IDIAP RESEARCH REPORT

WHAT IS BETTER: GMM OF TWO GAUSSIANS OR TWO CLUSTERS WITH ONE GAUSSIAN?

Itshak Lapidot

IDIAP-RR 02-56

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 02-56

WHAT IS BETTER: GMM OF TWO GAUSSIANS OR TWO CLUSTERS WITH ONE GAUSSIAN?

Itshak Lapidot

NOVEMBER 2002

Abstract. In this report, we provide a theoretical discussion on temporal data cluster analysis: does the data come from one source or two sources; is it better to cluster the data into two clusters or leave it as one cluster. Here we analyse only the simplest case: when the data comes from two symmetric Gaussian probability-density-functions (*pdf*s), i.e., with same variance and same absolute value of the mean, with the same prior probability per Gaussian. The data consists of segments with an *a-priori* known segment length. It will be shown that if the data belongs to two different Gaussian models, the likelihood of two clusters is always higher or equal than the one of a GMM with two Gaussians for any mean, variance, and segment length. If the data belongs to the GMM, the likelihood of two clusters might be either higher or less than the GMM one.

Key Terms: clustering, expectation maximization, Gaussian mixture model, temporal data clustering.

The author wants to thank the Swiss Federal Office for Education and Science (OFES) in the framework of both EC/OFES “MultiModal Meeting Manager (M4) project” and the Swiss National Science Foundation, through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)" for supporting this work.

1. Introduction

In temporal data clustering problems such as clustering of speakers, protein chains, music, etc, the real number of clusters is not always known. In the case where the number of clusters is unknown the data is frequently clustered first according to a relatively high number of clusters [1], [2]. Each cluster C_i is defined by a parametric *pdf* with estimated vector of parameters $\hat{\Theta}_i$. When the clusters are created the next step is to decide whether there are clusters to be merged. One way of doing so is to create a new cluster using the union of the data of both clusters. The log-likelihood of the new cluster can then be compared with the accumulated log-likelihood of the two clusters. According to Bayesian model selection [3] the comparison can be done only if the number of parameters in both cases are the same (similar conclusions about model comparison can be achieve, under some assumptions, according to information theory on two-part message length [4], [5]). Let $|\hat{\Theta}|$ be the size of parameters vector $\hat{\Theta}$ of the new cluster. If $|\hat{\Theta}| \neq |\hat{\Theta}_1| + |\hat{\Theta}_2|$ ($\hat{\Theta}_1$ and $\hat{\Theta}_2$ are the estimated parameter vectors of the original clusters) a difference in the complexity should be taken into account, e.g., AIC, BIC, MDL or MML [3]-[6]. Otherwise, the decision can be taken according to the comparison of the likelihoods of the two systems. The case where $|\hat{\Theta}| = |\hat{\Theta}_1| + |\hat{\Theta}_2|$ holds show very good results for speaker clustering and speaker change detection problems [2], [7]. In these problems the estimated parameters often come from GMMs with different number of Gaussian mixture components. Selecting the segment length is another important problem. This length should ensure a minimum duration to obtain sufficient statistics for cluster estimation. Another issue is the fact that segment lengths can vary from segment to segment, e.g., the duration of each speaker turn usually vary, although in several applications the length is known and constant [8]. Exactly the same logic can be used to ask if a given data should be split into two clusters or should remain as a single cluster. It seems very difficult to analyze the behavior of the general case, i.e., unknown density functions and segment length that is not constant or unknown parameter. An accurate analysis of the simplest case might give a significant insight on the behavior of such applications.

The question of how to define a source is an additional problem that is usually task dependent. Different sources can be defined as different phonemes or different speakers. Minimum segment duration can be a helpful parameter for source definition. We will assume that two discrete stochastic process Strict-Sense Stationary (SSS) have the same source if they have the same n-th order *pdf*s for any sample of the process and for any $n \in \mathbb{N}$.

In this work, we will assume that the data consist of segments with constant and *a-priori* known length. These segments are streams of data and each stream comes from a single source. The question is then: "are all the streams coming from the same source or from two different sources?" The sources can be speakers, images etc. This is actually a hypothesis test: H_0 – all the segments come from the same source, H_1 – the segments come from two sources. The complete data *pdf* $f_x(\alpha)$ should be approximated from a known parametric *pdf* model class $\mathcal{M} = \{ f_X(\alpha|\Theta) \}$. The goal is to estimate the parameters $\hat{\Theta}$, in order to maximize the likelihood. We would like to know whether the data comes from two different sources or from one single source. In this case the data is first clustered into two clusters C_1 and C_2, with two *pdf*s, $g_x(\alpha|\hat{\Theta}_1) \in \mathcal{M}$ and $g_x(\alpha|\hat{\Theta}_2) \in \mathcal{M}$. Then the likelihood using the estimated *pdf*s of the two clusters is compared with the likelihood using the estimated *pdf* of one large cluster, $f_x(\alpha|\hat{\Theta}) \in \mathcal{M}$, with $|\hat{\Theta}| = |\hat{\Theta}_1| + |\hat{\Theta}_2|$.

In this report we analyze the case where the data comes from two symmetrical one-dimensional Gaussians sources, i.e., the absolute value of their mean is the same, they have the same variance and the same prior probability for each Gaussian. This knowledge, of course, is not taken into consideration during the clustering process and parameter estimation. Additionally, we will assume that the segments (each source produces many streams, or segments, with fixed length) are statistically independent and the samples in each segment are independently and identically distributed (i.i.d.). The i.i.d. assumption is usually wrong but simplifies the problem to the estimation of one-dimensional *pdf* rather than a *pdf* of the dimension of the segment. It will be shown that in this simple case, it is always better to produce two clusters of one Gaussian each rather than one GMM with two Gaussian mixture components, in the case where the data comes from two different sources (Gaussians). It will be shown that the results are valid for any fixed *a-priori* known segment length. On the other hand, if the data comes from the same source modeling by a GMM with two Gaussian mixture components, the decision depends on the Gaussian parameters and the segments length.

Another important assumption in all the following discussion is that the number of data samples, N, tends toward infinity. Let N_s be the segment length. In several cases N_s will also tend towards infinity. It is always assumed that $N_s/N \to 0$, i.e., the number of data samples tends toward infinity, infinitely faster than the segment length (the number of segments in the data is infinite).

In real life applications, due to the fact that our assumptions of independence, knowledge about segment boundaries, and the assumption about the model are not always valid, and most importantly, the choice of the complete data *pdf* model is usually not exact, the problem is of course much more difficult.

The report is structured as follows: in section II the relations between Maximum-Likelihood (ML) and Minimum-Entropy (ME) will be presented. Section III describes the calculation of the likelihood of multiple cluster case. In section IV an example of the segment length importance for clustering performance is analyzed. Section V presents the simplest case when the length of the segments is equal to one and the mean of the Gaussians is equal to zero or tends toward infinity. Section VI extends the problem in section V to any segment length when the data comes from two Gaussian sources. Section VII extends section VI for any value of the Gaussian mean. Section VIII presents partial results for the case where all the segments come from the same source. Section IX presents simulation results with artificial and speech data, and section X concludes the report.

2. The Relations Between Maximum Likelihood and Minimum Entropy

In many applications, given a dataset, we need to estimate the probabilistic model from which the data was derived. The actual *pdf* of the data $f_x(\alpha)$ is usually unknown and a parametric model is assumed as a *pdf* $f_x(\alpha|\Theta)$ from a parametric model class $\mathcal{M}$ (assuming that $f_x(\alpha)\cup\mathcal{M}$ are a set of regular functions). The most common approach is maximizing the likelihood (or log-likelihood) [9]: given a dataset with N i.i.d. examples and a model class $\mathcal{M}$, we would like to estimate the parameters Θ that maximize the likelihood of the given data:

$$\hat{\Theta} = \arg\max_{\Theta\in\mathcal{M}} \left\{ l\left(X|\Theta\right) \right\} = \arg\max_{\Theta\in\mathcal{M}} \left\{ \prod_{n=1}^{N} p\left(x_n|\Theta\right) \right\}. \tag{1}$$

We can maximize the log-likelihood instead of the likelihood function and get:

$$\hat{\Theta} = \underset{\Theta \in \mathcal{M}}{\arg\max} \left\{ L\left(X \mid \Theta\right) \right\} = \underset{\Theta \in \mathcal{M}}{\arg\max} \left\{ \sum_{n=1}^{N} \log\left(p\left(x_n \mid \Theta\right)\right) \right\}. \tag{2}$$

As the dataset consists of i.i.d. examples then the dataset $Y = \left\{ y_n = f\left(x_n;\Phi\right)\right\}_{n=1}^{N}$ is also i.i.d. ($f\left(x_i;\Phi\right)$ is a parametric function of x_i, with parameter vector Φ). As Y is i.i.d. then according to the law of large numbers: $\frac{1}{N}\sum_{n=1}^{N} y_n \underset{N\to\infty}{\to} E\{Y\}$ if $Var\{Y\} < \infty$, and this is the typical case, otherwise it is true in probability only. Let us define $y_n = f\left(x_n;\Phi\right) = \log\left(p\left(x_n \mid \hat{\Theta}\right)\right)$. Let us define the normalize log-likelihood as $NL\left(X \mid \hat{\Theta}\right) = \frac{1}{N} L\left(X \mid \hat{\Theta}\right)$. Then for $N \to \infty$ we immediately get the next expression:

$$\mathrm{NL}\left(X \mid \hat{\Theta}\right) = \lim_{N\to\infty} \frac{1}{N} \mathrm{L}\left(X \mid \hat{\Theta}\right) = E\left\{\log\left(p\left(x \middle| \hat{\Theta}\right)\right)\right\} = \int_{-\infty}^{\infty} f_x(\alpha) \cdot \log\left(f_x\left(\alpha \middle| \hat{\Theta}\right)\right) d\alpha\,. \tag{3}$$

We will define:

1. $H\left(f_x(\alpha)\right)$ – the entropy of the random variable x according to *pdf* $f_x(\alpha)$.
2. Cross-entropy – $H\left(f_x(\alpha); f_x\left(\alpha|\hat{\Theta}\right)\right) = -E\left\{f_x\left(\alpha|\hat{\Theta}\right)\right\}$, expectation of the function $f_x\left(\alpha|\hat{\Theta}\right)$ according to the *pdf* $f_x(\alpha)$.

