UNSUPERVISED LEARNING OF VISUAL CONTEXT USING COMPLETED LIKELIHOOD AIC

Tao Xiang[1], Shaogang Gong[1]

[1] Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
Email:{txiang,sgg}@dcs.qmul.ac.uk, Fax: 020-8980-6533

Keywords: Learning for vision, visual context, model selection, dynamic scene modelling, mixture model.

Abstract

Learning visual context is a critical step of dynamic scene modelling. This paper addresses the problem of choosing the most suitable probabilistic model selection criterion for learning visual context of a dynamic scene. A Completed Likelihood Akaike's Information Criterion (CL-AIC) is formulated to estimate the optimal model order (complexity) for a given visual scene. CL-AIC is designed to overcome poor model selection by existing popular criteria when the data sample size varies from very small to large. Extensive experiments on learning visual context for dynamic scene modelling are carried out to demonstrate the effectiveness of CL-AIC, compared to that of BIC, AIC and ICL.

1 Introduction

The problem of dynamic scene understanding can be tackled based on building models for various activities occurring in the scene [3, 9, 12, 24, 16]. Learning scene-specific visual context is a critical step of this model-based dynamic scene understanding approach, which reduces the complexity of activity models and makes them tractable given limited visual observations. Visual context is scene specific. It is thus defined differently according to the nature of different visual tasks. For example, the visual context of a scene can be a semantically meaningful decomposition of spatial regions for human behaviour interpretation [16, 3], or a decomposition of prototypic facial expressions for facial expression recognition [22]. We consider the problem of learning visual context as modelling the underlying structure of activity captured in a dynamic scene. To this end, we model visual context using mixture models based on automatic model order selection.

In this paper, we address the problem of choosing the most appropriate probabilistic criteria for model selection according to the nature of visual data. Existing probabilistic model selection criteria can be classified into two categories: (1) methods based on approximating the Bayesian Model Selection criterion [17], such as Bayesian Information Criterion (BIC) [20], Laplace Empirical Criterion (LEC) [19], and the Integrated Completed Likelihood (ICL) [2]; (2) methods based on the information coding theory such as the Minimum Message Length (MML) [8], Minimum Description Length (MDL) [18], and Akaike's Information Criterion (AIC) [1]. The performance of various probabilistic model selection criteria has been studied intensively in the literature [19, 8, 2, 17, 4, 10], which motivated the derivation of new criteria. In particular, a number of previous works were focused on mixture models [19, 8, 2]. However, most previous studies assume the sample sizes of data sets to be sufficiently large in comparison to the number of model parameters [19, 8, 2], except for a few works that focused on linear autoregression models [4, 10]. This is convenient due to the fact that the derivations of all existing probabilistic model selection criteria involve approximations that can only be accurate when the sample size is sufficiently large, ideally approaching infinity. Existing criteria for mixture models are also mostly based on known model kernels, e.g. Gaussian. Realistically, visual data available for dynamic scene modelling are always sparse, incomplete, noisy and with unknown model kernels. Therefore, existing model selection criteria based on previous studies may not be suitable for learning visual context given the nature of visual observations commonly available.

In the rest of the paper, we propose a novel probabilistic model selection criterion to improve model estimation for data sets with unknown distribution kernel functions and severe overlapping among mixture components. Mixture models are briefly described in Section 2. Bayesian Information Criterion (BIC) is widely used for determining the model order of a mixture model [16, 9, 24], which is identical in formulation to Minimum Description Length (MDL). It is shown by our experiments (see Sections 4 and 5) that BIC tends to under-fit when the sample size is small and tends to over-fit when the sample size is large. Integrated Completed Likelihood (ICL) was proposed in [2] to solve this problem. Nevertheless, ICL performs poorly when data belonging to different mixture components are severely overlapped. We argue that to overcome these problems with the existing criteria, we need to optimise *explicitly* the explanation and prediction functionalities of a mixture model through a model selection criterion. To this end, we introduce in Section 3 a Completed Likelihood AIC (CL-AIC) criterion, which aims to give the optimal clustering of the given data set and best predict unseen data. In Section 4, we analyse through synthetic data experiments how the performance of CL-AIC are affected by two factors: (1) the sample size, and (2) whether and how the true kernel functions are different from the assumed ones. Extensive experiments are also presented in Section 5 to demonstrate the effectiveness of CL-AIC on learning visual context for dynamic scene understanding, compared to that of BIC, AIC and ICL. A conclusion is drawn in Section 6.

2 Mixture Models

Suppose a D-dimensional random variable $\mathbf{y}$ follows a K-component mixture distribution, the probability density function of $\mathbf{y}$ can be written as $p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k p(\mathbf{y}|\boldsymbol{\theta}_k)$, where w_k is the mixing probability for the kth mixture component with $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$, $\boldsymbol{\theta}_k$ is the internal parameters describing the kth mixture component, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K; w_1, \ldots, w_K\}$ is a C_K dimensional vector describing the complete set of parameters for the mixture model. Let us denote N independent and identically distributed samples of $\mathbf{y}$ as $\mathcal{Y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(N)}\}$. The log-likelihood of observing $\mathcal{Y}$ given a K-component mixture model is

$$\log p(\mathcal{Y}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(\log \sum_{k=1}^{K} w_k p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k) \right), \tag{1}$$

where $p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k)$ defines the model kernel for the k-th component. In this paper, the model kernel functions for different mixture components are assumed to have the same form. If the number of mixture components K is known, the Maximum Likelihood (ML) estimate of model parameters, given by $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{\log p(\mathcal{Y}|\boldsymbol{\theta})\}$, can be computed using the EM algorithm [6]. Therefore the problem of estimating a mixture model boils down to the estimation of K, known as the model order selection problem. A K-component mixture model is thereafter denoted as $\mathcal{M}_K$.

