Hidden Markov Models Based on Generalized Dirichlet Mixtures for Proportional Data Modeling

Elise Epaillard[1] and Nizar Bouguila[2]

[1] Department of Electrical and Computer Engineering
[2] Concordia Institute for Information Systems Engineering
Concordia University, Montreal, Quebec, Canada
e_epail@encs.concordia.ca
nizar.bouguila@concordia.ca

Abstract. Hidden Markov models (HMMs) are known for their ability to well model and easily handle variable length time-series. Their use in the case of proportional data modeling has been seldom mentioned in the literature. However, proportional data are a common way of representing large data in a compact fashion and often arise in pattern recognition applications frameworks. HMMs have been first developed for discrete and Gaussian data and their extension to proportional data through the use of Dirichlet distributions is quite recent. The Dirichlet distribution has its limitations and is a special case of the more general generalized Dirichlet (GD) distribution that suffers from less restrictions on the modeled data. We propose here to derive the equations and the methodology of a GD-based HMM and to assess its superiority over a Dirichlet-based HMM (HMMD) through experiments conducted on both synthetic and real data.

Keywords: Hidden Markov models, generalized Dirichlet, mixtures, machine learning, EM-algorithm.

1 Introduction

HMMs are probabilistic generative models used in various fields such as speech processing [18], object and gesture classification [3,9] or anomaly detection [2,14]. Their use has been popularized by [18], and numerous extensions and adaptations to specific applications have been developed along the years.

Among the extensions developed for HMMs, the study of time-series generated from multiple processes and/or involving dynamics at different scales led to the development of factorial HMMs [11]. In this framework, each state is decomposed into a collection of sub-states, often assumed independent at each time step in order to reduce algorithmic complexity.

Classic HMMs naturally embed a geometric distribution as for state duration, i.e. state self-transitioning, with parameter depending on the state transition matrix [10]. Variable Duration HMMs have been a first attempt to modify the state duration probability distribution [17]. At each state transition, the duration of the new state is drawn from a probability mass function and the corresponding number of observations is generated before drawing a new state accordingly to the state transition matrix. An alternate approach that explicitly introduces the time variable into the state transition matrix is

N. El Gayar et al. (Eds.): ANNPR 2014, LNAI 8774, pp. 71–82, 2014.

proposed in [10]. Known as Non-Stationary HMMs, they have been shown equivalent to Variable Duration HMMs, though allowing an easier and computationally more efficient parameter estimation [10].

The most widely used estimation algorithm for HMMs is the so-called Baum-Welch algorithm, though its iterative nature can be prohibitive in some applications. [12] proposed a non-iterative method for parameters estimation. Based on subspace estimation, the idea has been theoretically derived in [1] and provides, under few conditions, a computationally fast method to estimate HMMs with finite alphabet output.

HMMs have been initially developed for discrete and Gaussian data [18]. The multiplication of applications in domains such as weather forecast or medical studies raised the need to modify the original HMM algorithm so it can efficiently work with new data types [13,15]. Longitudinal or panel data are time-series collected from multiple entities. Example of these data in the context of a medical study could be the evolution of some disease characteristics evaluated every day for a given period of time on a number of patients (see [16] for concrete example). At the entity level, data heterogeneity is involved by the presence of multiple data sources. HMMs have been shown to be able to model this heterogeneity by introducing a random variable in the model, known as the *random effect*, that follows a predefined probability distribution. Doing so, the conditional independence of the observed data given the latent states assumption is relaxed. [15] provides a review of the use of these HMMs that are known in the literature as Mixed HMMs. [13] discusses circular data processing, i.e data taking cyclic values (e.g. directions, angles,...). Von Mises, Wrapped Normal and Wrapped Cauchy are proposed as state emission probability distributions to handle such data. A Maximum-Likelihood estimation algorithm is derived and applied to circular time-series.

Proportional data (i.e. positive data that sum up to 1) results from numerous pattern recognition pre-processing procedures, the most common being histograms. Their use in an HMM framework has been first studied in [8] where Dirichlet mixtures are used as emission probability functions, involving a deep modification of the M-step of the Expectation-Maximization algorithm (EM) for Dirichlet parameters estimation. The limitation of the Dirichlet distribution has been brought to light by [5] and we propose here to derive the equations of an HMM based on mixtures of GD distributions (HMMGD). This model is expected to be more general and versatile as the GD distribution embeds the Dirichlet distribution as a special case.

