International Journal of Computer Vision 69(2), 181–201, 2006

DOI: 10.1007/s11263-005-5024-8

Model Selection for Unsupervised Learning of Visual Context

TAO XIANG AND SHAOGANG GONG
Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK
txiang@dcs.qmul.ac.uk
sgg@dcs.qmul.ac.uk

Received February 26, 2005; Revised July 14, 2005; Accepted September 14, 2005

First online version published in May, 2006

Abstract. This study addresses the problem of choosing the most suitable probabilistic model selection criterion for unsupervised learning of visual context of a dynamic scene using mixture models. 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 a given visual scene. Both criteria are designed to overcome poor model selection by existing popular criteria when the data sample size varies from small to large and the true mixture distribution kernel functions differ from the assumed ones. Extensive experiments on learning visual context for dynamic scene modelling are carried out to demonstrate the effectiveness of BICr and CL-AIC, compared to that of existing popular model selection criteria including BIC, AIC and Integrated Completed Likelihood (ICL). Our study suggests that for learning visual context using a mixture model, BICr is the most appropriate criterion given sparse data, while CL-AIC should be chosen given moderate or large data sample sizes.

Keywords: learning for vision, visual context, model selection, dynamic scene modelling, clustering, Bayesian methods, mixture models

1. Introduction

The problem of dynamic scene understanding can be tackled based on building models for various activities occurring in the scene (Haritaoglu et al., 2000; McKenna, 2000; Stauffer and Grimson, 2000; Wada and Matsuyama, 2000; Johnson et al., 1998; Brand et al., 1996; Oliver et al., 2000; Hongeng and Nevatia, 2001; Gong and Xiang, 2003; Brand and Kettnaker, 2000; Cohen et al., 2003). Learning 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 (McKenna and Nait-Charif, 2004; Brand and Kettnaker, 2000), or a decomposition of prototypic facial expressions for facial expression recognition (Tian et al., 2001; Cohen et al., 2003). We consider the problem of learning visual context as modelling the underlying structure of activity captured in a dynamic scene. To this end, we propose to discover visual context based on unsupervised learning. Specifically, visual observations of activities are represented in a feature space, and the structure and complexity of the visual data distribution are profiled using a mixture model with the number of mixture components being determined automatically through model order selection.

Model selection is key to unsupervised statistical modelling of data. Suppose that a data set $\mathbf{D}$ arises from one of M candidate models, the problem is to choose the best candidate model for $\mathbf{D}$ following two

considerations. First, the measure of a good model can be based on how well a model explains the data. However, if 'explaining' or 'fitting' the data using the model is the only criterion for model selection, complex models will be favoured over simple models and, in the most extreme case, the 'best' model becomes the data set itself. This is clearly undesirable as in many cases a model is only useful when it can predict new data. Therefore, for better 'prediction' or 'generalization', a simpler model, i.e. a model with fewer parameters, is preferred. This is the second consideration for model selection. The principle of choosing a model that not only best fits a given data set but also satisfies simplicity is known as the Ockham' Razor principle after the 13th century philosopher William of Ockham, and is widely adopted for determining model complexity, especially in the form of probabilistic model selection criteria (Mclachlan and Peel, 1997). Other model selection criteria include heuristic methods such as Fuzzy Hyper-Volume (FHV) (Gath and Geva, 1989) and evidence density (Roberts, 1997), and cross-validation methods (Bishop, 1995; Lange et al., 2004).

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 (Raftery, 1995), such as Bayesian Information Criterion (BIC) (Schwarz, 1978), Laplace Empirical Criterion (LEC) (Roberts et al., 1998), and the Integrated Completed Likelihood (ICL) (Biernacki et al., 2000); (2) methods based on the information coding theory such as the Minimum Message Length (MML) (Figueiredo and Jain, 2002), Minimum Description Length (MDL)[1] (Rissanen, 1989), and Akaike's Information Criterion (AIC) (Akaike, 1973). The performance of various probabilistic model selection criteria has been studied intensively in the literature (Roberts et al., 1998; Figueiredo and Jain, 2002; Biernacki et al., 2000; Raftery, 1995; Chapelle et al., 2002; Hurivich et al., 1990; Cherkassky and Ma, 2003; Hastie et al., 2001), which motivated the derivation of new criteria. In particular, a number of previous works were focused on mixture models (Roberts et al., 1998; Figueiredo and Jain, 2002; Biernacki et al., 2000). However, most previous studies assume the sample sizes of data sets to be sufficiently large in comparison to the number of model parameters (Roberts et al., 1998; Figueiredo and Jain, 2002; Biernacki et al., 2000), except for a few works that focused on linear autoregression models (Cherkassky and Ma, 2003; Hastie et al., 2001; Chapelle et al., 2002; Hurivich et al., 1990). 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 discovering visual context given the nature of visual data commonly available.

In the rest of the paper, 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. Mixture models are briefly described in Section 2. In Section 3, 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 given small data 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 Biernacki et al. (2000) 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 through a model selection criterion. To this end, we introduce in Section 4 a Completed Likelihood AIC (CL-AIC) criterion, which aims to give the optimal clustering of a given data set and best predict unseen data. In Section 5, we analyse through synthetic data experiments how the performance of BICr and 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 6 to demonstrate the effectiveness of BICr and CL-AIC on learning visual context for dynamic scene understanding, compared to that of BIC, AIC and ICL. A conclusion is drawn in Section 7.

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:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k p(\mathbf{y} \mid \boldsymbol{\theta}_k), \tag{1}$$

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 $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}\}$. The log-likelihood of observing $\mathcal{Y}$ given a K-component mixture model is

$$\log p(\mathcal{Y} \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \left(\log \sum_{k=1}^{K} w_k p(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_k) \right), \tag{2}$$

where $p(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_k)$ defines the model kernels, i.e., the form of the probability distribution function 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, as given by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{\log p(\mathcal{Y} \mid \boldsymbol{\theta})\},$$

can be computed using the EM algorithm (Dempster et al., 1977). Therefore the problem of estimating a mixture model boils down to the estimation of K, known as the model order selection problem.

