Redefining the Bayesian Information Criterion for Speaker Diarisation

Themos Stafylakis\textsuperscript{1,2}, Vassilis Katsouros\textsuperscript{1}, George Carayannis\textsuperscript{1,2}

\textsuperscript{1}Institute for Language and Speech Processing, Athena Research Center, Greece
\textsuperscript{2}National Technical University of Athens, Greece
\{themosst,vsk,gcara\}@ilsp.gr

Abstract

A novel approach to the Bayesian Information Criterion (BIC) is introduced. The new criterion redefines the penalty terms of the BIC, such that each parameter is penalized with the effective sample size rather than the overall sample size. This change eliminates the dependency on the overall sample size and gives rise to partitions of higher cluster purity.

1. Introduction

Speaker diarisation is the problem of segmenting an audio document according to speaker turns and regrouping the derived segments, such that each cluster corresponds to a single speaker. The Global-BIC, proposed by Stafylakis et al.\textsuperscript{3}, is a widely adopted criterion deployed to detect (or verify) the boundaries of speaker turns. The Local-BIC, on the contrary, has no such dependency and is not restricted to hierarchical clustering algorithms. As analyzed in [5], the marginal likelihood (or evidence) pr(\(\mathcal{X}|\mathcal{H}\)) = \int pr(\mathcal{X}|\theta, \mathcal{H})pr(\theta|\mathcal{H})d\theta of a model \(\mathcal{H}\) given the data \(\mathcal{X} = [x_1, \ldots, x_i]^T, x_i \in \mathbb{R}^d\) is the quantity that naturally embodies the Occam’s razor, i.e. the factor that penalizes the over-complex models [5].

2. The BIC in clustering tasks

2.1. Model-Based Clustering and Evidence

Model comparison is the problem of assessing the parsimonious order of a family of models that best describes the data. The BIC was introduced as a means to infer the order of regression, multi-step Markov chains, normal linear models, etc. It approximates the logarithmic evidence of a model by expanding it as quadratic about its posterior mode [1]. This method is known as the Laplace approximation [5], and the regularity conditions that should be met in order to be valid can be found in [6]. Bayesian analysis is a powerful statistical paradigm that overcomes the limitations of maximum likelihood (ML) techniques. As analyzed in [5], the marginal likelihood (or evidence) pr(\(\mathcal{X}|\mathcal{H}\)) = \int pr(\mathcal{X}|\theta, \mathcal{H})pr(\theta|\mathcal{H})d\theta of a model \(\mathcal{H}\) given the data \(\mathcal{X} = [x_1, \ldots, x_i]^T, x_i \in \mathbb{R}^d\) is the quantity that naturally embodies the Occam’s razor, i.e. the factor that penalizes the over-complex models [5].

Model-based clustering can be divided into two distinct problems. The first problem is semi-parametric density estimation, i.e. infer the parsimonious order of a family of models that exhibit the optimal generalization performance to unseen data. Such problems are usually based on soft clustering techniques, where the ML estimate is evaluated as an average of all possible partitions of the data, weighted by their posterior probabilities. The second category is cluster analysis and is the one that speaker diarisation belongs to. The goal is to achieve a one-to-one mapping between an unknown number of sources and clusters. A crucial distinction between the two problems is how one evaluates the best-fit likelihood of a model \(\mathcal{H}\) of \(m\) clusters (i.e. how the hypothesis space is defined). In cluster analysis, the competing hypotheses are the partitions and not the number of clusters. Therefore, the log-likelihood is evaluated only on a single partition of the data, i.e. conditioned on the latent variables (or cluster indicators) \(\gamma_i \in \{1, \ldots, m\}\).

\[
\mathcal{L}(\hat{\theta}, \gamma|\mathcal{X}) = \sum_{i=1}^{n} \log f_{\gamma_i}(x_i|\hat{\theta}_{\gamma_i})
\] (1)

where \(\hat{\theta} = [\hat{\theta}_1^T, \ldots, \hat{\theta}_m^T]^T\). Note that \(\hat{\theta}_{\gamma_i}\) denotes the ML estimate of \(\theta_{\gamma_i}\) given \(\mathcal{X}_{\gamma_i} = [x_i: \gamma_i = k]|\mathcal{H}\) and is placed instead of the posterior estimate \(\hat{\theta}_{\gamma_i}\) as its large-sample limit [1]. The quantity in (1) is known as the classification or marginal log-likelihood, which can be derived using the Classification EM (CEM) [7].

