MIHMM: Mutual Information Hidden Markov Models

Nuria Oliver NURIA@MICROSOFT.COM
Adaptive Systems & Interaction, Microsoft Research, One Microsoft Way, Redmond, WA 99052 USA

Ashutosh Garg ASHUTOSH@IFP.UIUC.EDU
Beckman Institute, University of Illinois, Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801,USA

Abstract

This paper proposes a new family of Hidden Markov Models (HMMs) named Mutual Information Hidden Markov Models (MIHMMs). MIHMMs have the same graphical structure as HMMs. However, the objective function being optimized is not the joint likelihood of the observations and the hidden states. It is a convex combination of the mutual information between the hidden states and the observations, and the likelihood of the observations and the states. First, we present both theoretical and practical motivations for having such an objective function. Next, we derive the parameter estimation (learning) equations for both the discrete and continuous observation cases. Finally, we illustrate the superiority of our approach in different classification tasks by comparing the classification performance of our proposed Mutual Information HMMs with standard Maximum Likelihood HMMs, in the case of synthetic and real, discrete and continuous, supervised and unsupervised data. We believe that MIHMMs are a powerful tool to solve many of the problems associated with HMMs when used for classification and/or clustering.

1. Introduction

It has been claimed (Tishby et al., 1999) that a fundamental problem in formalizing our intuitive ideas about information is to provide a quantitative notion of "meaningful" or "relevant" information. These issues were missing in the original formulation of information theory, where the attention was focused on the problem of transmitting information rather than evaluating its value to the recipient. Information theory has therefore traditionally been seen as a theory of communication. However, in recent years there has been growing interest in applying information theoretic principles in machine learning and statistics (Cover & Thomas, 1991). It has been argued that information theory provides a natural quantitative approach to the question of "relevant" information.

There are many situations when we would like to compress or summarize dynamic time data (for example speech or video). One possible approach to solving that problem is having an additional "hidden" variable that determines what is relevant. In the case of speech, for example, it could be the transcription of the signal, if we are interested in the speech recognition problem, or it might be the speaker's identity if speaker identification is our goal. The formal underlying structure of such problems would consist of extracting the information from one variable that is relevant for the prediction of another variable.

In this paper we propose an approach for using information theory in the framework of Hidden Markov Models (HMMs), by enforcing the hidden state variables to capture relevant information about the observations. At the same time, we would like our models to explain the generative process of the data as accurately as possible. Therefore, we propose an objective function that combines both the information theoretic and the maximum likelihood (ML) criteria.

The paper is organized as follows: First, in Section 2 we review the most relevant previous work. Section 3 motivates and describes the proposed objective function to be maximized. The learning algorithms that estimate the parameters of the model (in the discrete and continuous, supervised and unsupervised cases) while optimizing such function is presented in Section 4. Experimental results are presented in Section 6. Finally, we summarize our work and discuss future directions of research in Section 7.

2. Previous Work

In this work we introduce a new algorithm for parameter estimation in Hidden Markov Models. Numerous variations of the standard formulation of Hidden Markov Models have been proposed in the past, such as Entropic-HMM (Brand & Kettnaker, 2000),

Variable-length HMM (Galata et al., 2001), Coupled-HMM (Brand et al., 1997; Oliver, 2000), Input-Output-HMM (Bengio & Frasconi, 1995), Factorial-HMM (Ghahramani & Jordan, 1996) and Hidden-Markov Decision Trees (Jordan et al., 1996), to name a few. Each of these approaches attempts to solve some of the deficiencies of standard HMMs given the particular problem or set of problems at hand. Most of them aim at modeling the data and learning the parameters using ML. In many cases their main differences lie in the conditional independence assumptions made while modeling the data, *i.e.* in their graphical structure. Conversely, the graphical structure of the model presented in this paper remains the same as that of a standard HMM, but the objective function is different. Note that although our analysis in this paper focuses soley on HMMs, the framework proposed here could be generalized to other graphical models (Buntine, 1994).

Tishby's et al. work on the Information Bottleneck (Tishby et al., 1999) method and its extensions has been one of the sources of inspiration for our work. The Information Bottleneck method is an unsupervised non-parametric data organization technique. Given a joint distribution $P(A, B)$, the method constructs, using information theoretic principles, a new variable T that extracts partitions, or clusters, over the values of A that are informative about B. In particular, consider two random variables X and Q with their joint distribution $P(X, Q)$, where X is the variable that we are trying to compress with respect to the "relevant" variable Q. They propose the introduction of a soft partitioning of X through an auxiliary variable T, and the probabilistic mapping $P(T|X)$, such that the mutual information (MI) $I(T, X)$ is minimized (maximum compression) while the relevant information $I(T, Q)$ is maximized.

