Factorial Hidden Markov Models

Emile Mathieu

Motivations

Hidden Markov model (HMM) is the most used tool for discrete time-series modelling. From a graphical model perspective, HMM is composed by multinomial latent variables modelled by a Markov chain and by observed variables linked to their associated hidden state. HMM's parameters can be estimated with the efficient Baum-Welch algorithm.

The multinomial assumption limits the representational capacity of HMMs, and that is why distributed state representations have been considered, such as in [1]. Such representations can be preferred because the model can automatically decompose the state space into features, and because a priori information about the process' generation can be used. An HMM with a distributed state representation can represent n bits of information with n binary state variables, whereas an HMM would need $K = 2^n$ distinct states. In [2], authors propose efficient learning algorithms for HMMs with a distributed state representation.

Probabilistic Model

Instead of considering a unique Markov chain for the state variables as in HMM, factorial HMM (FHMM) represents the state by a collection of M independent Markov chains, as shown in Figure 1. At time t, observation Y_t can depend on all states variables $S_t = S_t^{(1)}, \dots, S_t^{(M)}$. The joint probability can be factored as:

$$P(\{S_t, Y_t\}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t)$$

$$\text{with } P(S_t|S_{t-1}) = \prod_{m=1}^{M} P(S_t^{(m)}|S_{t-1}^{(m)})$$

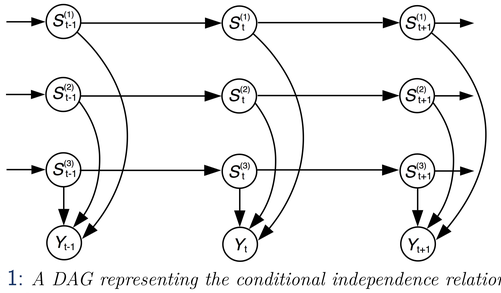


Figure 1: *A DAG representing the conditional independence relations in a factorial HMM*

In [2], for continuous observations $\{Y_t\}$, authors consider Gaussian model whose mean is a linear function of the state variables:

$$P(Y_t|S_t) = |C|^{-\frac{1}{2}}(2\pi)^{-\frac{D}{2}}\exp\left\{-\frac{1}{2}(Y_t-\mu_t)'C^{-1}(Y_t-\mu_t)\right\}$$

$$\text{where } \mu_t = \sum_{m=1}^{M} W^{(m)}S_t^{(m)}$$

Parameters Learning

As for HMM, an EM algorithm is used for learning parameters in FHMM. The maximization step is derived by computing MLE estimators of parameters. Setting the expectation of the complete likelihood's gradient to zero yields closed form solutions for these parameters. The expectation step which consists in computing posterior probabilities over the hidden state is more difficult.

Exact Inference

A forward-backward like algorithm that implement the exact E step has been derived in Appendix B of [2]. This computation is intractable for large K and M due to the unavoidable summation over all possible configurations of other hidden state variables within each time step t.

Gibbs Sampling Inference

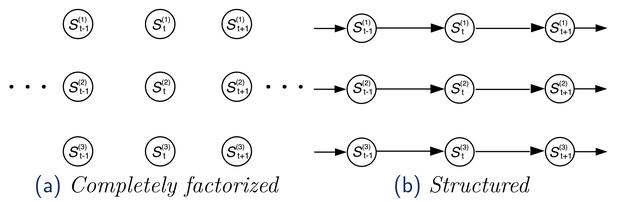
In order to avoid to sum over exponentially many states, one can relate on a Monte Carlo sampling procedure. Gibbs sampling iteratively sample state variables from their conditional distribution which is relatively simple because each node is conditionally independent of all other nodes given its Markov blanket:

$$\begin{aligned} S_t^{(m)} &\sim P(S_t^{(m)}|\{S_t^{(n)} : n \neq m\}, S_{t-1}^{(m)}, S_{t+1}^{(m)}, Y_t, \phi) \\ &\propto P(S_t^{(m)}|S_{t-1}^{(m)})P(S_{t+1}^{(m)}|S_t^{(m)})P(Y_t|S_t^{(1)}, \dots, S_t^{(M)}) \end{aligned}$$

First and second-order statistics needed for the M-step are computed using the states visited during the sampling process.

Variational methods

The main idea of variational methods is to approximate the posterior distribution over the hidden variables $p(\{S_t\}|\{Y_t\})$ by a tractable parametrized distribution $q(\{S_t\}|\theta)$. Parameters θ are then tuned so as to obtain the tightest bound between $p(\{S_t\}|\{Y_t\})$ and $q(\{S_t\}|\theta)$.



