

# Visual Learning Given Sparse Data of Unknown Complexity

Tao Xiang and Shaogang Gong  
Department of Computer Science  
Queen Mary, University of London, London E1 4NS, UK  
{txiang, sgg}@dcs.qmul.ac.uk

## Abstract

*This study addresses the problem of unsupervised visual learning. It examines existing popular model order selection criteria before proposes two novel criteria for improving visual learning given sparse data and without any knowledge about model complexity. In particular, a rectified Bayesian Information Criterion (BICr) and a Completed Likelihood Akaike's Information Criterion (CL-AIC) are formulated to estimate the optimal model order (complexity) for learning the dynamic structure of a visual scene. Both criteria are 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 a dynamic scene structure are carried out to demonstrate the effectiveness of BICr and CL-AIC, compared to that of BIC [15], AIC [1], ICL [3] and a MML based criterion [7].*

## 1. Introduction

We wish to learn the underlying visual structure of a given dynamic scene which can be considered as a semantically meaningful decomposition of spatial regions for human behaviour interpretation [11], or a decomposition of prototypic facial expressions for facial expression recognition [16]. We consider the problem of learning the underlying structural constraints for the activities captured in a visual scene. In particular, we address the problem of automatic model order selection for mixture models based visual structure learning given limited data of unknown complexity.

We aim to choose 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 [12], such as Bayesian Information Criterion (BIC) [15], Laplace Empirical Criterion (LEC) [14], and the Integrated Completed Likelihood (ICL) [3]; (2) methods based on the information coding theory such as the Minimum Message Length (MML) [7], Minimum Description Length (MDL) [13], and

Akaike's Information Criterion (AIC) [1]. The performance of various probabilistic model selection criteria has been studied intensively in the literature [14, 7, 3, 12, 4, 8], which motivated the derivation of new criteria. In particular, a number of previous works were focused on mixture models [14, 7, 3]. However, most previous studies assume the sample sizes of data sets to be sufficiently large in comparison to the number of model parameters [14, 7, 3], except for a few works that focused on linear autoregression models [4, 8]. 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. Existing criteria for mixture models are also mostly based on known model kernels, e.g. Gaussian. Realistically, however, visual data available for dynamic scene modelling are always sparse, incomplete, noisy and with unknown model kernels.

We propose two novel probabilistic model selection criteria to improve model estimation for sparse data sets, and with unknown kernels and severe overlapping among mixture components. In Section 2, we formulate a rectified Bayesian Information Criterion (BICr) which gives a more acceptable approximation to the Bayesian Model Selection (BMS) criterion compared to the conventional BIC, and rectifies the under-fitting tendency of BIC with small sample sizes. However, BICr is not able to rectify the over-fitting tendency of BIC when the true distribution kernel functions are very different from the assumed ones. Integrated Completed Likelihood (ICL) was proposed in [3] 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 capabilities of a mixture model. To this end, we introduce in Section 3 a Completed Likelihood AIC (CL-AIC) criterion, which aims to give the optimal clustering of a given data set and best predict unseen data. Extensive experiments are presented in Section 4 to demonstrate the effectiveness of BICr and CL-AIC on learning the dynamic structure of a visual scene, compared favourably to

the performance of a number of popular criteria including BIC, AIC, ICL and the MML based criterion proposed in [7]. Conclusions are drawn in Section 5.

## 2. Rectified BIC (BICr)

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  $k$ th 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  $k$ -th mixture component, and  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K; w_1, \dots, w_K\}$  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  $\mathcal{Y} = \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)}\}$ . The log-likelihood of observing  $\mathcal{Y}$  given a  $K$ -component mixture model is

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

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  $\mathcal{M}_K$ .

We formulate BICr to rectify the under-fitting tendency of BIC given small sample size. BIC was derived as an approximation of the Bayesian Model Selection (BMS) criterion [12]. This approximation is accurate only when the sample size is sufficiently large, ideally approaching infinity. It is shown by [14, 7] and also our experiments (see Sections 3 and 4) that BIC tends to underestimate the number of mixture components given sparse data. We suggest that the inaccurate approximation during the derivation of BIC based on BMS causes model under-fitting and propose a rectified BIC (BICr) to overcome it by providing more acceptable approximation. This introduces an extra penalty term in BICr which favours large  $K$  given sparse data.

