Variational Bayesian Speaker Change Detection

Fabio Valente, Christian Wellekens

Multimedia department
Institut Eurecom
Sophia-Antipolis BP 193 France
fabio.valente,christian.wellekens@eurecom.fr

Abstract

In this paper we study the use of Variational Bayesian (VB) methods for speaker change detection and we compare results with the classical BIC solution. VB methods are approximated learning algorithms for fully bayesian inference that cannot be achieved in an exact form. They embed in the objective function (also known as free energy) a term that penalizes more complex models. Experiments are run on the Hub4 1996 evaluation data set and show that the VB outperforms the BIC of almost 7%. Anyway as long as the decision must be taken on a limited amount of data the VB based method must be tuned as the BIC based method in order to produce reasonable results.

1. Introduction

In many speech processing systems, an important preliminary task is segmentation in blocks with the same acoustic properties i.e. detection of acoustic change points. The problem is generally formulated as a binary problem in which the competing hypothesis are change or non-change. The change hypothesis is modeled with two different gaussian components while the non-change hypothesis is modeled with a single gaussian. Anyway the two components model will in any case hold an higher likelihood score than the one gaussian model; here comes the need for a model selection criterion to select the best model.

Classical solutions to this problem consists in the use of the Log-Likelihood Ratio (LLR) ([8]) or the Bayesian Information criterion (BIC) ([5],[10]). The LLR criterion simply compare the ratio between likelihoods of one gaussian solution and two gaussian solution with a threshold. In the BIC an asymptotic approximation of the Bayesian integral is used to determine a penalized score. In real data problems anyway the BIC must be adapted to real conditions using an heuristic determined threshold as in the LLR. It makes the BIC in speech application basically a penalized LLR.

Variational Bayesian methods for model selection directly aims at estimating the bayesian integral even though in an approximated form. For this purpose the real posterior distributions over model parameters are substituted with approximated distributions referred as variational distributions that allow a tractable approximation.

In this paper we study the application of VB methods to speaker changing point detection. The paper is organized as follows: in section II we discuss the bayesian model selection and the variational bayesian methods, in section III we formulate the speaker change problem in terms of BIC and VB, in section IV we propose experimental results and finally we discuss them in the conclusions.

2. Bayesian Model selection

Let us consider a data set Y, a model m defined by some parameters θ, the marginal likelihood of the data is defined as:

$$p(Y) = \int p(Y,\theta|m)d\theta = \int p(Y|\theta,m)p(\theta|m)d\theta \qquad (1)$$

where $p(Y,\theta|m)$ is the joint probability of data and model parameters. It is straightforward to notice that the joint probability is proportional to the posterior parameter distribution $p(\theta|Y,m) \propto p(Y|\theta,m)p(\theta|m)$.

The marginal likelihood is itself a quantity that embeds information on the quality of the model i.e. can be used as a model selection criterion. In fact integral (1) benefits from the Occam's razor property (see [1]). In other words expression (1) can be written as the product of two terms: data likelihood computed on a MAP estimation of parameters θ times a penalty term referred to as *Occam factor* that penalizes more complex models.

For example bayesian integral (1) for a single gaussian under conjugate prior is a straight-forward task. Given a gaussian distribution $N(y|\mu,\Gamma)$ under a Normal-Wishart prior for μ and Γ i.e. $p(\mu|\beta\,\Gamma)$ and $p(\Gamma) = W(a,\Phi)$ and a training set y, we have $\int N(y|\mu,\Gamma)p(\mu|\beta\,\Gamma)p(\Gamma) = T(y|a-n+1)$ where T is a t-stud distribution with $a-n+1$ degree of freedom and $\{\beta, a, \Phi\}$ are distribution hyperparamters.