From (3) it can be seen that the estimation of the log-likelihood is equivalent to the estimation of the cross-entropy, i.e., maximizing the log-likelihood is the same as minimizing the cross-entropy between the real source *pdf* and the estimated source *pdf* ($\mathrm{NL}\left(X \mid \hat{\Theta}\right) = -H\left(f_x(\alpha); f_x\left(\alpha|\hat{\Theta}\right)\right)$). From the properties of the entropy we know that $H\left(f_x(\alpha)\right) < H\left(f_x(\alpha); g_x(\alpha)\right) \quad \forall f_x(\alpha) \neq g_x(\alpha)$ (the Kullback-Leibler Divergence is always greater than or equal to zero), so minimizing the cross-entropy will maximize the log-likelihood.

Let us examine a simple example. Let us assume that $\alpha \sim U(0,1)$ and that $f_x\left(\alpha \middle| \Theta = \{\mu,\sigma\}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$, i.e., in order to estimate the $f_x\left(\alpha \middle| \hat{\Theta}\right)$ function it is sufficient to estimate the mean, $\hat{\mu}$, and the variance, $\hat{\sigma}^2$, that will maximize the log-likelihood. The common way is to calculate the mean and the variance from the input data. To find the parameters using entropy minimization, we should solve $\frac{\partial}{\partial\Theta} H\left(f_x(\alpha); f_x(\alpha|\Theta)\right) = 0$:

$$\begin{cases} \dfrac{\partial}{\partial\mu} H\left(f_x(\alpha); f_x(\alpha|\Theta)\right) = -\displaystyle\int_{-\infty}^{\infty} f_x(\alpha) \cdot \frac{1}{\sigma^2}(\alpha-\mu)\, d\alpha = 0 \\ \qquad \Rightarrow \quad \hat{\mu} = \displaystyle\int_{-\infty}^{\infty} \alpha \cdot f_x(\alpha)\, d\alpha = E(\alpha) \\ \dfrac{\partial}{\partial\sigma} H\left(f_x(\alpha); f_x(\alpha|\Theta)\right) = -\displaystyle\int_{-\infty}^{\infty} f_x(\alpha) \cdot \left[\frac{1}{\sigma^3}(\alpha-\mu)^2 - \frac{1}{\sigma}\right] d\alpha \\ = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \displaystyle\int_{-\infty}^{\infty} (\alpha-\hat{\mu})^2 \cdot f_x(\alpha)\, d\alpha = Var(\alpha) \end{cases} \tag{4}$$

It is clear that in both cases we estimate the mean and the variance of the real distribution, $f_x(\alpha)$, and apply them to the assumed model.

3. Likelihood of K clusters

Assume that the data X consists of M statistically independent segments $X=\{X_m\}_{m=1}^{M}$, of length N_m. These segments are clustered into K clusters $\{C_k\}_{k=1}^{K}$. As a result of the clustering process we obtain: the estimated parameters of each cluster, $\{\hat{\Theta}_k\}_{k=1}^{K}$, and the estimated labels of each segment, $\{L_m\}_{m=1}^{M}$. Using the estimated parameters we want to estimate the likelihood of the data X. We must first define the *pdf* of one segment, X_m, using estimated parameters $\{\hat{\Theta}_k\}_{k=1}^{K}$ and $L_m = \underset{k=1,\dots,K}{\arg\max}\left\{p\left(X_m \middle| g_{X_m}\left(\boldsymbol{\alpha}_m \middle| \hat{\Theta}_k\right)\right)\right\} = \underset{k=1,\dots,K}{\arg\max}\left\{p\left(X_m \middle| \hat{\Theta}_k\right)\right\}$:

$$g_{X_m}(\boldsymbol{\alpha}_m) = \sum_{k=1}^{K} \delta(L_m = k)\, g_{X_m}\left(\boldsymbol{\alpha}_m \middle| \hat{\Theta}_k\right) \tag{5}$$

where $\delta(L_m = k)$ is an indicator function, i.e., equals one if $L_m = k$, zero otherwise.
As the segments are statistically independent, the *pdf* of the entire dataset will be a multiplication of the segment *pdf* functions. If the vectors in each segment are i.i.d., then each segment *pdf* is a multiplication as well:

$$\begin{aligned} g_X(\boldsymbol{\alpha}) &= \prod_{m=1}^{M} g_{X_m}(\boldsymbol{\alpha}_m) = \prod_{m=1}^{M}\left(\sum_{k=1}^{K} \delta(L_m = k)\, g_{X_m}\left(\boldsymbol{\alpha}_m \middle| \hat{\Theta}_k\right)\right) \\ &= \prod_{k=1}^{K} \prod_{X_m \in C_k} g_{X_m}\left(\boldsymbol{\alpha}_m \middle| \hat{\Theta}_k\right) = \prod_{k=1}^{K} \prod_{X_m \in C_k} \prod_{n=1}^{N_m} g_{x_{m,n}}\left(\alpha_{m,n} \middle| \hat{\Theta}_k\right) \end{aligned} \tag{6}$$

As a consequence, the likelihood of the data X will be as follows:

$$l(X) = \prod_{k=1}^{K} \prod_{X_m \in C_k} p\left(X_m \middle| \hat{\Theta}_k\right) = \prod_{k=1}^{K} \prod_{X_m \in C_k} \prod_{n=1}^{N_m} p\left(x_{m,n} \middle| \hat{\Theta}_k\right) \tag{7}$$

If the segments are all of the same size, , the number of segments in a cluster is M_k and the dataset size tends toward infinity ($\forall k : M_k \to \infty$), then the estimation of the normalized log-likelihood can be written as follows:

$$NL(X) = -\sum_{k=1}^{K} \frac{M_k}{M} H\left(f_x(\alpha | C_k); g_x\left(\alpha \middle| \hat{\Theta}_k\right)\right) \;\; ; \;\; M = \sum_{k=1}^{K} M_k \tag{8}$$

where $f_x(\alpha | C_k)$ is the true *pdf* of the cluster C_k. For simplicity we will use the following notation: $P_k \cdot H\left(f_x(\alpha|C_k); g_x\left(\alpha\middle|\hat{\Theta}_k\right)\right) = H\left(P_k \cdot f_x(\alpha|C_k); g_x\left(\alpha\middle|\hat{\Theta}_k\right)\right)$ $\quad P_k = \frac{M_k}{M}$. This term represents the negative of the contribution of the k-th cluster to the normalized log-likelihood.

4. Simple theoretical example

In order to show when temporal data clustering is superior to static clustering a simple theoretical example is presented. Assume that the data comes with equal probability from two Gaussian sources, $S_1 \sim N(\mu, \sigma^2)$ and $S_2 \sim N(-\mu, \sigma^2)$, and all data points are statistically independent. The data should be clustered into two clusters, C_1 and C_2. Each cluster will be modeled by only one Gaussian. We will consider two cases: first, assuming no knowledge about any dependency between the data points; and second, assuming that each pair of data points comes from the same source ($\forall m \in \mathbb{Z} \bullet x_{2m} \in S_i \Rightarrow x_{2m+1} \in S_i$, $i \in \{1,2\}$).

For the first case, because of the symmetry of the problem and without loss of generality the clusters will be: $C_1 \supseteq \{x | x > 0\}$ and $C_2 \supseteq \{x | x < 0\}$ ($x = 0$ can be attributed to any cluster). So the *pdf* given C_1 will be: $f_x(\alpha | C_1) = \frac{1}{\sqrt{2\pi}\sigma}\left[\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} + \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right]U_{-1}(\alpha)$ ($U_{-1}(x)$ is a step function), and because of the symmetry $f_x(\alpha | C_2) = f_x(-\alpha | C_1)$.

In the second case the decision will be taken according to two points, so without loss of generality $C_1 \supseteq \{x_{2m}, x_{2m+1} | x_{2m} + x_{2m+1} > 0\}$ and $C_2 \supseteq \{x_{2m}, x_{2m+1} | x_{2m} + x_{2m+1} < 0\}$ ($x_{2m} + x_{2m+1} = 0$ can be attributed to any cluster). The decision is taken according to two data points where each pair is i.i.d. In the two dimensional space the data will consist of two Gaussians with means at $[\mu, \mu]^T$ and $[-\mu, -\mu]^T$ and variances $[\sigma^2, \sigma^2]$. The clusters boundary will be on the line $x_1 + x_2 = 0$. In order to find the *pdf* of C_i we need to calculate the marginal distribution of x_j given C_i. The *pdf* of C_1 will be computed according to (9) and $f_x(\alpha | C_2) = f_x(-\alpha | C_1)$. Given the prior knowledge that both samples come from the same source (S_1 or S_2) only two terms of integration are present in (9). Without this knowledge two more terms of integration would be present for the cases where thet first input comes from S_1 and the second comes from S_2 and the case where the first input comes from S_2 and the second comes from S_1. $\Phi(\xi)$ is a distribution function (integral of the *pdf* function from zero to ξ) of a normally distributed variable z, $z \sim N(0,1)$.

Figure 1a shows the *pdf*s of two sources, with means $\mu_{1,2} = \pm 1$ and variance equals to one; the *pdf*s of the clusters in the first case are shown in 1b, and the *pdf*s of the second case are shown in 1c. It can be clearly seen that the additional information about the duration dependency of the data significantly improves the clustering performance even in the simplest case where segment length equals two and each source consists of i.i.d samples. In more difficult cases, more sophisticated models and longer segment length information should be applied to reach significant improvement.

$$
\begin{aligned}
f_{x_1}(\alpha | C_1) &= \int_{-\infty}^{\infty} f_{x_1,x_2}(\alpha, \beta | C_1)\, d\beta \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\int_{-\alpha}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\mu)^2\right\}d\beta \\
&+ \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\int_{-\alpha}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\beta+\mu)^2\right\}d\beta \qquad (9) \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\left(1-\Phi\left(\frac{-\alpha-\mu}{\sigma}\right)\right) + \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\left(1-\Phi\left(\frac{-\alpha+\mu}{\sigma}\right)\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\Phi\left(\frac{\alpha+\mu}{\sigma}\right) + \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\Phi\left(\frac{\alpha-\mu}{\sigma}\right)
\end{aligned}
$$