However, speaker diarisation severely differs from the class of problems that the CEM aims to deal with. In order to
achieve a speaker-oriented clustering, one should take into account the Markovian dynamics, that bound the desired partition of the data away from the unrestricted ML estimate given \( n \), which tends to be rather phone-level oriented. Compared to the phone-level inference, the intra-class covariance reduces in a much slower rate as \( m \) grows, which in turn means that the Hessian \( -\nabla^2 \log p(x|\theta, H) \) grows with \( m \) more moderately on average, i.e. the posterior distributions of \( \{\theta_k\}_{k=1}^m \) are not so sharply peaked around the ML estimates.

The above analysis shows that the straightforward use of the BIC in speaker diarisation may not be the most adequate solution. Furthermore, the Global-BIC [2] formulae

\[
BIC^{\text{GT}} = 2\mathcal{L}(\hat{\theta}; \gamma|X) - \lambda m \log n
\]

where \( m_H = \sum_{k=1}^m P_k \) and \( P_k = \dim(\theta_k) \), leads to ambiguities about \( n \) and might exhibits uncontrollable behavior.

### 2.2. Speaker diarisation and BIC

A distinction of the speaker diarisation systems arises from whether the segmentation and clustering stages are arranged in a chain or in a more coupled manner. In the former category, the segmentation stage generates an initial partition of the data which feeds a HC algorithm. Usually, the segmentation is performed with a two-pass algorithm, i.e. over-segmentation with a low-cost metric followed by the verification of change points with the BIC [8], [3]. Thus, estimating whether two segments belong to the same speaker or not is the core problem in such algorithms. Although the experiments are based on this scheme, we emphasize that the proposed criterion is neither restrictive to decoupled algorithms, nor to the HC approach, and a broad family of unsupervised clustering algorithms (Genetic, Simulated Annealing, E-HMMs, etc.) may utilize it as an objective function. For an overview of existing diarisation systems we refer to [9].

We further drop the subscripts from \( f_k(\cdot|\theta_k) \) and \( P_k \) and use the multivariate Gaussian model with full covariance matrix, i.e. \( \theta_k = (\mu_k, \Sigma_k) \) and \( P = d + (d + 1)/2 \). Let \( X_a, X_b \) be two segments (or clusters) of \( n_a \) and \( n_b \) sample sizes respectively. The Global \( \Delta \text{BIC} \) measure used amongst speaker and audio diarisation community is

\[
\Delta BIC^{\text{GT}} = 2\mathcal{L}(\hat{\theta}_a|X_a) + \mathcal{L}(\hat{\theta}_b|X_b) - \mathcal{L}(\hat{\theta}_{ab}|X_{ab}), \text{ i.e. the logarithmic generalized likelihood ratio.}
\]

### 2.3. The weaknesses of Global and Local BIC

We first examine the behaviour of Global-BIC. As one may note, Global \( \Delta \text{BIC} \) suffers from an incoherent incoherence. Two segments may be merged or not according to the overall sample size \( n \). Severe ambiguities arise from the above penalty term. Many algorithms categorize speaker utterances according to gender and bandwidth. Which sample size should be interpreted as \( n \), the category-dependent or the overall? How can we tune \( \lambda \) such that the desired trade-off is achieved? The tuning would always be conditioned on \( n \). What if \( n \) is unknown a priori, as happens with on-line algorithms? The questions above raise the need of an autonomous and more controllable dissimilarity measure, i.e. a \( \Delta \text{BIC} \) criterion that depends only on the sufficient statistics of the two segments (or clusters) and their sample sizes. The Local-BIC is such a criterion and is widely adopted as a measure for both segmentation and clustering tasks [3]. The corresponding formulae of the Local-BIC is derived by placing \( n_a + n_b \) instead of \( n \) in (3). Several experiment have shown the superiority of the local variant against the global one, [9], [3]. The problem with Local-BIC is that it only exists in \( \Delta \text{BIC} \) formule, and therefore it does not approximate the evidence of overall clustering hypotheses. The evidence maximization is restricted to score pairwise distances. Consequently, several optimization algorithms that require an objective function cannot make use of it. Therefore, the main criticism about the local variant is that by not defining a global statistical quantity to optimize, the derived algorithms do not make full use of the Bayesian paradigm as a means to infer the optimal partition of the data.

