Scoring unknown speaker clustering: VB vs. BIC

Fabio Valente, Christian Wellekens

Institut Eurecom, Sophia-Antipolis, France

Abstract

This paper aims at comparing the Bayesian Information Criterion and the Variational Bayesian approach for scoring unknown multiple speaker clustering. Variational Bayesian learning is a very effective method that allows parameter learning and model selection at the same time. The application we consider here consists in finding the optimal clustering in a conversation where the speaker number is not a priori known. Experiments are run on synthetic data and on the evaluation data set NIST-1996 HUB-4. VB learning achieves higher score in terms of average cluster purity and average speaker purity compared to ML/BIC.

1. Introduction

A main task in speech recognition systems consists in clustering speakers. Many techniques have been developed for this purpose: most popular approaches include vector quantization [11], Hidden Markov Models (HMM) [2] and Self-Organizing Maps (SOM) [3]. Anyway in many problems, exact speaker number is not *a priori* known. We will refer to this case as unsupervised learning and it is the issue we consider in this paper.

The most famous method for determining the optimal speaker number consists in using a model selection criterion in order to score the model. Generally the *Bayesian Information Criterion* is used to penalize too complex systems.

We consider here a relatively new approach to learning and model selection generally referred as *Variational Bayesian* (VB) learning or ensemble learning. Models like GMM and HMM can be learned using the VB framework (see [4],[10]). Even if VB learning is an approximated method, it has already been successfully applied in speech recognition problems for state clustering [5], dimension reduction [6], and GMM estimation [7]. The main interest of this technique consists in the possibility of doing model selection and parameter learning at the same time.

In [15] we introduced *Variational Bayesian* learning to unsupervised speaker clustering. We exploited VB capacity of self-pruning extra freedom degree that are not used, converging to a model with a smaller number of parameters than the initial model. In this paper we extend the work previously proposed using VB Free Energy (see section 3) for scoring different models and we compare results with those obtained using the BIC criterion.

2. HMM for speaker clustering

A popular approach for automatic speaker clustering uses Ergodic Hidden Markov Models. This method introduced in [1] consider a fully connected HMM in which each state represents a speaker and the state emission probability is the emission probability for each speaker. In order to obtain a non-spare solution, we use a duration constraint of 100 consecutive frames as proposed in [3] and [2] in order to model each speaker in a robust way.

Let us designate α_{rj} the transition probability from state r to state j. We make here the assumption that the probability of transition to state j is the same regardless the initial state i.e. $\alpha_{rj} = \alpha_{r'j} \quad \forall r, r'$, where $j = 1, ..., S$ with S the total number of states; in other words, under this assumption we can model the ergodic HMM as a simple mixture model. Let us designate $[O_1, ..., O_T]$ a sequence of T blocks of D consecutive frames $[O_{t1}, ..., O_{tD}]$ where D is the duration constraint. It is then possible to write the log-likelihood :

$$log\, P(O) = \sum_{t=1}^{T} log\, [\sum_{j=1}^{S} \alpha_j \{ \prod_{p=1}^{D} \sum_{i=1}^{M} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij}) \}] \quad (1)$$

where S represent the number of state (that represent a speaker), M gaussian component model each speaker, and $\{\beta_{ij}, \mu_{ij}, \Gamma_{ij}\}$ represent mixture model parameters (weights, means and gaussians) (for details about this model see [15]).

3. Variational Bayesian Learning

In this section we describe the Variational Bayesian framework that we use in our system. Let us consider a data set $Y = \{y_1, ..., y_n\}$ and a model m, learning algorithms aims at finding optimal model parameters θ that optimize some kind of criterion.

