HIDDEN MARKOV MODEL FRAMEWORK USING INDEPENDENT COMPONENT ANALYSIS MIXTURE MODEL

*Jian Zhou**

Pixelworks Inc., Toronto Office
38 Leek Crescent, Suite 200
Richmond Hill, Ontario, Canada, L4B 4N8
E-mail: jzhou@pixelworks.com

Xiao-Ping Zhang

Dept. of Electrical & Computer Engineering
Ryerson University 350 Victoria Street
Toronto, Canada, M5B 2K3
E-mail: xzhang@ee.ryerson.ca

ABSTRACT

This paper describes a novel method for the analysis of sequential data that exhibits strong non-Gaussianities. In particular, we extend the classical continuous hidden Markov model (HMM) by modeling the observation densities as a mixture of non-Gaussian distributions. In order to obtain a parametric representation of the densities, we apply the independent component analysis (ICA) mixture model to the observations such that each non-Gaussian mixture component is associated with a standard ICA. Under this new framework, we develop the re-estimation formulas for the three fundamental HMM problems, namely, likelihood computation, state sequence estimation, and model parameter learning. The simulations also validate the theoretical results.

1. INTRODUCTION

Hidden Markov models (HMMs) model non-stationary stochastic sequences by using distinct random transitions among a set of different stationary processes. An HMM model can be considered as a *stochastic finite state automation* [1], which generates a sequence of observations vectors from its hidden states. Continuous HMM assumes the observations at each hidden state are generated by the continuous probability density functions. Liporace [2] introduced maximum likelihood formulation and used Gaussian densities to model the observations. The observation model was later extended in [3] by using a mixture of Gaussian densities. Given a sufficient number of mixtures, the Gaussian mixture model can approximate arbitrarily closely any continuous density function. However, the results may highly depend on the number of mixtures and how the parameters are estimated. In some cases, it is desirable to use non-Gaussian densities to estimate the distributions if the observations show strong non-Gaussian characteristics.

In this paper, we introduce a new HMM modeling method using non-Gaussian mixtures. The major challenge of using non-Gaussian distributions is the estimation of the non-Gaussian densities in parametric form. We bring independent component analysis (ICA) mixture model into HMM framework and use ICA mixture as a density estimation tool. In the observation space, each non-Gaussian mixture is associated with a standard ICA. We called this new framework as ICA mixture HMM (ICAMHMM). ICA is a recently developed statistical technique which captures the high order statistics of input signals or data. We choose the *ICA mixture model* instead of the *ICA* model because observations can be categorized into mutually exclusive classes, similar to Gaussian mixture models. This allows modeling classes with non-Gaussian structures. The observation densities described by ICA mixture can model a broader range of probability density functions, and can be considered as a complement of Gaussian mixture modeling. When using ICA mixture to capture non-Gaussian structures and classify data, better results were reported in [4], compared with Gaussian mixture models. In HMM framework, we choose ICA mixture as the observation model such that each non-Gaussian mixture component is associated with a standard ICA. We then show that by optimizing the auxiliary function the re-estimation formulas for likelihood computation, hidden state sequence estimation, and model parameter learning are developed. Experimental results also verify the effectiveness of the proposed framework.

2. THE OBSERVATION MODEL

A continuous HMM is generally used to model continuous signals. We denote the given observation sequence as $\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T$, where T is the length of the sequence, and $\boldsymbol{o}_t, 1 \leq t \leq T$, is the D-dimensional feature vector for the t-th observation sample. Let $q = q_1, q_2, \cdots, q_T$ be the hidden state sequence. A continuous HMM model with N hidden states is determined by the parameters $\lambda = (A, B, \pi)$, where $A = \{a_{ij}\}, 1 \leq i, j \leq N$, is the state transition probability matrix, and $a_{ij} = P(q_{t+1} = j \mid q_t = i)$ is the probability of state j at time $t+1$ given the state is i at time t. The

*The first author performed the work while at Ryerson University.

parameter $B = \{b_j(\boldsymbol{o})\}$ is the continuous observation density in state j, and $\boldsymbol{o}$ is the vector being modeled. The parameter $\pi = \{\pi_i\}$, $1 \leq i \leq N$, is the initial state distribution where $\pi_i = P(q_1 = i)$.