To derive BICr, let us first briefly describe the general BMS criterion, which chooses a model that produces the Maximum a Posteriori (MAP) probability of observing a data set  $\mathcal{Y}$ :  $\hat{K} = \arg \max_K \{p(\mathcal{M}_K|\mathcal{Y})\}$ . Using Bayes' rule, the posterior probability is:

$$p(\mathcal{M}_K|\mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{M}_K)p(\mathcal{M}_K)}{p(\mathcal{Y})}, \quad (2)$$

where  $p(\mathcal{Y}|\mathcal{M}_K)$  is the marginal probability (likelihood) of the data and  $p(\mathcal{M}_K)$  is the *a priori* probability of model

$\mathcal{M}_K$ . If no *a priori* knowledge exists that favors any of the candidate models, the BMS method selects the model that yields the maximal marginal probability, given as:

$$p(\mathcal{Y}|\mathcal{M}_K) = \int p(\mathcal{Y}|\mathcal{M}_K, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{M}_K)d\boldsymbol{\theta}, \quad (3)$$

where  $p(\boldsymbol{\theta}|\mathcal{M}_K)$  is the *a priori* probabilistic density function of  $\boldsymbol{\theta}$  given  $\mathcal{M}_K$  and  $p(\mathcal{Y}|\mathcal{M}_K, \boldsymbol{\theta})$  is the probabilistic density function of  $\mathcal{Y}$  given  $\mathcal{M}_K$  and its parameters  $\boldsymbol{\theta}$ . For a simpler notation, we leave out the specific model label  $\mathcal{M}_K$  in the following derivations without losing generality.

Laplace approximation is adopted to compute the marginal probability  $p(\mathcal{Y}|\mathcal{M}_K)$  (see [15] for details), giving:

$$\begin{aligned} \log p(\mathcal{Y}) &= \log p(\mathcal{Y}|\hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + \frac{C_K}{2} \log(2\pi) \\ &\quad - \frac{C_K}{2} \log N - \frac{1}{2} \log |\mathbf{i}| + O(N^{-\frac{1}{2}}). \end{aligned} \quad (4)$$

where  $C_K$  is the dimensionality of the parameter space,  $N$  is the sample size,  $\hat{\boldsymbol{\theta}}$  is the ML estimate of  $\boldsymbol{\theta}$ ,  $\mathbf{i}$  is the expected Fisher information matrix for one observation [12],  $|\mathbf{i}|$  is its determinant, and  $O(N^{-\frac{1}{2}})$  represents any quantity such that  $N^{\frac{1}{2}}O(N^{-\frac{1}{2}})$  approaches a constant value as  $N$  approaches infinity. The first term on the right-hand side of Equation (4) is of order  $O(N)$ , the fourth term is of order  $O(\log N)$ , while all the other terms are of order  $O(1)$  or less. Eliminating those order  $O(1)$  or less terms gives:

$$\text{BIC} = -\log p(\mathcal{Y}) = -\log p(\mathcal{Y}|\hat{\boldsymbol{\theta}}) + \frac{C_K}{2} \log N. \quad (5)$$

The approximation error in BIC is thus of order  $O(1)$  which can be significant given small  $N$ . To have a more accurate approximation with small  $N$ , we keep the order  $O(1)$  terms in Equation (4) in the following derivation of BICr.

Assuming that the parameters for different mixture components are independent from each other and also from the mixing probabilities, the parameter priori  $p(\hat{\boldsymbol{\theta}})$  is computed as  $p(\hat{\boldsymbol{\theta}}) = p(\hat{w}_1, \dots, \hat{w}_K) \prod_{k=1}^K p(\hat{\boldsymbol{\theta}}_k)$ . The form and parameters of the prior distributions are determined according to three criteria: (1) They lead to an analytic solution; (2) They represent the common situation where a little, but not much, prior information is available; (3) The order  $O(1)$  terms in Equation (4) favour large  $K$  given small  $N$ , thus rectifying the under-fitting tendency of BIC. To this end, the Dirichlet prior [2] is employed for the mixing probabilities:

$$p(\hat{w}_1, \dots, \hat{w}_K) = \frac{\Gamma(\sum_{k=1}^K u_k)}{\prod_{k=1}^K \Gamma(u_k)} \prod_{k=1}^K \hat{w}_k^{u_k-1}, \quad (6)$$

where  $u_k$  are distribution parameters and  $\Gamma(\cdot)$  is the gamma function. Here we set  $u_k = \frac{1}{2}$  to reflect the lack of knowledge about the mixing probabilities. For the internal parameters  $\hat{\boldsymbol{\theta}}_k$ , independent flat priors are adopted. More

specifically, each element of the mean vector of each of the  $K$  components follows a flat distribution in the range of  $(-\alpha\sigma_{\mathcal{Y}}, \alpha\sigma_{\mathcal{Y}})$  and the diagonal covariance elements of each component follow a flat distribution in the range of  $(0, \beta\sigma_{\mathcal{Y}})$  where  $\sigma_{\mathcal{Y}}$  is the maximal diagonal element of the covariance matrix of the data set  $\mathcal{Y}$  and  $\alpha$  and  $\beta$  are scale parameters. We thus have:

$$\prod_{k=1}^K p(\hat{\boldsymbol{\theta}}_k) = \frac{1}{(2\alpha\beta\sigma_{\mathcal{Y}}^2)^{KD}}, \quad (7)$$

where  $D$  is the dimensionality of the data space. As pointed out by Roberts et al. [14], the scale parameters  $\alpha$  and  $\beta$  are essentially arbitrary. We thus set

$$\alpha = \beta = \frac{\Gamma(\frac{K}{2})^{\frac{1}{2KD}} (2\pi)^{\frac{CK}{4KD}}}{\sqrt{2}\sigma_{\mathcal{Y}}\Gamma(\frac{1}{2})^{\frac{1}{2D}} |\mathbf{i}|^{\frac{1}{4KD}}} \quad (8)$$

to satisfy the prior selection criteria (1) and (3). Replacing  $p(\boldsymbol{\theta})$  in Equation (4) using Equations (6)-(8) gives:

$$\log p(\mathcal{Y}) = \log p(\mathcal{Y}|\hat{\boldsymbol{\theta}}) - \frac{1}{2} \sum_{k=1}^K \log \hat{w}_k - \frac{C_K}{2} \log N + O(N^{-\frac{1}{2}}).$$

A rectified BIC is then derived as the negative of  $\log p(\mathcal{Y})$  with the order  $O(N^{-\frac{1}{2}})$  term being eliminated:

$$\text{BICr} = -\log p(\mathcal{Y}|\hat{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{k=1}^K \log \hat{w}_k + \frac{C_K}{2} \log N. \quad (9)$$

For the particular prior distributions we choose (Equations (6)-(8)), the error in the approximation of BICr is of order  $O(N^{-\frac{1}{2}})$  instead of  $O(1)$  in that of BIC. BICr is thus a more accurate approximation of Bayesian Model Selection and able to better select model in the sense of maximising  $p(\mathcal{Y}|\mathcal{M}_K)$ . Also importantly, the extra penalty term in BICr ( $\frac{1}{2} \sum_{k=1}^K \log \hat{w}_k$ ) has the following property:

$$-\infty < \frac{1}{2} \sum_{k=1}^K \log \hat{w}_k \leq -\frac{1}{2} K \log K \leq 0,$$

given  $0 \leq w_k \leq 1$ . It thus weakens the effect of the other penalty term  $\frac{C_K}{2} \log N$  especially when  $K$  becomes large with some mixture components only being poorly supported by the data. In other words, it favors larger  $K$  compared to BIC. Since the extra penalty term is of order  $O(1)$ , its effect is only significant given sparse data. *This extra penalty term in BICr thus rectifies the under-fitting tendency of BIC given sparse data and results in better model selection.* It is noted that the idea of integrating a priori knowledge of model parameters into an existing model selection criterion has been exploited previously in [14, 7]. However,

unlike our BICr, none of them were directly motivated by rectifying the weakness of existing criteria.

Even with BICr, the problem of BIC tending to overfit remains when the true model kernels are very different from the assumed ones (e.g. typically Gaussian). To solve this problem, we propose a Completed Likelihood Akaike's Information Criterion (CL-AIC).