Unfortunately marginal likelihood (1) cannot be computed in close form for complicated models like Gaussian Mixture Models where model contains hidden variables. In those cases approximation of the bayesian integral must be considered. The most simple approximation is the Bayesian Information Criterion (BIC) derived from the Laplace approximation under large data limit and regularity conditions. The BIC is:

$$BIC(Y,m) = log\,p(Y|\hat{\theta},m) - \frac{d}{2}log\,N \qquad (2)$$

where d is the number of free parameters in model m and N is the number of vector in the feature space. Asymptotically (i.e. $N \to \infty$) the BIC converges to the bayesian integral. BIC has a very intuitive explanation, in fact the penalty term becomes huger when the model contains more parameters. Compared to other more effective approximations like the Laplace approximation it has the appealing property of being independent from the basis representation (see [2]). In real data application generally the penalty term is multiplied by an heuristic threshold in order to compensate the lack of data and other weaknesses coming from the approximation.

2.1. Variational Bayesian Learning

In Variational Bayesian learning the goal is directly approximating the posterior distribution $p(\theta|Y,m)$ with a variational posterior distribution referred as $q(\theta)$. Applying Jensen inequality to log-marginal likelihood it is possible to write:

$$log \int p(Y,\theta|m)d\theta = log \int d\theta q(\theta|Y) \frac{p(\theta|m)p(Y,\theta|m)}{q(\theta|Y)} = \geq \int d\theta q(\theta|Y) log \frac{p(Y,\theta|m)}{q(\theta|Y)} = F_m \quad (3)$$

F_m is referred as free energy and constitutes a lower bound to the log-marginal likelihood. VB learning aims at maximizing the free energy w.r.t. the variational posterior distribution $q(\theta)$. Free energy (3) can be rewritten in the following form:

$$F_m = \int q(\theta) log\, p(Y|\theta,m) - KL(q(\theta)||p(\theta)) \quad (4)$$

where the first part is a likelihood term and the second is the KL divergence between variational posterior distributions and prior distributions. Because $KL(q(\theta)||p(\theta)) \geq 0$ by definition, this term behaves like a penalty term that becomes larger when the model contains more parameters but contrarily to the BIC it consider the divergence between prior and posterior distributions over parameters instead of the number of free parameters in the model.

If the model contains a hidden variable set X, a joint variational distribution $q(X,\theta)$ can be defined. In this form the optimal $q(X,\theta)$ cannot be derived in close form and a further approximation must be considered. Assuming the factorization $q(\theta,X) = q(\theta)q(X)$ the free energy can be rewritten as:

$$F_m = \int q(\theta)q(X) log \frac{p(Y|\theta,X,m)}{q(X)} - KL(q(\theta)||p(\theta)) \quad (5)$$

and it is possible to derive an EM-like algorithm (see [3]) for iteratively optimizing (5) w.r.t. $q(X)$ and $q(\theta)$ i.e.

$$q(X) \propto e^{<log\, p(Y,X|\theta)>_\theta} \quad (6)$$

$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta) \quad (7)$$

where $<\,.\,>_Z$ designate the expected value w.r.t. Z.

If prior distributions belong to a conjugate-exponential family, posterior distributions will have the same form with updated hyperparameters. This can be easily seen from the form of the M-like step (7).

As long as the free energy is an approximation of the bayesian integral it can be used as a model selection criterion. Mathematically speaking, a variational approximated distribution over model m referred as $q(m)$ can be defined; applying again Jensen inequality it is possible to write:

$$p(Y) \geq \sum_m [q(m)F_m + log \frac{q(m)}{p(m)}] \quad (8)$$

Optimizing w.r.t. $q(m)$ we obtain:

$$q(m) = exp(F_m)p(m) \quad (9)$$

If the prior $p(m)$ is uniform the decision on the best model is taken according to the best free energy F_m.