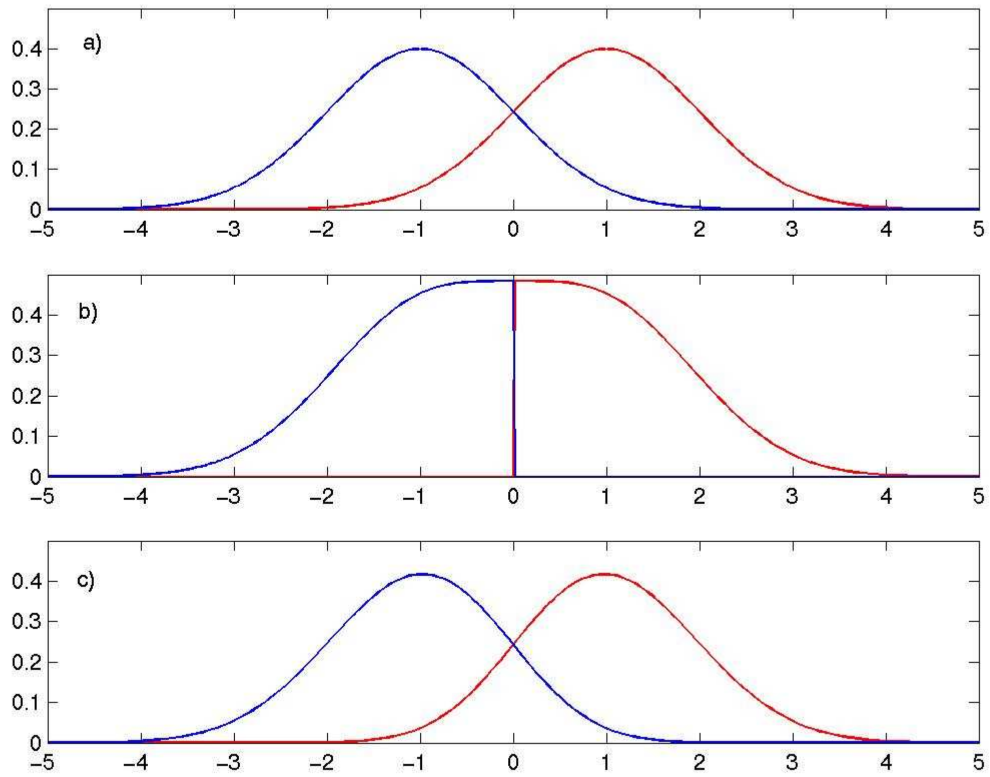

Fig. 1: Clustering of two Gaussian sources with means, $\mu_{1,2} = \pm 1$ and variances, $\sigma_{1,2}^2 = 1$ (red line corresponds to 1 and blue line to 2). a) *pdf*s of the sources, b) *pdf*s of the clusters without duration constrain, c) *pdf*s of the clusters with segment length equal 2.

Additional important characteristics of temporal data clustering can be observed from comparison between Figs. 1b and 1c. It is seen from the figures and it is easy to show that the variance of the data attributed to clusters of the first case is smaller than the variance in the second case. In the next section we will prove that the mean always become smaller and the variance larger as a function of the segment length. Moreover, in the limit the mean and the variance tend towards the values of the source mean and variance.

5. Segment length equals one

In this section, as segment length equals one, we will not assume that the data comes from two sources, each being a Gaussian, with same variances, and with prior probability of 0.5, or from one GMM of two Gaussian mixture components. When segment length is equal to one, both cases are exactly the same. The data samples are assumed to be statistically independent. The question is, again, whether it is better to consider the data belonging to one source or two sources containing one Gaussian each. An additional assumption is that the data size is infinite. Intuitively it seems that when the means are close relatively to the standard deviation they should be considered as one model while when they are far they would be considered as two models. Consider, without loss of generality, that the data is distributed as follows:

$$X \sim f_x(\alpha) = \frac{1}{2} N(\mu, \sigma^2) + \frac{1}{2} N(-\mu, \sigma^2) = \frac{1}{2} f_x^1(\alpha) + \frac{1}{2} f_x^2(\alpha) \tag{10}$$

where $f_x^1(\alpha)$ is a Gaussian distribution with mean μ, and $f_x^2(\alpha)$ is a Gaussian distribution with mean $-\mu$. The clusters with distribution $f_x^i(\alpha)$ will be called G_i. Each cluster contains all the data that was generated by $f_x^i(\alpha)$.

The objective is to cluster the data into two clusters $\{C_i\}_{i=1}^2$ such that $x_{ij} \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$ and $X_i = \{x_{ij} | x_{ij} \in C_i\}$ (X_i is the data attributed to cluster C_i). It is important to notice that C_i is an estimated cluster while G_i is the theoretical, true, cluster. Because of the symmetry it is obvious, without loss of generality that the stable clustering will be: $X_1 = \{x_{1j} | x_{1j} > 0\}$ $X_2 = \{x_{2j} | x_{2j} < 0\}$ ($x = 0$ can be attributed to any cluster), and $\hat{\mu}_1 = -\hat{\mu}_2$, $\hat{\sigma}_1 = \hat{\sigma}_2$. As each Gaussian is estimated from its cluster C_i, we will end up with the true cluster *pdf*s:

$$f_x(\alpha | C_i) = f_x\left(\alpha \middle| \alpha(-1)^{i+1} > 0\right) = 2 f_x(\alpha) U_{-1}\left(\alpha(-1)^{i+1}\right) \quad ; \quad U_{-1}(\alpha) = \begin{cases} 1 & \alpha > 0 \\ 0 & \alpha < 0 \end{cases} \tag{11}$$

It is difficult to analyze the general case for any μ, hence we will first analyze the extreme cases when $\mu \to \infty$ and $\mu = 0$.

5.1 Notations

For simplicity we will use the following notations:

1. $f_{i,x}(\alpha)$ – the true *pdf* of the i-th cluster C_i.
2. $g_{i,x}(\alpha)$ – estimated *pdf* of the i-th cluster, $g_x(\alpha | \hat{\Theta}_i)$.

5.2 μ tends toward infinity

Without loss of generality, we will only analyze the case where $\alpha > 0$. As $\mu \to \infty$ and σ is finite the estimated parameters of the mean and the standard deviation will not be affected by the Gaussian with the mean of $-\mu$ as for this Gaussian the probability of $\alpha > 0$ is equal to zero. From that it follows that: $\hat{\mu} = \mu$ and $\hat{\sigma} = \sigma$.

5.3 μ equals zero

In this case the GMM of two Gaussian mixture components with prior probability of 0.5 per Gaussian becomes a normal distribution with zero mean, and the *pdf* when $\alpha > 0$ is $f_{1,x}(\alpha) = 2\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\alpha^2\right\} U_{-1}(\alpha)$ (11). The estimation of the mean is:

$$\hat{\mu} = 2\int_0^\infty \alpha f_x(\alpha)\, d\alpha = \frac{2\sigma}{\sqrt{2\pi}} \tag{12}$$

and the estimated variance is:

$$\hat{\sigma}^2 = \mathrm{E}\{\alpha^2\} - \hat{\mu}^2 = \sigma^2 - \hat{\mu}^2 = \frac{(\pi - 2)\sigma^2}{\pi} \tag{13}$$

5.4 The normalized log-likelihood of the two cases of μ

As it was shown in section II, the normalized log-likelihood is equal to the negated cross-entropy. In this sub-section we will show that the cross-entropy of two clusters always lower than the entropy of the data. First let us find the real *pdf* of the data attributed to each C_i:

$$f_{i,x}(\alpha) = f_x\left(\alpha\middle|\alpha(-1)^{i+1} > 0\right) = \frac{f_x\left(\alpha \cap \left(\alpha(-1)^{i+1} > 0\right)\right)}{P\left(\alpha(-1)^{i+1} > 0\right)} = \frac{f_x\left(\alpha(-1)^{i+1} > 0\right)}{P\left(\alpha(-1)^{i+1} > 0\right)} = 2 f_x(\alpha) U_{-1}\left(\alpha(-1)^{i+1}\right)$$
$$= \frac{1}{\sqrt{2\pi}\sigma} U_{-1}\left(\alpha(-1)^{i+1}\right)\left[\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} + \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right] \quad .(14)$$

Remind that $g_{i,x}(\alpha)$ is the estimated *pdf* of $f_{i,x}(\alpha)$ with the assumption of Gaussian *pdf*. When μ tends toward infinity, $\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}$ tends to zero $\forall \alpha < 0$, and $\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}$ tends to zero $\forall \alpha > 0$, hence (14) can be rewritten as:

$$f_{i,x}(\alpha) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(\alpha + (-1)^i \mu\right)^2\right\}. \quad (15)$$

In V-A, $\mu \to \infty$, we showed that $g_{i,x}(\alpha) = f_{i,x}(\alpha)$. As each cluster has only half of the data, the normalized log-likelihood should be divided by two. The normalized log-likelihood of both clusters together is $-\frac{1}{2}\sum_{i=1}^{2} H\left(g_{i,x}(\alpha)\right)$ where $g_{i,x}(\alpha) = f_{i,x}(\alpha)$. Because of the symmetry, they are the same, i.e., the total normalized log-likelihood can be written as the entropy of $f_{1,x}(\alpha)$, $-H\left(f_{1,x}(\alpha)\right)$. On the other hand the normalized log-likelihood of $f_x(\alpha)$ is:

$$-H\left(f_x(\alpha)\right) = \int_{-\infty}^{\infty} f_x(\alpha)\log\left(f_x(\alpha)\right)d\alpha = \int_{-\infty}^{\infty} \frac{1}{2}\left[f_{1,x}(\alpha) + f_{2,x}(\alpha)\right]\log\left(\frac{1}{2}\left[f_{1,x}(\alpha) + f_{2,x}(\alpha)\right]\right)d\alpha$$
$$= \int_{-\infty}^{0} \frac{1}{2} f_{x_2}(\alpha)\log\left(\frac{1}{2} f_{2,x}(\alpha)\right)d\alpha + \int_{0}^{\infty}\frac{1}{2} f_{1,x}(\alpha)\log\left(\frac{1}{2} f_{1,x}(\alpha)\right)d\alpha \quad (16)$$
$$= 2\int_{0}^{\infty}\frac{1}{2} f_{x_1}(\alpha)\log\left(\frac{1}{2} f_{x_1}(\alpha)\right)d\alpha = -H\left(f_{x_1}(\alpha)\right) + \log\left(\frac{1}{2}\right)$$

As $\log(1/2) < 0$ the normalized log-likelihood of two clusters is higher than the one of a GMM of two Gaussian mixture components.

We will use the following notation to calculate the normalized log-likelihood:

1. $-H\left(P_i f_{i,x}(\alpha); g_{i,x}(\alpha)\right)$ – normalized log-likelihood of $g_{i,x}(\alpha)$.
2. $-H\left(f_x(\alpha)\| g_x(\alpha)\right) = -\sum_{i=1}^{2} H\left(P_i f_{i,x}(\alpha); g_{i,x}(\alpha)\right)$.