A related approach is the "infomax criterion", proposed in the neural network community, where the goal is to maximize the mutual information between the input and the output variables in a neural network. The biological relevance of miximizing the mutual information is discussed in (Atick, 1992).

Our work is also related to the recently popular debate of conditional versus joint density estimation (Caruana et al., 1998). The "conditional" approach (*i.e.* the maximization of the conditional likelihood of the variables of interest instead of the joint likelihood) is closely related to the use of discriminative approaches in learning theory. Jebara nicely summarizes in (Jebara, 1998) the advantages and disadvantages associated with joint and conditional density estimation. Standard HMM algorithms perform joint density estimation of the hidden state and observation random variables. However, in situations where the resources are limited (complexity, data, structures), the system has to handle very high dimensional spaces or when the goal is to classify or cluster with the learned models, a conditional approach is probably superior to the joint density approach. One can think of these two methods (conditional vs joint) as two extremes with our work providing a tradeoff between the two. Sections 3 and 5 analyze the properties of our approach and relate it to the purely probabilistic model more formally.

Finally, we would like to point out how our work is different to the Maximum Mutual Information Estimation (MMIE) approach that is so popular in the speech recognition community. In particular, Bahl et al. (Bahl et al., 1986) introduced the concept of Maximum Mutual Information Estimation (MMIE) for estimating the parameters of an HMM in the context of speech recognition, where typically a different HMM is learned for each possible class (*e.g.* one HMM for each word in the vocabulary). New waveforms are classified by computing their likelihood based on each of the models. The model with the highest likelihood is selected as the winner. However, in our approach, we learn a single HMM whose hidden states correspond to the different classes. The algorithm in (Bahl et al., 1986) attempts to maximize the mutual information between the choice of the HMM and the observation sequence to improve the discrimination across different models. In contrast, our algorithm aims at maximizing the mutual information between the observations and the hidden states, so as to minimize the classification error when the hidden states are used as the classification output.

3. Mutual Information, Bayes Optimal Error, Entropy and Conditional Probability

In the "generative approach" to machine learning, the goal is to learn a probability distribution that defines the process that generated the data. Such an approach is particularly good at modeling the general form of the data and can give some useful insights into the nature of the original problem. Recently, there has been an increasing focus on connecting the performance of these generative models to their classification accuracy when they are used for classification tasks. In particular, Garg and Roth develop an extensive analysis in (Garg & Roth, 2001) of the relationship between the Bayes optimal error[1] of a classification task using a probability distribution and the entropy between the random variables of interest. Consider the family of probability distributions over two random variables

[1] The Bayes optimal error is the error of a Bayes classifier. In the case of two equally probably classes, *i.e.* $P(y = 1) = P(y = 0) = .5$, it is given by $\epsilon = \frac{1}{2} P_0(\{x|P_1(x) > P_0(x)\}) + P_1(\{x|P_0(x) > P_1(x)\})$

(X, Q) denoted by $P(X, Q)$. The classification task is to predict Q after observing X. The relationship between the conditional entropy $H(X|Q)$ and the Bayes optimal error, ϵ, is given by

$$\frac{1}{2}H_b(2\epsilon) \leq H(X|Q) \leq H_b(\epsilon) + \log\frac{M}{2} \qquad (1)$$

with $H_b(p) = -(1-p)\log(1-p) - p\log p$ and M the dimensionality of the data.

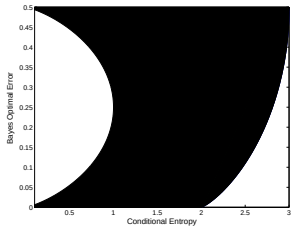


Figure 1. Bayes optimal error versus conditional entropy

Figure 1 illustrates this relationship between the conditional entropy and the Bayes optimal error. In Figure 1 the only realizable –and at the same time observable– distributions are those within the black region. One can conclude from Figure 1 that, if the data is generated according to a distribution that has high conditional entropy, the Bayes optimal error of any classifier for this data will be high. Even though this relationship is between the *true model* and the *Bayes optimal error*, it could also be applied to a model that has been estimated from data, –assuming a consistent estimator has been used, such as Maximum Likelihood (ML), and the model structure is the true one. As a result, when the learned distribution has high conditional entropy, it might not necessarily do well on classification. Therefore, if the final goal is classification, Figure 1 suggests that low entropy models should be preferred over high entropy ones. This result is related to *Fano's inequality* (Cover & Thomas, 1991): Suppose we know a random variable Q and we wish to guess the value of a correlated random variable X. Fano's inequality relates the probability of error in guessing the random variable X to its conditional entropy $H(X|Q)$. In particular, the conditional entropy of a random variable X given another random variable Q is zero if and only if X is a function of Q. Hence we can estimate X from Q with zero probability of error if $H(X|Q) = 0$. Extending this argument, we expect to be able to estimate X with a low probability of error only if the conditional entropy $H(X|Q)$ is small. Note that the objective function proposed in Eqn 2 favors low conditional entropy models to high entropy ones.