### 3. Completed Likelihood 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 a given data set, and (3) predicting unseen data. Objectives (1) and (2) emphasise data explanation while objective (3) is concerned with data prediction. Both BIC and BICr choose the model that maximises  $p(\mathcal{Y}|\mathcal{M}_K)$ . They thus enforce mainly objective (1). When the true mixture distribution kernel functions are very different from the assumed ones, both BIC and BICr tend to choose the model with its number of components larger than the true number of components in order to 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 completed data for a  $K$ -component mixture model is a combination of the data set and the labels of each data sample:

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

where  $\mathcal{Z} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}, \dots, \mathbf{z}^{(N)}\}$ , and  $\mathbf{z}^{(n)} = \{z_1^{(n)}, \dots, z_K^{(n)}\}$  is a binary label vector such that  $z_k^{(n)} = 1$  if  $\mathbf{y}^{(n)}$  belongs to the  $k$ th 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:

$$\begin{aligned} \text{CL}(K) &= \log p(\mathcal{Y}|\boldsymbol{\theta}) + \log p(\mathcal{Z}|\mathcal{Y}, \boldsymbol{\theta}) \\ &= \sum_{n=1}^N \log \sum_{k=1}^K w_k p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k) + \sum_{n=1}^N \sum_{k=1}^K z_k^{(n)} \log p_k^{(n)} \end{aligned} \quad (10)$$

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

$$p_k^{(n)} = \frac{w_k p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k)}{\sum_{i=1}^K w_i p(\mathbf{y}^{(n)}|\boldsymbol{\theta}_i)}. \quad (11)$$

In practice, the true parameters  $\boldsymbol{\theta}$  in Equation (11) is replaced using the ML estimate  $\hat{\boldsymbol{\theta}}$  and the completed log-

likelihood is rewritten as:

$$\text{CL}(K) = \sum_{n=1}^N \log \sum_{k=1}^K \hat{w}_k p(\mathbf{y}^{(n)} | \hat{\boldsymbol{\theta}}_k) + \sum_{n=1}^N \sum_{k=1}^K \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} \quad (12)$$

where

$$\hat{z}_k^{(n)} = \begin{cases} 1 & \text{if } \arg \max_j \hat{p}_j^{(n)} = k \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

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 [10]. Given a completed 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

$$d(\mathcal{M}_0, \mathcal{M}_K) = E \left[ \log \left( \frac{p(\bar{\mathcal{Y}} | \mathcal{M}_0, \boldsymbol{\theta}_{\mathcal{M}_0})}{p(\bar{\mathcal{Y}} | \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K})} \right) \right]. \quad (14)$$

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 approximates infinity. Our CL-AIC is thus derived as:

$$\text{CL-AIC} = -\log p(\bar{\mathcal{Y}} | \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}) + C_K, \quad (15)$$

where  $C_K$  is the dimensionality of the parameter space. The first term on the right hand side of (15) is the completed likelihood given by Equation (12). We thus have:

$$\begin{aligned} \text{CL-AIC} = & -\sum_{n=1}^N \log \sum_{k=1}^K \hat{w}_k p(\mathbf{y}^{(n)} | \hat{\boldsymbol{\theta}}_k) \\ & -\sum_{n=1}^N \sum_{k=1}^K \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} + C_K, \end{aligned} \quad (16)$$

The first and third terms on the right hand side of Equation (16) emphasise the prediction capability of the model. These two terms favour those candidate models that give small generalisation error. In the meantime, the second term favours well-separated mixture components through minimizing entropy of assigning data samples into different components. The second term has the effect of selecting

models that give small training error. It thus enforces the explanation capability of the model. This results in a number of important advantages compared to existing techniques: (1) Unlike previous probabilistic model selection criteria, our CL-AIC attempts to optimise *explicitly* the explanation and prediction capabilities of a model. This makes CL-AIC theoretically attractive. Its effectiveness in practice is demonstrated later in this paper. (2) Compared to a standard AIC, our CL-AIC has an extra penalty term (the second term on the right hand side of Equation (16)) which always assumes a non-negative 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 [9] and experimental studies [8]. The extra penalty term in our CL-AIC thus has the effect of rectifying the over-fitting tendency of AIC.