Denoting a K-component mixture model as $\mathcal{M}_K$, then $\mathcal{M}_K \subseteq \mathcal{M}_{K+1}$, i.e. the candidate mixture models are nested. To illustrate this, let us consider a K-component model described by

$$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K; w_1, \ldots, w_{K-1}, w_K\}$$

and a $K+1$-component model described by

$$\boldsymbol{\theta}' = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\theta}_{K+1}; w_1, \ldots, w_{K-1}, w'_K, w'_{K+1}\}.$$

These two models represent the same probability density function if $\boldsymbol{\theta}_{K+1} = \boldsymbol{\theta}_K$ and $w_K = w'_K + w'_{K+1}$. Consequently, $p(\mathcal{Y} \mid \boldsymbol{\theta})$ is a nondecreasing function of K and thus cannot be directly used for model order selection.

3. Rectified Bayesian Information Criterion (BICr)

We formulate BICr to rectify the under-fitting tendency of BIC given sparse data. BIC was derived as an approximation of the Bayesian Model Selection (BMS) criterion (Raftery, 1995). This approximation is accurate only when the sample size is sufficiently large, ideally approaching infinity. It is shown by our experiments (see Sections 5 and 6) and also those in (Roberts et al., 1998; Figueiredo and Jain, 2002) that BIC tends to underestimate the number of mixture components (i.e. under-fit) when the sample size is small. 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 $\mathcal{Y}$:

$$\hat{K} = \arg\max_{K} \{p(\mathcal{M}_K \mid \mathcal{Y})\}.$$

Using Bayes' rule, the posterior probability is:

$$p(\mathcal{M}_K \mid \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{M}_K) p(\mathcal{M}_K)}{p(\mathcal{Y})}, \tag{3}$$

where $p(\mathcal{Y} \mid \mathcal{M}_K)$ 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:

$$p(\mathcal{Y} \mid \mathcal{M}_K) = \int p(\mathcal{Y} \mid \mathcal{M}_K, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}_K)\, d\boldsymbol{\theta}, \tag{4}$$

where $p(\boldsymbol{\theta} \mid \mathcal{M}_K)$ is the *a priori* probabilistic density function of $\boldsymbol{\theta}$ given $\mathcal{M}_K$ and $p(\mathcal{Y} \mid \mathcal{M}_K, \boldsymbol{\theta})$ 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.

The analytic evaluation of the integral in Eq. (4) is only possible for the exponential family distributions. For a more general case, Laplace approximation is adopted to compute the marginal probability $p(\mathcal{Y})$ (see Schwarz, 1978 for details), giving:

$$\log p(\mathcal{Y}) = \log p(\mathcal{Y} \mid \hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + \frac{C_K}{2}\log(2\pi) - \frac{C_K}{2}\log N - \frac{1}{2}\log|\mathbf{i}| + O(N^{-\frac{1}{2}}). \quad (5)$$

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 (Raftery, 1995), $|\mathbf{i}|$ 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 Eq. (5) 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. BIC is derived as the negative of $\log p(\mathcal{Y})$ with those order $O(1)$ or less terms being eliminated:

$$\mathrm{BIC} = -\log p(\mathcal{Y}) = -\log p(\mathcal{Y} \mid \hat{\boldsymbol{\theta}}) + \frac{C_K}{2}\log N. \quad (6)$$

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 Eq. (5) 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:

$$p(\hat{\boldsymbol{\theta}}) = p(\hat{w}_1, \ldots, \hat{w}_K)\prod_{k=1}^{K} p(\hat{\boldsymbol{\theta}}_k). \quad (7)$$

The form and parameters of the prior distributions are determined according to four prior selection criteria: (1) They lead to an analytic solution; (2) They represents the common situation where a little, but not much, prior information is available; (3) They help eliminate as many terms in Eq. (5) as possible which are of order $O(1)$; and (4) The order $O(1)$ terms kept in the formulation of BICr favour large K given small N, thus rectifying the under-fitting tendency of BIC. To this end, the Dirichlet prior (Bernardo and Smith, 1994) is employed for the mixing probabilities:

$$p(\hat{w}_1, \ldots, \hat{w}_K) = \frac{\Gamma\left(\sum_{k=1}^{K} u_k\right)}{\prod_{k=1}^{K}\Gamma(u_k)}\prod_{k=1}^{K}\hat{w}_k^{u_k-1}, \quad (8)$$

where $u_k > 0$ are distribution parameters and $\Gamma(\cdot)$ is the gamma function. Here we set u_k to a constant value C_u for different k to reflect the lack of knowledge about the mixing probabilities,[2] thus satisfying the prior selection criteria (2). For the internal parameters $\hat{\boldsymbol{\theta}}_k$, independent flat priors are adopted which are independent from the parameter estimates. 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 α and β are scale parameters. We thus have:

$$\prod_{k=1}^{K} p(\hat{\boldsymbol{\theta}}_k) = \frac{1}{\left(2\alpha\beta\sigma_{\mathcal{Y}}^2\right)^{KD}}, \quad (9)$$

where D is the dimensionality of the data space. As pointed out by Fitzgerald (1996) and Roberts et al. (1998), the scale parameters α and β are essentially arbitrary. We thus set

$$\alpha = \beta = \frac{\Gamma\left(\frac{KC_u}{2}\right)^{\frac{1}{2KD}}(2\pi)^{\frac{C_K}{4KD}}}{\sqrt{2}\sigma_{\mathcal{Y}}\Gamma(C_u)^{\frac{1}{2D}}|\mathbf{i}|^{\frac{1}{4KD}}} \quad (10)$$

to satisfy the prior selection criteria (3) and (4). Replacing $p(\hat{\boldsymbol{\theta}})$ in Eq. (5) using Eqs. (7)–(10) gives:

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

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:

$$\mathrm{BICr} = -\log p(\mathcal{Y} \mid \hat{\boldsymbol{\theta}}) + (1 - C_u)\sum_{k=1}^{K}\log\hat{w}_k + \frac{C_K}{2}\log N. \quad (11)$$

For the particular prior distributions we choose (Eqs. (8)–(10)), 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 $p(\mathcal{Y} \mid \mathcal{M}_K)$. Also importantly, compared to the standard BIC formulation, BICr has an extra penalty term $\left((1 - C_u)\sum_{k=1}^{K} \log \hat{w}_k\right)$ derived from the *a priori* probability of the model parameters. Since $0 \leq \hat{w}_k \leq 1$, it is easy to show that by setting $C_u < 1$ we have:

$$(1 - C_u)\sum_{k=1}^{K} \log \hat{w}_k \leq -(1 - C_u)K \log K < 0.$$

This extra penalty term thus weakens the effect of the other penalty term $\frac{C_K}{2} \log N$ especially when K is large with some mixture components only being poorly supported by the data. In other words, it favors larger K compared to BIC (prior selection criterion (4)). Since the extra penalty term is of order $O(1)$, its effect is only significant when data set size is small. *This extra penalty term in BICr thus rectifies the under-fitting tendency of BIC given spare data and results in better model selection.*