3. Speaker change formulation

Let us consider now the speaker change problem formulation from a mathematical point of view. Let us consider a window on which we are interested in studying the changing point $Y = \{y_1, \ldots, y_N\}$ and an hypothesized changing point at y_t. From one side the non-changing point hypothesis consists in modeling Y with a single gaussian with parameters $\theta_{m0} = \{\mu_{m0}, \Gamma_{m0}\}$ while the changing hypothesis is modeled with two gaussians with parameters $\theta_{m1} = \{\mu_{m1}, \Gamma_{m1}\}$ and $\theta_{m2} = \{\mu_{m2}, \Gamma_{m2}\}$ estimated on $Y_1 = \{y_1, \ldots, y_t\}$ and $Y_2 = \{y_{t+1}, \ldots, y_N\}$. The LLR can be written as:

$$LLR = log\, p(Y_1|\theta_{m1}) + log\, p(Y_2|\theta_{m2}) - log\, p(Y|\theta_{m0}) \quad (10)$$

If $LLR > threeshold$ a changing point is detected otherwise there is no speaker change. The BIC can be written as:

$$BIC = log\, p(Y_1|\theta_{m1}) + log\, p(Y_2|\theta_{m2}) - log\, p(Y|\theta_{m0}) - \lambda * \frac{d}{2} * log(N) \quad (11)$$

If $BIC > 0$ a changing point is detected; the heuristic tuning of the parameter λ is essential for adapting the criterion to the real data conditions. Comparing expressions (10) and (11), it is evident that BIC is basically a penalized LLR.

An important issue must be pointed out: the model with two gaussians (one per window) does not have a valid probability density function as already pointed out in [7]. It means that the BIC approximation is not exact from a mathematical point of view because it is not the approximation of a valid pdf. Anyway it is applied with the two gaussian model as if it was a valid model.

In the Bayesian framework the speaker change detection problem may seem more simple. In fact fully bayesian treatment for single gaussians is possible: integration of a gaussian distribution under normal-Dirichlet priors results into a t-student distribution with updated hyperparameters. Anyway the problem requires estimation for a two-gaussian model that have no valid pdf. It means that we cannot simply decompose the two gaussian models as two different gaussians and give a bayesian solution for each single gaussian. Again an approximated solution must be considered.

Basically in the VB framework the situation is analogous to the BIC framework because the integral approximation needs as well a valid probability density function form. The EM-like algorithm described in section 2.1 is applied forcing hidden variables to be 0 or 1 if the data belongs to a gaussian or not. This is of course another approximation in the same fashion as the approximation realized with the BIC and it is not mathematically rigorous.

In the Variational Bayesian formulation at first prior probabilities (see [3]) over parameters must be defined, choosing probability in the conjugate-exponential family we define:

$$p(\mu_{m0}|\Gamma_{m0}) = N(\rho_0|\beta_0\Gamma_{m0}) \quad p(\Gamma_{m0}) = W(a_0,\Phi_0) \quad (12)$$

$$p(\mu_{m1}|\Gamma_{m1}) = N(\rho_0|\beta_0\Gamma_{m1}) \quad p(\Gamma_{m1}) = W(a_0,\Phi_0) \quad (13)$$

$$p(\mu_{m2}|\Gamma_{m2}) = N(\rho_0|\beta_0\Gamma_{m2}) \quad p(\Gamma_{m2}) = W(a_0,\Phi_0) \quad (14)$$

where $W()$ designate a Wishart distribution and $\{\rho_0, \beta_0, a_0, \Phi_0\}$ are distribution hyperparameters. Estimation of variational posterior distributions is easygoing because they have the same form as prior but with updated

hyperparameters. Let us define the following quantities:

$$\bar{\mu}_{m0} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \bar{\mu}_{m1} = \frac{1}{t}\sum_{i=1}^{t} y_i \quad \bar{\mu}_{m2} = \frac{1}{N-t}\sum_{i=t}^{N} y_i \tag{15}$$

$$\bar{\Sigma}_{m0} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{\mu}_{m0})(y_i - \bar{\mu}_{m0})^T \tag{16}$$

$$\bar{\Sigma}_{m0} = \frac{1}{t}\sum_{i=1}^{t}(y_i - \bar{\mu}_{m1})(y_i - \bar{\mu}_{m1})^T \tag{17}$$

$$\bar{\Sigma}_{m2} = \frac{1}{N-t}\sum_{i=t}^{N}(y_i - \bar{\mu}_{m2})(y_i - \bar{\mu}_{m2})^T \tag{18}$$