Let us consider the so called marginal likelihood defined as:

$$p(Y|m) = \int d\theta\, p(\theta|m) p(Y|\theta, m) \quad (2)$$

where $p(\theta|m)$ is parameter probability given the model and $p(Y|\theta, m)$ is data likelihood given model and parameters. A simple way to approximate integral in (2) is using a point estimation for parameters θ that gives classical *Maximum Likelihood* (ML) and *Maximum a Posteriori* (MAP) solutions:

$$\theta_{ML} = argmax_\theta\, p(Y|\theta, m) \quad (3)$$

$$\theta_{MAP} = argmax_\theta\, p(\theta|m) p(Y|\theta, m) \quad (4)$$

In other words ML and MAP do not consider the parameter density on all possible parameter domain (like in integral) but just in single point. $p(\theta|m)$ can be approximated with $p(\theta|m, Y)$ i.e. parameter distribution given observations and model. Expression (2) can be computed in an exact way using numerical methods (e.g. Monte-Carlo methods) but when parameter space is huge, the task can be computationally prohibitive. Variational Bayesian learning consists in approximating (2) with a lower bound that makes inference possible using an *Expectation-Maximization*-like (EM) algorithm.

Let us introduce an approximated parameter density (the variational posterior) $q(\theta|Y)$ and let us consider the log marginal-likelihood $log \int d\theta\, p(\theta|m) p(Y|\theta, m)$. Considering Jensen inequality it is possible to write:

$$log\, p(Y|m) = log \int d\theta q(\theta|Y) \frac{p(\theta|m) p(Y, \theta|m)}{q(\theta|Y)}$$

$$\geq \int d\theta q(\theta|Y) log \frac{p(Y, \theta|m)}{q(\theta|Y)} = F(\theta) \quad (5)$$

$F(\theta)$ is called *Free Energy* and it is a strict lower bound on the log marginal-likelihood. Variational Bayesian learning aims at optimizing $F(\theta)$ w.r.t. variational posterior distribution $q(\Theta|Y)$. It is possible to rewrite expression (5) as:

$$F(\theta) = \int d\theta\, q(\theta|Y) log\, p(Y|\theta, m) - D(q(\theta|Y)||p(\theta, m)) \quad (6)$$

Second term in expression (6) represent the KL divergence between variational posterior distributions and parameter prior distributions; it acts as a penalty term that becomes huger for more complex models. In this sense the free energy can be used as a model selection criterion (see section 3.2).

If the variational posterior distribution is constrained to be a delta distribution i.e. $q(\theta|Y) = \delta(\theta - \theta^{'})$, the free energy reduces to the MAP estimator:

$$\begin{aligned} max_{Q(\theta)} F(\theta) &= max_{\theta'} \int \delta(\theta - \theta^{'}) log[p(Y|\theta)p(\theta)]d\theta \\ &= max_{\theta'} log[p(Y|\theta^{'})p(\theta^{'}))] \end{aligned} \quad (7)$$

where the term $\int q(\theta) log\, q(\theta) d\theta$ has been dropped because it is constant. Actually in VB learning, integration is not done w.r.t true posterior distributions but w.r.t approximated variational posterior distributions. The difference between VB and MAP affect of course the way they can be applied ; in MAP a careful estimation for prior distributions must be done because MAP learning will somehow "adapt" prior distributions using current data. In VB framework prior are usually chosen as non-informative as possible because during the learning they are integrated out on all the current domain. For this reason VB should be less sensitive than MAP to prior distributions.

3.1. Variational Bayesian learning with hidden variables

Variational Bayesian learning can be extended to the incomplete data case. In many machine learning problems, algorithms must take care of hidden variables X as well as of parameters θ (see [4]). In the hidden variable case, the variational posterior becomes $q(X, \theta|Y)$ and a further simplification is assumed considering it factorizes as $q(X, \theta|Y) = q(X|Y)q(\theta|Y)$. Then the free energy to maximize is:

$$\begin{aligned} F(\theta, X) &= \int d\theta\, dX\, q(X) q(\theta) log[p(Y, X, \theta)/q(X)q(\theta)] \\ &= < log\, \frac{p(Y, X|\theta)}{q(X)} >_{X,\theta} - D[q(\theta)||p(\theta)] \end{aligned} \quad (8)$$