As mentioned above, by setting u_k to a constant C_u for different k, the Dirichlet prior becomes noninformative therefore reflecting the fact that there is little or no *a priori* knowledge about the distribution of the mixture probabilities. However, u_k cannot be assigned to arbitrary constant values. By the definition of a Dirichlet prior, we have $u_k = C_u > 0$. In order for the second term of BICr (Eq. (11)) to rectify the under-fitting tendency of BIC, we have $u_k = C_u < 1$. Therefore, we should have $0 < C_u < 1$. In our experiments to be presented in Sections 5 and 6, C_u was set to $\frac{1}{2}$ which resulted in satisfactory results.

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).

4. 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 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 $p(\mathcal{Y} \mid \mathcal{M}_K)$. 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 a 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) for a mixture model. The completed data for a K-component mixture model is a combination of the data set and the labels of each data sample:

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

where $\mathcal{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)}, \ldots, \mathbf{z}^{(N)}\}$, and $\mathbf{z}^{(n)} = \{z_1^{(n)}, \ldots, z_K^{(n)}\}$ 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:

$$\begin{aligned} CL(K) &= \log p(\bar{\mathcal{Y}}) \\ &= \log p(\mathcal{Y} \mid \boldsymbol{\theta}) + \log p(\mathcal{Z} \mid \mathcal{Y}, \boldsymbol{\theta}) \\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} w_k p(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_k) \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} z_k^{(n)} \log p_k^{(n)} \end{aligned} \tag{12}$$

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

$$p_k^{(n)} = \frac{w_k p(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_k)}{\sum_{i=1}^{K} w_i p(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_i)}. \tag{13}$$

The $N \times K$ matrix $\{p_k^{(n)}\}$ is known as the Fuzzy Classification Matrix (Celeux and Soromenho, 1996).

The difference between the log-likelihood of the observed data and the completed log-likelihood

$$-\sum_{n=1}^{N}\sum_{k=1}^{K} z_k^{(n)} \log p_k^{(n)} \geq 0$$

is a random variable whose mean is the entropy of the Fuzzy Classification Matrix:

$$E(K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} p_k^{(n)} \log p_k^{(n)} \geq 0. \tag{14}$$

$E(K)$ measures the goodness of a model in clustering the observed data. If the mixture components are well separated, $E(K)$ is close to zero. $E(K)$ assumes large value when the mixture components are poorly separated.

In practice, the true parameters $\boldsymbol{\theta}$ in Eq. (12) is replaced using the ML estimate $\hat{\boldsymbol{\theta}}$ and the completed log-likelihood is rewritten as:

$$\begin{aligned} CL(K) = &\sum_{n=1}^{N} \log \sum_{k=1}^{K} \hat{w}_k\, p\big(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_k\big) \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} \end{aligned} \tag{15}$$

where

$$\hat{p}_k^{(n)} = \frac{\hat{w}_k\, p\big(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_k\big)}{\sum_{i=1}^{K} \hat{w}_i\, p(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_i)}, \tag{16}$$

and

$$\hat{z}_k^{(n)} = \begin{cases} 1 & \text{if } \arg\max_j \hat{p}_j^{(n)} = k \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

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 (Kullback, 1968). 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

$$d(\mathcal{M}_0, \mathcal{M}_K) = E\left[\log\left(\frac{p(\bar{\mathcal{Y}} \mid \mathcal{M}_0, \boldsymbol{\theta}_{\mathcal{M}_0})}{p(\bar{\mathcal{Y}} \mid \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K})}\right)\right]. \tag{18}$$

Ranking the candidate models according to $d(\mathcal{M}_0, \mathcal{M}_K)$ is equivalent to ranking them according to

$$\delta(\mathcal{M}_0, \mathcal{M}_K) = E\big[-2\log p\big(\bar{\mathcal{Y}} \mid \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}\big)\big].$$

$\delta(\mathcal{M}_0, \mathcal{M}_K)$ cannot be computed directly since the unknown true model is required. However, it was noted by Akaike (1973) that $-2\log p(\bar{\mathcal{Y}} \mid \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K})$ can serve as a biased approximation of $\delta(\mathcal{M}_0, \mathcal{M}_K)$, and the bias adjustment

$$E\big[\delta(\mathcal{M}_0, \mathcal{M}_K) + 2\log p\big(\bar{\mathcal{Y}} \mid \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}\big)\big]$$

converges to $2C_K$ when the number of data sample approaches infinity. Our CL-AIC is thus derived as:

$$\text{CL-AIC} = -\log p(\bar{\mathcal{Y}} \mid \mathcal{M}_K, \hat{\boldsymbol{\theta}}_{\mathcal{M}_K}) + C_K. \tag{19}$$

where C_K is the dimensionality of the parameter space. The first term on the right hand side of (19) is the completed likelihood given by Eq. (15). We thus have:

$$\begin{aligned} \text{CL-AIC} = &-\sum_{n=1}^{N} \log \sum_{k=1}^{K} \hat{w}_k\, p\big(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_k\big) \\ &- \sum_{n=1}^{N}\sum_{k=1}^{K} \hat{z}_k^{(n)} \log \hat{p}_k^{(n)} + C_K, \end{aligned} \tag{20}$$

The first and third terms on the right hand side of Eq. (20) 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 differences 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. The effectiveness of CL-AIC in practice is demonstrated through experiments in Sections 5 and 6.
2. Compared to a standard AIC, our CL-AIC has an extra penalty term (the second term on the right

hand side of Eq. (20)) 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 (Dempster et al., 1979; Kass and Raftery, 1995) and experimental studies (Shibata, 1976; Hurivich and Tsai, 1976). The extra penalty term in our CL-AIC thus has the effect of rectifying the over-fitting tendency of AIC.

3. Another approach for combining completed likelihood with an existing model selection criterion (in this case, BIC) was proposed in Biernacki et al. (2000) known as an Integrated Completed Likelihood (ICL) criterion (Biernacki et al., 2000). However, experiments reported in (Biernacki et al., 2000) 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. In comparison, our CL-AIC integrates an explanation criterion with a prediction criterion. It is thus theoretically better justified than ICL.