Variational distributions have the following form:

$$q(\mu_{m0}|\Gamma_{m0}) = N(\rho_{m0}|\beta_{m0}\Gamma_{m0}) \quad q(\Gamma_{m0}) = W(a_{m0}, \Phi_{m0}) \tag{19}$$
$$q(\mu_{m1}|\Gamma_{m1}) = N(\rho_{m1}|\beta_{m1}\Gamma_{m1}) \quad q(\Gamma_{m1}) = W(a_{m1}, \Phi_{m1}) \tag{20}$$
$$q(\mu_{m2}|\Gamma_{m2}) = N(\rho_{m2}|\beta_{m2}\Gamma_{m2}) \quad q(\Gamma_{m2}) = W(a_{m2}, \Phi_{m2}) \tag{21}$$

where

$$\beta_{m0} = N + \beta_0 \quad a_{m0} = N + a_0$$
$$\rho_{m0} = \frac{N\bar{\mu}_{m0} + \beta_0\rho_0}{N + \beta_0}$$
$$\Phi_{m0} = \Phi_0 + T\,\bar{\Sigma}_{m0} + T\,\beta_0(\bar{\mu}_{m0} - \rho_0)(\bar{\mu}_{m0} - \rho_0)^T/(T + \beta_0)$$

Analogous update formula holds for $m1$ and $m2$ hyperparmeters with $N - t$ and t data.

The decision is taken on the difference of the free energies:

$$VB = F_{m12} - F_{m0} \tag{22}$$

where F_{m0} designates the free energy of the mono-gaussian model and F_{m12} designates the free energy of the two gaussian model. In this case both F_{m0} and F_{m12} embed a penalty term: if $VB > 0$ then a speaker change is detected. Actually we should consider the exponent of free energies time the prior over models but as long as prior is uniform and exponential is a monotonal function, simple difference can be considered.

The free energy difference can be written as follows:

$$\begin{aligned} F_{m12} - F_{m0} = &\int q(\mu_{m1}|\Gamma_{m1})q(\Gamma_{m1})\,log\,p(Y_1|\mu_{m1},\Gamma_{m1}) + \\ &+ \int q(\mu_{m2}|\Gamma_{m2})q(\Gamma_{m2})\,log\,p(Y_2|\mu_{m2},\Gamma_{m2}) + \\ &- \int q(\mu_{m0}|\Gamma_{m0})q(\Gamma_{m0})\,log\,p(Y|\mu_{m0},\Gamma_{m0}) + \\ &- KL(q(\mu_{m1}|\Gamma_{m1})||p(\mu_{m1}|\Gamma_{m1}))) - KL(q(\Gamma_{m1})||p(\Gamma_{m1})) \\ &- KL(q(\mu_{m2}|\Gamma_{m2})||p(\mu_{m2}|\Gamma_{m2}))) - KL(q(\Gamma_{m2})||p(\Gamma_{m2})) \\ &+ KL(q(\mu_{m0}|\Gamma_{m0})||p(\mu_{m0}|\Gamma_{m0}))) + KL(q(\Gamma_{m0})||p(\Gamma_{m0})) \end{aligned} \tag{23}$$

All elements in expression (23) admit a close form; for a generic gaussian with variational distributions $N(\mu_m, \beta_m\Gamma_m)$ and $W(a_m, \Phi_m)$ and an observation y it is possible to write:

$$\int q(\mu_m|\Gamma_m)q(\Gamma_m)\,log\,p(y|\mu_m,\Gamma_m) =$$
$$log\,\tilde{\Gamma} - (y - \mu_m)^T\bar{\Gamma}(y - \mu_m) - d/2\beta_m \tag{24}$$