### 3. The Segmental-BIC approach

#### 3.1. The proposed formulae with normal priors

The proposed criterion evaluates the likelihood in the same way as the Global BIC. The difference is on how we penalize the complexity of the model. By utilizing the fact that the likelihood is conditional on a fixed partition \( \gamma \) under \( H \), we penalize each \( \theta_k \) only with the corresponding effective sample size. Therefore, we end-up with a summation of penalty terms

\[
BIC^{S} = 2\mathcal{L}(\hat{\theta}; \gamma|X) - \lambda P \sum_{k=1}^m \log n_k
\]

rather than the Global-BIC penalty \( m \lambda P \log n \).

More formally, we approximate the evidence of a clustering hypothesis \( H \), using normal unit-information priors (see [4]), centered at the MLE \( \theta \) and having scale matrix equal to the inverse of the expected information arising from one observation per cluster. (Actually, we use \( n_k^{-1} \) observations to built the prior of the \( k \)th cluster). To emphasize the distinction with the current approaches, we name the proposed criterion the Segmental-BIC.

Contrary to the Global-BIC, the proposed criterion is cluster-oriented, in the sense that the ratio of the posterior to prior uncertainty (i.e. the Occam’s factor) about \( \theta_k \in \mathbb{R}^P \) is proportional to \( n_k^{-1} \lambda P \) and therefore independent of \( n \). Recall that in cluster analysis \( \theta \) is a ML estimate conditional on a single partitioning \( \gamma \) of the data. Consequently, the observed information \( I_\theta(\theta) = -\nabla^2 \log p(x|\theta, H) \) is block-diagonal, allowing us to decompose it as \( I_\theta(\theta) = \sum_{k=1}^m P_k \mathcal{H}_k(\hat{\theta}_k) \), where \( \mathcal{H}_k(\hat{\theta}_k) \) denotes the class-conditional expected information matrix

\[
(\mathcal{H}_k(\hat{\theta}_k))_{ij} = \int p(x|\hat{\theta}_k) \left( - \frac{\partial^2 \log p(x|\hat{\theta}_k)}{\partial \theta_i \partial \theta_j} \right) dx
\]

This gives a rationale for how the cluster-oriented normal unit-information priors

\[
\theta_k \sim \mathcal{N}(\hat{\theta}_k, n_k^{-1} \mathcal{H}_k^{-1}(\hat{\theta}_k))
\]

asymptotically cancel out the remaining terms of the Laplace approximation of the model’s evidence, leaving only those appearing in (4). Note also that \( \lambda \) becomes the hyperparameter that controls the rate to which the covariance of the prior grows with \( n_k \). By placing \( \lambda = 1 \), one assumes a fixed prior (i.e. independent of \( n_k \)), with covariance matrix equal to the inverse information carried in a single observation.

Moreover, note that even when \( \lambda = 1 \), the amount of information we use to form the priors is much more moderate, when
The criterion falsely judges the two sets as being from different speakers. The same process is repeated, but with utterances of different speakers between the two sets. A type-II error is occurred when the criterion judges the two sets as being from the same speaker. The feature space consists of 18-order static mfcc, while \( c_0 \) is discarded. The results are illustrated in Fig. 1 for various values of \( \lambda \). Clearly, the Segmental-BIC outperforms the dominant approach of Local-BIC. The Global-BIC performance is not evaluated, since \( n \) has no coherent meaning in the particular experiment.