A Hidden Markov Model (HMM) is a probability distribution over a set of random variables, some of which are referred to as the hidden states (as they are normally not observed and they are discrete) and others are referred to as the observations (continuous or discrete). Traditionally, the parameters of HMMs are estimated by maximizing the joint likelihood of the hidden states Q and the observations X, $P(X, Q)$. Conventional Maximum Likelihood techniques would be optimal in the case of very large datasets (so that the estimate of the parameters is correct) if the true distribution of the data was in fact an HMM. However none of the previous conditions is normally true in practice. The HMM assumption might be in many occasions highly unrealistic and the available data for training is normally very limited, leading to important problems associated with the ML criterion (such as overfitting). Moreover, ML estimated models are often used for clustering or classification. In these cases, the evaluation function is different to the objective function, which suggests the need of an objective function that correctly models the problem at hand. The objective function defined in Eqn 2 is designed to tackle some of these problems associated to ML estimation.

When formulating our objective function, we were inspired by the relationship between the conditional entropy of the data and the Bayes optimal error, as previously described. In the case of HMMs, the X variable corresponds to the observations and the Q variable to the hidden states. We would like to choose $P(Q, X)$ such that the likelihood of the observed data is maximized while forcing the Q variable to contain maximum information about the X variable (*i.e.* to maximize their mutual information or minimize the conditional entropy). In consequence, we propose to maximize a function that trades-off the joint likelihood and the mutual information (Cover & Thomas, 1991) between the hidden variables and the observations, parameterized by α. This leads to the following function to maximize[2]

$$F = (1-\alpha)I(Q, X) + \alpha \log P(X_{obs}, Q_{obs}) \qquad (2)$$

where $\alpha \in [0, 1]$, provides a way of deciding the appropriate weighting between the Maximum Likelihood ($\alpha = 1$) and Maximum Mutual Information (MMI) ($\alpha = 0$) criteria, and $I(Q, X)$ refers to the mutual information between the states and the observations. However, very often one does not observe the state sequence[3]. In such a scenario, the objective function reduces to

$$F = (1-\alpha)I(Q, X) + \alpha \log P(X_{obs}) \qquad (3)$$

The mutual information $I(Q, X)$ is the reduction in the uncertainty of Q due to the knowledge of X.

[2]To make more clear the distinction between "observed" (supervised) and "unobserved" (unsupervised) variables, we will use the underscript $(.)_{obs}$ to denote that the variables have been observed, *i.e.* X_{obs} for the observations and Q_{obs} for the states.

[3]We will refer to this case as the unsupervised case while referring to the former as the supervised case.

The mutual information is also related to the KL-distance or relative entropy between two distributions $P(X)$ and $P(Q)$. In particular, $I(Q,X) = KL(P(Q,X)||P(X)P(Q))$, *i.e.* the mutual information between X and Q is the KL-distance between the joint distribution and the factored distribution. It is therefore a measure of how conditionally dependent the two random variables are. The objective function proposed in Eqn 2 penalizes factored distributions, favoring distributions where Q and X are mutually dependent. This is in accordance with the graphical structure of an HMM where the observations are conditionally dependent on the states, *i.e.* $P(X,Q) = P(Q)P(X|Q)$.

Mutual information is also closely related to conditional likelihood. Learning the parameters of a graphical model (Buntine, 1994) is equivalent to learning the conditional dependencies between the variables (edges in the graphical model). The following theorem by Bilmes et al. (Bilmes, 2000) makes explicit the relationship between conditional likelihood and mutual information in graphical models:

Theorem 1 Mutual Information and Likelihood *Given three random variables X, Q^a and Q^b, where $I(Q^a,X) > I(Q^b,X)$, there is an n_0 such that if $n > n_0$, then $P(X^n|Q^a) > P(X^n|Q^b)$.*

This theorem also holds true for conditional mutual information, such as $I(X,Z|Q)$, or for a particular value of q, $I(X,Z|Q=q)$. Therefore, given a graphical model in general (and an HMM in particular) whose parameters have been learned by maximizing the joint likelihood $P(X,Q)$, if we were to add some edges according to maximum mutual information the resulting graphical model would yield higher conditional likelihood score than before the modification (Bilmes, 2000). In an HMM we are maximizing the joint likelihood of the hidden states and the observations, $P(X,Q)$. At the same time, when using the HMM for classification, it would be desirable to make sure that the states Q are good predictors of the observations X. According to Theorem 1, maximizing the mutual information between states and observations would increase the conditional likelihood of the observations given the states $P(X|Q)$. This justifies, to some extent, why the objective function defined in Eqn 2 combines the two desirable properties of maximizing the mutual information and joint likelihood of the states and the observations.