5. Experiments on Synthetic Data

In this section, we illustrate the effectiveness of BICr and CL-AIC, compared to that of BIC, AIC and ICL, using synthetic data. Experiments on discovering visual context of three different real scenarios are presented in Section 6. 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) whether and how the true kernel functions are different from the assumed ones. To this end, Gaussian mixture models were adopted while synthetic data sets were generated using either Gaussian or 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 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.

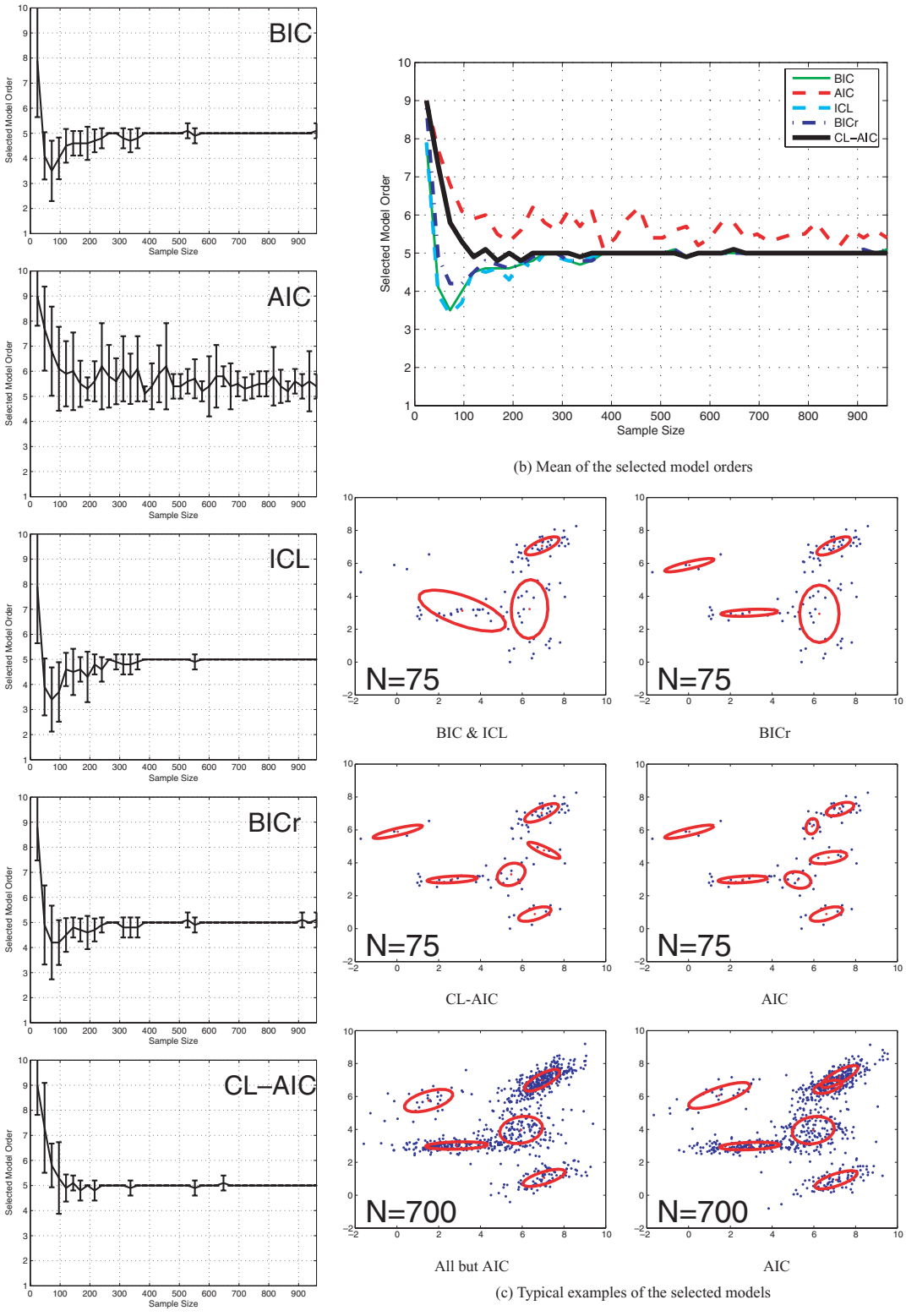
5.1. Gaussian Distributed Data

Let us first consider a data set generated using a 5-component bivariate Gaussian mixture. Modelled using a Gaussian Mixture model, this represents an ideal case where the true kernel function is identical to the assumed one. The parameters of the true mixture distribution are:

$$
\begin{aligned}
&w_1 = 0.05,\ w_2 = 0.10,\ w_3 = 0.20,\ w_4 = 0.40,\\
&w_5 = 0.25;\\
&\boldsymbol{\mu}_1 = [1.5, 6.0]^T, \quad \boldsymbol{\mu}_2 = [7.0, 1.0]^T,\\
&\boldsymbol{\mu}_3 = [6.0, 4.0]^T, \quad \boldsymbol{\mu}_4 = [7.0, 7.0]^T,\\
&\boldsymbol{\mu}_5 = [3.0, 3.0]^T;\\
&\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.89 & 0.25 \\ 0.25 & 0.50 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.72 & 0.14 \\ 0.14 & 0.34 \end{bmatrix},\\
&\boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.99 & 0.04 \\ 0.04 & 0.65 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 1.78 & 0.46 \\ 0.46 & 0.42 \end{bmatrix},\\
&\boldsymbol{\Sigma}_5 = \begin{bmatrix} 1.97 & 0.05 \\ 0.05 & 0.10 \end{bmatrix},
\end{aligned} \tag{21}
$$

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. Different model selection criteria were tested on the data set with sample sizes varying continuously from 25 to 1000 in increments of 25. The average number of mixture components determined by different criteria over 50 trails are plotted against the sample size in Figs. 1(a) and (b), which shows explicitly the under-fitting or under-fitting tendency of different criteria. Examples of models selected by different criteria are shown in Fig. 1(c). Table 1 shows examples of the percentage of correct model order selection (over 50 trials) by different criteria give small and large sample sizes.