Remind that $f_{i,x}(\alpha) = 2 f_x(\alpha) U_{-1}\left(\alpha(-1)^{i+1}\right)$, then in the case where $\mu = 0$:

$$-H\left(f_x(\alpha)\| g_x(\alpha)\right) = \int_{-\infty}^{0} f_x(\alpha)\log\left(g_{2,x}(\alpha)\right)d\alpha + \int_{0}^{\infty} f_x(\alpha)\log\left(g_{1,x}(\alpha)\right)d\alpha$$
$$= \log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right) + \int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\alpha^2\right\}\left[-\frac{1}{2\hat{\sigma}^2}(\alpha+\hat{\mu})^2\right]d\alpha \quad . \quad (17)$$
$$+ \int_{0}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\alpha^2\right\}\left[-\frac{1}{2\hat{\sigma}^2}(\alpha-\hat{\mu})^2\right]d\alpha = \log\left(\frac{1}{\sqrt{2(\pi-2)}\sigma}\right) - \frac{1}{2}$$

On the other hand the normalized log-likelihood of the real *pdf* is $-H\left(f_x(\alpha)\right) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}$. If we subtract it from equation (17) the result will be:

$$-H\left(f_x(\alpha)\|g_x(\alpha)\right)+H\left(f_x(\alpha)\right)=\frac{1}{2}\log\left(\frac{\pi}{\pi-2}\right)\approx 0.5062>0\,. \tag{18}$$

As the result is greater than zero it means that two clusters have a slightly higher normalized log-likelihood than a GMM of two identical Gaussian mixture components.

To conclude this section it could be said that for the two extremes of the mean value (zero and $+\infty$), two clusters are better than one GMM if segment length is set to one, $N_s=1$. In the next section we will see that the same happens for $N_s\geq 1$, and in section VII we will show that it is valid for all μ.

6. General case at the extremes when $N_s \geq 1,\ \mu\in\{0,\infty\}$

This section deals with the general case from the point of view of segment length only. We will still assume the extreme values of the mean, $\mu=0$ and $\mu\to\infty$. The data is assumed to come from two Gaussian sources, i.e., not from one GMM of two Gaussian mixture components. First, the *pdf* of each cluster will be found, for any $\mu\geq 0$, under the segment length restriction. Then the normalized log-likelihood will be found for $\mu\to\infty$ and it will be shown that it is the same as for $N_s=1$. At the last part of this section, a more difficult case of $\mu=0$ will be analyzed.

6.1 Notations

The following notations will be used:

1. $f_{i,x,N_s}(\alpha)$ – the true *pdf* of the i-th cluster C_i, given segment length N_s.
2. $g_{i,x,N_s}(\alpha)$ – estimated *pdf* of the i-th cluster, $g_x\left(\alpha\middle|\hat{\Theta}_i\right)$, given segment length N_s.

6.2 Cluster *pdf*s

In this section it is assumed that the data consists of segments with length N_s, i.e., data samples $\mathbf{X}_k=\{x_j\}_{j=N_sk+1}^{N_s(k+1)}$, $k\in\mathbb{Z}$, are generated from the same Gaussian. For the GMM of two Gaussian mixture components, the clustered data will be divided into two symmetrical clusters: C_1 is the union of all the segments such that $\sum_{j=N_sk+1}^{N_s(k+1)}x_j>0$, $k\in\mathbb{Z}$ (19), and C_2 is the union of all the segments such that $\sum_{j=N_sk+1}^{N_s(k+1)}x_j<0$, $k\in\mathbb{Z}$ ($\sum_{j=N_sk+1}^{N_s(k+1)}x_j=0$, $k\in\mathbb{Z}$ can be attributed to any cluster). Let us define a new random variable $y_k=\sum_{j=N_sk+1}^{N_s(k+1)}x_j$. As all the variables in the k-th segment are i.i.d. It means that $y_k\sim N\left(N_s\mu,N_s\sigma^2\right)$. From the point of view of y_k the clustering is identical to the analysis in section V. From now on, without loss of generality, the set of random variables belonging to a segment will always be $\{x_j\}_{j=1}^{N_s}$, instead of $\{x_j\}_{j=N_sk+1}^{N_s(k+1)}$, $k\in\mathbb{Z}$. To find the *pdf* of a random variable x_j that belongs to C_1 we need to find the marginal *pdf* of all the segments given $\sum_{n=1}^{N_s}x_n>0$:

$$\begin{aligned} f_{1,x,N_s}\left(\alpha_j\right) &= f_x\left(\alpha_j \left| \sum_{n=1}^{N_s}\alpha_n > 0\right.\right) = \iint_{-\infty}^{\infty}\cdots\int f_{x_1,x_2,\ldots,x_{N_s}}\left(\underbrace{\alpha_1,\alpha_2,\ldots,\alpha_{N_s}}_{\mathrm{A}} \left| \sum_{n=1}^{N_s}\alpha_n\right.\right)\prod_{\substack{n=1\\ n\neq j}}^{N_s} d\alpha_n \\ &= \iint_{-\infty}^{\infty}\cdots\int f_{x_1,x_2,\ldots,x_{N_s}}\left(\mathrm{A}\left|\sum_{n=1}^{N_s}\alpha_n\right.\right)\prod_{\substack{n=1\\ n\neq j}}^{N_s} d\alpha_n \end{aligned} \quad . \tag{19}$$

First let us find the conditional probability of all segments (20). The last form of (20) is correct only because of the assumption that all the samples of one segment belongs to the same Gaussian and they are all i.i.d.

To simplify equations the following notation will be used: $\prod_{n=1,\ n\neq j}^{N_s} d\alpha_n \triangleq \mathbf{d\alpha_{\text{-}j}}$ and $\sum_{n=1,\ n\neq j}^{N_s}\alpha_n \triangleq \mathbf{\Sigma_{\text{-}j}}$. Using this notation and the conditional *pdf* properties, (19) can be rewritten in a simpler way as in (21), where $h_{x_j}^i\left(\alpha_j\right)$ is the multiple integral of $\prod_{n=1}^{N_s} f_{x_n}^i\left(\alpha_n\right)$, and $\mathbf{\Sigma_{\text{-}m}}$ assumes that $m\neq j$. The second part in the third equality is correct only because each segment can belong either to f_x^1 or to f_x^2. The case where the segment belongs to one GMM will be discussed in section VIII.

$$\begin{aligned} f_{x_1,x_2,\ldots,x_{N_s}}\left(\alpha_1,\alpha_2,\ldots,\alpha_{N_s}\left|\sum_{n=1}^{N_s}\alpha_n\right.\right) &= \frac{f_x\left(\mathrm{A}\right)P\left(\sum_{n=1}^{N_s}\alpha_n > 0 \middle| \mathrm{A}\right)}{P\left(\sum_{n=1}^{N_s}\alpha_n > 0\right)} = \frac{f_x\left(\mathrm{A}\right)U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n > 0\right)}{\underbrace{P\left(\sum_{n=1}^{N_s}\alpha_n > 0\right)}_{\text{Equals } \frac{1}{2}}} \\ &= 2\left[f_x\left(\mathrm{A}\cap\mathrm{A}\in G_1\right)P\left(G_1\right) + f_x\left(\mathrm{A}\cap\mathrm{A}\in G_2\right)P\left(G_2\right)\right]U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n > 0\right) \\ &= \left[\prod_{n=1}^{N_s} f_{x_n}^1\left(\alpha_n\right) + \prod_{n=1}^{N_s} f_{x_n}^2\left(\alpha_n\right)\right]U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n > 0\right) \end{aligned} \quad . \tag{20}$$

$$\begin{aligned} f_{1,x_j,N_s}\left(\alpha_j\right) &= \iint_{-\infty}^{\infty}\cdots\int \frac{f_{x_1,x_2,\ldots,x_{N_s}}\left(\alpha_1,\alpha_2,\ldots,\alpha_{N_s}\right)U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n > 0\right)}{P\left(\sum_{n=1}^{N_s}\alpha_n > 0\right)}\mathbf{d\alpha_{\text{-}j}} \\ &= 2\iint_{-\infty}^{\infty}\cdots\int\int_{\mathbf{\Sigma_{\text{-}m}}}^{\infty} f_{x_1,x_2,\ldots,x_{N_s}}\left(\alpha_1,\alpha_2,\ldots,\alpha_{N_s}\right)\mathbf{d\alpha_{\text{-}j}} = \iint_{-\infty}^{\infty}\cdots\int\int_{\mathbf{\Sigma_{\text{-}m}}}^{\infty}\prod_{n=1}^{N_s} f_x^1\left(\alpha_n\right) + \prod_{n=1}^{N_s} f_x^2\left(\alpha_n\right)\mathbf{d\alpha_{\text{-}j}} = \\ &= h_{x_j}^1\left(\alpha_j\right) + h_{x_j}^2\left(\alpha_j\right) \end{aligned} \tag{21}$$

As $\left\{x_j\right\}_{j=1}^{N_s}$ are a sequence of Gaussian i.i.d. random variables, if we define $z=\sum_{n=1,n\neq j}^{N_s} x_n$, z will have a Gaussian distribution with $\mu_z=\left(N_s-1\right)\mu$ and $\sigma_z^2=\left(N_s-1\right)\sigma^2$. It is easy to show that $h_{x_j}^1\left(\alpha_j\right)$ can be written in terms of $f_z\left(\gamma\right)$ as:

$$
\begin{aligned}
h_x^1(\alpha) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\int_{-\alpha}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_z}\exp\left\{-\frac{1}{2\sigma_z^2}(\gamma-\mu_z)^2\right\}d\gamma \\
&= f_x^1(\alpha)\left[1-\Phi\left(\frac{-\alpha-\mu_z}{\sigma_z}\right)\right] = f_x^1(\alpha)\Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right)
\end{aligned} \tag{22}
$$

and in the same way:

$$
h_x^2(\alpha) = f_x^2(\alpha)\Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) \tag{23}
$$

and the final expression for (19) is:

$$
f_{1,x,N_s}(\alpha) = f_x^1(\alpha)\Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) + f_x^2(\alpha)\Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right). \tag{24}
$$

The same technique could be used for the second cluster and the final result is:

$$
f_{2,x,N_s}(\alpha) = f_x^1(\alpha)\Phi\left(\frac{-\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) + f_x^2(\alpha)\Phi\left(\frac{-\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right). \tag{25}
$$

To give a feeling of the correctness of the final results, assume $N_s = 1$ and zero mean. Then the $\Phi(\bullet)$ acts exactly as a step function (for negative α_j, the argument of $\Phi(\bullet)$ tends toward $-\infty$ and the result is zero, for positive α_j, the argument of $\Phi(\bullet)$ tends toward $+\infty$ and the result is one) and $f_x^1(\alpha) = f_x^2(\alpha)$, i.e., the *pdf* of C_1 is identical to the one found in sub-section V-C, $f_{1,x}(\alpha) = 2 f_x(\alpha) U_{-1}(\alpha)$.

