Draft version September 14, 2017
Typeset using LaTeX default style in AASTeX61

A SEARCH METHOD FOR GAUSSIAN MIXTURES USING MML

Aleti, Casey, Dowe, Lattanzio

(Received; Revised; Accepted)
Submitted to TBD

ABSTRACT

1. INTRODUCTION

There are N data points each with D dimensions, which are to be modelled by K mixture of D-dimensional gaussian distributions, each with a relative weighting w_k such that $\sum_{k=1}^{K} w_k = 1$. The data have the same error value, y_{err}, in all D dimensions, for all N observations.

The full expression for the message length of a given by,

$$I_K = K\log 2 + \frac{(K-1)}{2}\log N - \frac{1}{2}\sum_{k=1}^{K}\log w_k - \log|\Gamma(K)| + \mathcal{L}(\mathbf{y}|\boldsymbol{\theta}) - DN\log y_{err} \tag{1}$$

$$+\frac{1}{2}\sum_{k=1}^{K}\left[\frac{D(D+3)}{2}\log Nw_k - (D+2)\log|\mathbf{C}_k| - D\log 2\right] - \frac{Q}{2}\log(2\pi) + \frac{\log Q\pi}{2} \tag{2}$$

where $Q = \frac{1}{2}DK(D+3) + K - 1$, the number of free parameters, and $\mathcal{L}$ is the log-likelihood of a multivariate gaussian distribution.

Say we wanted to calculate whether another mixture was warranted. If another mixture were preferred then we would want:

$$\Delta I_{K+1} - I_K < 0 \tag{3}$$

The expression is given as:

$$\Delta I_{K+1} - I_K = (K+1)\log 2 - K\log 2 \tag{4}$$

$$+\frac{(K)}{2}\log N - \frac{1}{2}\sum_{k=1}^{K+1}\log {w_k}^{(new)} - \log|\Gamma(K+1)| \tag{5}$$

$$-\frac{(K-1)}{2}\log N + \frac{1}{2}\sum_{k=1}^{K}\log w_k + \log|\Gamma(K)| \tag{6}$$

$$+\mathcal{L}^{(new)} - DN\log y_{err} \tag{7}$$

$$-\mathcal{L}^{(old)} + DN\log y_{err} \tag{8}$$

$$+\frac{1}{2}\sum_{k=1}^{K+1(new)}\left[\frac{D(D+3)}{2}\log Nw_k - (D+2)\log|\mathbf{C}_k| - D\log 2\right] \tag{9}$$

$$-\frac{1}{2}\sum_{k=1}^{K(old)}\left[\frac{D(D+3)}{2}\log Nw_k - (D+2)\log|\mathbf{C}_k| - D\log 2\right] \tag{10}$$

$$-\frac{Q^{(new)}}{2}\log(2\pi) + \frac{\log Q^{(new)}\pi}{2} \tag{11}$$

$$+\frac{Q^{(old)}}{2}\log(2\pi) - \frac{\log Q^{(old)}\pi}{2} \tag{12}$$

By making use of $\log\Gamma(K) - \log\Gamma(K+1) = -\log K$ and re-arranging the expression:

$$\begin{aligned}\Delta I_{K+1} - I_K =& \log 2 + \frac{1}{2}\log N - \log K - \frac{1}{2}\left(\sum_{k=1}^{K+1}\log w_k^{(new)} - \sum_{k=1}^{K}\log w_k^{(old)}\right)\\ &+\mathcal{L}^{(new)} - \mathcal{L}^{(old)}\\ &+\frac{1}{2}\left[\frac{D(D+3)}{2}\left(\sum_{k=1}^{K+1}\log Nw_k^{(new)} - \sum_{k=1}^{K+1}\log Nw_k^{(old)}\right) - (D+2)\left(\sum_{k=1}^{K+1}\log|\mathbf{C}_k|^{(new)} - \sum_{k=1}^{K+1}\log|\mathbf{C}_k|^{(old)}\right)\right]\\ &+\frac{\log(2\pi)}{2}(Q^{(old)} - Q^{(new)}) + \frac{\pi}{2}\left(\log Q^{(new)} - \log Q^{(old)}\right)\end{aligned} \tag{13}$$