Figures. 1(a) and (b) show how the performance of different criteria were affected by the sample size of



(a) Selected Model orders

(b) Mean of the selected model orders

(c) Typical examples of the selected models

Figure 1. Model selection results for Gaussian distributed data.

Table 1. Percentage of correct model order selection (over 50 trials) by different criteria for synthetic Gaussian data with 75 and 700 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
75	8	12	6	14	**26**
700	100	44	100	100	100

Table 2. Percentage of correct model order selection (over 50 trials) by different criteria for synthetic Noisy Gaussian data with 100 and 725 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
100	0	10	0	4	**48**
725	88	64	82	90	**100**

the data set. When the data set was sampled extremely sparsely (e.g. $N < 50$), all 5 criteria tended to over-fit. As the sample size increased, the number of components determined by all the criteria decreased. In particular, BIC, BICr, ICL and CL-AIC all turned from over-fitting to under-fitting before converging to the true component number, with the number of components selected by CL-AIC being the closest to the true number given small, but not too small sample sizes. As expected, the number of components estimated by BICr was closer to the right number 5 compared to BIC. Overall, AIC appeared to favor larger number of components even when the sample size is large. It can be seen from Fig. 1(b) that 4 out of 5 criteria, except AIC, selected the right number of components when the sample size was large (e.g. $N > 400$). It is also noted that AIC exhibited large variations in the estimated model orders no matter what the sample size was, while other criteria had smaller variation given larger sample sizes. This experiment suggests that given an ideally distributed data set, our proposed criteria only outperform the existing criteria when the sample size is small, but not too small.

5.2. *Gaussian Distributed Data Perturbed with Random Noise*

Here we consider a situation under which the true kernel functions are slightly different from the assumed ones. Each data sample from the same data set used in Section 5.1 was perturbed with a 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 Fig. 2 and Table 2. The results shown in Fig. 2 are similar to those obtained using the noiseless Gaussian data set (see Fig. 1) except that BIC, BICr, ICL and CL-AIC all needed more data samples to converge to the true model and AIC suffered more severe over-fitting. It is also noted that when the sample size was large (e.g. $N > 500$), both BIC and BICr 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 Biernacki et al. (2000).

5.3. *Uniformly Distributed Data*

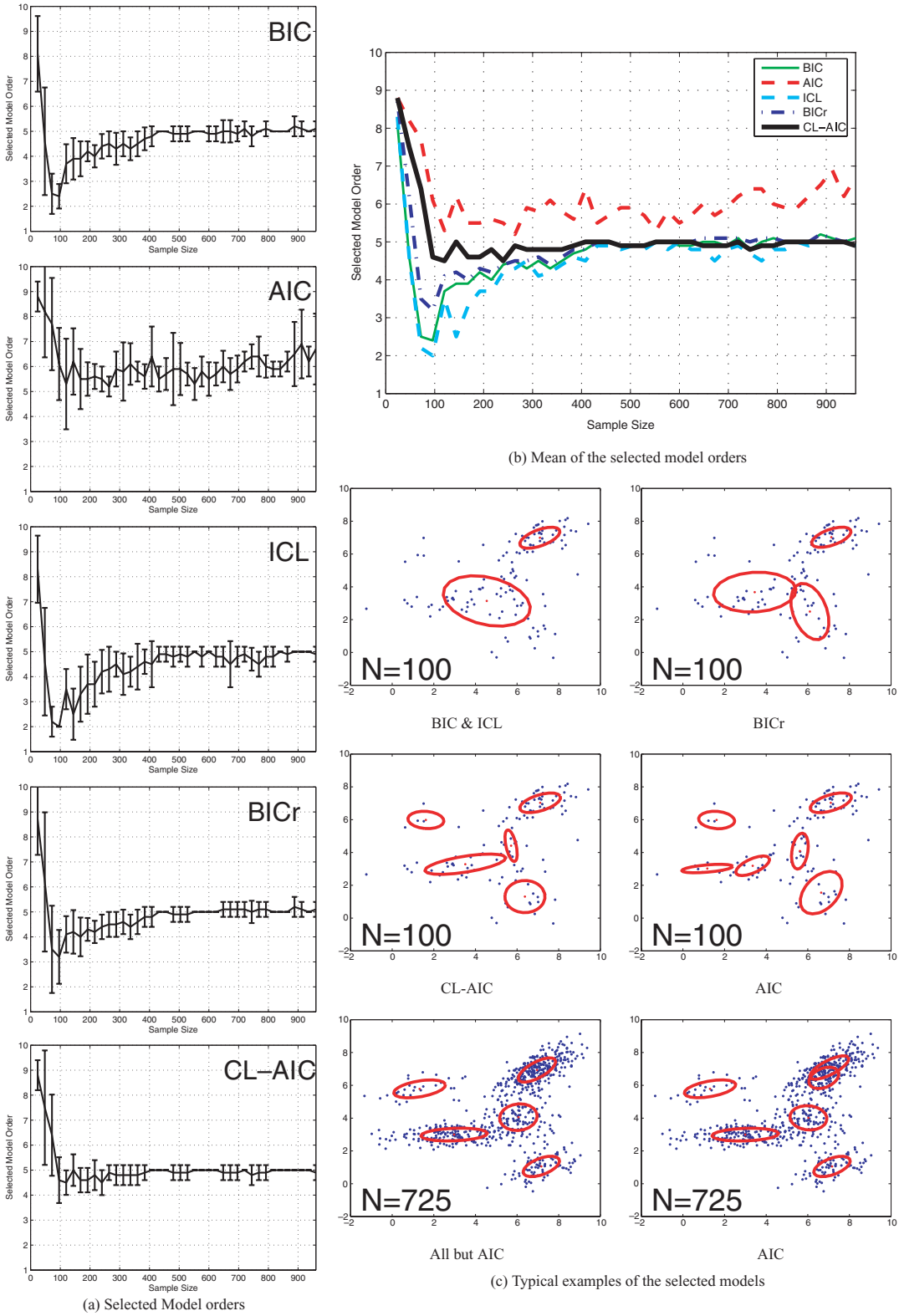
Now we consider a situation where the true kernel functions are very different from the assumed ones. A synthetic 2D data set were generated with data from each components following the uniform random distribution:

$$u_{\mathbf{r}}(y_1, y_2) = \begin{cases} \dfrac{1}{(r_2 - r_1) \times (r_4 - r_3)} & \text{if } r_1 \le y_1 \le r_2 \\ & \quad \text{and } r_3 \le y_2 \le r_4 \\ 0 & \text{otherwise} \end{cases}$$