A classical continuous HMM model assumes the observations are generated from a finite mixture of Gaussian densities. In this paper, we propose to extend the observation model as a mixture of non-Gaussian densities. In order to estimate the densities in a parametric form, we choose ICA mixture model as a density estimation tool to describe the densities such that each non-Gaussian mixture component is associated with a standard ICA. The ICA mixture model is first to divide the observed data into mutually exclusive classes, and then model each class as a linear combination of independent, non-Gaussian ICA sources.

Let q_t be the hidden state at time t, the proposed non-Gaussian mixture observation model that brings ICA mixture model into HMM framework to capture the non-Gaussian structures can be represented as follows

$$b_{q_t}(\boldsymbol{o}) = \sum_{k=1}^{K} p(\boldsymbol{o} \mid C_{q_t k}, \theta_{q_t k}) P(C_{q_t k}), \tag{1}$$

where $\boldsymbol{o}$ is the vector being modeled. $P(C_{q_t k})$ is the class probability to the k-th class. $p(\boldsymbol{o} \mid C_{q_t k}, \theta_{q_t k})$ is a non-Gaussian probability density function that describes the statistics of the observations for the k-th class at time t, given the state at time t is q_t, where $\theta_{q_t k}$ represents the parameters of the densities in state q_t. Note that (1) is very similar to the Gaussian mixture observation model since both are essentially mixture models. The only difference is that the mixture component $p(\boldsymbol{o} \mid C_{q_t k}, \theta_{q_t k})$ in (1) is a non-Gaussian density function instead of a Gaussian density. The challenges and difficulties reside in the inferring the non-Gaussian density analytically. However, the non-Gaussianities can be captured by ICA by seeking statistically independent sources. Thus, each non-Gaussian mixture component density in (1) can be further modeled as a standard ICA.

In classical ICA without considering the mixture modeling, the observation sequence $\boldsymbol{O} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T$ is modeled as an D-dimensional random variable $\boldsymbol{o}_t$ which is further modeled as a linear combination of D statistically independent sources $\boldsymbol{s}_t$ plus the bias $\boldsymbol{\mu}$, i.e.:

$$\boldsymbol{o}_t = \boldsymbol{M}\boldsymbol{s}_t + \boldsymbol{\mu}, \quad t = 1, \ldots, T. \tag{2}$$

where $\boldsymbol{M}$ $(D \times D)$ is known as the *mixing matrix* in other ICA literatures. To avoid the ambiguity of the terms, in this paper we always refer to $\boldsymbol{M}$ as *the basis matrix* to distinguish the word "mixture" in the mixture model. The ICA task is to find the filter matrix $\boldsymbol{W} \approx \boldsymbol{M}^{-1}$ using only the observed signals $\boldsymbol{O}$. Since the observed signals are a linear transformation of the sources, their multivariate probability density functions satisfy the following relationship:

$$p(\boldsymbol{o}) = \frac{p(\boldsymbol{s})}{|\boldsymbol{J}|} \tag{3}$$

where $|\cdot|$ denotes the absolute value and $\boldsymbol{J}$ is the Jacobian of the transformation determined by the basis matrix. The vector $\boldsymbol{o}$ and the vector $\boldsymbol{s}$ are the observation vector and ICA source vector being modeled, respectively. Therefore, the log likelihood can be written as:

$$\log p(\boldsymbol{o}) = \log p(\boldsymbol{s}) - \log(\det|\boldsymbol{M}|). \tag{4}$$