4. MIHMMs

We develop in this section the learning algorithms for discrete and continuous, supervised and unsupervised MIHMMs. For the sake of clarity and simplicity, we will start with the supervised case, where the "hidden" states are actually observed in the training data.

Consider an HMM with $\mathbf{Q}$ as the states and $\mathbf{X}$ as the observations. Let F denote the function to maximize, $F = (1-\alpha)I(Q,X) + \alpha \log P(X_{obs}, Q_{obs})$. The mutual information term $I(Q,X)$ can be expressed as $I(Q,X) = H(X) - H(X/Q)$, where $H(\cdot)$ refers to the entropy. Since $H(X)$ is independent of the choice of the model and is characteristic of the generative process, our objective function reduces to

$$F = -(1-\alpha)H(X/Q) + \alpha \log P(X_{obs}, Q_{obs}) = (1-\alpha)F_1 + \alpha F_2$$

In the following we will use the standard HMM notation for the transition a_{ij} and observation b_{ij} probabilities,

$$a_{ij} = P(q_{t+1} = j | q_t = i), \quad b_{ij} = P(x_t = j | q_t = i)$$

Expanding each of the terms F_1 and F_2 separately we obtain,

$$\begin{aligned} F_1 &= -H(X|Q) = \sum_X \sum_Q P(X,Q) \log \prod_{t=1}^{T} P(x_t|q_t) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{i=1}^{N} P(x_t = j|q_t = i) P(q_t = i) \log P(x_t = j|q_t = i) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{i=1}^{N} P(q_t = i) b_{ij} \log b_{ij} \end{aligned}$$

$$F_2 = \log \pi_{q_1^o} + \sum_{t=2}^{T} \log a_{q_{t-1}^o, q_t^o} + \sum_{t=1}^{T} \log b_{q_t^o, x_t^o}$$

Combining F_1 and F_2 and adding the appropriate Lagrange multipliers to ensure that the a_{ij} and b_{ij} coefficients sum to 1, we obtain:

$$\begin{aligned} F_L &= (1-\alpha) \sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{i=1}^{N} P(q_t = i) b_{ij} \log b_{ij} \\ &\quad + \alpha \log \pi_{q_1^o} + \alpha \sum_{t=2}^{T} \log a_{q_{t-1}^o, q_t^o} + \alpha \sum_{t=1}^{T} \log b_{q_t^o, x_t^o} \\ &\quad + \beta_i \left(\sum_j a_{ij} - 1 \right) + \gamma_i \left(\sum_j b_{ij} - 1 \right) \end{aligned} \tag{4}$$

Note that in the case of continuous observation HMMs, we can no longer use the concept of entropy as previously defined, but its counterpart differential entropy. Because of this important distinction, we will carry out the analysis for discrete and continuous observation HMMs separately.

4.1 Discrete MIHMMs

To obtain the parameters that maximize the function, we take the derivative of F_L from Eqn 4 and will equate

it to zero. First solving for b_{ij}, we obtain

$$\frac{\partial F_L}{\partial b_{ij}} = (1-\alpha)(1+\log b_{ij}) \left(\sum_{t=1}^{T} P(q_t = i) \right) + \frac{N_{ij}^b \alpha}{b_{ij}} + \gamma_i = 0 \tag{5}$$

where N_{ij}^b is the number of times one observes state j when the hidden state is i. Eqn 5 can be expressed as

$$\log b_{ij} + \frac{W_{ij}}{b_{ij}} + g_i + 1 = 0 \tag{6}$$

where

$$W_{ij} = \frac{N_{ij}^b \alpha}{(1-\alpha)\left(\sum_{t=1}^{T} P(q_t = i)\right)}$$

$$g_i = \frac{\gamma_i}{(1-\alpha)\left(\sum_{t=1}^{T} P(q_t = i)\right)}$$

The solution of Eqn 6 is given by

$$b_{ij} = -\frac{W_{ij}}{\mathsf{LambertW}(-\mathsf{W_{ij}} \mathsf{e}^{1+\mathsf{g_i}})}$$

where $\mathsf{LambertW(x)} = \mathsf{y}$ is the solution of $ye^y = x$.

