An E-M Algorithm for Joint Model Estimation

Dr. Paul M. Baggenstoss and T. E. Luginbuhl *
Naval Undersea Warfare Center
Newport RI, 02841
401-832-8240 (TEL)
p.m.baggenstoss@ieee.org (EMAIL)

November 24, 1999

Abstract

This paper describes an EM algorithm for jointly estimating the parameters of multiple models when the data is an unlabeled mixture of data from all models. It maximizes the likelihood function of all the parameters jointly, but does so without incurring the full dimensionality of the problem. The algorithm uses class-specific sufficient statistics.

Keywords: sufficient statistics, Gaussian mixtures, EM algorithm, expectation-maximization, parameter estimation, class-specific, classification.

1 Introduction

In many real-world problems, the data may contain one of a number of possible signal subclasses where the classification of the signals as they arrive at the input are unknown. This is sometimes known as the *unlabeled data* problem (i.e. Redner and Walker [1], Type I problem) and results in a mixture probability density function (PDF). The mixture PDF of the data $\mathbf{X}$ is written

$$p(\mathbf{X};\Lambda) = \sum_{m=1}^{M} p(\mathbf{X}|H_m;\lambda_m)\; p(H_m), \tag{1}$$

where H_m is the hypothesis that the data is from signal subclass m, and $p(H_m)$ is the *a priori* probability of subclass m. The PDF parameters, which we denote by Λ, consist of the subclass parameters as well as the mixing probabilities:

$$\Lambda = \{p(H_1),\ldots,p(H_M),\lambda_1,\ldots,\lambda_M\}\ ;$$

Maximum likelihood estimation involves maximizing $p(\mathbf{X};\Lambda)$ over Λ. To better understand the complexities of this problem, it is useful to examine the simpler *labeled* data problem. Parameter estimation involves a set of M smaller problems

$$\max_{\lambda_m}\ p(\mathbf{X}^m|H_m;\lambda_m) \tag{2}$$

*This work supported by Office of Naval Research and appeared in Proceedings of 1999 ICASSP, Phoenix

where $\mathbf{X}^m$ is labeled data from the class m. This is best accomplished using sufficient statistics. Let $\mathbf{Z}_m = T_m(\mathbf{X})$ be a sufficient statistic for λ_m, then we use

$$\max_{\lambda_m} p(\mathbf{Z}_m^m | H_m; \lambda_m), \tag{3}$$

where $\mathbf{Z}_m^m = T_m(\mathbf{X}^m)$ is obtained from labeled data from subclass m. By themselves, these smaller problems may be quite simple. In contrast, the *unlabeled data* problem is much more difficult because

1. it involves a high dimensional search over the combined parameter set, and
2. there may not be a single sufficient statistic for all signal classes.

In this paper, we solve both problems through the introduction of a "common" or "noise-only" class H_0.

2 Mathematical Results

2.1 Assumptions

We now introduce the "common" class H_0 through the following two assumptions.

Assumption 1 *Let the PDF $p(\mathbf{X}|H_0)$ be a member of each PDF family $p(\mathbf{X}|H_m; \lambda_m)$, $m = 1, \ldots, M$. More precisely, for each m, there exists a parameter value λ_m^0 such that*

$$\lim_{\lambda_m \to \lambda_m^0} p(\mathbf{X}|H_m; \lambda_m) = p(\mathbf{X}|H_0)$$

In most applications, we may think of H_0 as the *noise-only* conditon, i.e. where the signal amplitudes are zero and only noise remains. Another interpretation is that H_0 is the *normal* condition, such as in automatic fault localization.

Assumption 2 *Suppose for each Class PDF, $p(\mathbf{X}|H_m)$, there exists a sufficient statistic for the parameter λ_m. Denote this sufficient statistic by $\mathbf{Z}_m \triangleq T_m(\mathbf{X})$.*

A result from decision theory is that the likelihood ratio is unchanged when written as a function of a sufficient statistic (i.e. Kendall and Stuart [2], section 22.14), thus

$$\frac{p(\mathbf{X}|H_m; \lambda_m)}{p(\mathbf{X}|H_0)} = \frac{p(\mathbf{X}|H_m; \lambda_m)}{p(\mathbf{X}|H_m; \lambda_m^0)} = \frac{p(\mathbf{Z}_m|H_m; \lambda_m)}{p(\mathbf{Z}_m|H_m; \lambda_m^0)} = \frac{p(\mathbf{Z}_m|H_m; \lambda_m)}{p(\mathbf{Z}_m|H_0)}.$$

Therefore,

$$\frac{p(\mathbf{X}; \Lambda)}{p(\mathbf{X}|H_0)} = \sum_{m=1}^{M} \frac{p(\mathbf{Z}_m|H_m; \lambda_m)}{p(\mathbf{Z}_m|H_0)} p(H_m). \tag{4}$$

Define $L(\mathbf{X}; \Lambda)$ as

$$\begin{aligned} L(\mathbf{X}; \Lambda) &\triangleq \frac{p(\mathbf{X}; \Lambda)}{p(\mathbf{X}|H_0)} = \prod_{k=1}^{K} \frac{p(\mathbf{X}_k; \Lambda)}{p(\mathbf{X}_k|H_0)} \\ &= \prod_{k=1}^{K} \sum_{m=1}^{M} \frac{p(\mathbf{Z}_{m,k}|H_m; \lambda_m)}{p(\mathbf{Z}_{m,k}|H_0)} p(H_m), \end{aligned} \tag{5}$$

where $\mathbf{Z}_{m,k} \triangleq T_m(\mathbf{X}_k)$, $\Lambda = \left\{p(H_m); \{\lambda_m\}_{m=1}^M\right\}$. Now $L(\mathbf{X}; \Lambda)$ may be used for the likelihood function for estimation of Λ because the denominator is independent of Λ.