6.3 μ tends toward infinity

If we push μ towards infinity in (24) the $\Phi(\bullet)$ in the first term tends toward one and to zero in the second term, for any value of α_j, i.e., $f_{1,x}(\alpha) \underset{\mu\to\infty}{\to} f_x^1(\alpha)$. In this case, like when $N_s = 1$, the estimation of the mean and the variance will be: $\hat{\mu}_\infty = \mu$ and $\hat{\sigma}_\infty^2 = \sigma$. Similarly, for $f_{2,x}(\alpha)$ the estimation of the mean and the variance will be: $\hat{\mu}_\infty = -\mu$ and $\hat{\sigma}_\infty^2 = \sigma$.

As in the case $N_s = 1$, the normalized log-likelihood will be higher by $\log(2)$.

6.4 μ equals zero

For $\mu = 0$ the equality holds for $f_x(\alpha) = f_x^1(\alpha) = f_x^2(\alpha)$ and the *pdf* of C_1 is:

$$
f_{1,x,N_s}(\alpha) = 2 f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right). \tag{26}
$$

First let us remind the following properties: $f_x(\alpha) = f_x(\alpha)$ and $\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right) = 1-\Phi\left(\frac{-\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)$. These will help us understand the properties of the mean and variance of C_1.

$$\begin{aligned}\hat{\mu}_{N_s} &= \int_{-\infty}^{\infty} \alpha f_{1,x,N_s}(\alpha)\,d\alpha = \int_{-\infty}^{\infty} \alpha 2 f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)d\alpha \\ &= 2\int_0^{\infty}\alpha f_x(\alpha)\left[\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)-\Phi\left(\frac{-\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\right]d\alpha\,. \\ &= 2\int_0^{\infty}\alpha f_x(\alpha)\underbrace{\left[2\cdot\underbrace{\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)}_{\alpha\ge 0\Rightarrow\Phi(\bullet)\ge 0.5}-1\right]}_{\text{Greater or equal to 0 }\forall\alpha\ge 0}d\alpha > 0\end{aligned} \tag{27}$$

For $N_s = 1$, $\Phi(\bullet)$ becomes a step function and we get exactly (12). For $N_s \to \infty$, the value of $\Phi(\bullet)$ is zero for any α and the integral is zero, i.e., $\hat{\mu}_\infty = 0$. Now let us show that $\hat{\mu}_{N_s}$ is a monotonically decreasing function of N_s, bounded by $\hat{\mu}_0$ and $\hat{\mu}_\infty$. As $\Phi\left(\frac{\alpha}{\sqrt{N_s+1}\cdot\sigma}\right) > \Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)$ for $\forall\alpha > 0$. It leads to:

$$\hat{\mu}_{N_s+1} = 2\int_0^{\infty}\alpha f_x(\alpha)\left[2\cdot\Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)-1\right]d\alpha < 2\int_0^{\infty}\alpha f_x(\alpha)\left[2\cdot\Phi\left(\frac{\alpha}{\sqrt{N_s+1}\cdot\sigma}\right)-1\right]d\alpha = \hat{\mu}_{N_s}\,. \tag{28}$$

Now we can show that the variance is a monotonically increasing function of N_s (29). As σ^2 is the variance of the zero mean Gaussian data, it is constant. In (28) we showed that $\hat{\mu}_{N_s}$ is monotonically decreasing. From these two facts we obtain that $\hat{\sigma}^2_{N_s}$ is a monotonically increasing function of N_s. $\hat{\sigma}^2_{N_s}$ is bounded by $\hat{\sigma}^2_1 = \frac{(\pi-2)\sigma^2}{\pi}$ according to (13), and $\hat{\sigma}^2_\infty = \sigma^2$.

$$\begin{aligned}\hat{\sigma}^2_{N_s} &= \int_{-\infty}^{\infty}\alpha^2 f_{1,x,N_s}(\alpha)\,d\alpha - \hat{\mu}^2_{N_s} = \int_{-\infty}^{\infty}\alpha^2 2 f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)d\alpha - \hat{\mu}^2_{N_s} \\ &= 2\int_0^{\infty}\alpha^2 f_x(\alpha)\underbrace{\left[\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)+1-\Phi\left(\frac{-\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\right]}_{\text{Equals 1}}d\alpha - \hat{\mu}^2_{N_s} = \sigma^2 - \hat{\mu}^2_{N_s}\end{aligned}\,. \tag{29}$$

We obtain the same result for the mean and variance of C_2 with the exception that the mean will have a minus sign (but the absolute value of the mean is still a decreasing function of N_s, and $\hat{\sigma}^2_{N_s}$ an increasing function of N_s).

We showed that the mean is a monotonically decreasing function of N_s, the variance is a monotonically increasing function of N_s and we gave the normalized log-likelihood for $N_s = 1$ and $N_s \to \infty$. We now need to show that the normalized log-likelihood is a monotonically decreasing function of N_s. Let us first define $g_{1,x,N_s}(\ )$ as a Gaussian *pdf* for cluster C_1 with mean $\hat{\mu}_{N_s}$ and variance $\hat{\sigma}^2_{N_s}$; in the same way $g_{2,x,N_s}(\ )$ is a Gaussian *pdf* for cluster C_2. As each cluster has only half of the data the normalized log-likelihood is:

$$\begin{aligned}NL_{N_s}(\alpha) &= \frac{1}{2}\int_{-\infty}^{\infty} f_{1,x,N_s}(\alpha)\log\left[g_{1,x,N_s}(\alpha)\right]d\alpha + \frac{1}{2}\int_{-\infty}^{\infty} f_{2,x,N_s}(\alpha)\log\left[g_{2,x,N_s}(\alpha)\right]d\alpha \\ &= -H\left(f_{1,x,N_s}(\alpha)\middle\| g_{1,x,N_s}(\alpha)\right)\end{aligned}\,. \tag{30}$$

Given the symmetry, in order to show that $NL_{N_s}(\alpha)$ is a monotonically decreasing function of N_s, the sufficient condition for any N_s is to prove that $\Delta_{N_s} = -H\left(1/2\, f_{1,x,N_s}(\alpha), g_{1,x,N_s}(\alpha)\right) + H\left(1/2\, f_{2,x,N_s+1}(\alpha), g_{2,x,N_s+1}(\alpha)\right) > 0$ is according to (31).

Until now we have shown that the normalized log-likelihood of two clusters is higher than the one of a GMM of two Gaussian mixture components for any segment length when $\mu = 0$ or $\mu \to \infty$. In the next section we shall extend this result for any value of the mean.

7. General case – $N_s \geq 1$ and $0 < \mu < \infty$

As in section V, it is obvious that when N_s tends toward infinity, the means and the variances will be identical to the true Gaussian mean and variances, i.e. $\hat{\mu}_\infty = \mu$ and $\hat{\sigma}^2_\infty = \sigma^2$. We will show that $\hat{\mu}_1 > \mu$ and $\hat{\sigma}^2_1 < \sigma^2$. We will then show that the mean is a decreasing function of N_s and the variance is an increasing function of N_s. These facts will help us to show that the normalized log-likelihood is an increasing function of N_s, while the normalized log-likelihood for $N_s = 1$ is greater than the GMM normalized log-likelihood. This will conclude the proof that for any mean and any segment length, two single Gaussian clusters are better than one GMM of two Gaussian mixture components, if the data really comes from two Gaussian sources.

First let us show that $\hat{\mu}_1 > \mu$. The value of the real mean can be written as a difference of two integrals as in (32).

$$
\begin{aligned}
\Delta_{N_s} &= \int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\log\left[g_{1,x,N_s}(\alpha)\right]d\alpha - \int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)\log\left[g_{1,x,N_s+1}(\alpha)\right]d\alpha \\
&= \int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\left[\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{N_s}}\right) - \frac{1}{2\pi}\left(\alpha-\hat{\mu}_{N_s}\right)^2\right]d\alpha \\
&- \int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)\left[\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{N_s+1}}\right) - \frac{1}{2\pi}\left(\alpha-\hat{\mu}_{N_s+1}\right)^2\right]d\alpha \\
&= \int_{-\infty}^{\infty} f_x(\alpha)\left[\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\log\left(\frac{\hat{\sigma}_{N_s+1}}{\hat{\sigma}_{N_s}}\right)\right] - \underbrace{\left[\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right) - \Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)\right]}_{\text{Odd function}}\cdot\underbrace{\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{N_s}}\right)}_{\text{Constant}}d\alpha \\
&- \frac{1}{2\hat{\sigma}^2_{N_s}}\underbrace{\int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\left(\alpha-\hat{\mu}_{N_s}\right)^2 d\alpha}_{\hat{\sigma}^2_{N_s}\text{ by definition}} + \frac{1}{2\hat{\sigma}^2_{N_s+1}}\underbrace{\int_{-\infty}^{\infty} f_x(\alpha)\Phi\left(\frac{\alpha}{\sqrt{N_s}\cdot\sigma}\right)\left(\alpha-\hat{\mu}_{N_s+1}\right)^2 d\alpha}_{\hat{\sigma}^2_{N_s+1}\text{ by definition}} \\
&= \int_{-\infty}^{\infty} f_x(\alpha)\left[\Phi\left(\frac{\alpha}{\sqrt{N_s-1}\cdot\sigma}\right)\underbrace{\log\left(\frac{\hat{\sigma}_{N_s+1}}{\hat{\sigma}_{N_s}}\right)}_{>0\ \forall N_s}\right]d\alpha > 0
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
\mu &= \int_{-\infty}^{\infty}\alpha\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)\right\}d\alpha \\
&= \int_0^{\infty}\alpha\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)\right\}d\alpha - \int_0^{\infty}\alpha\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)\right\}d\alpha = \underbrace{I_1}_{>0} - \underbrace{I_2}_{>0}
\end{aligned}
\tag{32}
$$

while the estimated one is a sum of the same integrals:

$$
\begin{aligned}
\hat{\mu}_1 &= \int_{-\infty}^{\infty} \alpha f_x(\alpha) U_{-1}(\alpha)\, d\alpha_j = \int_{0}^{\infty} \alpha f_x(\alpha)\, d\alpha \\
&= \int_{0}^{\infty} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)\right\} d\alpha + \int_{0}^{\infty} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)\right\} d\alpha . \\
&= I_1 + I_2 > I_1 - I_2 = \mu
\end{aligned}
\tag{33}
$$