Section 2 fully develops the HMMGD derivation, Section 3 presents experimental work on synthetic data, and Section 4, on real data. We conclude and explain our future work in Section 5.

2 HMM Based on Generalized Dirichlet Mixtures

Based on [18], a first-order HMM is a probabilistic model assuming an ordered observation sequence $O = \{O_1, ..., O_T\}$ to be generated by some hidden states, each of them associated with a probability distribution governing the emission of the observed data. The hidden states $H = \{h_1, ..., h_T\}$, $h_j \in [1, K]$, with K the number of states, are assumed to form a Markov chain.

At each time *t*, a new state is entered based on a transition matrix $B = \{B_{ij} = P(h_t = j|h_{t-1} = i)\}$ that specifies the transition probabilities between states. Once in

the new state, an observation is made following its associated probability distribution. For discrete observation symbols taken from a vocabulary $V = \{v_1, ..., v_S\}$, the emission matrix is defined as $D = \{D_j(k) = P(O_t = v_k|h_t = j)\}, [t, k, j] \in [1, S] \times [1, M] \times [1, K]$. For continuous observation vectors, emission probability distributions are usually taken as Gaussian mixtures [2,3,18] defined by their mean and covariance matrices, denoted θ. In the latter case, a matrix $C = \{C_{i,j} = P(m_t = i|h_t = j)\}$, $i \in [1, M]$, is defined with M the number of mixture components associated with state j (which can be assumed to be the same for all states without loss of generality). An initial probability distribution π controls the initial state. We denote an HMM as $\lambda = \{B, D, \pi\}$ or $\{B, C, \theta, \pi\}$.

HMMs are well fit for classification tasks that rely on the probability of an observation sequence given a model λ, computed using a forward-backward procedure [18]. Model training consists in the estimation of the parameters that maximize the probability of a given set of observations and is addressed with the Baum-Welch algorithm, an Expectation-Maximization process [18]. Finally finding the most probable sequence of states and mixture components that generated a series of observations can be solved with the Viterbi algorithm [18].

The number of hidden states and the parameters initial values have to be a priori set. Both are strongly linked to model's performance. Indeed, the former is a trade-off between performance and complexity [9], while the latter leads the Baum-Welch procedure to converge towards the closest local maximum of the likelihood function, not guaranteed to be the global one given its high modality [3].

In this paper we propose to develop HMMs with mixtures of GD as emission probability distributions. [8] derived the equations for HMMs with Dirichlet mixtures, yet these distributions have one main limitation residing in the fact that data covariance is always negative. Therefore, they might not be adapted to model all types of proportional data. The GD distribution overcomes this limitation and embeds the Dirichlet distribution as a special case.

2.1 Expected Complete-Data Log-Likelihood Equation Setting

An N-dimensional generalized Dirichlet distribution is defined as

$$GD(\mathbf{x}|\alpha, \beta) = \prod_{n=1}^{N} \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} x_n^{\alpha_n - 1} \left(1 - \sum_{r=1}^{n} x_r\right)^{\nu_n}, \quad (1)$$

where Γ denotes the Gamma function and $\alpha = [\alpha_1, ..., \alpha_N]$ and $\beta = [\beta_1, ..., \beta_N]$ the GD parameters, with $\alpha \in \mathbb{R}_+^N$, $\beta \in \mathbb{R}_+^N$, $\mathbf{x} \in \mathbb{R}_+^N$, and $\sum_{n=1}^{N} x_n < 1$. For $n \in [1, N-1]$, $\nu_n = \beta_n - \alpha_{n+1} - \beta_{n+1}$, and $\nu_N = \beta_N - 1$.

This change of probability distribution involves modifications in the EM parameters estimation process. The rest of the HMM algorithm is unchanged. We set notations for the quantities $\gamma_{h_t,m_t}^t \triangleq p(h_t, m_t|x_0, ..., x_T)$ and $\xi_{h_t,h_{t+1}}^t \triangleq p(h_t, h_{t+1}|x_0, ..., x_T)$, that represent the estimates of the states and mixture components, and of the local states sequence given the whole observation set, respectively. The E-step leads to γ_{h_t,m_t}^t and $\xi_{h_t,h_{t+1}}^t$ estimates for all $t \in [1, T]$. These two quantities are obtained using the initial

parameters at step 1 and the result of the last M-step then. They are computed using a forward-backward procedure (not detailed here) as in HMM with mixtures of Gaussian.