2.2 E-M Algorithm

The objective is to estimate Λ using the E-M algorithm where $L(\mathbf{X};\Lambda)$ is used for the likelihood function. The key to the E-M algorithm is the auxiliary function which, if increased or maximized is guaranteed to result in an increase in the likelihood function. It is shown in section 5.1 that the auxiliary function is

$$Q(\Lambda;\Lambda') = \sum_{k=1}^{K}\sum_{m=1}^{M} \left[\log p(H_m) + \log p(\mathbf{Z}_{m,k}|H_m;\lambda_m) - \log p(\mathbf{Z}_{m,k}|H_0)\right]\ \gamma_{mk}(\Lambda'), \tag{6}$$

where

$$\gamma_{mk}\left(\Lambda\right) = \frac{\frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0)}p(H_m)}{\sum_{l=1}^{M}\frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)}{p(\mathbf{Z}_{l,k}|H_0)}p(H_l).} \tag{7}$$

The physical interpretation of $\gamma_{mk}\left(\Lambda\right)$ is the probability that sample $\mathbf{X}_k$ is from model m given the model parameters Λ and the data $\mathbf{X}_k$. From hereafter, we simplify the notation to $\gamma_{mk} \triangleq \gamma_{mk}(\Lambda')$. Computing γ_{mk} is part of the E-step. The M-step consists of maximizing (6) over Λ. The estimates of $p(H_m)$ are

$$p(H_m) = \frac{1}{K}\sum_{k=1}^{K}\gamma_{mk}. \tag{8}$$

Notice that in (6), the maximization (or increase) of $Q(\Lambda;\Lambda')$ with respect to Λ requires the functions

$$Q_m(\lambda_m;\Lambda') = \sum_{k=1}^{K}\ \log p(\mathbf{Z}_{m,k}|H_m;\lambda_m)\ \gamma_{mk} \tag{9}$$

to be *independently* maximized (or increased) over λ_m, for each m. This is in contrast to (5), which contains mixed terms and requires joint maximization. Equation (9) is, in effect, a probabilistic weighting of each data sample, a minor modification of individual maximum likelihood estimators represented by (3). If an existing algorithm exists for maximization (or increase) of $\sum_{k=1}^{K}\log p(\mathbf{Z}_{m,k}|H_m;\lambda_m)$, then this algorithm may be used with a minor modification. One way to do it, albeit impractical, would be to scale γ_{mk} by a large constant C, round to an integer $n_{km} = \lfloor C\ \gamma_{mk}\rfloor$, then form a larger data set created by replicating each data sample $\mathbf{Z}_{m,k}$ by the corresponding integer:

$$\mathbf{Z}' \triangleq \left\{\{\mathbf{Z}_{m,1}\cdots\mathbf{Z}_{m,1}\},\{\mathbf{Z}_{m,2}\cdots\mathbf{Z}_{m,2}\},\ldots,\{\mathbf{Z}_{m,K}\cdots\mathbf{Z}_{m,K}\}\right\},$$

then maximize the PDF of $\mathbf{Z}'$ under the assumption of independence of the samples, i.e.,

$$\sum_{k=1}^{K'}\log p(\mathbf{Z}'_k|H_m;\lambda_m)$$

where $K' = \sum_{k=1}^{K} n_{km}$. A more practical, yet suboptimal method would be to threshold γ_{mk} and include only those data samples in $\mathbf{X}$ that exceed the threshold. Of course, the best approach would be to integrate the weighting directly in the algorithm.

2.2.1 Summary of the E-M algorithm

Prior to beginning the algorithm, it is assumed that initial values of $P(H_m)$ and γ_{mk} are available (uniform constants can be used for initialization). The algorithm proceeds as follows:

1. E-Step: Use (7) to update γ_{mk} for $1 \leq k \leq K$, $1 \leq m \leq M$.
2. M-step.
   (a) Maximize (or increase) (9) over λ_m for $1 \leq m \leq M$.
   (b) Use (8) to update $P(H_m)$ for $1 \leq m \leq M$.
3. Repeat steps 1 and 2 until convergence.

2.3 EM algorithm for a Non-Homogenious Gaussian Mixture

Suppose that the sufficient statistics $\mathbf{Z}_m$ are known but parametric forms for the PDFs are not known. Now, if $p(\mathbf{Z}_m|H_m)$ are continuous, they may be approximated to arbitrary accuracy by any kernel-based estimator [3], such as the method of Gaussian Mixtures [4]. We now show how the E-M algorithm is changed when the individual class PDFs where $p(\mathbf{Z}_m|H_m)$ are Gaussian Mixtures. Consider a Gaussian mixture for $\mathbf{Z}_m \in \mathcal{R}^{N_m}$ under class m

$$p(\mathbf{Z}_m|H_m) = \sum_{i=1}^{L_m} \alpha_{mi}\, \mathcal{N}(\mathbf{Z}_m, \boldsymbol{\mu}_{mi}, \boldsymbol{\Sigma}_{mi}) \tag{10}$$