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:

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

The model selection results are presented in Fig. 3 and Table 3. It can be seen from Figs. 3(a) and (b) that with a small sample size (e.g. $50 < N < 200$), BIC, ICL and BICr tended to under-fit while AIC and CL-AIC tended to over-fit. 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

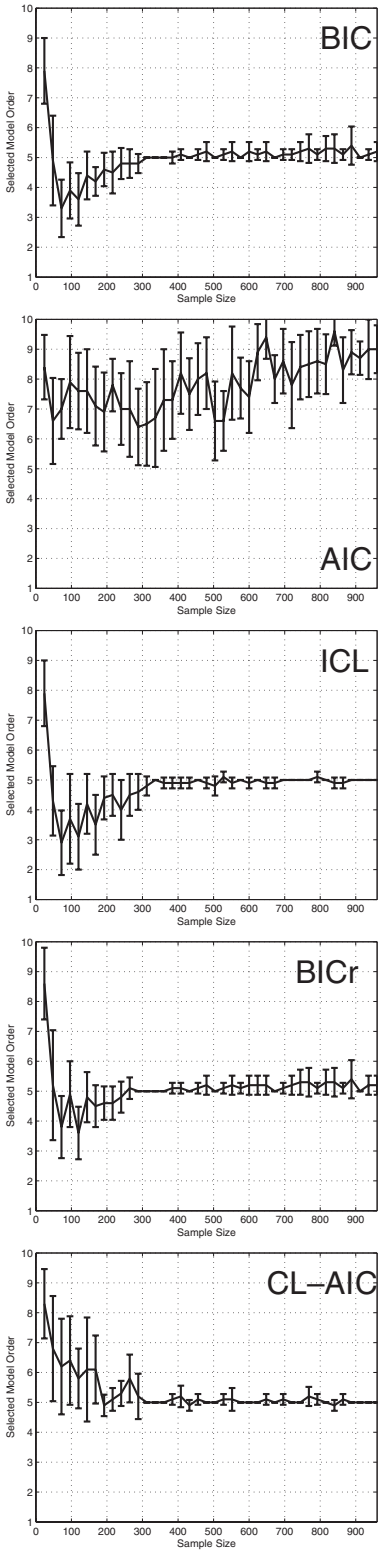


(a) Selected Model orders

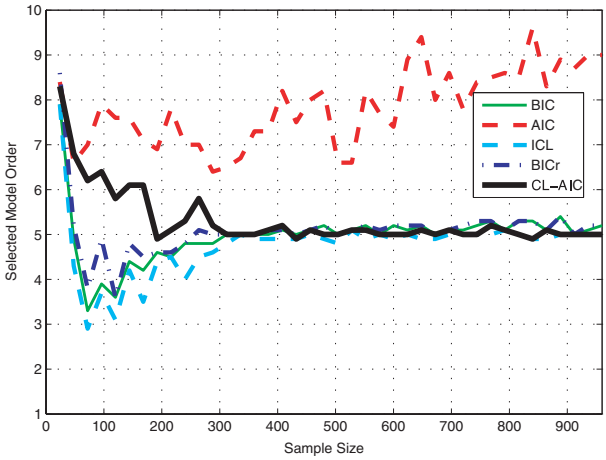
(b) Mean of the selected model orders

(c) Typical examples of the selected models

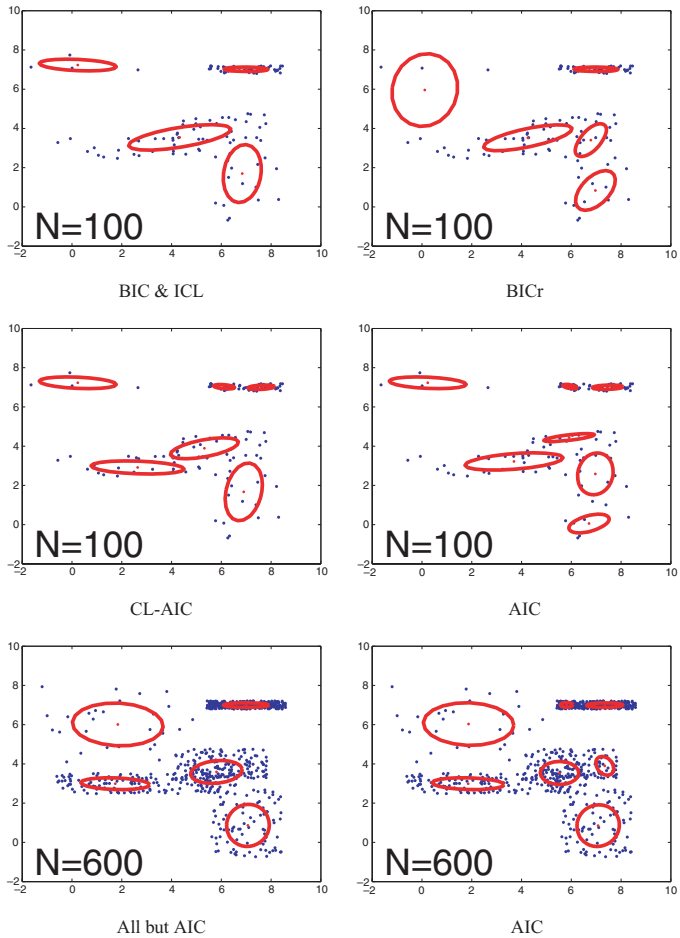
Figure 2. Model selection results for synthetic Gaussian data perturbed with uniformly distributed random noise.



(a) Selected Model orders



(b) Mean of the selected model orders



BIC & ICL BICr

CL-AIC AIC

All but AIC AIC

(c) Typical examples of the selected models

Figure 3. Model selection results for synthetic data of uniform distribution.

Table 3. Percentage of correct model order selection (over 50 trials) by different criteria for synthetic uniform data with 100 and 600 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
100	4	2	8	**54**	10
600	86	4	94	86	**100**

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.