Now we are going to solve for a_{ij}. Let's first look at the derivative of F_1 with respect to a_{lm}

$$\frac{\partial F_1}{\partial a_{lm}} = \sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{i=1}^{N} b_{ij} \log b_{ij} \frac{\partial P(q_t = i)}{\partial a_{lm}}$$

To solve the above equation, we need to compute $\frac{\partial P(q_t=i)}{\partial a_{lm}}$, which can be computed using the following iteration

$$\frac{\partial P(q_t = i)}{\partial a_{lm}} = \begin{cases} \sum_j \frac{\partial P(q_{t-1}=j)}{\partial a_{lm}} a_{ji} & \text{if } m \neq i, \\ \sum_j \frac{\partial P(q_{t-1}=j)}{\partial a_{lm}} a_{ji} + P(q_{t-1} = l) & \text{if } m = i \end{cases} \tag{7}$$

with the initial conditions

$$\frac{\partial P(q_2 = i)}{\partial a_{lm}} = \begin{cases} 0 & \text{if } m \neq i, \\ \pi_l & \text{if } m = i \end{cases}$$

Taking the derivative of F_L, with respect to a_{lm}, we obtain,

$$\frac{\partial F}{\partial a_{lm}} = (1-\alpha) \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{M} b_{ik} \log b_{ik} \frac{\partial P(q_t = i)}{\partial a_{lm}} + \alpha \frac{N_{lm}}{a_{lm}} + \beta_l$$

where N_{lm} is the count of the number of occurrences of $q_{t-1} = l, q_t = m$ in the data set. The update equation for a_{lm} is obtained by equating this quantity to zero and solving for a_{lm}

$$a_{lm} = \frac{-\alpha N_{lm}}{(1-\alpha) \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{M} b_{ik} \log b_{ik} \frac{\partial P(q_t=i)}{\partial a_{lm}} + \beta_l} \tag{8}$$

where β_l is chosen so that $\sum_m a_{lm} = 1, \forall l$.

4.2 Continuous MIHMMs

For the sake of clarity, we will restrict our attention to the case when the $P(x|q)$ is a single Gaussian[4]. Under this assumption, the HMM is characterized by the following parameters

$$P(q_t = j | q_{t-1} = i) = a_{ij}$$

$$P(x_t | q_t = i) = \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp(-\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1} (x_t - \mu_i))$$

where Σ_i is the covariance matrix when the hidden state is i and $|\Sigma_i|$ is the determinant of the covariance matrix. Now, for the objective function given in Eqn 2, F_1 and F_2 can be written as

$$\begin{aligned} F_1 &= -H(X|Q) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N} \int P(q_t = i) \log P(x_t | q_t = i) dP(x_t | q_t = i) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N} P(q_t = i) \int (-\frac{1}{2} \log(2\pi|\Sigma_i|) \\ &\quad -\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1} (x_t - \mu_i)) dP(x_t | q_t = i) y \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N} P(q_t = i)(-\frac{1}{2} \log(2\pi|\Sigma_i|) - \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} F_2 &= \log P(Q_{obs}, X_{obs}) \\ &= \sum_{t=1}^{T} \log P(x_t | q_t) + \log \pi_{q_1^o} + \sum_{t=2}^{T} \log a_{q_{t-1}^o, q_t^o} \end{aligned}$$

Following the same steps as for the discrete case, we again form the Lagrange F_L, take its derivative with respect to each of the unknown parameters and obtain the corresponding update equations. First the means of the Gaussians

$$\mu_i = \frac{\sum_{t=1, q_t=i}^{T} x_t}{N_i}$$

where N_i is the number of times $q_t = i$ in the observed data. Note that this is the standard update equation for the mean of a Gaussian, and it is the same as for ML estimation in HMMs. This is because the conditional entropy is independent of the mean.

Next, the update equation for a_{lm} is same as in Eqn 8 except for replacing $\sum_k b_{ik} \log b_{ik}$ by $-\frac{1}{2}\log(2\pi|\Sigma_i|) - \frac{1}{2}$. Finally, the update equation for Σ_i is

$$\Sigma_i = \frac{\sum_{t=1, q_t=i}^{T} (x_t - \mu_i)(x_t - \mu_i)^T}{N_i + \frac{(1-\alpha)}{\alpha} \sum_{t=1}^{T} P(q_t = i)} \tag{9}$$

[4] We could extend our reasoning to other distributions and in particular to other members of the exponential family.

It is interesting to note that Eqn 9 is very similar to the one obtained when using ML estimation, except for the term in the denominator $\frac{(1-\alpha)}{\alpha}\sum_{t=1}^{T} P(q_t = i)$, which can be thought of as a regularization term. Because of this positive term, the covariance Σ_i is smaller than what it would have been otherwise. This corresponds to lower conditional entropy, as desired.

4.3 Unsupervised Case

The above analysis can easily be extended to the unsupervised case, *i.e.* when only X_{obs} is given and Q_{obs} is not available. In this case, we use the objective function given in Eqn 3. The update equations for the parameters are very similar to those obtained in the supervised case. The only difference is that now we replace N_{ij} in Eqn 5 by $\sum_{t=1,x_t=j}^{T} P(q_t = i|X_{obs})$, N_{lm} is replaced in Eqn 8 by $\sum_{t=2}^{T} P(q_{t-1=l}, q_t = m|X_{obs})$, and N_i is replaced in Eqn 9 by $\sum_{t=1}^{T} P(q_t = i|X_{obs})$. These quantities can be easily computed using the Baum-Welch algorithm by means of the forward and backward variables.