where

$$\mathcal{N}(\mathbf{Z}_m, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-N_m/2}\, |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\,(\mathbf{Z}_m - \boldsymbol{\mu})'\ \boldsymbol{\Sigma}^{-1}\ (\mathbf{Z}_m - \boldsymbol{\mu})\right\},$$

where N_m is the dimension of $\mathbf{Z}_m$. By substituting (10) into (4), we have an expression for $p(\mathbf{X}|\Lambda)$ that is a mixture of mixtures, with each sub-mixture a function of a different sufficient statistic. We call this a *non-homogeneous* Gaussian mixture. Deriving the E-M algorithm requires regarding not only the class-assignments as missing information, but the the Gaussian mixture mode assignments as well. This is analogous to the incorporation of Gaussian mixtures in a Baum-Welch algorithm for HMMs [5] where the Markov state assignments take the part of the class assignments. The auxiliary function incorporating the mixture mode assignments is derived in section 5.2 and is given below.

$$\begin{aligned} Q(\Lambda;\Lambda') \;=\; & \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{L_m} [\log p(H_m) + \log \alpha_{mi} + \log \mathcal{N}(\mathbf{Z}_{m,k}, \boldsymbol{\mu}_{mi}, \boldsymbol{\Sigma}_{mi}) - \log p(\mathbf{Z}_{m,k}|H_0)] \\ & \cdot\, \xi_{mik}(\Lambda') \end{aligned} \tag{11}$$

where

$$\xi_{mik}\,(\Lambda) \triangleq \frac{\alpha_{mi}\ \dfrac{\mathcal{N}(\mathbf{Z}_{m,k}, \boldsymbol{\mu}_{m,i}, \boldsymbol{\Sigma}_{m,i})}{p(\mathbf{Z}_{m,k}|H_0)}\ p(H_m)}{\displaystyle\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)}{p(\mathbf{Z}_{l,k}|H_0)} p(H_l)} \tag{12}$$

The physical interpretation of $\xi_{mik}\,(\Lambda)$ is the probability that sample $\mathbf{X}_k$ is from model m and mixture component i given the model parameters Λ and the data $\mathbf{X}_k$. From hereafter, we simplify the notation

to $\xi_{mik} \triangleq \xi_{mik}(\Lambda')$. Computing ξ_{mik} is part of the E-step. In the M-step, we first update the class probabilities

$$p(H_m) = \frac{1}{K} \sum_{k=1}^{K} \gamma_{mk}, \tag{13}$$

where as before,

$$\gamma_{mk} = \frac{\frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0)} p(H_m)}{\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)}{p(\mathbf{Z}_{l,k}|H_0)} p(H_l)} = \sum_{i=1}^{L_m} \xi_{mik}.$$

Next, update mode weights

$$\alpha_{mi} = \frac{\sum_{k=1}^{K} \xi_{mik}.}{\sum_{k=1}^{K} \gamma_{mk}}. \tag{14}$$

Lastly, we update $\boldsymbol{\mu}_{mi}, \boldsymbol{\Sigma}_{mi}$. Notice that (11) may be maximized by maximizing

$$Q_{mi}(\Lambda; \Lambda') \triangleq \sum_{k=1}^{K} \log \mathcal{N}(\mathbf{Z}_{m,k}, \boldsymbol{\mu}_{mi}, \boldsymbol{\Sigma}_{mi})\, \xi_{mik}(\Lambda') \tag{15}$$

over $\boldsymbol{\mu}_{mi}, \boldsymbol{\Sigma}_{mi}$ independently for pair (m, i). This is accomplished by

$$\boldsymbol{\mu}_{mi} = \frac{\sum_{k=1}^{K} \xi_{mik}\, \mathbf{Z}_{m,k}}{\sum_{k=1}^{K} \xi_{mik}} \tag{16}$$

and

$$\boldsymbol{\Sigma}_{mi} = \frac{\sum_{k=1}^{K} \xi_{mik}\, (\mathbf{Z}_{m,k} - \boldsymbol{\mu}_{mi})\, (\mathbf{Z}_{m,k} - \boldsymbol{\mu}_{mi})'}{\sum_{k=1}^{K} \xi_{mik}}, \tag{17}$$

where ξ_{mik} is a shorthand notation for $\xi_{mik}(\Lambda')$.

Because of possible numerical issues, it may be necessary to add a constant to the diagonal elements of $\boldsymbol{\Sigma}_{mi}$ at each iteration. These constants may be regarded as prior knowledge in the form of independent measurement error variances, and should be chosen carefully.

2.3.1 Summary of E-M algorithm for non-homogenious mixture

1. E-Step: Use (12) to update ξ_{mik} for all k, i, m.
2. M-step.
   (a) Use (13) to update $P(H_m)$ for all m.
   (b) Use (14) to update α_{mi} for all m, i.

(c) Use (16) to update $\boldsymbol{\mu}_{mi}$ for all m, i.

(d) Use (17) to update $\boldsymbol{\Sigma}_{mi}$ for all m, i.

3. Repeat steps 1 and 2 until convergence.

3 Simulation Results (7-class example)

The example problem to be discussed here is a subset of the 9-class synthetic problem discused in a previous paper [6]. We consider the following 7 data classes denoted $H_1, \ldots, H_7$.