3 Completed Likelihood Akaike's Information Criterion (CL-AIC)

Given a data set $\mathcal{Y}$, a mixture model $\mathcal{M}_K$ can be used for three objectives: (1) estimating the unknown distribution that most likely generates the observed data, (2) clustering the given data set, and (3) predicting unseen data. Objectives (1) and (2) emphasise data explanation while objective (3) is concerned with data prediction. Model selection criteria based on approximating the Bayesian Model Selection criterion [17], such as Bayesian Information Criterion (BIC) [20] and Laplace Empirical Criterion (LEC) [19], choose the model that maximises $p(\mathcal{Y}|\mathcal{M}_K)$, the probability of observing a data set $\mathcal{Y}$ given a candidate model $\mathcal{M}_K$. They thus enforce mainly objective (1). When the true distribution kernel functions are very different from the assumed ones, all these criteria tend to choose models with the number of components larger than the true number of clusters in order to approximate approximate the unknown distribution more accurately. To better balance the explanation and prediction capabilities of a mixture model, we derive a novel model selection criterion, referred as CL-AIC. CL-AIC utilises Completed Likelihood (CL), which makes explicit the clustering objective of a mixture model, and follows a derivation procedure similar to that of AIC, which chooses the model that best predict unseen data.

Let us first formulate Completed Likelihood (CL). The complete data for a K-component mixture model is a combination of the data set and the labels of each data sample:

$$\bar{\mathcal{Y}} = \{\mathcal{Y}, \mathcal{Z}\} = \left\{ (\mathbf{y}^{(1)}, \mathbf{z}^{(1)}), ..., (\mathbf{y}^{(N)}, \mathbf{z}^{(N)}) \right\},$$

where $\mathcal{Z} = \left\{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(n)}, ..., \mathbf{z}^{(N)}\right\}$, and $\mathbf{z}^{(n)} = \left\{ z_1^{(n)}, ..., z_K^{(n)} \right\}$ is a binary label vector such that $z_k^{(n)} = 1$ if $\mathbf{y}^{(n)}$ belongs to the kth mixture component and $z_k^{(n)} = 0$ otherwise. $\mathcal{Z}$ is normally unknown, and must be inferred from $\mathcal{Y}$. The completed log-likelihood of $\bar{\mathcal{Y}}$ is:

$$\begin{aligned} CL(K) &= \log p(\mathcal{Y}|\boldsymbol{\theta}) + \log p(\mathcal{Z}|\mathcal{Y}, \boldsymbol{\theta}) \qquad (2)\\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} w_k p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} z_k^{(n)} \log p_k^{(n)} \end{aligned}$$

where $p_k^{(n)}$ is the conditional probability of $\mathbf{y}^{(n)}$ belonging to the kth component and can be computed as:

$$p_k^{(n)} = \frac{w_k p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k)}{\sum_{i=1}^{K} w_i p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_i)}. \tag{3}$$

In practice, the true parameters $\boldsymbol{\theta}$ in Equation (3) is replaced using the ML estimate $\hat{\boldsymbol{\theta}}$ and the completed log-likelihood is rewritten as:

$$CL(K) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \hat{w}_k p(\mathbf{y}^{(n)}|\hat{\boldsymbol{\theta}}_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} \tag{4}$$

where

$$\hat{z}_k^{(n)} = \begin{cases} 1 & \text{if} \arg\max_j \hat{p}_j^{(n)} = k \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

CL-AIC aims to choose the model that gives the best clustering of the observed data and has the minimal divergence to the true model, which thus best predicts unseen data. The divergence between a candidate model and the true model is measured using the Kullback-Leibler information [14]. Given a complete data set $\bar{\mathcal{Y}}$, we assume that $\bar{\mathcal{Y}}$ is generated by the unknown true model $\mathcal{M}_0$ with model parameter $\boldsymbol{\theta}_{\mathcal{M}_0}$. For any given model $\mathcal{M}_K$ and the Maximum Likelihood Estimate $\hat{\boldsymbol{\theta}}_{\mathcal{M}_K}$, the Kullback-Leibler divergence between the two models is computed as

$$d(\mathcal{M}_0, \mathcal{M}_K) = E\left[\log \left(\frac{p(\bar{\mathcal{Y}}|\mathcal{M}_0, \boldsymbol{\theta}_{\mathcal{M}_0})}{p(\bar{\mathcal{Y}}|\mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K})} \right) \right]. \tag{6}$$

Ranking the candidate models according to $d(\mathcal{M}_0, \mathcal{M}_K)$ is equivalent to ranking them according to $\delta(\mathcal{M}_0, \mathcal{M}_K) = E\left[-2 \log p(\bar{\mathcal{Y}}|\mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}) \right]$. $\delta(\mathcal{M}_0, \mathcal{M}_K)$ cannot be computed directly since the unknown true model is required. However, it was noted by Akaike [1] that $-2 \log p(\bar{\mathcal{Y}}|\mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K})$ can serve as a biased approximation of $\delta(\mathcal{M}_0, \mathcal{M}_K)$, and the bias adjustment $E\left[\delta(\mathcal{M}_0, \mathcal{M}_K) + 2 \log p(\bar{\mathcal{Y}}|\mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}) \right]$ converges to $2C_K$

when the number of data sample approaches infinity. Our CL-AIC is thus derived as:

$$CL\text{–}AIC = -\log p(\bar{\mathcal{Y}}|\mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}) + C_K. \qquad (7)$$

where C_K is the dimensionality of the parameter space. The first term on the right hand side of (7) is the completed likelihood given by Equation (4). We thus have:

$$\begin{aligned} CL\text{–}AIC &= -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \hat{w}_k p(\mathbf{y}^{(n)}|\hat{\boldsymbol{\theta}}_k) \\ &\quad - \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} + C_K. \end{aligned} \qquad (8)$$

The first and third terms on the right hand side of Equation (8) emphasise the prediction capability of the model, while the second term, favouring well separated mixture components, enforces the explanation capability of the model. This formulation results in a number of important differences compared to existing techniques:

1. Unlike previous probabilistic model selection criteria, CL-AIC attempts to optimise *explicitly* the explanation and prediction capabilities of a model. This makes CL-AIC theoretically attractive. The effectiveness of CL-AIC in practice is demonstrated through experiments in Sections 4 and 5.