The M-step aims at maximizing the data log-likelihood by maximizing its lower bound. If Z represents the hidden variables and X the data, the data likelihood $\mathcal{L}(\theta|X) = p(X|\theta)$ can be expressed as

$$\begin{aligned} E(X,\theta) - R(Z) &= \sum_Z p(Z|X)\ln(p(X,Z)) - \sum_Z p(Z|X)\ln(p(Z|X)) \\ &= \sum_Z p(Z|X)\ln(p(X)) \quad \text{(Bayes' rule)} \\ &= \ln(p(X))\sum_Z p(Z|X) = \ln(p(X)) = \mathcal{L}(\theta|X)\ , \end{aligned} \tag{2}$$

with θ, representing all the HMM parameters, omitted on the given variables side of all the quantities involved. $E(X,\theta)$ is the value of the complete-data log-likelihood with the true/maximized parameters θ. $R(Z)$ is the log-likelihood of the hidden data given the observations and has the form of an entropy representing the amount of information brought by the hidden data itself (see eq. (12) for the detailed form of $R(Z)$). As we estimate the complete-data log-likelihood using non-optimized parameters, we have $E(X,\theta,\theta^{old}) \leq E(X,\theta)$, and hence $E(X,\theta,\theta^{old}) - R(Z)$ is a lower bound of the data likelihood.

The key quantity for data likelihood maximization is the expected complete-data log-likelihood which directly depends on the data and is written as

$$E(X,\theta,\theta^{old}) = \sum_Z p(Z|X,\theta^{old})\ln(p(X,Z|\theta))\ . \tag{3}$$

The complete-data likelihood of an observation (the case of multiple observation sequences is addressed later) can be expanded as (eq. 4) that leads by identification to eq. (5).

$$p(X,Z|\theta) = p(h_0)\prod_{t=0}^{T-1} p(h_{t+1}|h_t) \times \prod_{t=0}^{T} p(m_t|h_t)p(x_t|h_t,m_t)\ , \tag{4}$$

$$p(X,Z|\theta) = \pi_{h_0}\prod_{t=0}^{T-1} B_{h_t,h_{t+1}} \prod_{t=0}^{T} C_{h_t,m_t} GD(x_t|h_t,m_t)\ . \tag{5}$$

We substitute eq. (1) into eq. (5) and take the logarithm of the expression. Using the logarithm sum-product property the complete-data log-likelihood is split up into eight terms:

$$\ln(p(X,Z|\theta)) = \ln(\pi_{h_0}) + \sum_{t=0}^{T} \ln(C_{h_t,m_t}) + \sum_{t=0}^{T-1} \ln(B_{h_t,h_{t+1}}) + \sum_{t=0}^{T} \sum_{n=1}^{N} \left\{ \ln(\Gamma(\alpha_{h_t,m_t,n} + \beta_{h_t,m_t,n})) + (\alpha_{h_t,m_t,n} - 1) \ln(x_n^t) + \nu_{h_t,m_t,n} \ln(1 - \sum_{r=1}^{n} x_r^t) - \ln(\Gamma(\alpha_{h_t,m_t,n})) - \ln(\Gamma(\beta_{h_t,m_t,n})) \right\} . \quad (6)$$

Using eq. (6) into eq. (3), the expected complete-data log-likelihood can then be written:

$$E(X,\theta,\theta^{old}) = \sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_{k,m}^{0} \ln(\pi_k) + \sum_{t=0}^{T} \sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_{k,m}^{t} \ln(C_{k,m}) + \sum_{t=0}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi_{i,j}^{t} \ln(B_{i,j}) + L(\alpha,\beta) , \quad (7)$$

with,

$$L(\alpha,\beta) = \sum_{t=0}^{T} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{M} \left\{ \gamma_{k,m}^{t} \ln(\Gamma(\alpha_{k,m,n} + \beta_{k,m,n})) + \gamma_{k,m}^{t} (\alpha_{k,m,n} - 1) \ln(x_n^t) + \gamma_{k,m}^{t} (\nu_{k,m,n} \ln(1 - \sum_{r=1}^{n} x_r^t)) - \gamma_{k,m}^{t} \ln(\Gamma(\alpha_{k,m,n})) - \gamma_{k,m}^{t} \ln(\Gamma(\beta_{k,m,n})) \right\} . \quad (8)$$