- Class H_0: Noise only
- Class H_1: Long Sinewave
- Class H_2: Medium Sinewave
- Class H_3: Short Sinewave
- Class H_4: Long Gaussian Signal
- Class H_5: Short Gaussian Signal
- Class H_6: Short Impulse Signal
- Class H_7: Long Impulse Signal

Details of the signals and how they are generated may be found in the reference. For convenience, we have tabulated the sufficient statistics as well as the distributions under H_0 in tables 1, 2.

3.1 Data Set

To simulate a data set from a mixture of the seven data classes, an equal share of 1024 samples from each data class were created. Each input data sample was a time-series of 256 data points. The true class index of each sample was not used by the algorithm, but was remembered for use later in validation. From each time series, features $\mathbf{Z}_1$ through $\mathbf{Z}_7$ were calculated.

3.2 Algorithm Initialization

Initial values of $\boldsymbol{\mu}_{mi}$ were set equal to randomly chosen (unlabeled) input data samples. Initial values of $\boldsymbol{\Sigma}_{mi}$ were set equal to the sample covariance of the entire data set. Initial values of α_{mi} were all equal, as were the initial values of $P(H_m)$. The number of Gaussian mixture components per data class was 10.

3.3 Algorithm Performance

Algorithm performance may be measured by monitoring the likelihood function (5). Notice also that γ_{mk} in (9) acts as a probabilistic data weighting for each sample. It is in effect an estimate of the probability that data sample k is from class m. If the algorithm is working properly and $p(Z_m|H_m)$ are converging to the true PDFs, γ_{mk} should act as data classifiers. Thus, for a given sample k, maximizing over m will produce a guess as to the class index of the sample. But this will not work in general. Specifically, if two data classes have the same or equivalent sufficient statistics, the algorithm has no way to make the separation between the classes except perhaps as different Gaussian Mixture

$\mathbf{Z}_1 = \left[\sum_{i=1}^{N} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_2 = \left[\sum_{i=1}^{N/2} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/2} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_3 = \left[\sum_{i=1}^{N/4} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/4} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_4 = \sum_{i=1}^{N} x_i^2$
$\mathbf{Z}_5 = \sum_{i=1}^{N/2} x_i^2$
$\mathbf{Z}_6 = \log(x_1^2)$
$\mathbf{Z}_7 = \log(x_1^2 + x_2^2)$

Table 1: Class-Specific Statistics

components within a fixed m. This shortcoming of the algorithm is expected since it is designed only to estimate the PDF of the overall non-homogenious mixture (Type I problems). The separation of the subclasses is irrelevant to its operation. However, if all the sufficient statistics are different, it has a chance of accomplishing this goal (Type II problems). In this example, we monitor the algorithm performance as a Type II problem by determining the probability of correct classification (P_{cc}). P_{cc} was determined by determining what percentage of the data was classified correctly (i.e. when $\arg\max_m \gamma_{mk}$ was equal to the true class index).

The algorithm was allowed to iterate 380 times. At each iteration, the total likelihood as well as the probability of correct classification (P_{cc}) were determined. These quantities are plotted in Figure 1. Notice that the likelihood was monotonic increasing as expected.

The algorithm was re-run using labeled data, i.e. only data from class m was used to train $p(\mathbf{Z}_m|H_m)$. The result is plotted in Figure 2. As would be expected, P_{cc} is higher, but the likelihood is lower (not discernible on the graph). Using labeled data does not necessarily maximize (5).

The PDF estimate of $p(\mathbf{Z}_4|H_4)$ from unlabeled data is plotted in Figure 3. Superimposed on the graph are the histograms of $\mathbf{Z}_4$ for all data classes and for just class 4. The fact that the PDF estimate matches the histogram for class 4 illustrates the fact that PDF estimates may be obtained from unlabeled data.

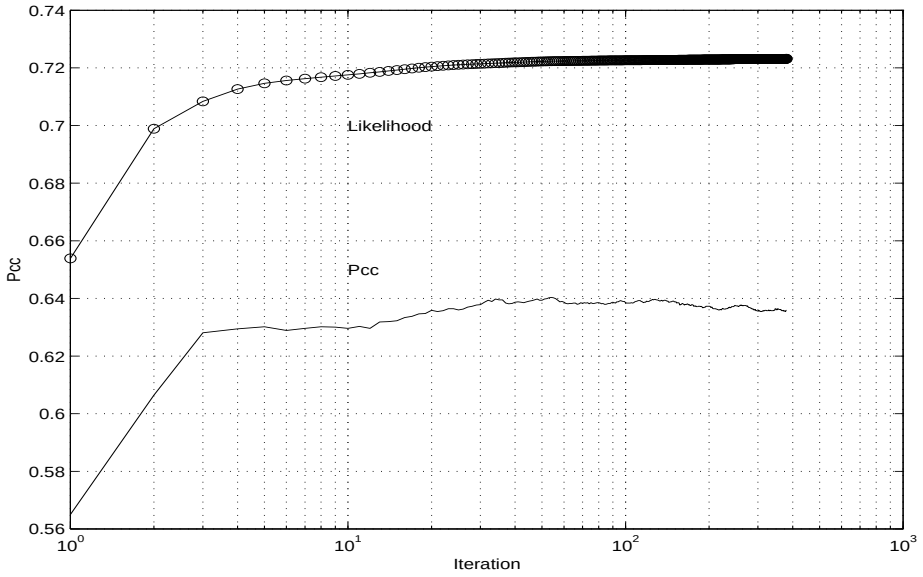

Figure 1: Algorithm Convergence Properties. Scaled log likelihood values superimposed on a plot of P_{cc}.