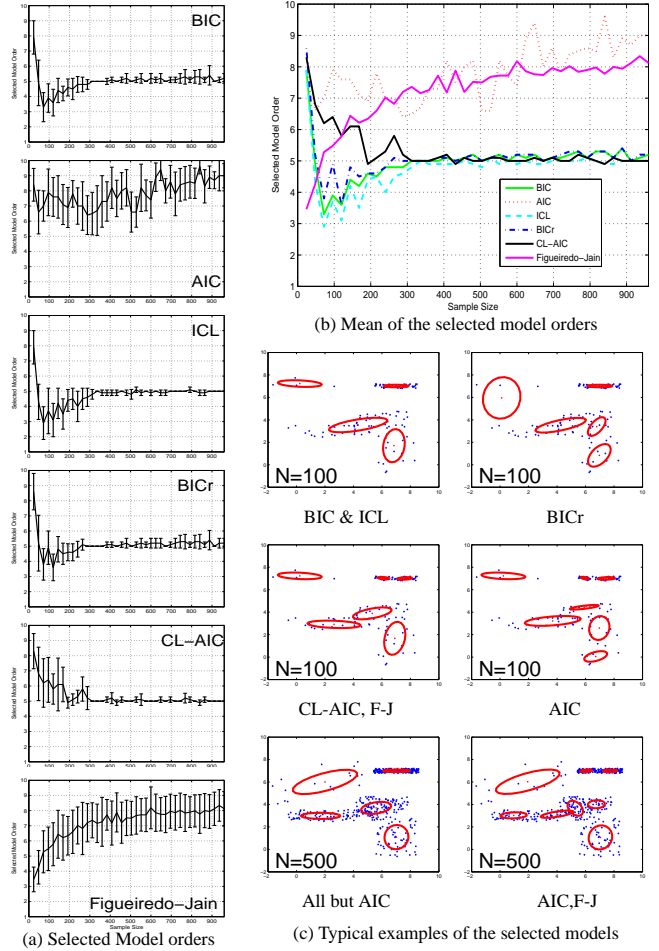


Figure 1: Model selection results for a toy problem.

A toy problem is used here to illustrate the effectiveness of CL-AIC and BICr, compared to that of AIC, BIC, ICL and a MML based criterion proposed by Figueiredo and

Jain [7] (referred as F-J hereafter<sup>1</sup>). Experiments on learning the visual structure of three different real scenarios are presented in Section 4. We consider a synthetic 2D data set where data from each cluster follow the uniform random distribution:

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

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:

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

Gaussian mixture models were adopted to illustrate the situation where the unknown kernel functions are very different from the assumed ones. Models with the number of components  $K$  varying from 1 to  $K_{max}$ , a number that is considered to be safely larger than the true number  $K_{true}$ , were evaluated.  $K_{max}$  was set to 10 in this case. To avoid being trapped at local maxima, the EM algorithm used for estimating model parameters  $\theta$  was randomly initialised for 20 times and the solution that yielded the largest observation likelihood after 30 iterations were chosen. Different model selection criteria were tested on the data set with sample sizes varying from 25 to 1000 in increments of 25. The mean and  $\pm 1$  standard deviation of the model order selection results over 50 trials are plotted against sample size in Figure 1(a), with each trial having a different random number seed. Figure 1(b) shows only the mean of the model order selected by different criteria in a single plot. Examples of models selected by different criteria are shown in Figure 1(c). It can be seen from Figure 1(b) that with a small sample size (e.g.  $50 < N < 200$ ), the number of components selected by BICr was the closest to the true number 5. As the sample size increased, both BIC and BICr slightly over-fitted and ICL slightly under-fitted, while CL-AIC yielded the most accurate results. F-J and 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 F-J and AIC suffered from severe over-fitting and failed to converge.

## 4. Experiments

Experiments were conducted on learning the dynamic structures of three different visual scenes. Gaussian mixture models were adopted in our experiments while the true

<sup>1</sup>Courtesy of M. Figueiredo for providing the code for implementing F-J.

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 toy problem experiment presented in the preceding section, unless otherwise specified.