5.4. Discussions

Overall, our experiments show that BICr is capable of rectifying the under-fitting tendency of BIC given small sample size. When the true mixture component kernel functions are different from the assumed ones, CL-AIC outperforms BIC, BICr, AIC and ICL given moderate to large sample sizes. A number of issues deserve further discussion, concerning the merits of different criteria:

1. Our experiments also show that all criteria tend to over-fit given extremely sparse data (e.g. $N < 2C_{K_{\text{true}}}$ where $C_{K_{\text{true}}}$ 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.
2. Given sparse data sets, all the model selection criteria are sensitive to the initialization of the EM algorithm for model parameter estimation, even though multiple initialisation strategy has been adopted to avoid this problem. This is illustrated by the large error bars in Figs. 1(a), 2(a), and 3(a). This problem is caused by the fact that each mixture component is weakly supported by the data samples given a sparse data set.
3. As the sample size increases, some of the mixture components are well supported while others are not. Our results show that those poorly supported mixture components are treated as noise by BIC and ICL (see examples of small sample model estimation results in Figs. 1(c), 2(c), and 3(c)). This explains the under-fitting tendency of BIC and ICL given small sample. As stated earlier in Section 3, the extra penalty term in the formulation of BICr ($\frac{1}{2}\sum_{k=1}^{K}\log\hat{w}_k$) favours those model candidates with weakly supported mixture components. BICr thus rectifies this under-fitting tendency given small, but not too small samples.
4. The superiority of CL-AIC over AIC is illustrated clearly by Figs. 1(b), 2(b), and 3(b). These figures highlight the contribution of the extra penalty term (the second term on the right hand side of Eq. (20)) on making CL-AIC a better model selection criterion than AIC. In particular, the absolute value of this extra penalty term grows with the sample size. Therefore, unlike AIC, the model orders selected by CL-AIC converge to a constant number as the sample size increases.
5. The more the true kernel functions differ from the assumed ones, the more likely it is for BIC and BICr to over-estimate the number of mixture components in order to better explain the data. On the other hand, CL-AIC utilises both the explanation and prediction capabilities of a mixture model. It is thus able to yield better model estimation, especially given moderate or large sample sizes.

We shall further demonstrate the above observations through experiments for unsupervised learning of visual context of three different dynamic scenes in the next section.

6. Discovering Visual Context

Experiments were conducted on discovering 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.

6.1. 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

(a) A typical scene (b) Motion trajectories (c) Inactivity points

Figure 4. Detecting inactivity points in a tearoom scenario.

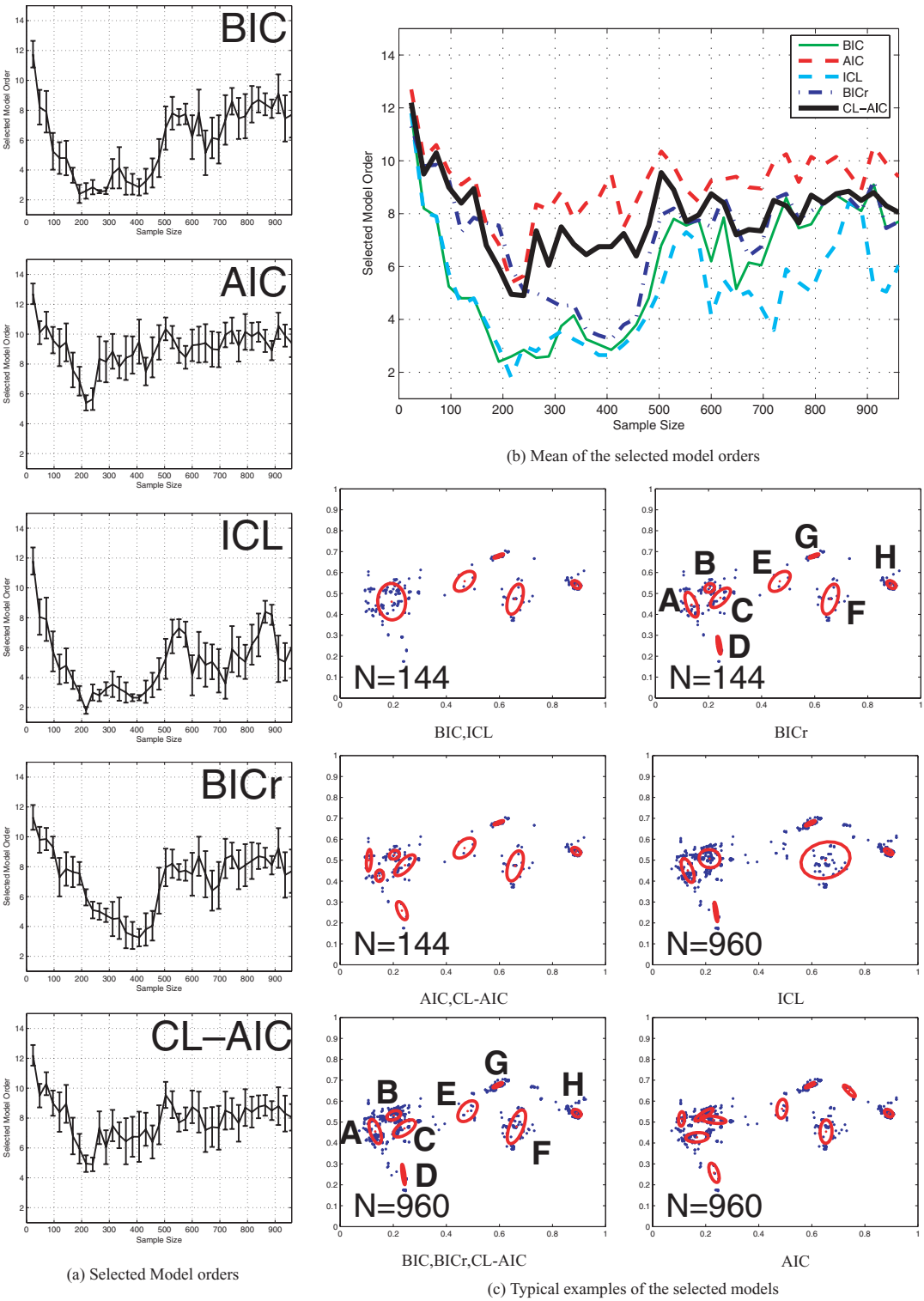
consists of a kitchenette on the top right hand side of the view and two dining tables located on the middle and left side of the view respectively (see Fig. 4(a)). Typical activities occurring in the kitchenette area included people making tea or coffee at the work surface, and people filling the kettle 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.