First moment and first log moment for Wishart distribution can be written as:

$$log\,\tilde{\Gamma} = \sum_{i=1}^{d}\Psi((a_m + 1 - i)/2) - log|\Phi_m| + d\,log2 \tag{25}$$

$$\bar{\Gamma} = < \Gamma_m >= a_m\Gamma_m^{-1} \tag{26}$$

where $< . >$ designates the expected value, d is the dimension of the acoustic vector and Ψ is the digamma function.

KL divergences for Wishart and Normal distributions admits a close form (see for instance [9])

In the initialization step the VB method needs prior distributions i.e. an initial value for hyperparameters. It has been found that when amount of data is large, final result is not sensitive to the initial hyperparameters (see [4]). Anyway speaker change is generally estimated on extremely limited amount of data and for this reason we expect a dependency on the prior distribution.

4. Experimental framework

Experiments are run on the HUB-4 1996 evaluation set that consists of 4 files of almost half an hour each for a total of more than 500 speaker changing points from different speakers and in different conditions. Feature extractions consists in 12 MFCC coefficients estimated with a sliding window of 20ms shifted each 10ms.

4.1. Search algorithm

In order to compare the BIC and the VB solutions in the fairest way a common experimental framework is fixed. The search algorithm used in our experiments is the same of [5]: it consists in two neighboring window shifted and resized.

1 Initialize the window $[a, b]$ where $a = 0$ and $b = MIN - WINDOW$

2 Find the changing point in $[a, b]$

for BIC find the point of local maxima of $BIC(m) \geq 0$

for VB find the point of local maxima of $VB(m) \geq 0$

3 if no change is detected in $[a, b]$ then $b = b + MORE - FRAMES$

else if t is the detected changing point in $[a, b]$ then $a = t + 1, b = a + MORE - FRAMES$

4 if $b - a > MAX - WINDOW$ then $a = b - MAX - WINDOW$

5 go to point 2

Value of $MORE - FRAMES$ is experimentally set to 1 second and value of $MAX - WINDOW$ is set to 10 seconds. The changing point is determined using a BIC or a VB method and dependency on the parameter λ and on the prior distribution is studied.

The prior distribution is initialized thought tying different hyperparameters as it is proposed in [6]:

$$a_0 = \beta_0 = \tau \quad \Phi_0 = \tau\,I \quad \rho_0 = \bar{y} \tag{27}$$

where I designates the identity matrix and $\bar{y}$ is the mean of the observation vector. Performances are plotted as function of the τ.

The evaluation metric is the very classical metric in those cases that consider a type I error also referred as precision (PRC) and a type II error also referred as recall (RCL) and their