Equations (2)-(4) describe the case when all the observations are assumed to be generated from one class (i.e. $K = 1$). The observation model introduced in (1) requires a mixture modeling since the observations are assumed to be generated from multiple classes. The standard ICA can be generalized into ICA mixture model that allows modeling of classes with non-Gaussian structures. The ICA mixture model, originally proposed by Lee and Lewicki [4], assumes that the observed data $\boldsymbol{O} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T$ are drawn independently and generated by a mixture density model. The likelihood of the data is given by the joint density:

$$p(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T \mid \Theta) = \prod_{t=1}^{T} p(\boldsymbol{o}_t \mid \Theta), \tag{5}$$

The mixture density is:

$$p(\boldsymbol{o} \mid \Theta) = \sum_{k=1}^{K} p(\boldsymbol{o} \mid C_k, \theta_k) p(C_k) \tag{6}$$

where $\Theta = (\theta_1, \ldots, \theta_k)$ are the unknown parameters for each $p(\boldsymbol{O}_t \mid C_k, \theta)$. And C_k denotes the class k and K is the number of classes. Then the data in each class are described by:

$$\boldsymbol{o}_t = \boldsymbol{M}_k \boldsymbol{s}_{k,t} + \boldsymbol{\mu}_k, \quad \boldsymbol{o}_t \in C_k, \tag{7}$$

where M_k is a square basis matrix for the k-th class, and $\boldsymbol{\mu}_k$ is the bias vector for class k. The task is first to classify the unlabeled data points into one of the K classes, and then to determine the parameters for each classes. The parameters include $(\boldsymbol{M}_k, \boldsymbol{\mu}_k)$ for the k-th class, and the class probability $p(C_k \mid \boldsymbol{o}_t, \theta)$ for each data point. Within each class, the data points are modeled by the standard ICA. Therefore, considering (4), we rewrite the total likelihood of the data based on the ICA mixture model as:

$$\log p(\boldsymbol{o} \mid \Theta) = \sum_{t=1}^{T} \log\Big(\sum_{k=1}^{K} p(C_k)(\log p(\boldsymbol{s}_{k,t}) - \log(\det|\boldsymbol{M}_k|))\Big). \tag{8}$$

Note that we intentionally drop the state index j in (8). The reason is that we choose the same observation model for all the hidden states for simplicity of implementation.

3. MODEL PARAMETERS

In the previous section, we integrate the ICA mixture model into HMM framework to model the observation densities as the mixture of non-Gaussians. We call this new proposed framework as *ICA Mixture Hidden Markov Model* (ICAMHMM). Based on the observation model described in the previous section, the ICAMHMM framework now has following model parameters:

$$\lambda = (A, C, \mu, M, \pi), \tag{9}$$

where $A = \{a_{ij}\}, 1 \leq i, j \leq N$, is the transition matrix. The parameter $C = \{P(C_{jk})\}$, $1 \leq j \leq N$, $1 \leq k \leq K$, is the mixture matrix. The parameter $\mu = \{\boldsymbol{\mu}_{jk}\}$, $1 \leq j \leq N$, $1 \leq k \leq K$, is the mean coefficients for the mixture densities, where $\boldsymbol{\mu}_{jk}$ is the D-dimensional mean vector for the k-th mixture component at state j. The parameter $M = \{\boldsymbol{M}_{jk}\}$, $1 \leq j \leq N, 1 \leq k \leq K$, is the ICA basis coefficients, where $\boldsymbol{M}_{jk}$ is the $D \times D$ basis matrix for the k-th ICA source at state j. The parameter $\pi = \{\pi_i\}$, $1 \leq i \leq N$, is the initial state distribution. Note that C, μ, and M are essentially the parameters for the modeling of the observation distributions in parametric form.

4. LEARNING THE MODEL

To apply the maximum likelihood estimation, we define the auxiliary function (Q-function) with similar form as [2] [3]:

$$Q(\lambda, \lambda^*) = \sum_{q \in \ell} \sum_{\boldsymbol{K} \in \hbar} L_\lambda(\boldsymbol{O}, q, \boldsymbol{K}) \log L_{\lambda^*}(\boldsymbol{O}, q, \boldsymbol{K}), \tag{10}$$

where λ is the current parameter set, and λ^* is the new parameter set. The likelihood function $L_\lambda(\boldsymbol{O}, q, \boldsymbol{K})$ is the joint probability of the observation sequence $\boldsymbol{O}$, a particular state sequence q over the set ℓ, and a particular mixture sequence $\boldsymbol{K}$ over the set $\hbar$, given the model parameters are λ. The Q-function defined in (10) can be considered as a partition in the state space, and then a further partition in the mixture space.