The next level of experiments is based on two Speaker Diarization Benchmark tests. The features used are 18-order mfcc, augmented by the log-energy. The metric scores are the average cluster purity (acp) and the average speaker purity (asp) as in [10]. Both feature extraction and algorithm implementation are based on the open-source software provided by the LIUM Laboratory and analyzed in [3] and [11]. The algorithm used is the agglomerative HC. No Viterbi resegmentation has been applied. Note also that many systems append a MAP-adapted GMM scheme to improve the asp derived from the HC stage, that aims to merge those clusters being from the same speaker but recorded under different acoustic environment [3], [9]. Such implementations require long cluster durations (asp) that are as pure as possible (acp). Since false cluster mergings are irreversible errors, the tuning of \( \lambda \) should be biased in favour of asp, i.e. of models that overfit the data. A similar biased tuning is applied to diarisation systems that operate as a module of an open-set speaker identification system. Therefore, a fair comparison is to measure the relative increase in asp for fixed asp, where \( \text{asp} > \text{asp} \).

The first data-set we examine is the 2002-Rich Transcription of NIST. The Broadcast News data is composed of six approximately 10-minute excerpts from six different broadcasts. The result are shown in Fig. 2 for various values of \( \lambda \) (ranging from 1 to 9).

The second benchmark is the ESTER speaker diarisation corpus [11]. The corpus is divided into the development and test set. The former consists of about 8 hours audio material (14 French Radio shows), ranging from 8 to 60 minutes). The asp - asp curves are illustrated in Fig. 3. The test set consists of 18 broadcasts and the overall duration is approximately 10 hours. The asp - asp curves are illustrated in Fig. 4. Note that the performance of the Segmental-BIC with Jeffreys’s priors is not illustrated in Fig. 4, due to the overlap that exhibits with the Segmental-BIC with normal ones.

| Table 1: Minimum Overall Speaker diarization error (%) for the three sets we examined |
|---------------------------------|-----------------|-----------------|
|                                | NIST-02         | ESTER-DEV       | ESTER-TEST |
| Local-BIC                      | 12.99           | 18.29           | 19.47      |
| Segm-BIC                       | 12.71           | 17.90           | 20.05      |
| Segm-BICc                      | 12.71           | 17.65           | 20.05      |

The minimum overall diarisation error rates (DER, %) for each criterion are shown in Table 1. However, we strongly suggest to tune \( \lambda \) via the asp - asp trade-off; DER is a 1-D lossy score and, therefore, suboptimal for tuning the parameters (especially those of intermediate stages, [10]). Furthermore, the DER counts only the overlap between a reference and the system speaker that matched best. Thus, the DER is invariant to the excessive fragmentation of a reference speaker into more
than one system speakers. However, the excessive fragmentation causes objectively further degradation of the system’s performance because it corresponds to lower levels of coverage. The \( \text{aep} \times \text{asp} \) curves show that the Segmental-BIC outperforms the Global-BIC for the entire range of operational points that corresponds to \( \text{aep} > \text{asp} \). The Local-BIC exhibits inferior performance to the Segmental-BIC on both the NIST RT-02 and the ESTER development set, while the two criteria have comparable performance on the ESTER test set.

5. Conclusions

We introduce a new version of the BIC, which aims to score competing partitions of the data with respect to speaker diarization. What motivated us to define such a criterion was to discard the dependency posed by the Global-BIC, that is attached to the model via the rather conservative implied priors. We believe that such a document-oriented prior is more adequate when inferring the natural (phoneme-level) partitions of the data and assuming a fixed loss for guessing the wrong one. Instead of defining a local dissimilarity measure as the Local-BIC suggests, we adopt a more informative strategy and use cluster-oriented unit-information priors. The experiments show significant improvement in performance, especially when the decision-theoretic utility is asymmetric and favours the purity of the partitions more than their coverage.

6. Acknowledgements

This work is funded by the Greek General Secretariat of Research and Technology under the program PENED-03/251.

7. References