2. Compared to a standard AIC, our CL-AIC has an extra penalty term (the second term on the right hand side of Equation (8)) which always assumes a positive value. This extra penalty term makes CL-AIC in favour of smaller K compared to AIC given the same data set. It has been shown that AIC tends to over-fit by both theoretical [7, 13] and experimental studies [21, 11]. The extra penalty term in CL-AIC thus has the effect of rectifying the over-fitting tendency of AIC.

3. Completed likelihood has been combined with BIC which leads an Integrated Completed Likelihood (ICL) criterion [2]. However, reported experiments in [2] indicated that ICL performs poorly when data belonging to different mixture components are severely overlapped. We suggest this is caused by the factor that ICL is a combination of two explanation oriented criteria without considering the prediction capability of a mixture model. To that end, CL-AIC integrates an explanation criterion and a prediction criterion. It is thus theoretically better justified than ICL.

4 Experiments on Synthetic Data

In this section, we illustrate the effectiveness of our CL-AIC, compared to that of existing popular model selection criteria including AIC, BIC and ICL, using synthetic data. Experiments on learning visual context of three different real scenarios are presented in Section 5. The experiments presented in this section aim to examine how the performance of different criteria is affected by the following two factors: (1) the sample size and (2) how different the true kernel functions are from the assumed ones. To this end, Gaussian mixture models were adopted while synthetic data sets were generated using non-Gaussian kernels with sample size varying from very small to large in comparison to the number of model parameters. To simulate the real world data, data belonging to different mixture components were severely overlapped. Moreover, our synthetic data were unevenly distributed among different mixture components.

Models with the number of components K varying from 1 to K_{max}, a number that is considered to be safely larger than the unknown true number K_{true}, were evaluated. In our experiments, K_{max} was 10 unless otherwise specified. To avoid being trapped at local maxima, the EM algorithm used for estimating model parameters $\boldsymbol{\theta}$ was randomly initialized for 20 times and the solution that yielded the largest observation likelihood after 30 iterations were chosen. Each Gaussian component was assumed to have full covariance. Different model selection criteria were tested on the data sets with sample sizes varying from 25 to 1000 in increments of 25. The final model selection results are illustrated using the mean and ± 1 standard deviation of the selected number of components over 50 trials, with each trial having a different random number seed.

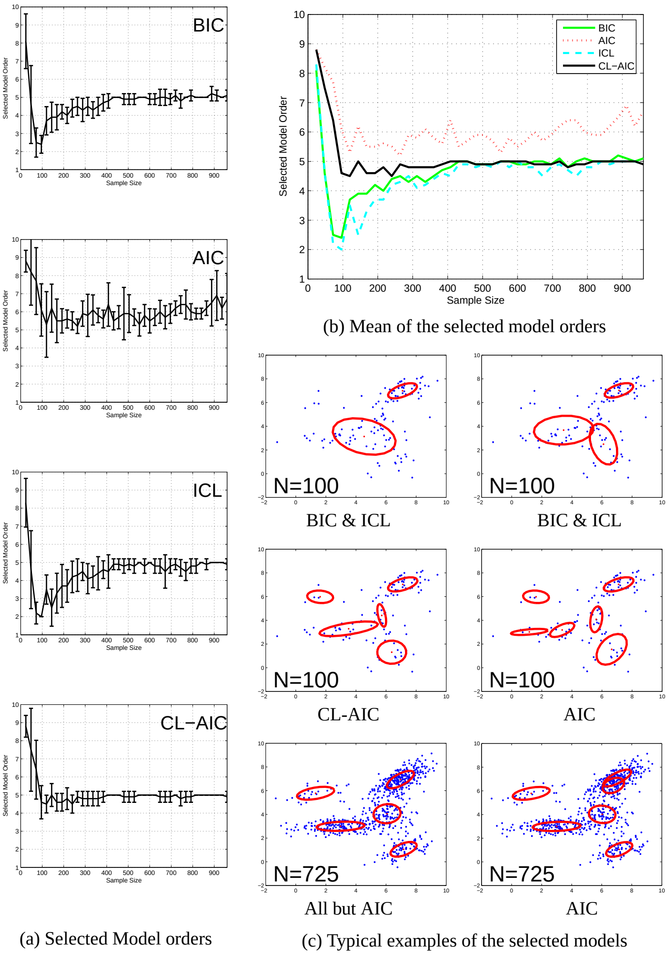
We first consider a situation under which the assumed kernel functions are different from the true one, but not by too much. A data set was firstly generated according to a Gaussian mixture distribution whose parameters are:

$$\begin{array}{l} w_1 = 0.05, w_2 = 0.10, w_3 = 0.20, w_4 = 0.40, w_5 = 0.25; \\ \boldsymbol{\mu}_1 = [1.5, 6.0]^T, \boldsymbol{\mu}_2 = [7.0, 1.0]^T, \boldsymbol{\mu}_3 = [6.0, 4.0]^T, \\ \boldsymbol{\mu}_4 = [7.0, 7.0]^T, \boldsymbol{\mu}_5 = [3.0, 3.0]^T; \boldsymbol{\Sigma}_1 = \left[\begin{array}{cc} 1.89 & 0.25 \\ 0.25 & 0.50 \end{array} \right], \\ \boldsymbol{\Sigma}_2 = \left[\begin{array}{cc} 0.72 & 0.14 \\ 0.14 & 0.34 \end{array} \right], \boldsymbol{\Sigma}_3 = \left[\begin{array}{cc} 0.99 & 0.04 \\ 0.04 & 0.65 \end{array} \right], \\ \boldsymbol{\Sigma}_4 = \left[\begin{array}{cc} 1.78 & 0.46 \\ 0.46 & 0.42 \end{array} \right], \boldsymbol{\Sigma}_5 = \left[\begin{array}{cc} 1.97 & 0.05 \\ 0.05 & 0.10 \end{array} \right], \end{array} \qquad (9)$$

where w_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mixing probability, mean vector and covariance matrix for the kth Gaussian component respectively. The data were then perturbed with uniformly distributed random noise. The noise had a range of $[-0.5 \ \ 0.5]$ in each dimension of the data distribution space. The model selection results are presented in Figure 1 and table 1.

	BIC	AIC	ICL	CL-AIC
100	0	10	0	**48**
725	88	64	82	**100**

Table 1: Percentage of correct model order selection (over 50 trials) by different criteria for synthetic Noisy Gaussian data with 100 and 725 samples respectively.



(a) Selected Model orders

(b) Mean of the selected model orders

BIC & ICL

BIC & ICL

CL-AIC

AIC

All but AIC

AIC

(c) Typical examples of the selected models

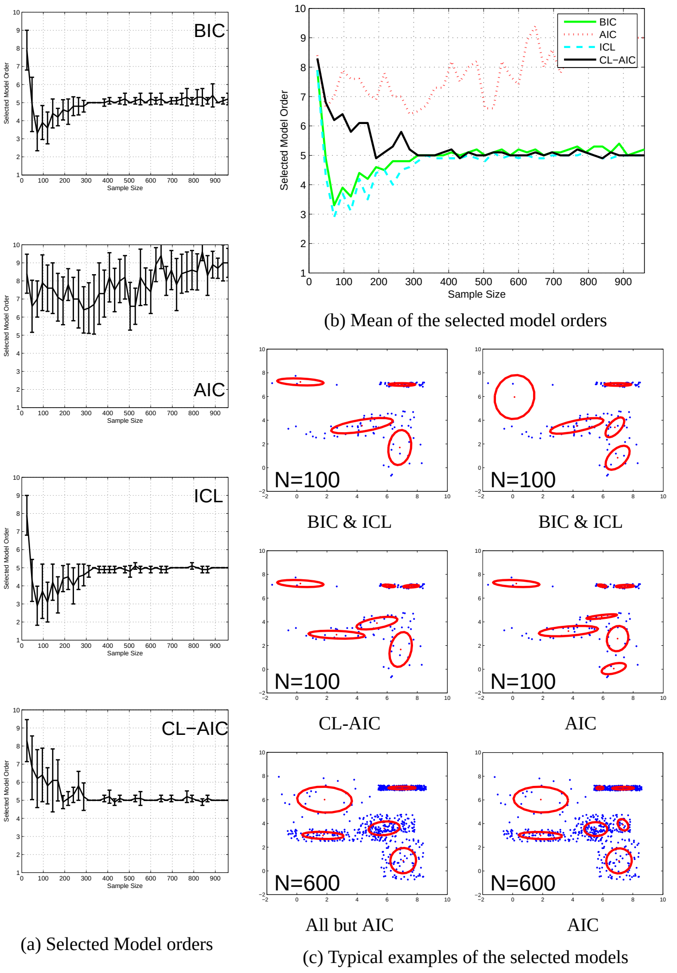
Figure 1: Model selection results for synthetic Gaussian data synthetic Gaussian data perturbed with uniformly distributed random noise.

Figures 1(a) and (b) show how the performance of different criteria were affected by the sample size of the data set. When the data set was sampled extremely sparsely (e.g. $N < 50$), all 5 criteria tend to over-fit. As the sample size increased, the number of components determined by all the criteria decreased. In particular, BIC, ICL, and CL-AIC all turned from over-fitting to under-fitting before converging towards the true component number, with the number of components selected by CL-AIC being the closest to the true number 5. It is noted that when the sample size is large (e.g. $N > 500$), BIC tended to over-fit slightly. The over-fitting tendency of BIC when the assumed kernels are different form the true ones was also reported in [2]. Overall, AIC appears to favor larger number of components even when the sample size is large. It is also noted that AIC exhibited large variations in the estimated model order no matter what the sample size was, while other criteria had smaller variation given larger sample sizes.

We then consider an extreme case where the true kernel functions are completely different from the assumed ones. A synthetic 2D data set were generated with data from each components following the uniform random distribution:

	BIC	AIC	ICL	CL-AIC
100	4	2	8	**10**
600	86	4	94	**100**

Table 2: Percentage of correct model order selection (over 50 trials) by different criteria for synthetic uniform data with 100 and 600 samples respectively.