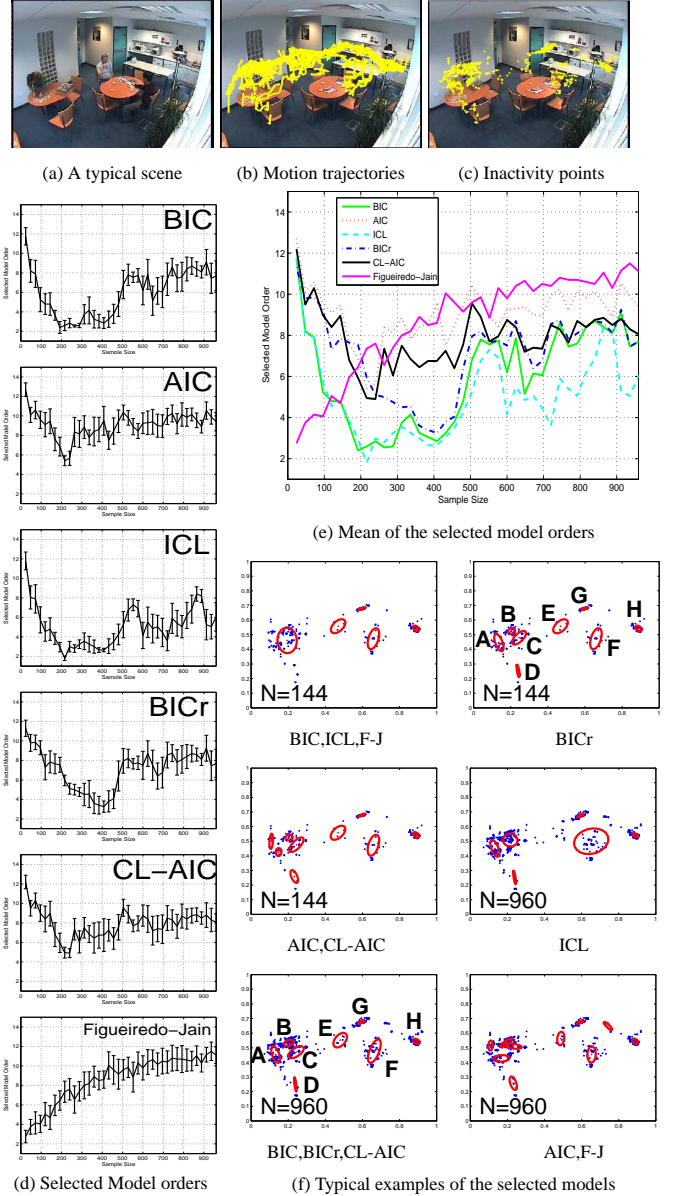


Figure 2: Model selection for learning inactivity zones. The inactivity zones in the tearoom scene included “A”; “B”: standing spots around the left table, “C”; “D”: two chairs around the left table, “E”; “F”: two chairs around the right table, “G”: work surface, and “H”: sink area. They were labelled in (f) when estimated correctly.

### 4.1. Learning Inactivity Zones

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  $320 \times 240$  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 2(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.

In this tearoom scenario, the dynamic structure of the visual scene includes 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 yielded temporally discretised motion trajectories (see Figure 2(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 2(c)). As can be seen in Figure 2(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.

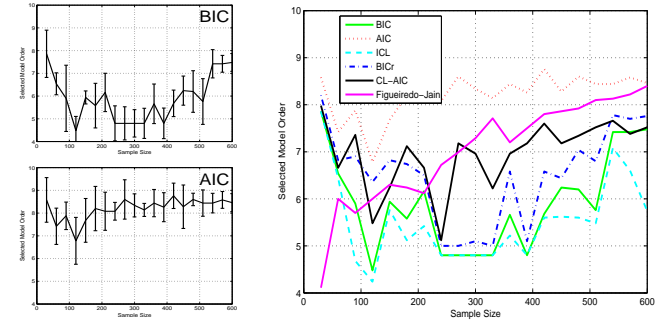
Finally, inactivity points were clustered using a Gaussian Mixture Model with each of the learned mixture components specifying one inactivity zone. The total number of mixture components, corresponding to the total number of inactivity zones, was determined using a model selection criterion. Through 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. 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 2. 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 BICr outperforming the other five. 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. Figure 2(f) demonstrates that each estimated cluster corresponded to one inactivity zone when the model order was selected correctly.