To set eq. (7) we make use of the two following properties, in which we omit the mention θ^{old} in the given variables side of the probabilities involved. Using the independence of h_t and m_t from h_{t+1}, we get $p(Z|X) = p(h_t = k, m_t = m|X)p(h_{t+1} = k')$ with $\sum_{k'=1}^{K} p(h_{t+1} = k') = 1$. Similar steps bring $p(Z|X) = p(h_t = k, h_{t+1} = k'|X, m_t = m)p(m_t = m)$, with $\sum_{m=1}^{M} p(m_t = m) = 1$.

Furthermore, if $D \geq 1$ observations are available, all can be used to avoid overfitting. In (7), a sum over $d \in [1, D]$ has to be added in front of the entire formula. The sum over time goes then from 0 to T_d, the length of the d-th observation sequence.

2.2 Update Equations of HMM and GD Parameters

Maximization of the expectation of the complete-data log-likelihood with respect to π, B, and C is solved introducing Lagrange multipliers in order to take into account

the constraints due to the stochastic nature of these parameters. The resulting update equations are:

$$\pi_k^{new} \propto \sum_{d=1}^{D}\sum_{m=1}^{M}\gamma_{k,m}^{0,d}\ , \quad B_{i,j}^{new} \propto \sum_{d=1}^{D}\sum_{t=0}^{T_d-1}\xi_{k,k'}^{t,d}\ , \quad C_{k,m}^{new} \propto \sum_{d=1}^{D}\sum_{t=0}^{T_d}\gamma_{k,m}^{t,d}\ , \tag{9}$$

where k and k' are in the range $[1, K]$, and m, in the range $[1, M]$.

GD distributions parameters update is less straightforward. Indeed, a direct method would lead to maximize $L(\alpha, \beta)$. Instead of going through heavy computations, we propose to use a practical property of the GD distribution that reduces the estimation of a N-dimensional GD to the estimation of N Dirichlet distributions. The latter is a known problem and can be solved using a Newton method [8,19]. Using this property calls the need for the problem to be expressed in a transformed space that we refer to as the W-space. The data is transformed from its original space into its W-space by [5,20]:

$$W_l = \begin{cases} x_l & \text{for } l = 1 \\ x_l \big/ \left(1 - \sum_{i=1}^{l-1} x_i\right) & \text{for } l \in [2, N]\ . \end{cases} \tag{10}$$

In the transformed space, each W_l follows a Beta distribution with parameters (α_l, β_l), which is a 2-dimensional Dirichlet distribution. The estimation of the N Beta distributions governing the N W_l clearly leads to the complete characterization of the GD distribution governing the observation vector $\mathbf{x}$. In the M-step of the HMMGD algorithm, the update of the GD distribution parameters can thus be done using N times a process similar to the one used in [8], considering the transformed data instead of the original one. Other parameters (B, C, π, γ, ξ) are estimated from the original data.

The initialization of the HMM parameters has been shown in [8] to be intractable as soon as the product KM grows up, if computed accurately. Following their framework, KM single Generalized Dirichlet distributions are initialized with a method of moments that uses the transformed data (detailed in [6]) and are then assigned to the HMM states. The parameters π, C, and B, are randomly initialized. Any EM-algorithm is iterative and thus needs a stop parameter. As the data log-likelihood is maximized by the means of its lower bound, convergence of this bound can be used as such. This lower bound is given by $E(X, \theta, \theta^{old}) - R(Z)$ (see eqs. (2) and (7)) and $R(Z)$ is derived using Bayes' rule:

$$\begin{aligned} p(Z|X) &= p(h_0)p(m_0|h_0)\prod_{t=1}^{T} p(h_t|h_{t-1})p(m_t|h_t) \\ &= p(h_0)\frac{p(m_0, h_0)}{p(h_0)}\prod_{t=1}^{T}\frac{p(h_t, h_{t-1})p(m_t, h_t)}{p(h_{t-1})p(h_t)}\ . \end{aligned} \tag{11}$$