(a) *Completely factorized* (b) *Structured*

Figure 2: *DAGs representing the conditional independence relations in variational approximation models*

Completely factorized variational approximation

Mean field considers a completely factorized approximation where there is no more edge between $S_t^{(m)}$ state variables as shown in Figure 2a. Hence, this variational distribution can be factorized as:

$$Q(\{S_t\}|\theta) = \prod_{t=1}^{T}\prod_{m=1}^{M} Q(S_t^{(m)}|\theta_t^{(m)})$$

Structured variational approximation

The structured mean field method relaxes the extreme independence assumption of mean field. Within this scheme, FHMM is approximated by M uncoupled HMM as shown in Figure 2b. Hence, this variational distribution can be factorized as:

$$Q(\{S_t\}|\theta) = \prod_{m=1}^{M} Q(S_1^{(m)}|\theta)\prod_{t=2}^{T} Q(S_t^{(m)}|S_{t-1}^{(m)}, \theta)$$

Experiments - Synthetic data

Goal This experiment aims at comparing different approximate and exact methods of inference on a likelihood basis.

Synthetic data is generated from a FHMM. Parameters are uniformly sampled and normalized when necessary, except the covariance matrix which is set to $C = 0.0025I$. Several values of K and M are considered, and for each problem size, 15 sets of parameters are sampled. For each randomly sampled set of parameters, a separate training set and test set of 20 sequences of length 20 were generated. Algorithms were run on those sets and log likelihoods were computed and averaged over the 15 runs. Results are shown in Table 1.

M	K	Algorithm	Training Set	Test Set
3	2	True	0.00 ± 0.23	0.00 ± 0.25
		HMM	2.16 ± 0.72	2.39 ± 0.66
		Exact	0.90 ± 0.86	0.79 ± 0.79
		SVA	0.69 ± 0.89	0.70 ± 0.92
5	2	True	0.00 ± 0.30	0.00 ± 0.29
		HMM	3.96 ± 0.85	4.01 ± 0.76
		Exact	1.06 ± 1.25	1.36 ± 1.28
		SVA	1.53 ± 1.30	1.09 ± 1.09

Table 1: *Comparison of inference algorithms*

Experiments - Real data

Goal This experiment intends to determine whether the decomposition of the state space in FHMMs can present any advantage.

The dataset consisted of discrete event sequences encoding melody lines of J.S. Bach's Chorales, obtained from the UCI Repository for Machine Learning Databases. Sixty-six chorales truncated to 40 events each, were divided into a training set of 30 chorales and a test set of 36. HMMs with state space ranging from 2 to 100 states were trained until convergence. FHMMs of varying sizes were also trained on the same data with a structured variational approximation for the E step. Test set log likelihoods are shown in Figure 3.

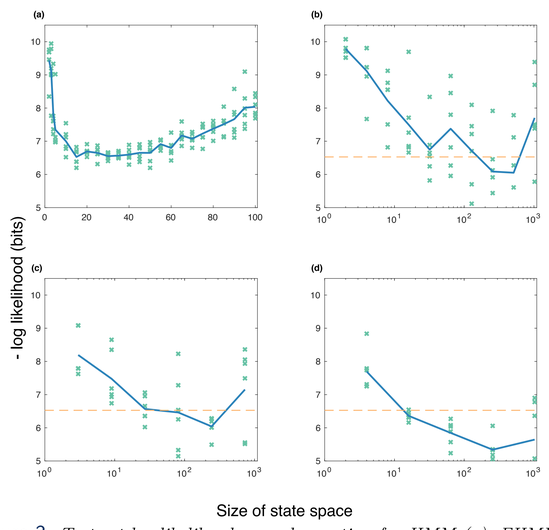


Figure 3: *Test set log likelihoods per observation for HMM (a), FHMM with $K = 2$ (b), $K = 3$ (c) and $K = 4$ (d)*

Conclusion

We have shown that hidden Markov models with distributed state representations provides a richer, and still efficient, modelling tool than classical HMMs.

References

[1] CKI Williams and Geoffrey E. Hinton. Mean field networks that learn to discriminate temporally distorted strings. *Connectionist Models: Proceedings of the 1990 Summer School*, pages 18–22, 1990.

[2] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Mach. Learn.*, 29(2-3):245–273, November 1997.