4.1. Model Learning

To derive the re-estimation formulas, we can calculate the Q-function and split the results into three terms such that the parameters π_i^*, a_{ij}^*, and $P(C_{jk}^*)$ are independently separated in the sum. Then each term can be optimized individually. The procedures are very similar to the one used to derive the classical HMM model parameters. The final results are presented as follows.

To efficiently calculate the HMM model parameters, some intermediate variable defined following the conventions in [5]. The HMM forward variable $\alpha_t(i)$ and the backward variable $\beta_t(i)$ are defined as:

$$\alpha_t(i) = P(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_t, q_t = i \mid \lambda), \tag{11}$$

$$\beta_t(i) = P(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, \ldots, \boldsymbol{o}_T, \mid q_t = i, \lambda). \tag{12}$$

Two probabilities of the joint event can then be defined as:

$$\xi_t(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(\boldsymbol{o}_{t+1})\, \beta_{t+1}(j)}{P(\boldsymbol{O} \mid \lambda)}, \tag{13}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\, \beta_t(i)}{P(\boldsymbol{O} \mid \lambda)}, \tag{14}$$

where (13) defines the probability of the joint event: a path passes through state i at time t *and* through state j at time $t+1$, given the available sequence of observations $\boldsymbol{O}$ *and* the parameters of the model λ. The (14) defines the probability of being in state i at time t, given the observation sequence $\boldsymbol{O}$, and the model λ.

The re-estimation formulas for ICAMHMM, derived using the optimization of Q-function, are listed as follows:

$$\pi_i^* = \gamma_1(i) \tag{15}$$

$$P(C_{jk}^*) = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(j,k)} \tag{16}$$

$$\boldsymbol{\mu}_{jk}^* = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{17}$$

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}. \tag{18}$$

Note that when calculating the intermediate variables like $\alpha_t(i)$, $\beta_t(i)$, $\xi_t(i,j)$, $\gamma_t(i,j)$, the observation densities $b_j(\boldsymbol{o}_t)$ is computed as

$$b_j(\boldsymbol{o}_t) = \sum_{k=1}^{K} P(C_{jk}) \cdot b_{jk}(\boldsymbol{o}_t), \tag{19}$$

where

$$b_{jk}(\boldsymbol{o}_t) = \exp(\log p(\boldsymbol{s}_{j,k,t}) - \log(\det|\boldsymbol{M}_{j,k}|)). \tag{20}$$

Equation (20) can be interpreted as the k-th component density, given the state j. Thus, the weighted summation over k gives the overall observation density at state j. During the practical implementation, the log is used to avoid the precision issues.

4.2. Likelihood Evaluation

Given any observation sequence $\boldsymbol{O}$, the likelihood of producing the observation sequence is given by $P(\boldsymbol{O} \mid \lambda)$. The calculation of the likelihood is very similar to the classical *Forward-Backward Procedure*. The major difference is that the observation densities are computed from the sources and basis matrices. The likelihood $P(\boldsymbol{O} \mid \lambda)$ is inductively solved as follows:

1)Initialization:

$$\alpha_1(i) = \pi_i \sum_{k=1}^{K} P(C_{ik}) \exp(\log p(\boldsymbol{s}_{i,k,1}) - \log(\det|\boldsymbol{M}_{ik}|)); \tag{21}$$

2)Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \cdot \sum_{k=1}^{K} P(C_{jk}) \exp(\log p(\boldsymbol{s}_{j,k,t+1}) - \log(\det|\boldsymbol{M}_{jk}|)); \tag{22}$$

3)Termination:

$$P(\boldsymbol{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i); \tag{23}$$

where $1 \leq i \leq N$, and $1 \leq t \leq T-1$.