Denoting $\eta_t \triangleq p(h_t|X)$ and using the independence properties set earlier, the following expression is derived (see detail in [8], this expression is valid for any type of emission function):

$$R(Z) = \sum_{k=1}^{K} \left[\eta_k^0 \ln(\eta_k^0) + \eta_k^T \ln(\eta_k^T) - 2 \sum_{t=0}^{T} \eta_k^t \ln(\eta_k^t) \right] + \sum_{t=0}^{T} \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{k,m}^t \ln(\gamma_{k,m}^t) + \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{k'=0}^{K} \xi_{i,j}^t \ln(\xi_{i,j}^t) \; . \quad (12)$$

This stands for a unique observation sample, if more are used, a summation over them has to be added in front of the whole expression and the index T has to be adapted to the length of each sequence. At each iteration, the difference between the former and current data likelihoods is computed. Once it goes below a predefined threshold, the algorithm stops and the current parameters values are kept to define the HMMGD. This threshold, empirically fixed to 10^{-6} in our experiments, is a trade-off between estimates precision and computational time.

3 Experiments on Synthetic Data

3.1 Process Description

We propose here to assess the superiority of HMMGD over HMMD with synthetic data. 1000 observations sequences of length randomly taken in the range $[10, 20]$ are generated from a known HMMGD with randomly chosen parameters. The generation of GD samples is described in [20]. The generative state and mixture component are recorded for each sample. As in [8], performance is computed as the proportion of states and mixture components correctly retrieved by an HMM trained on the generated data. Multiple experiments are run varying the number of states K, the number of mixture components M, and the data dimension N. The study of the influence of N is of particular importance as with proportional data, the greater N, the smaller the observation values. Too small values, through numerical processing, can lead to matrices invertibility issues which is not desirable for accurate estimation.

As stated earlier, the GD distribution relaxes the constraint on the sign of the data correlation coefficients. The proposed model is then expected to give a more accurate representation of the data in the case of data mostly positively correlated. On the other hand, with mostly negatively correlated data, HMMD should provide as good results with a reduced complexity. To verify this, we generate data from known HMMGDs and attempt to retrieve the state and mixture component that generated every sample using an HMMGD and an HMMD. We noticed that data generaed from HMMGDs with parameters randomly and uniformly drawn in the range $[1, 60]$, are quasi-automatically mostly positively correlated. To overcome this point we imposed some of the HMM parameters to follow a Dirichlet distribution expressed in the form of a GD distribution. We used the three following scenarii: 1- Data generated from HMMDs only, 2- Data generated from an hybrid HMM with on each state half of the components being Dirichlet and half GD distributions, 3- Data generated from HMMGDs only. Extensive testing confirmed our expectations. Results are illustrated in Figure 1 using a *correlation ratio* which is the number of positively correlated variables (minus the autocorrelations) over the number of negatively correlated ones. A ratio greater than 1 means the variables are mostly positively correlated and vice versa.

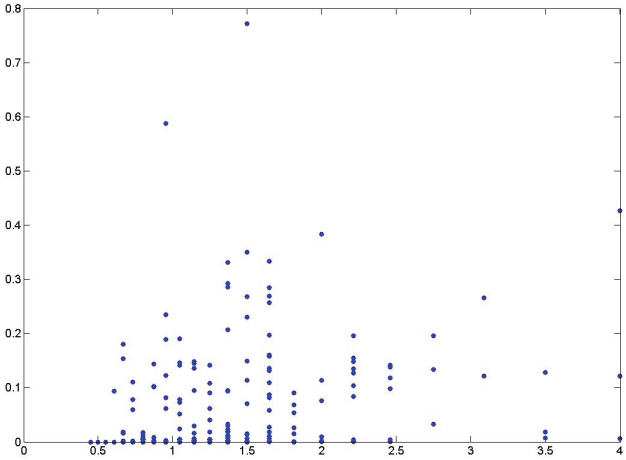

Fig. 1. Gain of accuracy using HMMGD compared to HMMD in function of the variables correlation ratio. The gain of accuracy is computed as the difference between the two models' performance.