(b) Mean of the selected model orders

BIC & ICL

BIC & ICL

CL-AIC

AIC

All but AIC

AIC

(a) Selected Model orders

(c) Typical examples of the selected models

Figure 2: Model selection results for synthetic data of uniform distribution.

$$u_{\mathbf{r}}(y_1, y_2) = \begin{cases} \frac{1}{(r_2 - r_1)\cdot(r_4 - r_3)} & \text{if } r_1 \leq y_1 \leq r_2 \\ & \& \ r_3 \leq y_2 \leq r_4 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{r} = [r_1, r_2, r_3, r_4]$ are the parameters of the distribution. Our data set was generated using a 5-component uniform mixture model. Its parameters are:

$$w_1 = 0.05, w_2 = 0.10, w_3 = 0.20, w_4 = 0.40, w_5 = 0.25;$$
$$\boldsymbol{r}_1 = [-1.89, 4.07, 4.89, 7.94], \boldsymbol{r}_2 = [5.58, 8.42, -0.77, 2.77],$$
$$\boldsymbol{r}_3 = [4, 17, 7.83, 2.23, 5.77], \boldsymbol{r}_4 = [5.41, 8.59, 6.79, 7.21],$$
$$\boldsymbol{r}_5 = [-0.61, 6.61, 2.47, 3.53].$$

The model selection results are presented in Figure 2 and Table 2. It can be seen from Figure 2(b) that with a small sample size (e.g. $50 < N < 200$), BIC and ICL tended to under-fit while AIC and CL-AIC tended to over-fit. As the sample size increased, BIC slightly over-fitted and ICL slightly under-fitted, while CL-AIC yielded the most accurate results. Again, AIC exhibited large variations in the estimated model order no matter what the sample size was, while other criteria had smaller variation given larger sample sizes. It is also noted that AIC suffered from severe over-fitting and failed to converge.

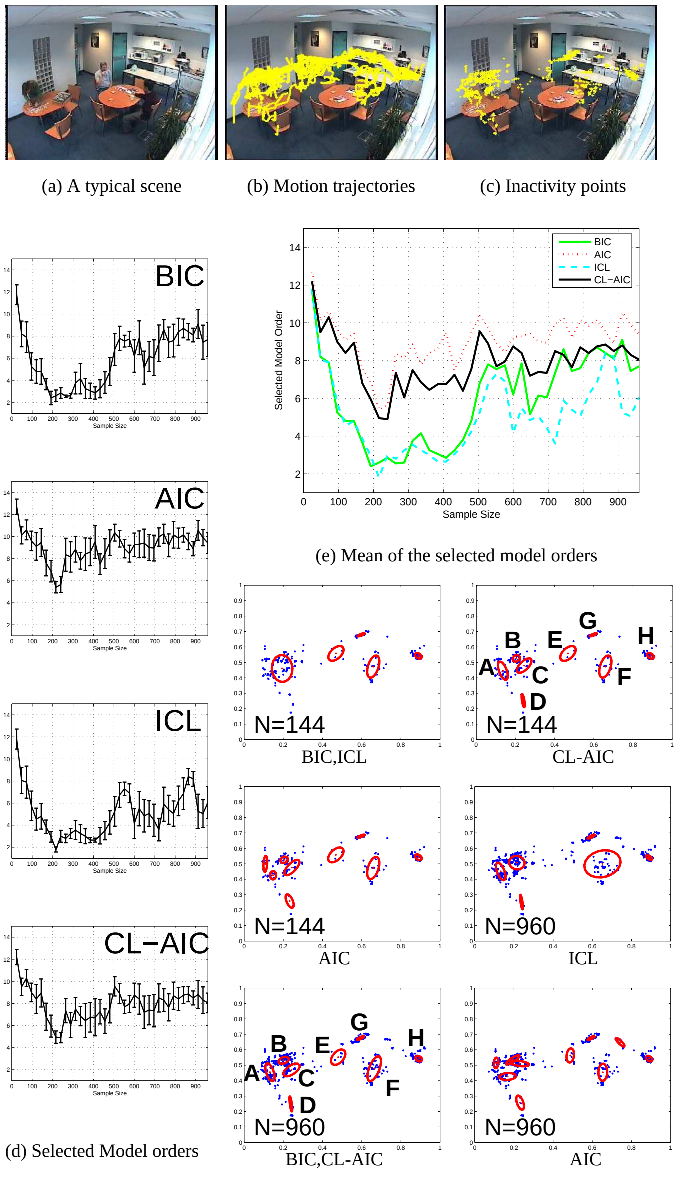
Our experiments show that CL-AIC outperforms BIC, AIC and ICL when the true kernel functions are different from the assumed ones and the sample size varies from small to large. Our experiments also indicate that all criteria tend to over-fit given extremely sparse data (e.g. $N < 2C_{K_{true}}$ where $C_{K_{true}}$ is the number of parameters of the true model). Given a very small sample size, none of the mixture components is supported well by the data. Data samples belonging to the same mixture component tend to be interpreted as being drawn from different mixture components. This explains the over-fitting tendency for all the model selection criteria. Our experiments suggest that the more the true kernel functions differ from the assumed ones, the more likely it is for BIC to over-fit and ICL to under-fit even with large sample size. On the other hand, CL-AIC utilises both the explanation and prediction capacities of a mixture model. It is thus able to yields better model estimation especially when the sample size is moderate or sufficiently large.

5 Learning Visual Context

Experiments were conducted on learning visual context of three different dynamic scene modelling problems. Gaussian mixture models were adopted in our experiments while the true model kernels were unknown and clearly non-Gaussian by observation. The model estimation results were obtained by following the same procedure as that of the synthetic data experiments presented in the preceding section, unless otherwise specified.

5.1 Learning Spatial Context

A tearoom scenario was captured at 8Hz over three different days of changeable natural lighting, giving a total of 45 minutes (22430 frames) of video data. Each image frame has a size of 320×240 pixels. The scene consists of a kitchenette on the top right hand side of the view and two dining tables located on the middle and left side of the view respectively (see Figure 3(a)). Typical activities occurring in the kitchenette area included people making tea or coffee at the work surface, and people filling the kettle or washing up in the sink area. Other activities taking place in the scene mainly involved people sitting or standing around the two dining tables while drinking, talking or doing the puzzle. In total 66 activities were captured, each of them lasting between 100 and 650 frames. It is noted that the same activities performed by different people can differ greatly.