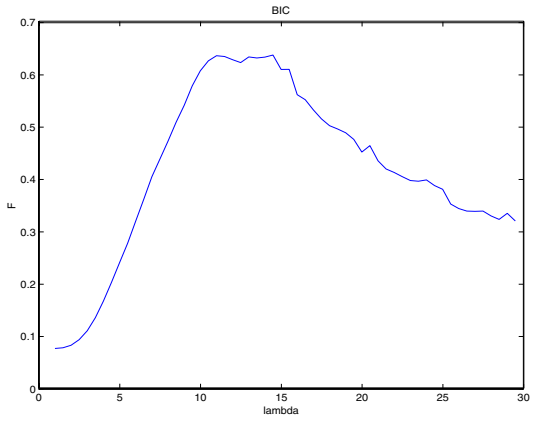

Figure 1: BIC/F score versus λ

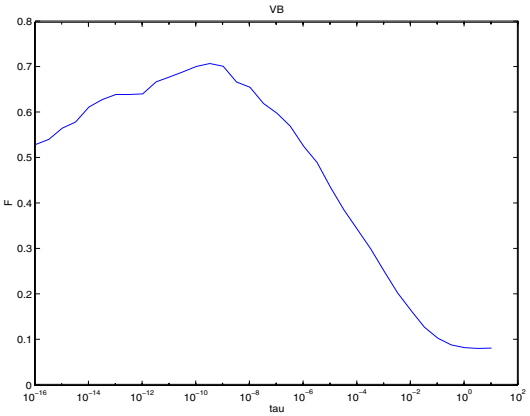

Figure 2: VB/F score versus τ

global measure F defined as:

$$PRC = \frac{\text{number of correctly found changes}}{\text{total number of changes}} \quad (28)$$

$$RCL = \frac{\text{number of correctly found changes}}{\text{total number of correct changes}} \quad (29)$$

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \quad (30)$$

As long as there is a consistent portion of non-speech data in those files, in the final score we do not consider changes that have place in the middle of non-speech segments

	PRC	RCL	F
BIC $\lambda = 14.5$ (best)	0.64	0.63	0.63
VB $\tau = 1E-10$ (best)	0.75	0.66	0.70

Table 1: Value of PRC, RCL and F for the best tuned BIC and the best tuned VB

Figure 1 plots the F value for the BIC system w.r.t. λ on a linear scale while figure 2 plots the F value for the VB system w.r.t. τ on a logarithmic scale. Table 1 shows the best result for the VB and for the BIC methods.

A low value of λ results in a large number of speaker change points providing an high RCL and a low PRC; when the value of λ increases the score result in high PRC and low RCL. The situation is inverted in the case of hyperparameter τ: a high τ produces more false alarms and on the other side when τ is reduced many real speaker change points are missed.

As first remark we can notice that the VB method for speaker change detection is extremely sensitive to the prior distribution initialization. Anyway the best VB system clearly outperforms the best BIC system of almost 7% in absolute value.

This is due to the more efficient approximation the VB method can do compared to the BIC. In fact even if VB is not an exact method, it is still bayesian and offers definitely a finer tuning at the model level. Furthermore the embedded penalty term in the VB is not simply a modified threshold like in the BIC but explicitly consider of the divergence between posterior and prior distributions.

5. Conclusions

In this paper we have formulated the speaker changing point problem as a model selection problem based on fully approximated bayesian method referred as Variational Bayesian learning. Speaker changing score evaluated in term of the F function is increased of about 7% in absolute value. The VB method also needs to be finely tuned as the BIC, but the tuning is done at the prior distribution level ensuring a better approximation of the bayesian integral contrary to to BIC that offers a very rough approximation.

6. References

[1] McKay D.J.C. “Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks”, Network:Computation in Neural System 6,1995

[2] MacKay D. J. C. ”Choice of basis for Laplace approximation” Machine Learning 33(1) 1998

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

[4] Watanabe S. et al. “Application of the Variational Bayesian approach to speech recognition” NIPS’02. MIT Press.

[5] Chen S. and Gopalakrishnan P. “Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion”Proceedings of the DARPA Workshop 1998.

[6] Gauvain J. and Lee C. “Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains”,IEEE Transactions SAP volume 2 pag. 291–298 1994.

[7] Ajmera J., “Robust audio segmentation”, PhD thesis, EPFL - IDIAP 2004

[8] Gish H., Siu M.H. and Rohlicek R. ”Segregation of speakers for speech recognition and speaker identification” Proceedings of ICASSP’91 1991

[9] Penny W.,”Kullback-Liebler divergences of Normal,Gaussian, Dirichlet and Wishart densities”,Wellcome Department of Cognitive Neurology, 2001

[10] Delacourt P. and Wellekens C. “DISTBIC : A speaker-based segmentation for audio data indexing “ Speech Communication, Volume 32 N1-2 -2000 , pp 111-126