Now we can find the sign of the value $\hat{\mu}_{N_s} - \hat{\mu}_{N_s+1}$. To find this we first need to find the next values:

$$
\left\{\Phi\left(\frac{\alpha - N_s\mu}{\sqrt{N_s}\cdot\sigma} \Leftrightarrow \frac{\alpha - N_s\mu}{\sqrt{N_s}\cdot\sigma} > \frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) > \Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right)\right\} \Rightarrow \alpha > \underbrace{-\mu\sqrt{N_s}\sqrt{N_s-1}}_{<0}
\tag{34a}
$$

$$
\left\{\Phi\left(\frac{\alpha + N_s\mu}{\sqrt{N_s}\cdot\sigma}\right) < \Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) \Leftrightarrow \frac{\alpha + N_s\mu}{\sqrt{N_s}\cdot\sigma} < \frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right\} \Rightarrow \alpha < \underbrace{\mu\sqrt{N_s}\sqrt{N_s-1}}_{>0}
\tag{34b}
$$

then $\Delta_{\mu_{N_s}} = \hat{\mu}_{N_s} - \hat{\mu}_{N_s+1}$:

$$
\begin{aligned}
\Delta_{\mu_{N_s}} &= \underbrace{\int_{-\infty}^{0} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} \underbrace{\left[\Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) - \Phi\left(\frac{\alpha+N_s\mu}{\sqrt{N_s}\cdot\sigma}\right)\right]}_{<0} d\alpha}_{I_1>0} \\
&\quad \underbrace{+\int_{-\infty}^{0} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\} \left[\Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) - \Phi\left(\frac{\alpha-N_s\mu}{\sqrt{N_s}\cdot\sigma}\right)\right] d\alpha}_{I_2} \\
&\quad \underbrace{+\int_{0}^{\infty} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} \left[\Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) - \Phi\left(\frac{\alpha+N_s\mu}{\sqrt{N_s}\cdot\sigma}\right)\right] d\alpha}_{I_3} \\
&\quad \underbrace{+\int_{0}^{\infty} \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\} \underbrace{\left[\Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) - \Phi\left(\frac{\alpha-N_s\mu}{\sqrt{N_s}\cdot\sigma}\right)\right]}_{>0} d\alpha}_{I_4>0} \\
&= I_1 + \underbrace{I_2 + I_3}_{=0} + I_4 > 0
\end{aligned}
\tag{35}
$$

and this means that $\hat{\mu}_{N_s}$ is decreasing and bounded between $\hat{\mu}_1 > \mu$ and $\hat{\mu}_\infty = \mu$.

Then the variance can be found as:

$$
\begin{aligned}
\hat{\sigma}^2_{N_s} &= \int_{-\infty}^{\infty} \alpha^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} \Phi\left(\frac{\alpha+(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) d\alpha \\
&\quad + \int_{-\infty}^{\infty} \alpha^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\} \Phi\left(\frac{\alpha-(N_s-1)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) d\alpha - \hat{\mu}^2_{N_s} \\
&= \underbrace{\int_{-\infty}^{\infty} \alpha^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} d\alpha}_{\sigma^2+\mu^2} - \hat{\mu}^2_{N_s} = \sigma^2 + \mu^2 - \hat{\mu}^2_{N_s} < \sigma^2
\end{aligned}
\tag{36}
$$

As the mean is a decreasing function, the variance is an increasing function of N_s and bounded between $\hat{\sigma}_\infty^2 = \sigma^2$ and $\hat{\sigma}_1^2 < \sigma^2$.

To show that the normalize log-likelihood of two clusters with $N_s = 1$ is higher than the one of a GMM of two Gaussian mixture components, let us first choose a sub-optimal estimator of $f_{1,x,1}(\alpha)$. Assume that the values of the estimator are $\hat{\mu}_1 = \mu$ and $\hat{\sigma}_1^2 = \sigma^2$. As this is not the best estimator, if the normalized log-likelihood with this parameters is higher than the log-likelihood of a GMM of two Gaussian mixture components, then the normalized log-likelihood using ML estimator will be higher for sure. Since $\forall \alpha_j \geq 0$:

$$\begin{aligned} f_x(\alpha) &= \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} + \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\right] \\ &\leq \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\} \end{aligned} \tag{37}$$

and the same can be said with the second cluster for $\forall \alpha \leq 0$, then:

$$\int_0^\infty f_x(\alpha)\log\left(f_x(\alpha)\right)d\alpha < \int_0^\infty f_x(\alpha)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\right)d\alpha \tag{38a}$$

$$\int_{-\infty}^0 f_x(\alpha)\log\left(f_x(\alpha)\right)d\alpha < \int_{-\infty}^0 f_x(\alpha)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right)d\alpha \tag{38b}$$

For the case where $N_s \to \infty$, the estimated *pdf* of the cluster $g_{i,x,\infty}(\alpha)$ is equal to the real *pdf* of the cluster $f_{i,x,\infty}(\alpha)$, and we can write:

$$\begin{aligned} -H\left(f_x(\alpha)\right) &= -H\left(\frac{1}{2}f_{1,x,\infty}(\alpha) + \frac{1}{2}f_{2,x,\infty}(\alpha)\right) \\ &= -H\left(\frac{1}{2}f_{1,x,\infty}(\alpha); f_x(\alpha)\right) - H\left(\frac{1}{2}f_{2,x,\infty}(\alpha); f_x(\alpha)\right) \\ &< -\frac{1}{2}\left[H\left(f_{1,x,\infty}(\alpha); f_{1,x,\infty}(\alpha)\right) + H\left(f_{2,x,\infty}(\alpha); f_{2,x,\infty}(\alpha)\right)\right] \\ &= -\frac{1}{2}\left[H\left(f_{1,x,\infty}(\alpha)\right) + H\left(f_{2,x,\infty}(\alpha)\right)\right] \end{aligned} . \tag{39}$$

The normalized log-likelihood for any value of N_s is then:

$$\begin{aligned} &\int_{-\infty}^{\infty}\frac{1}{2}f_{1,x,N_s}(\alpha)\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{N_s}}\exp\left\{-\frac{1}{2\hat{\sigma}_{N_s}^2}\left(\alpha-\hat{\mu}_{N_s}\right)^2\right\}\right)d\alpha \\ &\quad + \int_{-\infty}^{\infty}\frac{1}{2}f_{2,x,N_s}(\alpha)\log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{N_s}}\exp\left\{-\frac{1}{2\hat{\sigma}_{N_s}^2}\left(\alpha+\hat{\mu}_{N_s}\right)^2\right\}\right)d\alpha = -\frac{1}{2}\log(2\pi) - \log\left(\hat{\sigma}_{N_s}\right) - 1 \end{aligned} . \tag{40}$$

Since $\hat{\sigma}_{N_s}^2$ is an increasing function, the normalized log-likelihood is a decreasing function of N_s. As for $N_s \to \infty$, the normalized log-likelihood of two clusters is higher than the one of a GMM of two Gaussian mixture components, consequently it holds for any N_s.

8. Single Source Data

When the data comes from a single source, (20) cannot be applied. Now there are much more options than the data only belongs to G_1 or G_2. In this case it is possibly that one part of the variables in the segments comes from G_1 and the other part from G_2. It makes the problem combinatorial:

$$f_{x_1,x_2,\ldots,x_{N_s}}\left(\alpha_1,\alpha_2,\ldots,\alpha_{N_s}\left|\sum_{n=1}^{N_s}\alpha_n>0\right.\right)=\frac{f_X(\mathrm{A})P\left(\sum_{n=1}^{N_s}\alpha_n>0\middle|\mathrm{A}\right)}{P\left(\sum_{n=1}^{N_s}\alpha_n>0\right)}=\frac{f_X(\mathrm{A})U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n>0\right)}{\underbrace{P\left(\sum_{n=1}^{N_s}\alpha_n>0\right)}_{\text{Equals }\frac{1}{2}}}$$

$$=2\sum_{n=0}^{N_s}\binom{N_s}{n}f_X\left(\mathrm{A}\cap\left\{\alpha_j\left|\alpha_j\in G_1\vee\left|\left\{\alpha_j\in G_1\right\}\right|=n\right.\right\}\right)P(G_1)^n\,P(G_2)^{N_s-n}\,U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n>0\right). \quad (41)$$

$$=2\,f_{x_{N_s}}\left(\alpha_{N_s}\right)\left(\frac{1}{2}\right)^{N_s-1}U_{-1}\left(\sum_{n=1}^{N_s}\alpha_n>0\right)\cdot\sum_{n=0}^{N_s-1}\binom{N_s-1}{n}\prod_{k=1}^{n}f_{x_k}^1(\alpha_k)\prod_{k=n+1}^{N_s-1}f_{x_k}^2(\alpha_k)$$

In the same way as was done in (21)-(25) the probabilities of the two clusters will be:

$$f_{1,x,N_s}(\alpha)=2\,f_x(\alpha)\left(\frac{1}{2}\right)^{N_s-1}\sum_{n=0}^{N_s-1}\binom{N_s-1}{n}\Phi\left(\frac{\alpha+(N_s-1-2n)\mu}{\sqrt{N_s-1}\cdot\sigma}\right) \quad (42a)$$

$$f_{2,x,N_s}(\alpha)=2\,f_x(\alpha)\left(\frac{1}{2}\right)^{N_s-1}\sum_{n=0}^{N_s-1}\binom{N_s-1}{n}\Phi\left(\frac{-\alpha+(N_s-1-2n)\mu}{\sqrt{N_s-1}\cdot\sigma}\right). \quad (42b)$$

From (42a) and (42b) it can be seen that for $\mu=0$ we will get the same result as in the two-source case. In this case $\Phi(\bullet)$ gets out from the sum and $\sum_{n=0}^{N_s-1}\binom{N_s-1}{n}=2^{N_s-1}$ and the result is as in (24) and (25) for $\mu=0$. When $N_s=1$ the results are the same. This is also intuitively expected. In both cases it will be estimated that two clusters are better.