(a) A typical scene
(b) Motion trajectories
(c) Inactivity points
(d) Selected Model orders
(e) Mean of the selected model orders

(f) Typical examples of the selected models

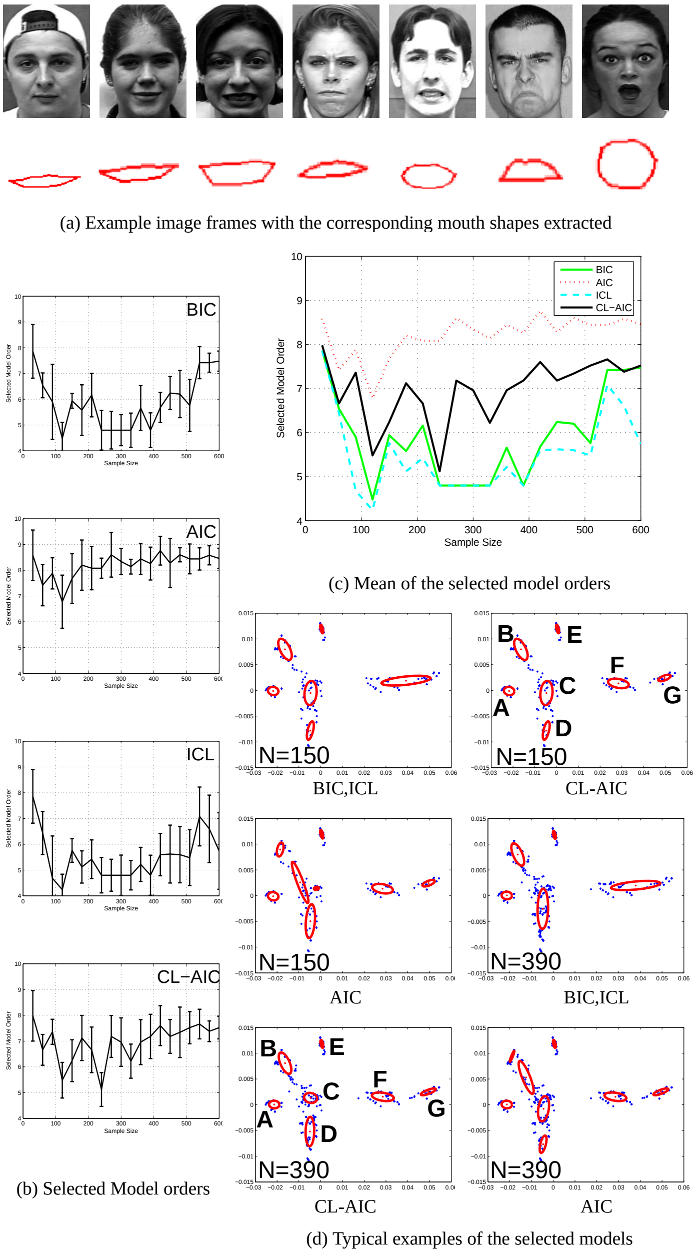
Figure 3: Model selection for learning spatial context. The visual context of spatial regions in the tearoom scene included "A","B": standing spots around the left table, "C","D": two chairs around the left table, "E","F": two chairs around the right table, "G": work surface, and "H": sink area. They were labelled in (f) only when estimated correctly.

In this tearoom scenario, the spatial context refers to semantically meaningful spatial regions, especially inactivity zones where people typically remain static or exhibit only localised movements (e.g. sink area and chairs). The problem of learning inactivity zones was tackled by performing unsupervised clustering of the inactivity points detected on motion trajectories. Firstly, a tracker based on blob matching matrix [15] was employed which yielded temporally discretised motion trajectories (see Figure 3(b)). The established trajectories were then smoothed using an averaging filter and the speed of each person tracked on the image plane was estimated. Secondly, inactivity points on the motion trajectories were detected when the speed of the tracked people was below a threshold. This inactivity threshold was set to the average speed of people walking slowly across the view. A total of 962 inactivity points were detected over the 22430 frames (see Figure 3(c)). As can be seen in Figure 3(c)), these inactivity points were mainly distributed around the semantically meaningful inactivity zones, although they were also caused by errors in the tracker and the fact that people can exhibit inactivity anywhere in the scene.

	BIC	AIC	ICL	CL-AIC
144	4	14	2	**24**
960	34	8	12	**58**

Table 3: Percentage of correct model order selection (over 50 trials) by different criteria for learning spatial context with 144 and 960 samples respectively.

Finally, inactivity points were clustered using a Gaussian Mixture Model with each of the learned mixture components specifying one inactivity zone. The total number of mixture components, corresponding to the total number of inactivity zones, was determined using a model selection criterion. Through observation of the captured video data, 8 inactivity zones can be identified which correspond to the left side of the work surface, the sink area, 4 of the chairs surrounding the two dining tables, and 2 spots near the left dining table where people stand while doing the puzzle. The correct number of mixture components was thus set to 8. In our experiments, the sample size of the data set varied from 24 to 962 in increments of 24. The maximum number of components K_{max} was set to 15. The model selection results are shown in Figure 3 and Table 3. It can be seen that when the sample size was small but not too small compared to the number of model parameters (e.g. $100 < N < 250$), all criteria tended to under-fit, with CL-AIC outperforming the other three. As the sample size increased, all criteria turned towards slightly over-fitting except ICL, with the model orders selected by CL-AIC being the closest to the true model order of 8. Examples of the estimated models shown in Figure 3(f) demonstrate that each estimated cluster corresponded to one inactivity zone when the model order was selected correctly.