5. Discussion

5.1 Convexity

From the asymptotic equipartition property (Cover & Thomas, 1991), it is known that, in the limit (*i.e.* as the number of samples approaches infinity), the likelihood of the data tends to the negative of the entropy, $P(X) \approx -H(X)$. Therefore and in the limit, the negative of our objective function for the supervised case can be expressed as

$$-F = (1-\alpha)H(X|Q) + \alpha H(X,Q) = H(X|Q) + \alpha H(Q) \tag{10}$$

Note that $H(X|Q)$ is a strictly concave function of $P(X|Q)$, and $H(X|Q)$ is a linear function of $P(Q)$. Consequently, in the limit, the objective function from Eqn 10 is strictly convex (its negative is concave) with respect to the distributions of interest.

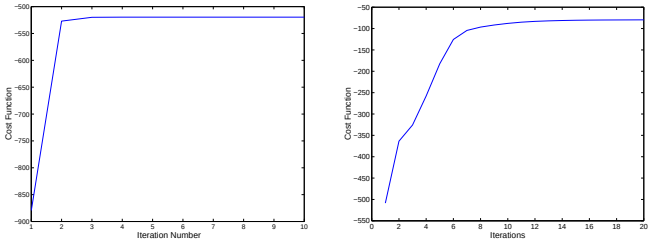
In the unsupervised case and in the limit again, our objective function can be expressed as

$$\begin{aligned} F &= -(1-\alpha)H(X|Q) - \alpha H(X) \\ &= -H(X) + (1-\alpha)(H(X) - H(X|Q)) \\ &= -H(X) + (1-\alpha)I(Q,X) \approx P(X) + (1-\alpha)I(Q,X) \end{aligned}$$

The unsupervised case thus reduces to the original case with α replaced by $1-\alpha$. Maximizing F is, in the limit, the same as maximizing the likelihood of the data and the mutual information between the hidden and the observed states, as expected. The above analysis shows that in the asymptotic case, the objective function is strictly convex and as such a unique solution exists. However, in the case of finite amount of data, local maxima could be a problem (as has been observed in case of standard ML for HMM). We have not observed a severe local maxima problem in any of our experiments.

5.2 Convergence

We will discuss next the convergence of the MIHMM learning algorithm in the supervised and unsupervised cases. In the supervised case, the HMM parameters are directly learned without any iteration. However, we do not have a closed form solution for estimating the parameters (b_{ij} and a_{ij}) in MIHMMs. These parameters are inter-dependent (*i.e.* in order to compute bij, we need to compute $P(q_t = i)$, which requires the knowledge of a_{ij}). Therefore an iterative solution is needed. Fortunately, the convergence of the iterative algorithm is extremely fast, as Figure 2 illustrates. This figure shows the objective function with respect to the iterations for a particular case of the speaker detection problem (a) (see Section 6), and for synthetically generated data in an unsupervised situation (b). From Figure 2 it can be seen that the algorithm typically converges after a few (5-6) iterations.



(a) (b)

Figure 2. Objective function with respect to the iteration number in (a) the speaker detection experiment; (b) a continuous unsupervised case with synthetic data.

5.3 Computational Complexity

The MIHMM algorithms proposed in this paper are computationally more expensive than standard ML estimation in HMMs. The main additional complexity is due to the computation of the derivative of probability of state with respect to the transition probabilities, *i.e.* $\frac{\partial P(q_t=i)}{\partial a_{lm}}$ in Eqn. 7. Let us consider an HMM with N states and M observation values –or dimensions in the continuous case– and sequences of length T. The complexity of Eqn. 7 in MIHMMs is $O(TN^4)$. Besides this term, the computation of a_{ij} adds TN^2 computations. The computation of b_{ij} (*i.e.* the observation probabilities) requires solving for the Lambert function, which is done iteratively. However, this normally requires a small number of iterations that we will ignore in this analysis. Consequently, the computational complexity for the discrete supervised case for MIHMMs is $O(TN^4+TNM)$. In contrast, ML for HMMs is $O(TN^2+TNM)$. In the unsupervised case, the counts are replaced by probabilities, which can be estimated using the forward-backward algorithm and whose com-

putation is of the order of $O(TN^2)$. Hence the overall order remains the same. Note that there is an additional incurred penalty because of the cross-validation computations to estimate the optimal value of α. However, if the number of cross-validation rounds and the number of α's tried are fixed, the order remains same even though the actual numbers will increase.

A similar analysis for the continuous case reveals that, when compared to standard HMM, the additional cost is $O(TN^4)$. Once the parameters have been learned, inference is carried out in exactly the same fashion and with the same complexity as with HMMs.

