
$$

Hence we obtain an approximation of the normalization term $C(\mathcal{M})$ for generalized logistic distributions in an analytical way.

6   Acknowledgements

References

[1] S. Hirai and K. Yamanishi. Efficient computation of normalized maximum likelihood codes for gaussian mixture models with its applications to clustering. *JOURNAL OF IEEE INFORMATION THEORY*, 2013.

[2] Arakaparampil M Mathai. *Jacobians of matrix transformations and functions of matrix arguments*. World Scientific Publishing Co Inc, 1997.