(a) Example image frames with the corresponding mouth shapes extracted

(b) Selected Model orders

(c) Mean of the selected model orders

(d) Typical examples of the selected models

Figure 4: Model selection for learning facial expression categories. The visual context of facial expressions included “A”: sad, “B”:smile, “C”:neutral, “D”:anger, “E”:grin,“F”:fear, and “G”:surprise. They were labelled in (d) only when estimated correctly.

5.2 Learning Facial Expression Context

The visual task of modelling the dynamics of facial expressions and performing robust recognition becomes easier if key facial expression categories can be discovered and modelled. In this experiment, we aim to learn this important visual context using the shape of mouth. A face was modeled using the Active Appearance Model (AMM) [5]. The face model was learned using 1790 images sized 320×240 pixels, capturing people exhibiting different facial expression continously. Firstly, the jaw outline and the shapes of eye, eyebrow and mouth were manually labeled and represented using 74 landmarks during training. Secondly, the trained model was employed to track face and extract the shape of mouth (represented using 12 landmarks) from the test data which consisted of 613 image frames. Both the training and test data included seven different expression categories: neutral, smile, grin, sadness, fear, anger and surprise. Some example test frames are shown in Figure 4(a). Thirdly, the mouth shape data extracted from the test frames were projected onto a Mixture of Probabilistic Principal Component Analysis (MPPCA) space [23] which was learned using the mouth shape data labeled manually from the training data. It was identified that only the second and third principal components of the learned MPPCA sub-space corresponded to facial expression changes. Facial expressions were thus represented using a 2D feature vector comprising the second and third MPPCA components of the mouth shape data.

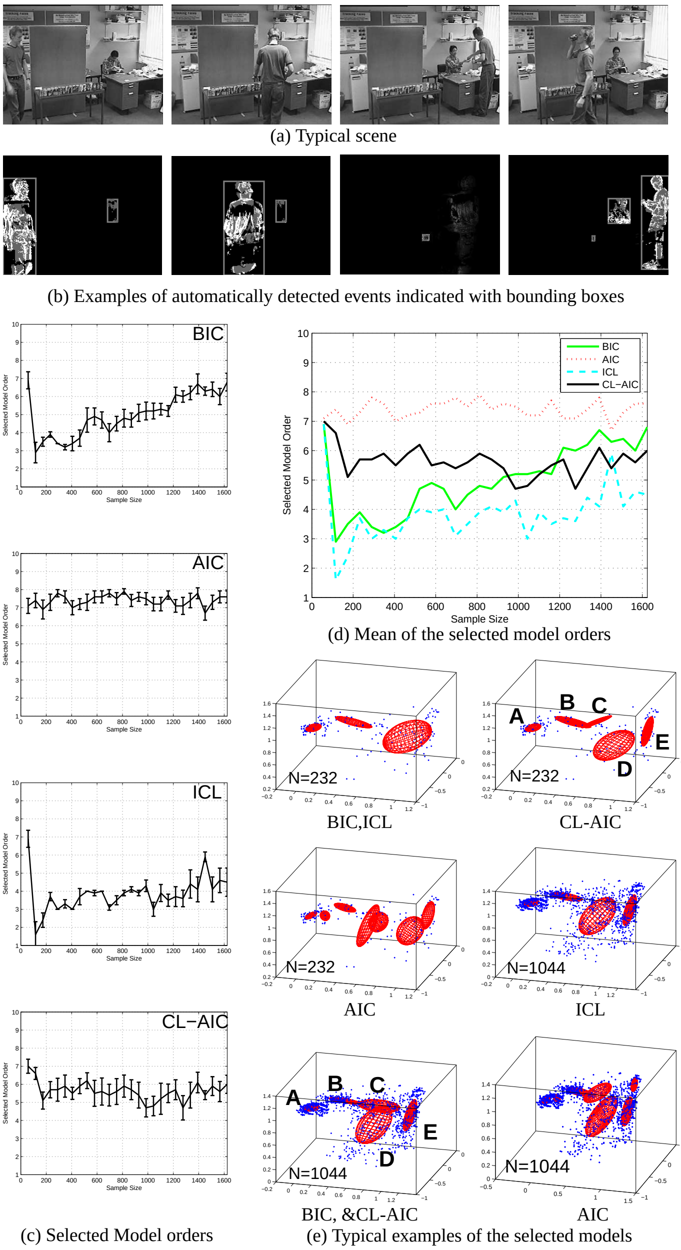
	BIC	AIC	ICL	CL-AIC
150	6	14	8	**24**
390	4	6	4	**40**

Table 4: Percentage of correct model order selection (over 50 trials) by different criteria for learning facial expression context with 150 and 390 samples respectively.

Finally, unsupervised clustering was performed using a Gaussian Mixture Model in the 2D feature space with the number of clusters automatically determined by a model selection criterion. Ideally, each cluster corresponds to one facial expression category and the right model order is 7. The data set was composed of 613 2D feature vectors obtained from the testing data set. Different model selection criteria were tested with sample sizes varying from 30 to 600 in increments of 30. The maximum number of components K_{max} was set to 15. The model selection results are shown in Figure 4 and Table 4. It can be seen that all criteria except AIC tended to under-estimate the number of components when the sample size was small but not too small (e.g. $50 < N < 200$) with CL-AIC outperforming BIC, ICL and AIC. With an increasing sample size, the models selected by BIC and CL-AIC turned towards slightly over-fitting with CL-AIC performing better than BIC, while those selected by ICL remained under-fitting. It is also noted that AIC suffered from over-fitting whatever the sample size was. Figure 4(d) shows that, when the model order was selected as 7, each learned cluster corresponded correctly to each of the 7 facial expression categories.