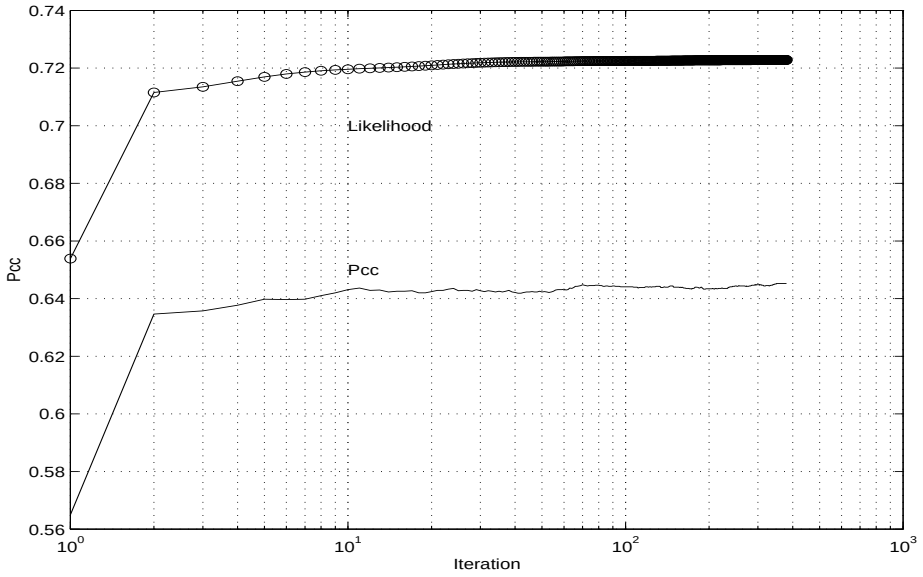

Figure 2: Repeat of Figure 1 with labeled data used to train $p(\mathbf{Z}_m|H_m)$.

$p(\mathbf{Z}_1\|H_0) = \left(\frac{e^{\mathbf{Z}_1}}{N\sigma^2}\right) \exp\left\{-\frac{e^{\mathbf{Z}_1}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_2\|H_0) = \left(\frac{2e^{\mathbf{Z}_2}}{N\sigma^2}\right) \exp\left\{-\frac{2e^{\mathbf{Z}_2}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_3\|H_0) = \left(\frac{4e^{\mathbf{Z}_3}}{N\sigma^2}\right) \exp\left\{-\frac{4e^{\mathbf{Z}_3}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_4\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{2})\ 2^{-\frac{N}{2}}(\frac{\mathbf{Z}_4}{\sigma^2})^{\frac{N}{2}-1}\ \exp\left\{-\frac{\mathbf{Z}_4}{2\sigma^2}\right\}$
$p(\mathbf{Z}_5\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{4})\ 2^{-\frac{N}{4}}(\frac{\mathbf{Z}_5}{\sigma^2})^{\frac{N}{2}-1}\ \exp\left\{-\frac{\mathbf{Z}_5}{2\sigma^2}\right\}$
$p(\mathbf{Z}_6\|H_0) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{e^{\mathbf{Z}_6}}{2\sigma^2}\right\} e^{\mathbf{Z}_6/2}$
$p(\mathbf{Z}_7\|H_0) = (4\pi\sigma^2)^{-1/2} \exp\left\{-\frac{e^{\mathbf{Z}_7}}{4\sigma^2}\right\} e^{\mathbf{Z}_7/2}$

Table 2: Distributions of Class-Specific Statistics

4 Conclusions

An E-M algorithm has been derived for the case when the input data is a mixture density of several data classes, with each data class dependent on a different set of parameters. By taking advantage of different sufficient statistics for each data class, it is possible to jointly estimate the parameters efficiently and with low dimensionality.