Intuitively we can say that if $\mu>0$ and the segments are sufficiently long then one GMM will be better. It is very difficult to find this sufficient length but the case of $N_s\to\infty$ can be analyzed. The analysis will be only for C_1, but the same analysis is valid for C_2 as well. Two cases can be observed and in both $\Phi(\bullet)=const$.

$$\forall\varepsilon>0\vee\forall\frac{2n}{N_s}<\frac{1}{2}-\varepsilon\Rightarrow\forall\alpha\bullet\Phi\left(\frac{\alpha-(N_s-1-2n)\mu}{\sqrt{N_s-1}\cdot\sigma}\right)=\Phi(-\infty)=0 \quad (43a)$$

$$\forall\varepsilon>0\vee\forall\frac{2n}{N_s}-1>\frac{1}{2}+\varepsilon\Rightarrow\forall\alpha\bullet\Phi\left(\frac{\alpha-(N_s-1-2n)\mu}{\sqrt{N_s-1}\cdot\sigma}\right)=\Phi(\infty)=1. \quad (43b)$$

The probability of each case is half, so equation (42a) becomes:

$$f_{1,x,\infty}(\alpha) = 2 f_x(\alpha) \cdot \lim_{N_s \to \infty} \left[P\left(\Phi\left(\frac{\alpha - (N_s - 1 - 2n)\mu}{\sqrt{N_s - 1} \cdot \sigma} \right) = 1 \right) \right]$$
$$= f_{2,x,\infty}(\alpha) = 2 f_x(\alpha) \lim_{N_s \to \infty} \left[P\left(\Phi\left(\frac{-\alpha + (N_s - 1 - 2n)\mu}{\sqrt{N_s - 1} \cdot \sigma} \right) = 1 \right) \right] = f_x(\alpha) \tag{44}$$

The Gaussian estimation of such *pdf* gives $\hat{\mu}_\infty = 0$ and $\hat{\sigma}_\infty^2 = \sigma^2 + \mu^2$. In this case the normalized log-likelihood will be the negative cross-entropy between the GMM and one Gaussian, $-H\left(f_x(\alpha); g_{x_1,\infty}(\alpha)\right) < -H\left(f_x(\alpha)\right)$, $\forall \mu > 0$.

$$-H\left(f_x(\alpha); g_{1,x,\infty}(\alpha)\right) = \log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_\infty}\right) - \frac{1}{2} = \log\left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 + \mu^2}}\right) - \frac{1}{2}. \tag{45}$$

In this section we have shown that in the one-source case the decision about one or two clusters is not clearly defined like in the two-source case. While we are sure that for $N_s = 1$, two clusters are better than one and for $N_s \to \infty$ one cluster is better, for any other value of N_s it will depend on μ and σ.

The following experiments also support the fact that a sufficient segment length is required to reach the decision that one GMM is better than two clusters.

9. Experiments

Simulations were performed with artificial data to demonstrate the correctness of the theoretical analysis. Then real speech data from two speakers was used for a speaker clustering application.

9.1 Artificial data

In the simulations the data was derived from two one-dimensional Gaussians with unit variance and with equal prior probability for each Gaussian. The mean varied from zero to four by steps of 0.4 (taking 11 different values). For each value, 11 datasets where created. Each dataset has a constrain about the length of a segment from the same Gaussian. It simulates data coming from two sources. Segment length was $N_s = 2^k \big|_{k=0,\ldots,10}$. For each segment length N_s, two tests were performed: first, assuming data coming from two sources and second assuming data coming from one source (a GMM of two Gaussian mixture components).

Figure 2 shows the results from clustering the data with the segment length constraint. As was proved it can be seen that two clusters are always better than one GMM. When the mean becomes sufficiently large (about three times the variance) segment length does not make any difference due to the fact that the probability of negative data attributed to the Gaussian with positive mean, tends to zero and vice versa.

It can also be seen that for $\mu = 0$ and $N_s = 1$ the normalized log-likelihood of two clusters is higher by about a half, as expected from (18). When $\mu = 4$ the difference is about $\log(2) \approx 0.7$. This is due to the fact that for $\sigma^2 = 1$, $\mu = 4$ can be approximated by $\mu \to \infty$. In this case the normalized log likelihood of the GMM can be approximated as:

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}+\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right] \\
&\quad\cdot\log\left(\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}+\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right]\right)d\alpha_i \\
&\quad\approx \int_{-\infty}^{\infty}\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\log\left(\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha-\mu)^2\right\}\right) \qquad (46)\\
&\quad+\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\log\left(\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\alpha+\mu)^2\right\}\right)d\alpha \\
&\quad=\log\left(\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{1}{2}=\log\left(\frac{1}{2}\right)+\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{1}{2}
\end{aligned}
$$

and the normalized log-likelihood of two clusters is according to (40). If we apply $\hat{\sigma}^2=1$ the difference is exactly $\log(2)\approx 0.7$.

Figure 3 shows the results for the one source case. As was shown in section VII, when the mean equals zero the results of two clusters are always better. The same is true for clustering without constraint (one segment length constraint), see section V. When the mean equals 0.4 segments having long duration $N_s \geq 32$ are the only ones to have a lower log-likelihood than the GMM. When the mean becomes equal or higher to 2.4, the GMM always give a higher log-likelihood than two clusters with duration constraint $N_s \geq 2$. It means that when the distance between the Gaussians is higher than one standard deviation it is very easy to decide whether the data comes from one source or two.

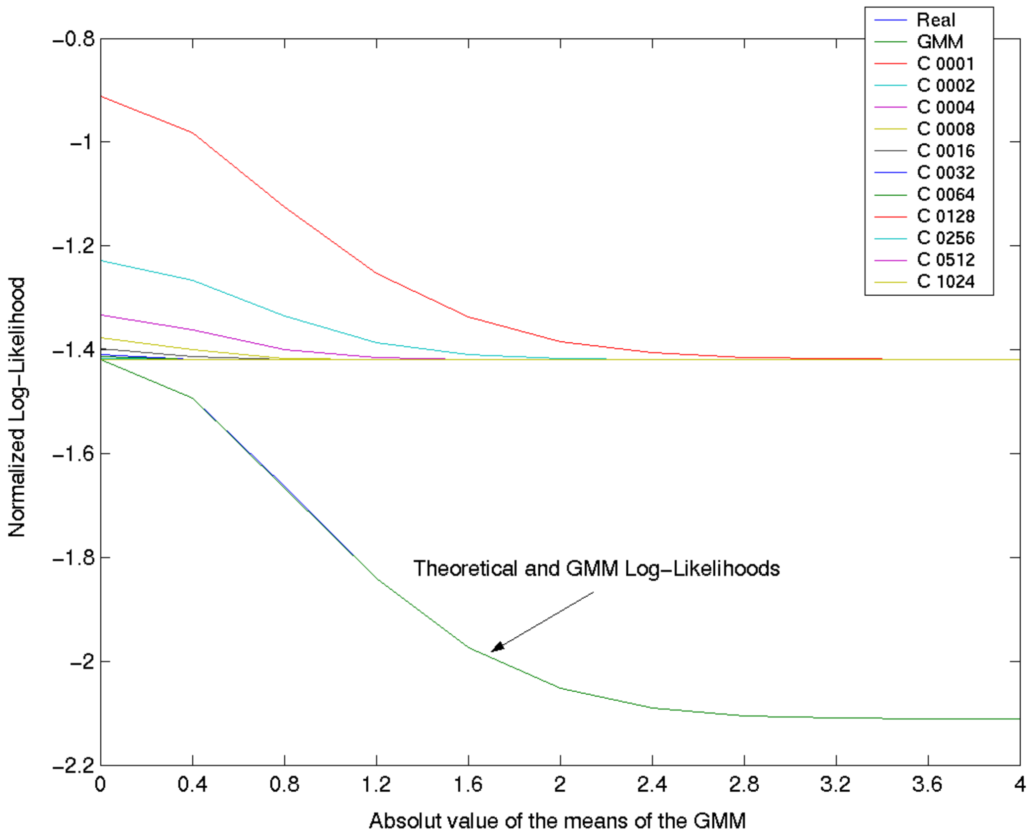

Fig. 2: Clustering data comes from two sources.

Looking at the curve for $N_s = 1024$ it can be seen that it follows exactly the result of (45) for infinite segment length, starting with -1.42 for $\mu = 0$ and finishing at -2.84 for $\mu = 4$.

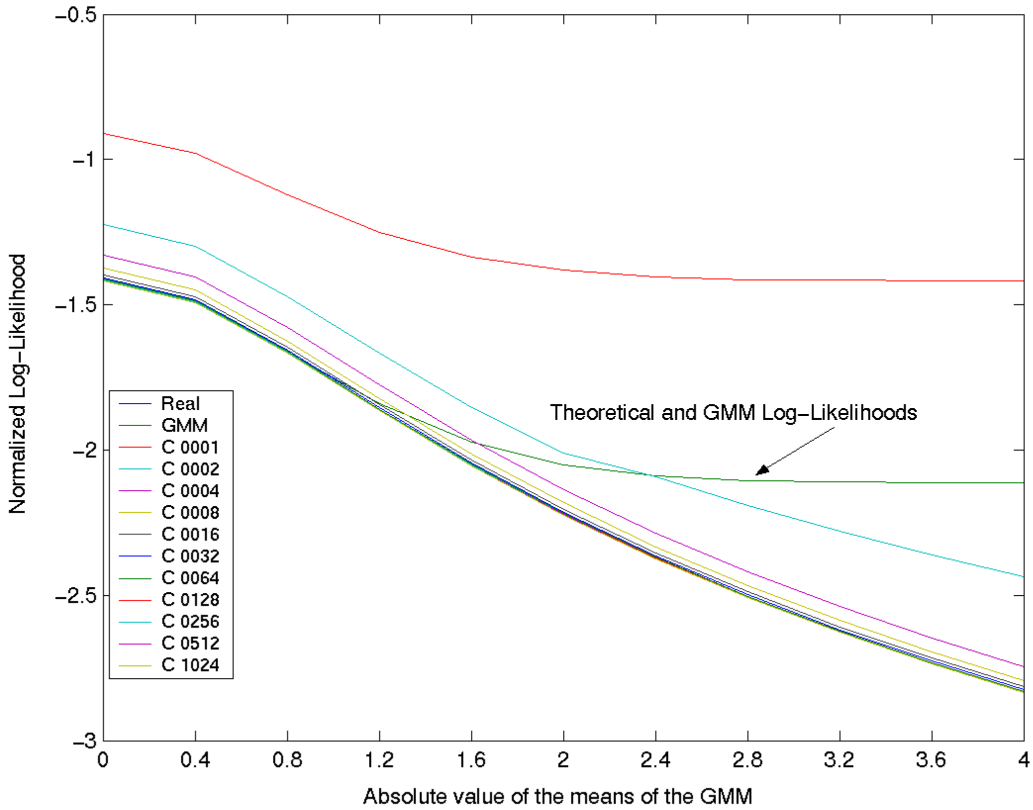

Fig. 3: Clustering data coming from one source.