6. Experimental Results

In this section we describe the set of experiments that we carried out to obtain quantitative measures of the performance of MIHMMs when compared to standard HMMs in various classification tasks. We conducted experiments with synthetic and real, discrete and continuous, supervised and unsupervised data. In all the experiments, the optimal value for alpha, $\alpha_{optimal}$, was estimated using k-fold cross-validation (Kohavi, 1995) on a validation set. In our experiments k was either 10 or 12. We randomly divided the given dataset into two groups, one for training $\mathcal{D}^{tr}$ and the other for testing $\mathcal{D}^{te}$. The size of the test dataset was typically $20-50\%$ of the training dataset. For crossvalidation – to choose the best α– the training set $\mathcal{D}^{tr}$ was furhter subdivided into k mutually exclusive subsets (folds) $\mathcal{D}_1^{tr}, \mathcal{D}_2^{tr}, ..., \mathcal{D}_k^{tr}$ of the same size ($1/k$ of the training data size). The models were trained k times; each time $t \in \{1, ..., k\}$ we trained on $\mathcal{D}^{tr} \backslash \mathcal{D}_t^{tr}$ and tested on $\mathcal{D}_t^{tr}$. We chose the alpha, $\alpha_{optimal}$, that provided the best performance and used it on the testing data $\mathcal{D}^{te}$.

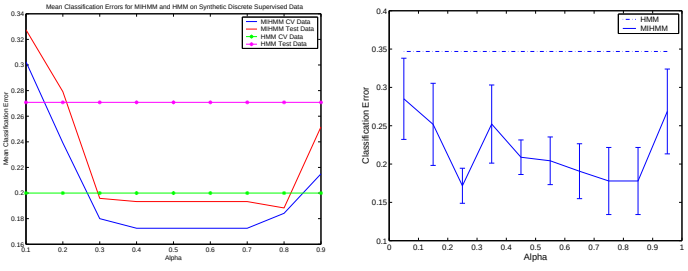
1. Synthetic Discrete Supervised Data:

We generated a 10 datasets of randomly sampled synthetic discrete data with 3 hidden states, 3 observation values and random additive observation noise. We used 120 samples per dataset for training, 120 per dataset for testing and 10-fold crossvalidation to estimate α. The training was supervised for both HMMs and MIHMMs. MIHMMs had an average improvement over the 10 datasets of **11**%, when compared to HMMs of exactly the same structure. The $\alpha_{optimal}$ was 0.5[5]. The mean classification error over the ten datasets for HMMs and MIHMMs with respect to α is depicted in Figure 3 (a). A summary of the mean accuracies of HMMs and MIHMMs is shown in Table 1.

2. Speaker Detection:

An estimate of the person's state is important for the reliable functioning of any interface that uses speech communication. In particular, detecting when users are speaking is a central component of open mike speech-based user interfaces, specially given their need to handle multiple people in noisy environments. We carried out some experiments in a speaker detection task. The speaker detection dataset was the same that appears in (Garg et al., 2000). It consisted of five sequences of one user playing blackjack in a simulated casino setup using CRL's Smart Kiosk (Rehg et al., 1997). The sequences were of varying duration from 2000 to 3000 samples, with a total of 12500 samples. The original feature space had 32 dimensions that resulted from quantizing five binary features (skin color presence, face texture presence, mouth motion presence, audio silence presence and contextual information). Only the 14 most significant dimensions were selected out of the original 32-dimensional space.

The learning task in this case was supervised for both HMMs and MIHMMs. Three were the variables of interest: the presence/absence of a speaker, the presence/absence of a person facing frontally, and the existence/absence of an audio signal or not. The goal was to identify the correct state out of four possible states: (1) no speaker, no frontal, no audio; (2) no speaker, no frontal and audio; (3) no speaker, frontal and no audio; (4) speaker, frontal and audio. Figure 3 (b) illustrates the classification error for HMMs (dotted line) and MIHMMs (solid line) with α varying from 0.05 to 0.95 in .1 increments. In this case, instead of displaying the results for the optimal α, Figure 3 (b) displays the results for all α. Note how in this case MIHMMs outperformed HMMs for all the values of α. The optimal alpha using cross-validation was $\alpha_{optimal} = 0.75$. The accuracies of HMMs and MIHMMs are summarized in table 1. The accuracy reported in (Garg et al., 2000) using a bi-modal (audio and video) DBN was of $\approx 80\%$.



(a) (b)

Figure 3. (a) Mean Classification Errors with respect to alpha for MIHMM and HMM (star-line) on Synthetic Discrete Supervised data; (b) Error bars for the Speaker Detection data in MIHMMs and HMMs.

3. Gene Data: Gene identification and gene discovery in new genomic sequences is an important computational question addressed by scientists working in the domain of bioinformatics. In this example, we tested

[5]Note that from the cross-validation results, any alpha in [.3 .8] would be equally acceptable. Among all these values, we chose 0.5.