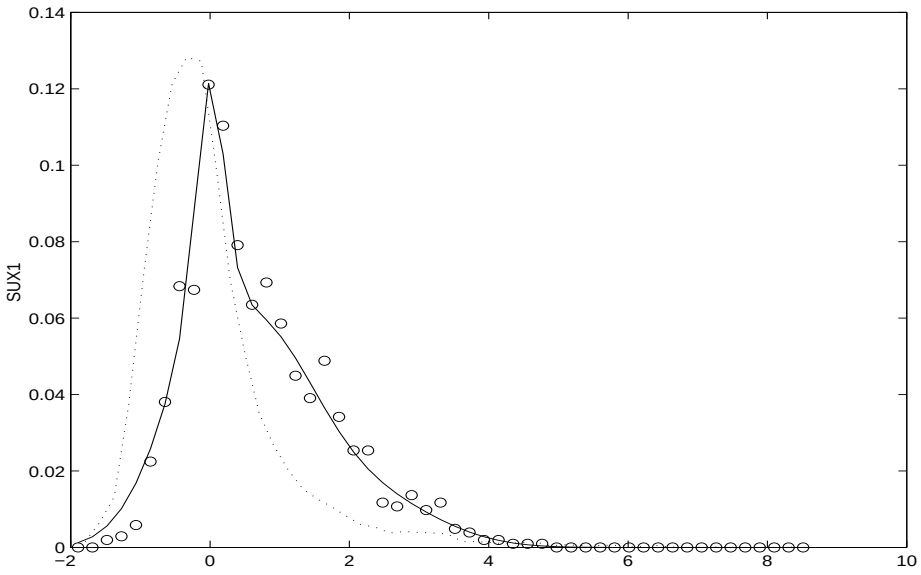

Figure 3: PDF estimation results for feature $\mathbf{Z}_4$ using unlabeled data. Graph includes histogram of data from all 7 classes (dotted), histogram of data from class 4 (circles), PDF estimate for $p(\mathbf{Z}_4|H_4)$ (solid). Data is plotted on a normalized axis. The designation "SUX1" is the name used to identify feature $\mathbf{Z}_4$.

5 Derivation of E-M Algorithm for Joint Model Estimation

5.1 Arbitrary PDFs

This derivation is similar to that found in Redner and Walker [1]. The missing information in this problem are the class assignments

$$\mathbf{I} = \{i_k\}_{k=1}^{K}$$

of each $\mathbf{X}_k$ to a PDF (class). Each i_k is an integer between 1 and M. The complete information is then

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \bigcup \mathbf{I} \\ &= \{\mathbf{X}_k; i_k\}_{k=1}^{K} \\ &= \{\mathbf{Y}_k\}_{k=1}^{K} \end{aligned}$$

The likelihood function of the complete data is defined as

$$\begin{aligned} L(\mathbf{Y};\Lambda) &\triangleq \frac{p(\mathbf{Y};\Lambda)}{p(\mathbf{X}|H_0)} \\ &= \prod_{k=1}^{K} \frac{p(\mathbf{X}_k|i_k;\Lambda)\; p(i_k;\Lambda)}{p(\mathbf{X}_k|H_0)} \\ &= \prod_{k=1}^{K} \frac{p(\mathbf{Z}_{i_k,k}|H_{i_k};\lambda_m)}{p(\mathbf{Z}_{i_k,k}|H_0)}\; p(H_{i_k}) \end{aligned}$$

where $\mathbf{Z}_{m,k} \triangleq T_m(\mathbf{X}_k)$. Hence, the PDF of the missing data conditioned on the observed data is given by

$$\begin{aligned} p(\mathbf{I}|\mathbf{X};\Lambda) &= \frac{p(\mathbf{Y};\Lambda)}{p(\mathbf{X};\Lambda)} = \frac{L(\mathbf{Y};\Lambda)}{L(\mathbf{X};\Lambda)} \\ &= \prod_{k=1}^{K} \frac{\frac{p(\mathbf{Z}_{i_k,k}|H_{i_k};\lambda_{i_k})}{p(\mathbf{Z}_{i_k,k}|H_0)}\, p(H_{i_k})}{\sum_{m=1}^{M} \frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0)} p(H_m)} \\ &= \prod_{k=1}^{K} \gamma_{i_k k}\left(\Lambda\right) \end{aligned}$$

where

$$\gamma_{mk}\left(\Lambda\right) \triangleq \frac{\frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0)} p(H_m)}{\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)}{p(\mathbf{Z}_{l,k}|H_0)} p(H_l)}$$

Now the auxiliary function can be defined. The auxiliary function of the E.M. algorithm $Q(\Lambda;\Lambda')$ is defined as

$$\begin{aligned} Q(\Lambda;\Lambda') &\triangleq \mathbf{E}_{\{\mathbf{I}|\mathbf{X};\Lambda'\}} \log L(\mathbf{Y};\Lambda) \\ &= \sum_{\mathbf{I}} \log L(\mathbf{Y};\Lambda)\; p(\mathbf{I}|\mathbf{X};\Lambda') \\ &= \sum_{\mathbf{I}} \sum_{k=1}^{K} \left[\log p(H_{i_k}) + \log p(\mathbf{Z}_{i_k,k}|H_{i_k};\lambda_{i_k}) - \log p(\mathbf{Z}_{i_k,k}|H_0)\right] \prod_{j=1}^{K} \gamma_{i_j j}(\Lambda') \end{aligned} \tag{18}$$

where

$$\sum_{\mathbf{I}} \triangleq \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_K=1}^{M}$$

First, it is necessary to prove that increasing $Q(\Lambda;\Lambda')$ also increases $L(\mathbf{X};\Lambda)$ with respect to Λ. The following proof follows the form of Streit [7] of Baum's inequality [8].

Proof: Suppose $Q(\Lambda;\Lambda') \geq Q(\Lambda';\Lambda')$, then

$$\begin{aligned}
0 &\leq Q(\Lambda;\Lambda') - Q(\Lambda';\Lambda') \\
&= \sum_{\mathbf{I}} \{\log L(\mathbf{Y};\Lambda)\; p(\mathbf{I}|\mathbf{X};\Lambda')\} - \sum_{\mathbf{I}} \{\log L(\mathbf{Y};\Lambda')\; p(\mathbf{I}|\mathbf{X};\Lambda')\} \\
&= \sum_{\mathbf{I}} \left\{\log \tfrac{L(\mathbf{Y};\Lambda)}{L(\mathbf{Y};\Lambda')}\; p(\mathbf{I}|\mathbf{X};\Lambda')\right\} \\
&\leq \sum_{\mathbf{I}} \left\{\left(\tfrac{L(\mathbf{Y};\Lambda)}{L(\mathbf{Y};\Lambda')} - 1\right) \tfrac{L(\mathbf{Y};\Lambda')}{L(\mathbf{X};\Lambda')}\right\} \\
&= \tfrac{1}{L(\mathbf{X};\Lambda')} \sum_{\mathbf{I}} \{L(\mathbf{Y};\Lambda) - L(\mathbf{Y};\Lambda')\} \\
&= \tfrac{L(\mathbf{X};\Lambda)}{L(\mathbf{X};\Lambda')} - 1 \\
&\longrightarrow L(\mathbf{X};\Lambda) \geq L(\mathbf{X};\Lambda')
\end{aligned}$$