where $< . >_z$ means average w.r.t. z. Note that q is always understood to be conditioned on Y and m. It can be shown (see [4]) that when $N \rightarrow \infty$ the penalty term reduce to $(|\theta_0|/2) log\, N$ where θ_0 is the number of parameters i.e. the free energy becomes the Bayesian Information Criterion (BIC). To find the optimum $q(\theta)$ and $q(X)$ an EM-like algorithm is proposed in [4] based on the following steps:

$$q(X) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \quad (9)$$

$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta) \quad (10)$$

By iteratively applying eq.(9) and eq.(10) it is possible to estimate variational posteriors for parameters and hidden variables. If $p(\theta)$ belongs to a conjugate family, posterior distribution $q(\theta)$ will have the same form as $p(\theta)$.

An interesting property of VB learning is that extra degrees of freedom are not used i.e. the model prunes itself. There are two possible opinions about the correctness of model self pruning: on the one hand it is not satisfactory because prediction will not take into account uncertainty that models with extra parameters can provide (see [8]), on the other hand it can be used to find the optimal model while learning the model itself, initializing it with a lot of parameters and letting the model prune parameters that are not used.

Let us consider now model in expression (1) and let us define following probability distributions over parameters:

$$\begin{aligned} P(\alpha_j) = Dir(\lambda_{\alpha\,0}) \;\; P(\beta_{ij}) = Dir(\lambda_{\beta\,0}) \\ P(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0\Gamma_{ij}) \;\; P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \end{aligned}$$

where $Dir()$, $N()$, $W()$ are respectively Dirichlet, Normal, Wishart distributions and $\{\lambda_{\alpha\,0}, \lambda_{\beta\,0}, \rho_0, \xi_0, \nu_0, \Phi_0\}$ are hyperparameters.

3.2. Model selection: VB versus BIC

An extremely interesting property of the Variational Bayesian learning is the possibility of doing model selection while training the model. As it was outlined in the previous section, the free energy (6) can be used as a model selection criterion because the KL distance between parameter posterior distributions and parameter prior distributions acts as a penalty term similar to the BIC criterion penalty. We will now consider a more rigorous framework for model selection.

Let us introduce the model posterior probability $q(m)$ on a given model m. It can be shown (see [4]) that optimal $q(m)$ can be written as:

$$q(m) \propto exp\{F(\theta, X, m)\}\, p(m) \quad (11)$$

where $p(m)$ is the model priors. In absence of any prior information on model, $p(m)$ is uniform and optimal $q(m)$ will simply depend on the term $F(\theta, X, m)$ i.e. since higher free energies will result in higher $q(m)$ free energy can be used as model selection criterion. An important advantage is that no threshold must be manually set (as for example in the BIC criterion) but on the other hand in real data problems prior distributions can affect final result. For the model considered here, it is possible to obtain a closed form for the free energy (6) (see [15] for details).

As previously outlined, another interesting point in using Variational Bayesian learning is the capacity of pruning extra freedom degrees. It means that it is possible to initialize the system with a high number of clusters and with a high number of gaussians per speaker and let the system eliminate clusters and gaussians that are not used. In gaussian based model the capacity of pruning extra parameters is somehow regulated by the prior parameter on covariance matrix Φ_0 that seems to be the more sensitive parameter w.r.t. clustering result (see e.g. [6]). In other words, we observe that large values of Φ_0 will result in smaller number of final clusters or smaller number of final gaussians.

In [13] an important point is outlined: when different speakers speak for a different amount of time, it is reasonable to model them with different models. Authors propose to use a BIC criterion to determine the best model between a GMM (that performs better when a lot of training data are available) and a VQ (that performs better when few training data are available). The use of Variational Bayesian learning allows a somehow similar effect; if we initialize speaker models with an initial high gaussian number, VB automatically prunes together with the cluster number, the best gaussian model at the same time, resulting in smaller models where few observations are available and in bigger models where more observations are available.