Table 1. Classification accuracies for HMMs and MIHMMs on different datasets

DATASET	HMM	MIHMM
SYNTDISC	73%	**81**%($\alpha_{\mathbf{optimal}} = .\mathbf{5}$)
SPEAKERID	64%	**88**%($\alpha_{\mathbf{optimal}} = .\mathbf{75}$)
GENE	51%	**61**%($\alpha_{\mathbf{optimal}} = .\mathbf{35}$)
EMOTION	47%	**58**%($\alpha_{\mathbf{optimal}} = .\mathbf{49}$)

both HMMs and MIHMMs in the analysis of part of an annotated sequence (7000 data points on training and 2000 on testing) of the Adh region in Drosophila (Drosophila, 1999). The task was to annotate the sequence into exons and introns and compare the results with the ground truth. 10-fold cross-validation was used to estimate the best value of α, which was $\alpha_{optimal} = 0.35$. The improvement of MIHMMs over HMMs on the testing sequence was of about **19**%, as Table 1 reflects.

4. Real-time Emotion Data: Finally, we carried out an emotion recognition task using the emotion data described in (Cohen et al., 2000). The data had been obtained from a video database of five people that had been instructed to display facial expressions corresponding to the following six basic emotions: anger, disgust, fear, happiness, sadness and surprise. The data collection method is described in detail in (Cohen et al., 2000). The database consisted of six sequences of each facial expression for each of the five subjects. In the experiments reported here, we used unsupervised training of continuous HMMs and MIHMMs. We used this time 12-fold cross-validation to select the optimal value of α, which led to $\alpha_{optimal} = 0.49$. The mean accuracies for both types of models are displayed in Table 1.

7. Summary and Future Work

We have presented a new framework for estimating the parameters of Hidden Markov Models. We have motivated, proposed and justified a new objective function that is a convex combination of the mutual information and the likelihood of the hidden states and the observations in an HMM. We have derived the parameter estimation equations in the discrete and continuous, supervised and unsupervised cases. Finally, we have shown the superiority of our approach in a classification task when compared to standard HMMs in different synthetic and real datasets.

Future lines of research include automatic estimation of the optimal α, extension of our approach to other graphical models with different structures, and better understanding of the connection between MIHMMs and other information theoretic and discriminative approaches. We are also exploring how to apply our framework to a number of applications and real-life problems.

References

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network, 3,* 213–251.

Bahl, Brown, de Souza, & Mercer (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. ICASSP *'86* (pp. 999–999).

Bengio, Y., & Frasconi, P. (1995). An input output HMM architecture. NIPS *'95* (pp. 427–434).

Bilmes, J. (2000). Dynamic bayesian multinets. *UAI'00.*

Brand, M., & Kettnaker, V. (2000). Discovery and segmentation of activities in video. *IEEE* TPAMI *'00, 22(8).*

Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hidden markov models for complex action recognition. *CVPR'97* (pp. 994–999).

Buntine, W. (1994). Operations for learning with graphical models. *JAIR'94.*

Caruana, R., de Sa, V., Kearns, M., & McCallum, A. (1998). Integrating supervised and unsupervised learning. NIPS'98 workshop.

Cohen, I., Garg, A., & Huang, T. S. (2000). Emotion recognition using multilevel hmms. *Workshop on Affective Computing in* NIPS *'00.*

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* Wiley Series in Telecommunications. New York, NY, USA: John Wiley & Sons.

Drosophila (1999). *Berkeley Drosophila Genome Project.* http://www.fruitfly.org/about/pubs/ashburner99.html.

Galata, A., Johnson, N., & Hogg, D. (2001). Learning variable length markov models of behaviour. *IJCV'01,* 398–413.

Garg, A., Pavlovic, V., Rehg, J., & Huang, T. S. (2000). Audio-visual speaker detection using dynamic bayesian networks. *FG'00.*

Garg, A., & Roth, D. (2001). Understanding probabilistic classifiers. *ECML'01.*

Ghahramani, Z., & Jordan, M. I. (1996). Factorial hidden Markov models. NIPS *'96.*

Jebara, T. (1998). Action-reaction learning: Analysis and synthesis of human behaviour. Master's thesis, MIT Media Lab.

Jordan, M. I., Ghahramani, Z., & Saul, L. K. (1996). Hidden Markov decision trees. NIPS'96.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95.*

Oliver, N. (2000). *Towards perceptual intelligence: Statistical modeling of human individual and interactive behaviors.* Doctoral dissertation, Massachusetts Institute of Technology, MIT.

Rehg, J., Loughlin, M., & Waters, K. (1997). Vision for a smart kiosk. *CVPR'97* (pp. 690–696).

Silva, L. D., Miyasato, T., & Nakatsu, R. (1997). Facial emotion recognition using multimodal information. *ICICS'97)* (pp. pp. 397–401).

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. *ACCC'99.*