Experiments have been led with $K = 3$ and $M = 2$. For scenario 1, HMMGD has a 85.3 % accuracy and HMMD 84.9%, confirming that both work equally well. For scenarii 2 and 3, HMMGD has an accuracy of 81.2% and 89.7%, respectively, and HMMD of 77.6% and 80.1%, respectively. As soon as some data are positively correlated, HMMGD outperforms HMMD. We observe that in scenario 2 (correlation ratio close to 1), for unclear reasons, it is more difficult for the HMMs to retrieve the correct state and component the sample comes from. Finally, the retrieval rate for data with a correlation ratio greater than 1 is of 86.1% for HMMGD and of 78.4% for HMMD, and of 84.8% and 83.2%, respectively, for correlation ratios smaller than 1. This shows HMMGDs overcome the weakness of HMMDs for positively correlated data.

Table 1 reports the results of experiments led fixing $N = 10$, generating 100 sequences only (because of time constraint), and letting K and M vary. According to the previous results, we only consider here mostly positively correlated data. For any combination (K, M), HMMGD achieves better results than HMMD showing the benefit of using HMMGD when proportional data is processed. As the product KM increases, the retrieval rate decreases which can be explained considering that the more distributions, the closer to each other they are, and the more difficult it is to clearly assign a sample to a distribution.

Table 1. HMMGD and HMMD retrieval rates with various (K, M) combinations

Parameters (K,M)	(2,2)	(2,3)	(3,2)	(2,4)	(4,2)	(3,3)	(3,4)	(4,3)	(4,4)	(5,5)	(10,5)
Product KM	4	6	6	8	8	9	12	12	16	25	50
HMMD retrieval rate (%)	84.2	75.9	82.0	82.0	86.2	81.2	72.9	73.3	61.9	66.0	52.4
HMMGD retrieval rate (%)	90.9	92.8	87.8	89.8	91.5	89.1	88.9	85.2	76.6	68.8	62.2

A bad initialization of the distribution parameters can give low retrieval rates. It can find its origin in the convergence of the clustering algorithm, used as the first step of the

method of moments, towards local extrema. To overcome this issue, the initialization process can be run several times and the comparison of the lower bound of the data likelihood with these initial parameters be used to choose the best ones. However, this requires extra computations and does not guarantee a good convergence of the clustering procedure, even within several attempts. As we are only interested here in the relative performance of HMMGD compared to HMMD we did not use this option. Instead, in order not to introduce any bias from this issue, a unique clustering algorithm is used for both initializations.

Figure 2 reports the results of experiments in which we fixed $K = 3$ and $M = 2$ and let N increase until retrieval rates degrade dramatically. For scenario 1, equivalent results are obtained with both HMMs, HMMGD giving sometimes slightly better results at the cost of extra computations (not reported on Figure 2). In other cases, HMMGD systematically outperforms HMMD up to the point data dimension is too high to perform calculations accurately (intermediate matrices become singular). Fluctuations in the overall results are due to bad initializations that involve retrieval rates to dramatically drop on some isolated runs. The general shape of the curves and their relative distance clearly shows that, within an HMM framework, mixtures of GD distributions give the best results and allow working with data of higher dimension than Dirichlet ones. This performance improvement is obtained at the cost of a more complex model involving $(2N - 2)$ parameters to be estimated for every GD distribution compared to only N parameters for a Dirichlet one. These results are essential to target real applications for which HMMGD could be a potentially efficient tool.

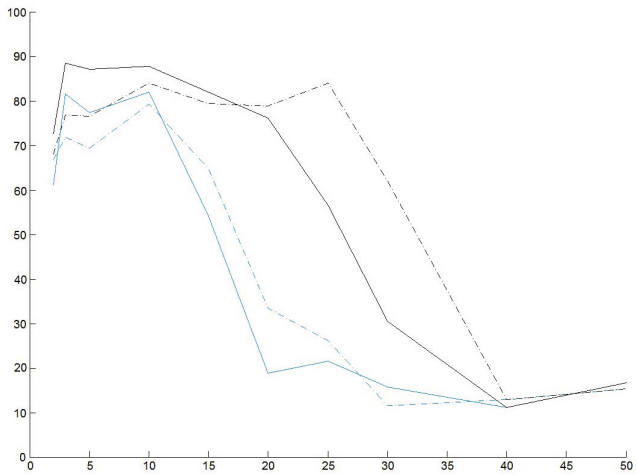

Fig. 2. Retrieval rate (%) of HMMGD (in black) and HMMD (in blue) against data dimension for scenarii 2 (dash lines) and 3 (solid lines)