On the other hand the $BayesianInformationCriterion$ consists of two well separated terms: the first one related to data likelihood and the second one as penalty term. It was shown in

[14] that the following approximation is valid under regularity conditions:

$$log\, p(Y|m) = log\, p(Y|m, \hat{\theta}) - \frac{\nu}{2} log\, N \quad (12)$$

where $\hat{\theta}$ is the Maximum Likelihood estimation for model parameters θ, ν is the free parameter number and N is the observation number. In real data applications, penalty terms is generally multiplied by a threshold value λ heuristically determined.

It is important to notice that in BIC there are two well separated terms, while in the VB the "penalty term" is somehow trained together with the "likelihood term". Now that VB details have been introduced, it is possible to model (1) using the described framework.

4. Experiments

In order to compare the VB model selection and the ML/BIC model selection we run experiments on the evaluation data set NIST-1996 HUB-4 and on some simple synthetic conversation we generated concatenating speech from the TIMIT database. All files are processed in order to obtain 12 LPCC coefficients.

The training procedure uses the following algorithm: the system is initialized with a huge speaker number $M_{initial}$ then optimal parameters are learned using both procedure (VB and ML). Initial speaker number is then reduced progressively from $M_{initial}$ to 1 and parameter learning is done for each new initial speaker number. Optimal speaker number is estimated scoring the different models with VB free energy (that was used as objective function in the training step) and with BIC criterion. It is important to outline that when $M_{initial}$ is big VB prunes to a smaller number of final speaker. Details about estimation formula for the ML and VB learning applied to model (1) can be found in [15]. Results are provided in terms of average cluster purity (*acp*) and average speaker purity (*asp*) and $K = \sqrt{acp \cdot asp}$ (for details see [15]).

4.1. Synthetic conversation

To test our model in a simple framework we generated two artificial conversation concatenating speech from the TIMIT database. In the first conversation (File 1), we concatenated speech from 4 different speakers speak for 10 seconds each. In the second conversation (File 2) 7 speakers speak for a different amount of time resulting in an asymmetric amount of data for each speaker. The system was initialized with 15 speakers modeled with a 6 component GMM each. Duration constraint is 100 frames (1 second). In this preliminary experiment we studied at the same time the dependency of VB from prior distributions and more specifically from parameter Φ_0 and dependence of ML/BIC from the value of λ.

Table 1 shows results on file 1 and file 2. Line (a) shows ML results when the speaker number is a priori known, line (b) shows the best score obtained by the ML system changing speaker number from $M_{initial} = 10$. Line (c) shows results for ML system with BIC selection. Lines (d),(e) and (f) are analogous to lines (a), (b) and (c) but model learning and model selection is done using VB learning.

On file 1, given the simple structure to learn, both algorithms determine the correct speaker number with a $K = 1$ for $\lambda = \{1, ..., 2\}$ and $b_0 = 1, ..., 1000$ where $\Phi_0 = b_0 \cdot I$. On file 2 the strong asymmetry between amount of data for each speaker makes clustering more difficult. The ML/BIC system detects an optimal speaker number of 4 for all $\lambda = \{1, ..., 2\}$ with a $K = 0.81$. On the other hand, this time VB seems to be sensible to prior distribution; the best result is achieved for $b_0 = 200$: 7 speakers and $K = 0.89$. In this case we can observe that the VB selected solution is near to best VB result contrarily to the ML/BIC solution. It is important to notice that the VB framework is very robust to non-informative prior distributions under ideal conditions (i.e. gaussian distribution of data) but in real life applications we tested where this hypothesis is not met, prior choice can heavily affect the final result.