5.3 Learning Scene Event Context

A simulated 'shopping scenario' was captured at 25Hz, giving a total of 19 minutes of video data. The video data was sampled at 5 frames per second with a total number of 5699 frames of images sized 320×240 pixels. Some typical scenes are shown in Figure 5. The scene consists of a shopkeeper sat behind a table on the right side of the view. A large number of drink cans were laid out on a display table. Shoppers entered from the left and either browsed without paying or took a can and paid for it.


(a) Typical scene

(b) Examples of automatically detected events indicated with bounding boxes

(d) Mean of the selected model orders

(c) Selected Model orders

(e) Typical examples of the selected models

Figure 5: Model selection for learning scene event context. The estimated models are shown using the first 3 principal component of the feature space. The visual context of scene events in the shopping scene included "A": shopkeeper moving, "B":can being taken, "C":shopper entering/leaving, "D":shopper browsing, and "E":shopper paying. They were labelled in (e) only when estimated correctly.

Interpreting the shopping behaviour requires not only the understanding of the behaviour of shoppers and shopkeeper in isolation, but also the interactions between them. Detecting whether a drink can is taken by the shopper is also a key element to shopping behaviour interpretation. To build such a complex behaviour model, it is important to learn the visual context which, in this case, corresponds to significant and semantically meaningful scene changes characterised by the location, shape and direction of the change. These significant scene changes, referred to as scene events, are detected and clustered with the number of clusters being determined using model selection criteria. It was observed and labeled manually that there were largely 5 different types of scene events captured in this scenario, caused by 'shopper entering/leaving the scene', 'shopper browsing', 'can being taken', 'shopper paying', and 'shopkeeper moving' respectively. Firstly, events were automatically detected as groups of accumulated local pixel changes occurred in the scene. An event was represented by a group of pixels in the image plane (see Figure 5) and defined as a 7D feature vector (see [24] for details). A total of 1642 scene events were detected from the 19 minutes of video.

	BIC	AIC	ICL	CL-AIC
232	4	2	2	**32**
1044	54	2	6	**56**

Table 5: Percentage of correct model order selection (over 50 trials) by different criteria for learning scene event context with 232 and 1044 samples respectively.

Secondly, unsupervised clustering was performed in the 7D feature space. A Gaussian Mixture Model was adopted. Model selection was conducted using a data set consisting of 1642 scene events. In our experiments, the sample size of the data set varied from 58 to 1624 in increments of 58. The model selection results are presented in Figures 5 and Table 5. Note that in Figures 5(e) only the first 3 principal components of the feature space are shown for visualisation. It can be seen that when the sample size was small but not too small (e.g. $100 < N < 800$), BIC and ICL tended to under-fit while AIC and CL-AIC tended to over-fit. In comparison, CL-AIC gave the best performance. As the sample size increased, model orders selected BIC and CL-AIC were getting closer to the true model order of 5 with CL-AIC performing slightly better than BIC. In the meantime, ICL remained under-fitting and AIC remained over-fitting. Examples of estimated models shown in Figure 5(e) demonstrate that each estimated cluster corresponded to one scene event class when the model order was selected correctly.

6 Conclusion

Our experiments demonstrate the effectiveness of the proposed CL-AIC on learning visual context. It is worth pointing out that the noise inevitably contained in the visual data can be distributed in a very complex manner. For instance, we notice that the noise formed additional cluster in the spatial context learning case, whereas the noise appeared to be distributed randomly over the whole feature space in the scene event context learning case. Due to the existence of noise, most model selection criteria tend to over-fit even when sufficiently large data samples are available.

In conclusion, a novel probabilistic model selection criterion was proposed to improve existing model selection criteria for variable data sample sizes. The effectiveness of CL-AIC were demonstrated on learning visual context information for dynamic scene modelling. Finally, it is worth pointing out that CL-AIC can be readily extended to select models for data generated by many other real world problems which have the similar characteristics to the visual data.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, 1973.

[2] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *PAMI*, 22(7):719–725, 2000.

[3] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *PAMI*, 22(8):844–851, August 2000.

[4] O. Chapelle, V. Vapnik, and Y. Bengio. Model selection for small sample regression. *Machine Learning*, 48(1):9–23, 2002.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, pages 484–498, 1998.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[7] A. Dempster, N. Laird, and D. Rubin. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society B*, 41:276–278, 1979.

[8] M. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.

[9] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, pages 742–749, 2003.

[10] C. Hurivich, R. Shumway, and C. Tsai. Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika*, 77(4):709–719, 1990.

[11] C. Hurivich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1976.

[12] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *CVPR*, pages 866–871, Santa Barbara, USA, 1998.

[13] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:377–395, 1995.

[14] S. Kullback. *Information theory and statistics*. Dover: New York, 1968.

[15] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking group of people. *CVIU*, 80:42–56, 2000.

[16] S. McKenna and H. Nait-Charif. Learning spatial context from tracking using penalised likelihoods. In *ICPR*, 2004.

[17] A. Raftery. Bayes model selection in social research. *Sociological Methodology*, 90:181–196.

[18] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scentific, 1989.

[19] S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modelling. *PAMI*, 20(11):1133–1142, 1998.

[20] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[21] R. Shibata. Selection of the order of an autoregressive model by Akaike's Information Criterion. *Biometrika*, 63:117–126, 1976.

[22] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *PAMI*, 23:97–115, 2001.

[23] M. Tipping and C. Biship. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11:443–482, 1999.

[24] T. Xiang, S. Gong, and D. Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. In *BMVC*, pages 233–242, 2002.