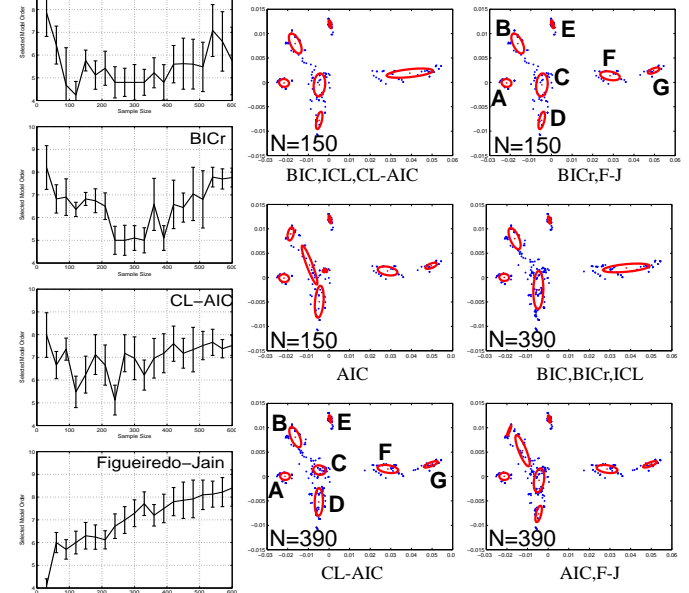
## 4.2. Learning Facial Expression Categories



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



(c) Mean of the selected model orders



(b) Selected Model orders

(d) Typical examples of the selected models

Figure 3: Model selection for learning facial expression categories. The visual structure 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.

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

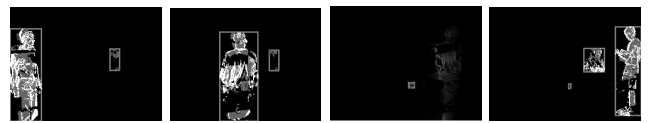
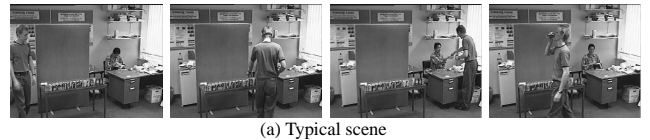
modelled. In this experiment, we aim to learn the range of mouth shape change caused by variation in expression. A face was modeled using the Active Appearance Model (AMM) [5]. The face model was learned using 1790 images sized  $320 \times 240$  pixels, capturing people exhibiting different facial expression continuously. 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 3(a). Thirdly, the mouth shape data extracted from the test frames were projected onto a Mixture of Probabilistic Principal Component Analysis (MPPCA) space [17] 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.

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 3. 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 BICr outperforming the other five. With an increasing sample size, the models selected by BIC, BICr and CL-AIC turned towards slightly over-fitting with CL-AIC performing better than the other two, while those selected by ICL remained under-fitting. Figure 3(d) shows that, when the model order was selected as 7, each learned cluster corresponded correctly to each of the 7 expression categories.

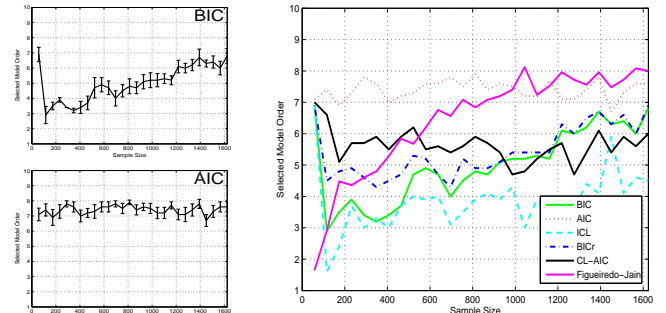
### 4.3. Learning Visual Event Classes

A simulated ‘shopping scenario’ was captured at 5Hz, giving a total of 19 minutes of video data with 5699 frames of images sized  $320 \times 240$  pixels (see 4(a)). 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.