4.2. NIST data set

Analogous experiments were run on the evaluation data set NIST-1996 HUB-4. It consists in 4 files of half an hour each in which speech and non-speech events occur together (music, noise, etc.). Furthermore in some file there is a big difference in amount of speech provided by each speaker that makes the unsupervised learning very difficult. The system is initialized with $M_{initial} = 35$ speakers modeled by a 15 components GMM. Results are shown in table 2 with the same meaning for lines (a)-(f). First of all, VB baseline and best results (lines d-e) are higher than the ML/BIC results (lines a-b) on the first three files while they are almost similar on the last one. It is very important to notice that on the first three files the VB selected model corresponds to the best model; this shows the fact that the VB bound is a very effective metrics for performing model selection. Figures 1 and 2 shows respectively optimal speaker number and optimal K value selected by BIC criterion w.r.t. the BIC threshold λ. Results in 2 refers to values selected using $\lambda = 2$: for this threshold value BIC selected model is near to the best ML model (even if its K score is lower compared to the VB score). In File 1 inferred speaker number is far away from the real speaker number probably because of the fact that a big part of the file is non-speech events that are clustered in many different clusters: anyway final K is high. In File 2 and File 3 inferred speaker number is near to real speaker number (File 2 contains very few non-speech parts). Finally in File 4 BIC infers the right cluster number while VB does not: anyway final K score is the same for BIC and ML. As we outlined in section 3.2, VB should infer the best gaussian component number per cluster together with the best speaker number. Figure 3 plots on a double Y axis graph final gaussian components (left Y axis) and observation number assigned to a cluster (right Y axis). It is easy to notice that small data amount assigned to a cluster results in a smaller number of final gaussian components; on the other hand a huge amount to data results in a model that keep all gaussian components (15 in our case).

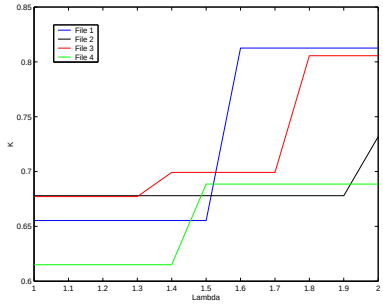

Figure 1: K values inferred by BIC criterion w.r.t λ

5. Conclusion

In this paper we compared Variational Bayesian learning vs. ML/BIC for scoring unknown speaker clustering. Results on artificial conversation and the NIST 1996 HUB-4 evaluation test outlined that VB can outperform ML/BIC. VB has many advantages; first of all together with the optimal speaker number, it can infer the

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
(a) ML	4	1	1	1	7	0.73	0.74	0.73
(b) Best ML	4	1	1	1	9	0.90	0.87	0.88
(c) BIC-ML	4	1	1	1	4	0.65	1	0.81
(d) VB (known)	4	1	1	1	7	0.83	0.95	0.89
(e) VB (best)	4	1	1	1	7	0.87	0.91	0.89
(f) VB (selected)	4	1	1	1	7	0.83	0.95	0.89

Table 1: Results on artificial conversation generated with the TIMIT database data

File	File 1				File 2				File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K
(a) ML (known)	8	0.60	0.84	0.71	14	0.76	0.67	0.72	16	0.75	0.74	0.75	21	0.72	0.65	0.68
(b) ML (best)	10	0.80	0.86	0.83	9	0.72	0.77	0.74	15	0.77	0.83	0.80	12	0.63	0.80	0.71
(c) ML (selected)	13	0.80	0.86	0.83	16	0.84	0.63	0.73	15	0.77	0.83	0.80	21	0.76	0.60	0.68
(d) VB (known)	8	0.70	0.91	0.80	14	0.75	0.82	0.78	16	0.68	0.86	0.76	21	0.60	0.80	0.69
(e) VB (best)	12	0.85	0.89	0.87	14	0.84	0.81	0.82	14	0.75	0.90	0.82	13	0.63	0.80	0.71
(f) VB (selected)	15	0.85	0.89	0.87	14	0.84	0.81	0.82	14	0.75	0.90	0.82	13	0.64	0.72	0.68