4 Application to Real Data

We now compare the results of HMMGD and HMMD on real data. We base our experiments on the Weizmann Action Recognition data set [4] which is composed of video sequences representing 10 different actions (such as walk, run, jump,...) performed by 9 subjects. The features we use are Histograms of Oriented Optical Flow [7] and 10-bin histograms are built, with each bin representing a range of optical flow angles with respect to the horizontal axis. The optical flow magnitude weights the contribution of each pixel to the histogram. [7] showed that good classification results could be obtained with features of dimension higher than 30 however, we choose to use features of dimension 10 as, within our HMM-based framework, we did not find any improvement when using more bins. Finally, time savings, we divided the frame rate of the video sequences by 2.

Experiments are led using a Leave-One-Out cross validation, the results are averaged over 10 runs, and analyzed in terms of rank statistics. We empirically determined the optimal values $K = M = 4$ for both HMMs. With these parameters, the HMMD method achieves a 44.0% accuracy while the HMMGD achieves 54.8%. Though these results are low [7], they show the out-performance of HMMGD over HMMD. The rank statistics of order 2 are 71.3% and 82.0% for HMMD and HMMGD, respectively. Here again it is clear that the use of the GD model leads to higher likelihood than the Dirichlet one and is thus much more adapted for real proportional data modeling. Given the small size of the feature vectors (dimension 10) and the huge gap between the rank statistics of order 1 and 2, HMMGD seems to have the potential to perform accurate classification with a parameters fine tuning and the addition of a well-chosen prior.

This last point is supported by the results of the following experiment: we added a very simple prior over the actions of the data set and combined the prior with the already obtained HMMGD results. For each video sequence, the greatest optical flow magnitude is computed. The prior is then based on the average μ_{OF} and standard deviation σ_{OF} of the optical flow magnitude maximum values of the set of video sequences available for each class (i.e. action type). Its computation is totally data-driven, calculated from the training videos available. We make the assumption that, for a given class, this maximal value follows a Gaussian distribution of parameters μ_{OF} and σ_{OF}. As a new video sequence has to be classified, its optical flow maximum magnitude m is computed. The prior is computed as a distance with the following expression:

$$d_{prior} = |\text{CDF}(m, \mu_{OF}, \sigma_{OF}) - 0.5| \,, \tag{13}$$

where $\text{CDF}(m, \mu_{OF}, \sigma_{OF})$ denotes the cumulative distribution function of the Gaussian with parameters μ_{OF} and σ_{OF}. The smallest the value, the highest the prior. The classification is obtained combining this prior result with the HMMGD ones.

Therefore, for a new video sequence, the quantity d_{prior} is computed for each class and a first classification result is obtained and stored. Then, a second classification result is obtained from the HMMGD method described in Section 2. For each class, we add up its rank in the HMMGD and prior results. We then assign the video sequence to the class with the lowest score (i.e. best cumulative rank). This simple prior used alone leads to a classification accuracy less than 50% however, combined with HMMGD results, the

algorithm ends up with a 72.6% accuracy. The rank statistics of order 2 shows an even greater potential as it reaches 91.9%. Better results could be undoubtedly obtained with a more complex prior. However, the study of the best tuning and prior choice is out of the scope of this work that strives at showing the superior performance of HMMGD over HMMD. Figure 3 reports the rank statistics for the three studied methods.

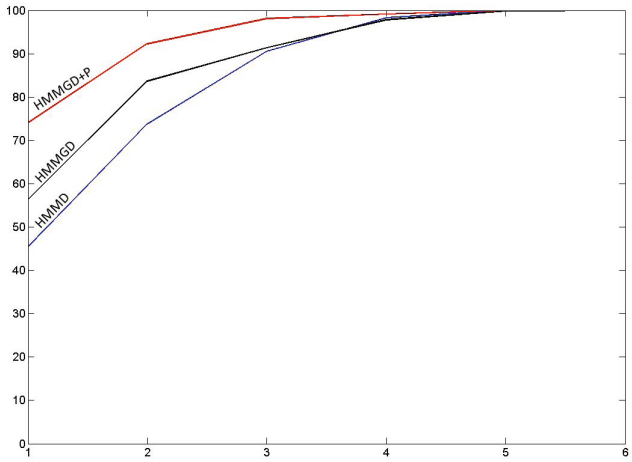

Fig. 3. Rank statistics of HMMD, HMMGD, and of the combination of HMMGD with a prior