where we use the fact that $\log(\mathbf{X}) \leq \mathbf{X} - 1$ and that

$$\sum_{\mathbf{I}} L(\mathbf{Y};\Lambda) = \frac{\sum_{\mathbf{I}} p(\mathbf{Y};\Lambda)}{p(\mathbf{X}|H_0)} = \frac{p(\mathbf{X};\Lambda)}{p(\mathbf{X}|H_0)} = L(\mathbf{X};\Lambda),$$

□

Having proved this, $Q(\Lambda;\Lambda')$ now can be obtained. First note that due to the independence of the samples,

$$\sum_{i_k} p(i_k|\mathbf{X};\Lambda) = \sum_{i_k} p(i_k|\mathbf{X}_k;\Lambda) = 1.$$

As a result,

$$\sum_{\mathbf{I}|i_k} p(\mathbf{I}|\mathbf{X};\Lambda) = \frac{\frac{p(\mathbf{Z}_{i_k,k}|H_{i_k};\lambda_{i_k})\; p(H_{i_k})}{p(\mathbf{Z}_{i_k,k}|H_0)}}{\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)\; p(H_l)}{p(\mathbf{Z}_{l,k}|H_0)}} = \gamma_{i_k k}(\Lambda), \tag{19}$$

where $\mathbf{I}|i_k$ is the summation over all indices except i_k,

$$\sum_{\mathbf{I}|i_k} \triangleq \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_{k-1}=1}^{M} \sum_{i_{k+1}=1}^{M} \cdots \sum_{i_K=1}^{M}.$$

Next, note that equation (18) can be rewritten as

$$\begin{aligned} Q(\Lambda;\Lambda') &= \sum_{\mathbf{I}}\sum_{k=1}^{K} A(i_k)\ B(\mathbf{I}) \\ &= \sum_{k=1}^{K}\sum_{\mathbf{I}} A(i_k)\ B(\mathbf{I}) \\ &= \sum_{k=1}^{K}\sum_{i_k}\sum_{\mathbf{I}|i_k} A(i_k)\ B(\mathbf{I}) \\ &= \sum_{k=1}^{K}\sum_{i_k} A(i_k) \sum_{\mathbf{I}|i_k} B(\mathbf{I}) \end{aligned}$$

Substituting (19) for the term $\sum_{\mathbf{I}|i_k} B(\mathbf{I})$, yields (6). This completes the derivation of the E-step of the E.M. algorithm. The estimate of $p(H_m)$ is obtained by maximizing $Q(\Lambda;\Lambda')$ with respect to $p(H_m)$. It is straight forward to show that the estimate of $p(H_m)$ is equal to (8).

5.2 Gaussian Mixtures

We now derive the E-M algorithm for estimating the parameters of the PDFs of the sufficient statistics when the PDFs are approximated by a mixture of the form (10). The missing information in this problem is the assignments

$$\mathbf{I} = \{i_k\}_{k=1}^{K}$$

of each $\mathbf{X}_k$ to a PDF (class) and the assignments

$$\mathbf{J} = \{j_k\}_{j=1}^{K}$$

of each $\mathbf{Z}_{m,k}$ to a mixture component (mode). Each i_k is an integer between 1 and M and each j_k is an integer between 1 and L_{i_k}. The complete information is then

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \bigcup \mathbf{I} \bigcup \mathbf{J} \\ &= \{\mathbf{X}_k; i_k; j_k\}_{k=1}^{K} \\ &= \{\mathbf{Y}_k\}_{k=1}^{K} \end{aligned}$$

The likelihood function of the complete data is defined by

$$\begin{aligned} L(\mathbf{Y};\Lambda) &\triangleq \frac{p(\mathbf{Y};\Lambda)}{p(\mathbf{X}|H_0)} \\ &= \prod_{k=1}^{K} \frac{\mathcal{N}(\mathbf{Z}_{i_k,k}, \boldsymbol{\mu}_{i_k j_k}, \boldsymbol{\Sigma}_{i_k j_k})}{p(\mathbf{Z}_{i_k,k}|H_0)}\ p(H_{i_k})\ \alpha_{i_k j_k} \end{aligned}$$

where $\mathbf{Z}_{m,k} \triangleq T_m(\mathbf{X}_k)$. Hence, the likelihood function of the missing data conditioned on the observed data is given by