Table 2: Results on NIST 1996 HUB-4 evaluation test for speaker clustering

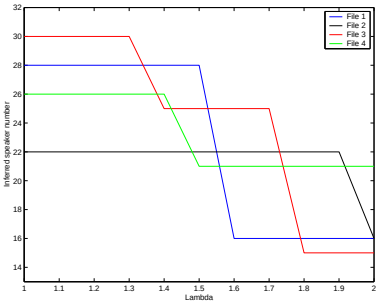

Figure 2: Speaker number inferred by BIC criterion w.r.t. λ

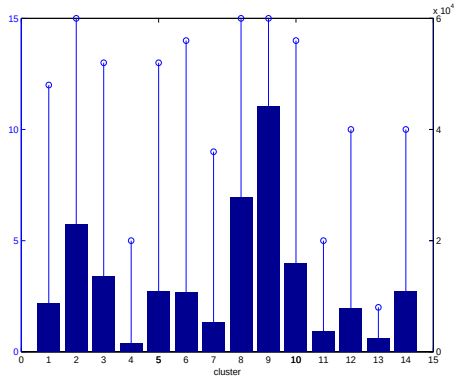

Figure 3: Thick line (left Y axis): final gaussian components vs. cluster number; big line (right Y axis): observation number assigned to a cluster vs. cluster number.

optimal gaussian component number thanks to its property of pruning out extra freedom degree. In this way overfitting problems that may occur using ML learning are avoided. Then "penalty term" for final model decision is trained together with the model without any need for a hand set threshold. On the other hand VB learning on real data shows a certain sensibility to prior that should be set as non-informative as possible. If data are actually distributed following a gaussian distribution, clustering is very robust to priors; if this hypothesis is not met, final result may be very sensitive to priors value.

6. References

[1] Olsen J. O., "Separation of speaker in audio data", EUROSPEECH 1995, pp. 355-358.

[2] Ajmera J., "Unknown-multiple speaker clustering using HMM",ICSLP 2002.

[3] Lapidot I. "SOM as Likelihood Estimator for Speaker Clustering",EUROSPEECH 2003.

[4] Attias, H., "A Variational Bayesian framework for graphical models", Adv. in Neural Inf. Proc. Systems 12, MIT Press,Cambridge, 2000.

[5] Watanabe S. et al. "Application of the Variational Bayesian approach to speech recognition" NIPS'02. MIT Press.

[6] O.-W. Kwon, T.-W. Lee, K. Chan, "Application of variational Bayesian PCA for speech feature extraction," Proc. ICASSP 2002, Orlando, FL, pp. I-825–I-828, May 2002.

[7] Somervuo P., "Speech modeling using Variational Bayesian mixture of gaussians",Proc ICSLP 2002.

[8] MacKay D.J.C. "Local Minima, symmetry breaking and model pruning in variational free energy minimization"

[9] Solomonoff A., Mielke A., Schmidt, Gish H.," Clustering speakers by their voices", ICASSP 98, pp. 557-560

[10] MacKay D.J.C., "Ensemble Learning for Hidden Markov Models"

[11] Cohen A. et Lapidus V. " Unsupervised text independent speaker classification", Proc. of the Eigtheenth Convention of Electrical and Electronics Engineers in Israel 1995, pp. 3.2.2 1-5

[12] Dempster A.P. , Laird N.M. , and Rubin D.B. , "Maximum Likelihood from Incomplete Data via theEM algorithm". Journal of the Royal statistical Society, Series B, 39(1): 1-38, 1977

[13] Nishida M. et Kawahara T. "Unsupervised speaker indexing using speaker model selection based on bayesian information criterion" Proc. ICASSP 2003

[14] Schwartz G. "Estimation of the dimension of a model", Annals of Statistics, 6, 1978

[15] Valente F., Wellekens C. "Variational Bayesian Speaker Clustering", Proc. Odyssey 2004