9.2 Speaker clustering application

Two Australian English speakers were recorded, $110\,\text{sec}$ each speaker. The data was sampled at $48KHz$ and down sampled to $16KHz$. 12^{th} order LPCC features were extracted from $20\,\text{msec}$ windows at $10\,\text{msec}$ frame rate. Three datasets were used. First, the data of the two speakers was combined such that speaker turns occur each two seconds. The second dataset was the data from the first speaker only and the third dataset was from the second speaker only. For each dataset a comparison between two clusters with one Gaussian each and a GMM of two Gaussian mixture components were performed. Segment lengths were $\{1,2,5,10,20,50,100,200\}$. As the goal is to compare the best results and since the EM algorithm is sensitive to initial conditions, each test was conducted 20 times and the system with the highest likelihood was chosen. The results are shown in Fig. 4. They show the normalized log-likelihood of two-cluster system minus the normalized log-likelihood of the GMM. The values above zero indicate that two clusters are better, otherwise one GMM is better. It can be seen that for all three datasets the behavior is similar. The likelihood always decreases as the segment length increase. The decision changes at the segment lengths between 20 and 50. While in the theoretical case two clusters were always better in real applications it is not always the case.

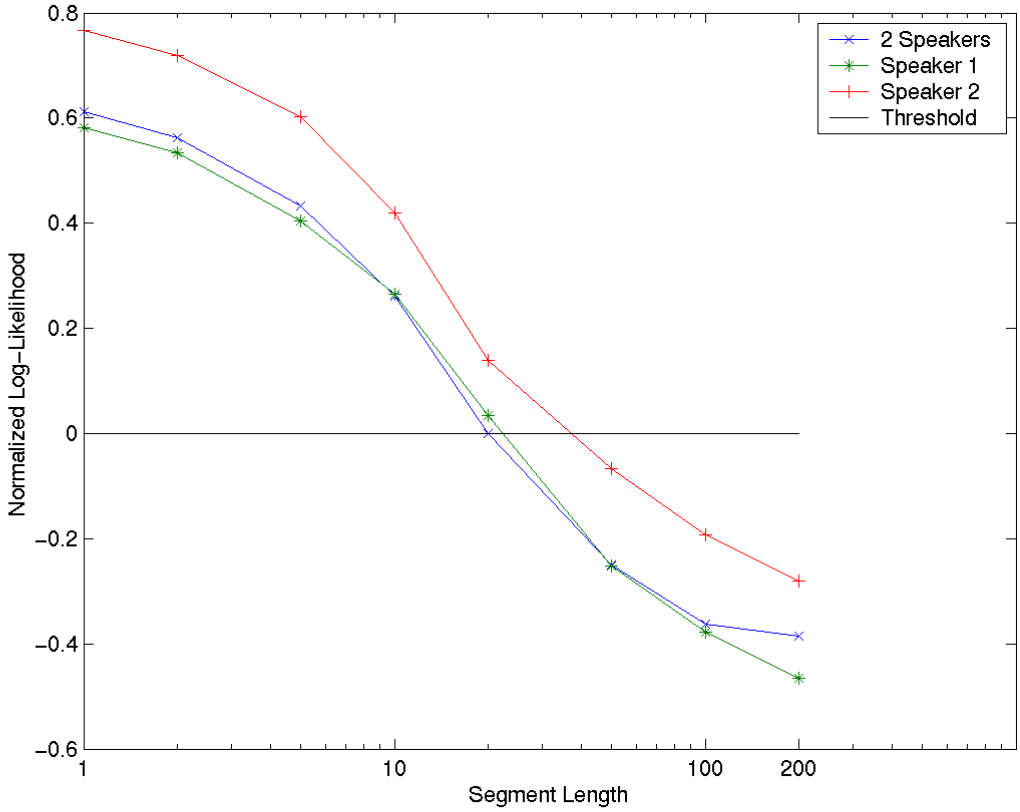

Fig. 4: Clustering with on Gaussian per cluster versus one GMM of two Gaussian mixture components for speaker clustering.

As the true clustering of the first database is known, Fig. 5 shows the clustering performances in terms of percentage of correct classification as a function of the segment length for the first dataset. In this experiment was assumed that we know that there are two speakers. The clustering performance becomes better as the segment length is bigger, i.e., the clustering becomes meaningful. For segments shorter than 50 vectors the clustering is about 50% correctness, which mean random data assignment between two clusters whit respect to two speaker sources. The clusters may have some meaning in some phonetic level sense but not according to speakers. For segments of 200 vectors (2 sec in duration) the results were over 87% correctness. Although in real applications the precise segment boundaries are unknown this simulation show the importance of the knowledge about the maximum available segment length for meaningful clustering performance. If we would not use the knowledge of two speakers, according to the results shown in Fig. 4 one GMM would be chosen. Usually the models that are used for speaker clustering have much more Gaussians for each cluster so more accurate clustering can be performed [2], [10]. When the cluster models are more accurate, it was shown experimentally that two clusters may have higher likelihood than one GMM with the same number of parameters [2].

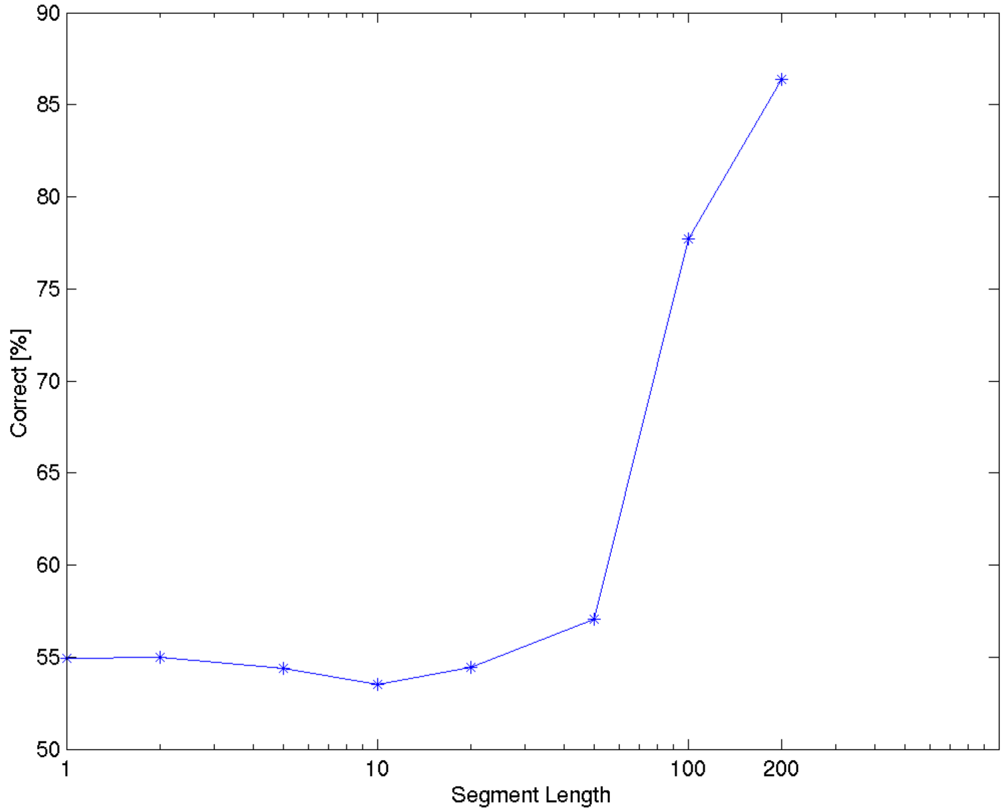

Fig. 5: Clustering results as a function of the segment length.

10. Conclusions

In this work we presented a theoretical analysis of temporal data clustering. Only a very simple case, two Gaussians with symmetrical mean, equal prior per Gaussian and with constant segment length that is known, was presented. It was shown that if the data comes from two different sources, and the length of each data stream is known and constant, two clusters will always be better than one GMM of two Gaussian mixture components. Although the decision is correct the clusters become close to the real sources only for large streams (large segments). For short streams the clusters are not very accurate. When either the data comes from the same Gaussian source with zero mean or segment length equals one the decision is always false, i.e., the result was that, two clusters should be better than one GMM. When the mean is not equal to zero the result depends on the segments length. It was not shown theoretically what should be the precise length of the segment but from simulation results it seems that for mean greater than twice the standard deviation, a short segment, even with length only of two, is already enough to obtain the right decision.

For real life applications such as speaker clustering it was shown that the likelihood decreases as a function of the segment length. As the source *pdf*s are not known the Gaussian approximation per cluster does not lead always to a higher likelihood. As it was shown, segment length has a very high importance for meaningful clustering. The more data we have for each segment the more meaningful clustering can be performed even when the cluster model is not accurate (usually speakers are modeled using several tenths or hundreds of Gaussians).

Acknowledgment

Many thanks to Dr. Samy Bengio for his discussions concerning the experiments that were included and suggestions for improving clarity of presentation.

References

[1] M. Cettolo, "Segmentation, classification and clustering of an Italian broadcast news corpus," *Proc. 6th RIAO Conf.*, April 2000, pp. 372-381.

[2] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved unknown-multiple speaker clustering using HMM," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR02-23, August 2002.

[3] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.

[4] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416-431, 1983.

[5] J. Oliver and R. Baxter, "MML and Bayesianism: similarities and differences: introduction to minimum encoding inference – Part II," Dep. Of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Tech. Rep. TR-206, December 1994.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Inform.Automat. Contr.*, vol. AC-19, no. 6, pp. 716-722, December 1974.

[7] J. Ajmera, I. McCowan, and H. Bourlard, "BIC revisited for speaker change detection," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR02-39, October 2002.

[8] O. A. S. Carpinteiro, "A hierarchical self-organising map model for sequence recognition," *Pattern Analysis and Applications*, vol. 3, pp. 279-287, 2000.

[9] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, USA, 1973.

[10] J. McLaughlin and D. A. Reynolds, "Speaker detection and tracking for telephone transcription," *Proc. ICASSP'02*, vol. 1, May 2002, pp. 129-132.