$$\begin{aligned} p(\mathbf{I}|\mathbf{X};\Lambda) &= \frac{p(\mathbf{Y};\Lambda)}{p(\mathbf{X};\Lambda)} = \frac{L(\mathbf{Y};\Lambda)}{L(\mathbf{X};\Lambda)} \\ &= \prod_{k=1}^{K} \frac{\dfrac{\mathcal{N}(\mathbf{Z}_{i_k,k}, \boldsymbol{\mu}_{i_k j_k}, \boldsymbol{\Sigma}_{i_k j_k})}{p(\mathbf{Z}_{i_k,k}|H_0)}\, p(H_{i_k})\, \alpha_{i_k j_k}}{\displaystyle\sum_{m=1}^{M} \frac{p(\mathbf{Z}_{m,k}|H_m)}{p(\mathbf{Z}_{m,k}|H_0)} p(H_m)} \\ &= \prod_{k=1}^{K} \xi_{i_k j_k k}(\Lambda) \end{aligned}$$

where $\xi_{mik}(\Lambda)$ is defined in (12). Now the auxiliary function can be defined. The auxiliary function of the E.M. algorithm $Q(\Lambda;\Lambda')$ is defined as

$$\begin{aligned} Q(\Lambda;\Lambda') &\triangleq \mathbf{E}_{\{\mathbf{I},\mathbf{J}|\mathbf{X};\Lambda'\}} \log L(\mathbf{Y};\Lambda) \\ &= \sum_{\mathbf{IJ}} \log L(\mathbf{Y};\Lambda)\, p(\mathbf{I},\mathbf{J}|\mathbf{X};\Lambda') \\ &= \sum_{\mathbf{IJ}} \sum_{k=1}^{K} \Big[\log p(H_{i_k}) + \log \alpha_{i_k j_k} + \log \mathcal{N}(\mathbf{Z}_{i_k,k}, \boldsymbol{\mu}_{i_k j_k}, \boldsymbol{\Sigma}_{i_k j_k}) - \log p(\mathbf{Z}_{i_k,k}|H_0)\Big] \\ &\quad \cdot \prod_{l=1}^{K} \xi_{i_l j_l l}(\Lambda') \end{aligned} \tag{20}$$

where

$$\sum_{\mathbf{IJ}} \triangleq \sum_{\mathbf{I}} \sum_{\mathbf{J}} \triangleq \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_K=1}^{M} \cdot \sum_{j_1=1}^{L_{i_1}} \sum_{j_2=1}^{L_{i_2}} \cdots \sum_{j_K=1}^{L_{i_K}}$$

First, it is necessary to prove that increasing $Q(\Lambda;\Lambda')$ also increases $L(\mathbf{X};\Lambda)$ with respect to Λ. This was proved for the general PDF case in section 5.1. The proof for the Gaussian mixture case is identical with the exception that the summation $\sum_{\mathbf{I}}$ is replaced with the summation $\sum_{\mathbf{IJ}}$, therefore it is not repeated. We may now obtain $Q(\Lambda;\Lambda')$. First note that due to the independence of the samples,

$$\sum_{i_k} \sum_{j_k} p(i_k, j_k|\mathbf{X};\Lambda) = \sum_{i_k} \sum_{j_k} p(i_k, j_k|\mathbf{X}_k;\Lambda) = 1.$$

As a result,

$$\sum_{\mathbf{IJ}|i_k,j_k} p(\mathbf{I},\mathbf{J}|\mathbf{X};\Lambda) = \frac{\dfrac{\mathcal{N}(\mathbf{Z}_{i_k,k}, \boldsymbol{\mu}_{i_k j_k}, \boldsymbol{\Sigma}_{i_k j_k})\, p(H_{i_k})\, \alpha_{i_k j_k}}{p(\mathbf{Z}_{i_k,k}|H_0)}}{\displaystyle\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)\, p(H_l)}{p(\mathbf{Z}_{l,k}|H_0)}} = \xi_{i_k j_k k}(\Lambda), \tag{21}$$

where $\sum_{\mathbf{IJ}|i_k,j_k}$ is the summation over all indices except i_k and j_k,

$$\sum_{\mathbf{IJ}|i_k,j_k} \triangleq \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_{k-1}=1}^{M} \sum_{i_{k+1}=1}^{M} \cdots \sum_{i_K=1}^{M} \cdot \sum_{j_1=1}^{L_{i_1}} \sum_{j_2=1}^{L_{i_2}} \cdots \sum_{j_{k-1}=1}^{L_{i_{k-1}}} \sum_{j_{k+1}=1}^{L_{i_{k+1}}} \cdots \sum_{j_K=1}^{L_{i_K}}$$

In a similar manner as section 5.1, the auxiliary function is derived from equation (20) using (21). The resulting auxiliary function is given in (11).

References

[1] R. A. Redner and H. F. Walker, "Mixture densities maximum likelihood, and the EM algorithm," *SIAM Review*, vol. 26, April 1984.

[2] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, Vol. 2.* London: Charles Griffin, 1979.

[3] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[4] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis Of Finite Mixture Distributions.* John Wiley & Sons, 1985.

[5] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, 1985.

[6] P. M. Baggenstoss, "Class-specific features in classification.," *IEEE Trans Signal Processing*, December 1999.

[7] R. L. Streit, "Introduction to hidden markov models and tracking," *Naval Undersea Warfare Center, Newport, RI, Technical Memorandum TM 10204*, May 1993.

[8] L. E. Baum, "An inequality and associated maximization technique in statisticsl estimation for probabilistic functions of Markov processes," in *Inequalities III* (O. Shisha, ed.), (New York), pp. 1–8, Academic Press, 1972.