5 Conclusion and Future Work

In this paper we theoretically derived a new HMM model for proportional data modeling based on mixtures of GD distributions. We then illustrated how this new model overcomes the limitations of Dirichlet-based HMMs in the case of positively correlated data using synthetic data. An extensive study of the impact of a number of parameters on the model's performance have been presented. Finally, we attempted to use this model on real data for action recognition. Though the first rank classification results are quite low, the study of rank statistics show a certain potential if a fine tuning is found and an appropriate prior used. The dramatic increase in classification accuracy observed when adding a very simple data-driven prior to the HMMGD framework reinforces this assessment. The HMMGD constitutes a new promising alternative when working with proportional data and has definitely to be used over HMMD methods for optimal results. Future work includes the study of HMMGD tuning for better performance and its application to other real-world tasks such as anomaly detection in crowded environment or texture classification.

References

1. Andersson, S., Ryden, T.: Subspace estimation and prediction methods for hidden Markov models. The Annals of Statistics 37(6B), 4131–4152 (2009)

2. Andrade, E.L., Blunsden, S., Fisher, R.B.: Hidden Markov models for optical flow analysis in crowds. In: ICPR (1), pp. 460–463. IEEE Computer Society (2006)
3. Bicego, M., Castellani, U., Murino, V.: A hidden Markov model approach for appearance-based 3d object recognition. Pattern Recogn. Lett. 26(16), 2588–2599 (2005)
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, pp. 1395–1402. IEEE Computer Society (2005)
5. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(10), 1716–1731 (2007)
6. Chang, W.Y., Gupta, R.D., Richards, D.S.P.: Structural properties of the generalized dirichlet distributions. Contemporary Mathematics 516, 109–124 (2010)
7. Chaudhry, R., Ravichandran, A., Hager, G.D., Vidal, R.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: CVPR, pp. 1932–1939. IEEE (2009)
8. Chen, L., Barber, D., Odobez, J.M.: Dynamical dirichlet mixture model. IDIAP-RR 02, IDIAP (2007)
9. Cholewa, M., Glomb, P.: Estimation of the number of states for gesture recognition with hidden Markov models based on the number of critical points in time sequence. Pattern Recognition Letters 34(5), 574–579 (2013)
10. Djuric, P.M., Chun, J.H.: An mcmc sampling approach to estimation of nonstationary hidden Markov models. IEEE Transactions on Signal Processing 50(5), 1113–1123 (2002)
11. Ghahramani, Z., Jordan, M.I.: Factorial hidden Markov models. Machine Learning 29(2-3), 245–273 (1997)
12. Hjalmarsson, H., Ninness, B.: Fast, non-iterative estimation of hidden Markov models. In: Proc. IEEE Conf. Acoustics, Speech and Signal Process, vol. 4, pp. 2253–2256. IEEE (1998)
13. Holzmann, H., Munk, A., Suster, M., Zucchini, W.: Hidden Markov models for circular and linear-circular time series. Environmental and Ecological Statistics 13(3), 325–347 (2006)
14. Jiang, F., Wu, Y., Katsaggelos, A.K.: Abnormal event detection from surveillance video by dynamic hierarchical clustering. In: ICIP (5), pp. 145–148. IEEE (2007)
15. Maruotti, A.: Mixed hidden Markov models for longitudinal data: An overview. International Statistical Review 79(3), 427–454 (2011)
16. Maruotti, A., Rocci, R.: A mixed nonhomogeneous hidden Markov model for categorical data, with application to alcohol consumption. Statist. Med. (31), 871–886 (2012)
17. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
18. Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. IEEE ASSP Magazine 3(1), 4–16 (1986)
19. Ronning, G.: Maximum-likelihood estimation of dirichlet distribution. Journal of Statistical Computation and Simulation 32, 215–221 (1989)
20. Wong, T.T.: Parameter estimation for generalized dirichlet distributions from the sample estimates of the first and the second moments of random variables. Computational Statistics and Data Analysis 54(7), 1756–1765 (2010)