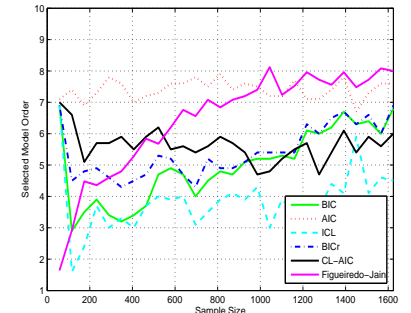
Interpreting the shopping behaviour requires not only



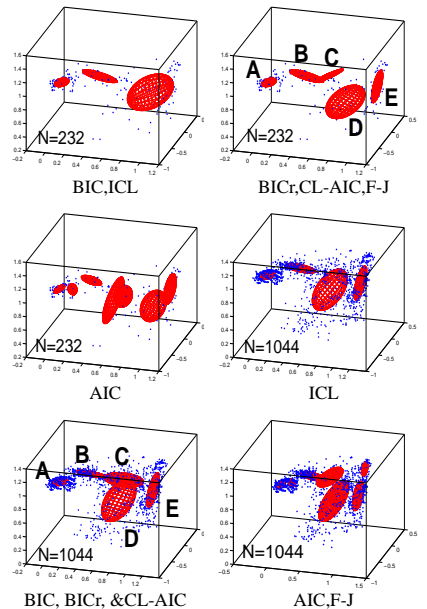
(b) Examples of automatically detected events indicated with bounding boxes



(c) Selected Model orders



(d) Mean of the selected model orders



(e) Typical examples of the selected models

Figure 4: Model selection for learning scene event classes. The estimated models are shown using the first 3 principal component of the feature space. The scene event classes therefore the underlying visual structure of 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) when estimated correctly.

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 underlying visual structure 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 4(b)) and defined as a 7D feature vector [18]. A total of 1642 scene events were detected.

Secondly, unsupervised clustering was performed in the 7D feature space. A Gaussian Mixture Model was adopted. 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 4. Note that in Figures 4(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, BICr, F-J and ICL all tended to under-fit while AIC and CL-AIC tended to over-fit. In comparison, BICr gave the best performance. As the sample size increased, model orders selected by BIC, BICr and CL-AIC were getting closer to the true model order of 5 with CL-AIC performing slightly better than the other two. Figure 4(e) demonstrates that each estimated cluster corresponded to one scene event class when the model order was selected correctly.

## 5. Conclusions

Our experiments demonstrate that the proposed BICr and CL-AIC outperform BIC, ICL, F-J, and AIC given severely overlapped data sets of different sizes arisen from both synthetic and real-world visual structure learning problems. This result was obtained using mixture models in the realistic situation where the true kernel distribution functions are very different from the assumed ones. Specifically, given sparse data, BICr rectifies the under-fitting tendency of BIC and also outperforms ICL, AIC, F-J, and CL-AIC. Given moderate to large data sample sizes, CL-AIC appears to be the best choice among the 6 criteria being compared.

Similar to BICr, the F-J criterion exploits the idea of integrating a priori knowledge of model parameters into an existing model selection criterion. ICL also aims to improve existing criteria by combining them together, which is the same motivation behind the derivation of CL-AIC. However, both F-J and ICL failed to produce satisfactory results

for the challenging problem of learning the dynamic structure of a visual scene. Our study thus highlights the importance of analysing the strength and weakness of existing criteria and the nature of data distribution for deriving a better criterion.

## 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] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley and Sons, 1994.
- [3] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *PAMI*, 22(7):719–725, 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] M. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.
- [8] C. Hurvich, R. Shumway, and C. Tsai. Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika*, 77(4):709–719, 1990.
- [9] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:377–395, 1995.
- [10] S. Kullback. *Information theory and statistics*. Dover: New York, 1968.
- [11] S. McKenna and H. Nait-Charif. Learning spatial context from tracking using penalised likelihoods. In *ICPR*, 2004.
- [12] A. Raftery. Bayes model selection in social research. *Sociological Methodology*, 90:181–196.
- [13] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.
- [14] S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modelling. *PAMI*, 20(11):1133–1142, 1998.
- [15] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [16] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *PAMI*, 23:97–115, 2001.
- [17] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11:443–482, 1999.
- [18] 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.