4.3. State Sequence Estimation

The algorithm to compute the hidden state sequence is based on the Viterbi algorithm [5]. When calculating the observation densities, Equations (19) and (20) are used.

5. EXPERIMENTAL RESULTS

The proposed ICAMHMM framework is implemented in Matlab to verify its effectiveness. We manually create a two-state sequence as the ground truth data. Two classes of distributions, as shown in Figure 1, are mapped to the two states. We randomly generate the samples based on the state sequence. Each sample is a two-dimensional point in the observation space. The total sample size is 1300. The results are shown in Figure 2. As it can be seen that the proposed algorithm is effective to capture the temporal dynamics for the mixture of non-Gaussian observation densities. During a ten-trial experiment, the number of correctly estimated states using ICAMHMM is 99.15%. Among the 1300 samples, there are only 11 samples whose states are incorrectly estimated. The overall performance is comparable to the classical continuous HMM model using mixture of Gaussians.

6. CONCLUSIONS

In this paper, we propose a novel framework that combines both ICA mixture and HMM to model strong non-Gaussian observation densities. In particular, the ICA mixture model is used to capture the spatial characteristics, and the HMM is used to capture the temporal characteristics of the signal. Under this new framework, we develop the updating rules for the three fundamental HMM problems. Experimental results show that the proposed framework can effectively model the data. However, we also notice that the ICAMHMM may not converge to a desired result occasionally because of the introduction of ICA mixture model. As part of our future work, we will be focusing on the re-estimation algorithms to improve the stability of model learning.

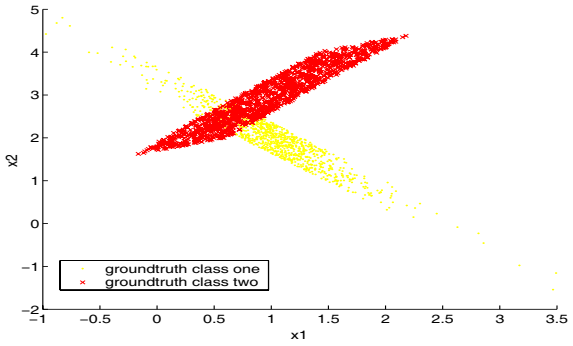

Figure 1: Sample distributions in observation space.

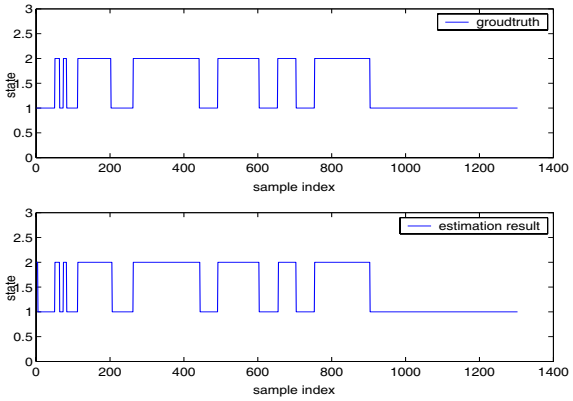

Figure 2: HMM state sequence learnt by ICAMHMM.

7. REFERENCES

[1] S. Theodoridis and K. Koutroumbas, *Pattern Recognition (2nd eds.)*, Academic Press, San Diego, CA, USA, 2003.

[2] L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. 28, no. 5, pp. 729–734, Sept. 1982.

[3] B.-H. Juang, S.E. Levinson, and M.M. Sondhi, "Maximum likelihood estimation for multivariate Mixture observations of Markov Chains," *IEEE Trans. Inform. Theory*, vol. 32, no. 2, pp. 307–309, Mar. 1986.

[4] T.-W. Lee and M.S. Lewicki, "Unsupervised image classification, segmentation, and enhancement using ICA mixture models," *IEEE Trans. on Image Processing*, vol. 11, no. 3, pp. 270–279, Mar. 2002.

[5] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.