Expanding the Q terms:

$$Q^{(old)} - Q^{(new)} = \frac{1}{2}DK(D+3) + K - 1 - \frac{1}{2}D(K+1)(D+3) + (K+1) - 1$$
$$Q^{(old)} - Q^{(new)} = -\frac{1}{2}D(D+3) + 2K - 1 \tag{14}$$

And making use of the following logarithmic identities,

$$\begin{aligned}\log Q^{(new)} &= \log\left(\frac{1}{2}D(K+1)(D+3) + K\right)\\ &= \log\left(\frac{1}{2}D(K+1)(D+3)\right) + \log\left(1 + \frac{K}{\frac{1}{2}D(K+1)(D+3)}\right)\end{aligned} \tag{15}$$

$$\begin{aligned}\log Q^{(old)} &= \log\left(\frac{1}{2}DK(D+3) + K - 1\right)\\ &= \log\left(\frac{1}{2}DK(D+3)\right) + \log\left(1 + \frac{K-1}{\frac{1}{2}DK(D+3)}\right)\end{aligned} \tag{16}$$

gives us,

$$\begin{aligned}\log Q^{(new)} - \log Q^{(old)} = &\log\left(\frac{1}{2}D(K+1)(D+3)\right) - \log\left(\frac{1}{2}DK(D+3)\right)\\ &+\log\left(1 + \frac{K}{\frac{1}{2}D(K+1)(D+3)}\right) - \log\left(1 + \frac{K-1}{\frac{1}{2}DK(D+3)}\right) \quad .\end{aligned} \tag{17}$$
$$\tag{18}$$

The second row of terms can be ignored because they are very small (typically less than 1 bit). This is because as $K \to \infty$, $2K/D(K+1)(D+3) \to 1$, thus $\log\left(1 + \frac{K}{\frac{1}{2}D(K+1)(D+3)}\right) \to \log 2$. Similarly as $D \to \infty$, $2K/D(K+1)(D+3) \to 0$.

As $K \to \infty$ then $2(K-1)/DK(D+3) \to 1$ and as $D \to \infty$ then $2(K-1)/DK(D+3) \to 0$ and thus $\log\left(1 + \frac{K-1}{\frac{1}{2}DK(D+3)}\right) \approx 0$.

Ignoring these minor terms:

$$\log Q^{(new)} - \log Q^{(old)} \approx \log\left(\frac{1}{2}D(K+1)(D+3)\right) - \log\left(\frac{1}{2}DK(D+3)\right)$$
$$\log Q^{(new)} - \log Q^{(old)} \approx \log(K+1) - \log K \tag{19}$$

Substituting Eqs. 19 and 14 into 13 yields:

$$\begin{aligned}\Delta I_{K+1} - I_K \approx &\log 2 + \frac{1}{2}\log N - \log K - \frac{1}{2}\left(\sum_{k=1}^{K+1}\log w_k^{(new)} - \sum_{k=1}^{K}\log w_k^{(old)}\right)\\ &+\mathcal{L}^{(new)} - \mathcal{L}^{(old)}\\ &+\frac{1}{2}\left[\frac{D(D+3)}{2}\left(\sum_{k=1}^{K+1}\log Nw_k^{(new)} - \sum_{k=1}^{K+1}\log Nw_k^{(old)}\right) - (D+2)\left(\sum_{k=1}^{K+1}\log|\mathbf{C}_k|^{(new)} - \sum_{k=1}^{K+1}\log|\mathbf{C}_k|^{(old)}\right)\right]\\ &+\frac{\log(2\pi)}{2}\left(-\frac{1}{2}D(D+3) + 2K - 1\right) + \frac{\pi}{2}\left(\log(K+1) - \log K\right)\end{aligned} \tag{20}$$

REFERENCES