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 (McKenna, 2000) was employed which yielded temporally discretised motion trajectories (see Fig. 4(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 Fig. 4(c)). As can be seen in Fig. 4(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.

Table 4. Percentage of correct model order selection (over 50 trials) by different criteria for learning spatial context with 144 and 960 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
144	4	14	2	**38**	24
960	34	8	12	36	**58**

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 Fig. 5 and Table 4. It can be seen that all five criteria tended to over-fit given extremely sparse data sample (e.g. $N < 100$). When the sample size was small but not too small compared to the number of model parameters (e.g. $100 < N < 250$), all criteria turned into under-fitting, with BICr outperforming the other four. 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 Fig. 5(c) demonstrate that each estimated cluster corresponded to one inactivity zone when the model order was selected correctly.



(a) Selected Model orders

(b) Mean of the selected model orders

(c) Typical examples of the selected models

Figure 5. 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 (c) only when estimated correctly.

6.2. *Learning Facial Expression Context*

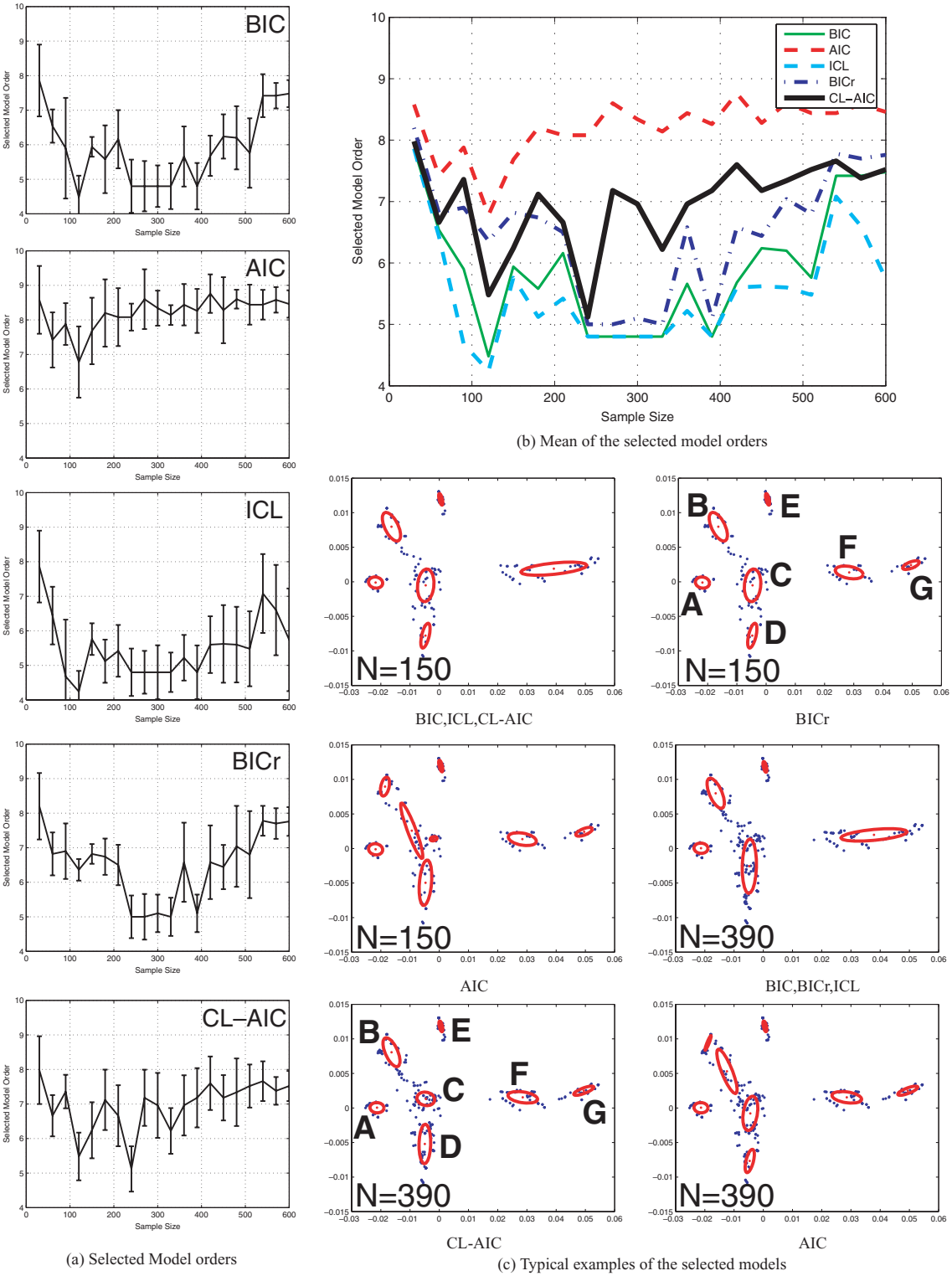
The visual task of modelling the dynamics of facial expressions and performing robust recognition becomes easier if key facial expression categories can be discovered and modelled. In this experiment, we aim to learn this important visual context using the shape of mouth. A face was modeled using the Active Appearance Model (AMM) (Cootes et al., 1998). The face model was learned using 1790 images sized 320×240 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 Fig. 6. Thirdly, the mouth shape data extracted from the test frames were projected onto a Mixture of Probabilistic Principal Component Analysis (MPPCA) space (Tipping and Bishop, 1999) 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. Details of data collection can be found in Zalewski and Gong (2004).

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 Fig. 7 and Table 5. It can be seen that the performance of different model selection criteria has similar characteristics to that demonstrated in the spatial context learning experiment. In particular, 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 BIC, ICL and CL-AIC. 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. It is also noted that AIC suffered from over-fitting whatever the sample size was. Figure 7(c) shows that, when the model order was selected as 7, each learned cluster corresponded correctly to each of the 7 facial expression categories.

Table 5. Percentage of correct model order selection (over 50 trials) by different criteria for learning facial expression context with 150 and 390 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
150	6	14	8	**54**	24
390	4	6	4	4	**40**

Figure 6. Top row: examples of image frames from the test data. From left to right, the facial expressions are neutral, smile, grin, sadness, fear, anger and surprise respectively. Bottom row: the corresponding mouth shapes extracted from the images.



(a) Selected Model orders

(b) Mean of the selected model orders

(c) Typical examples of the selected models

Figure 7. 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.

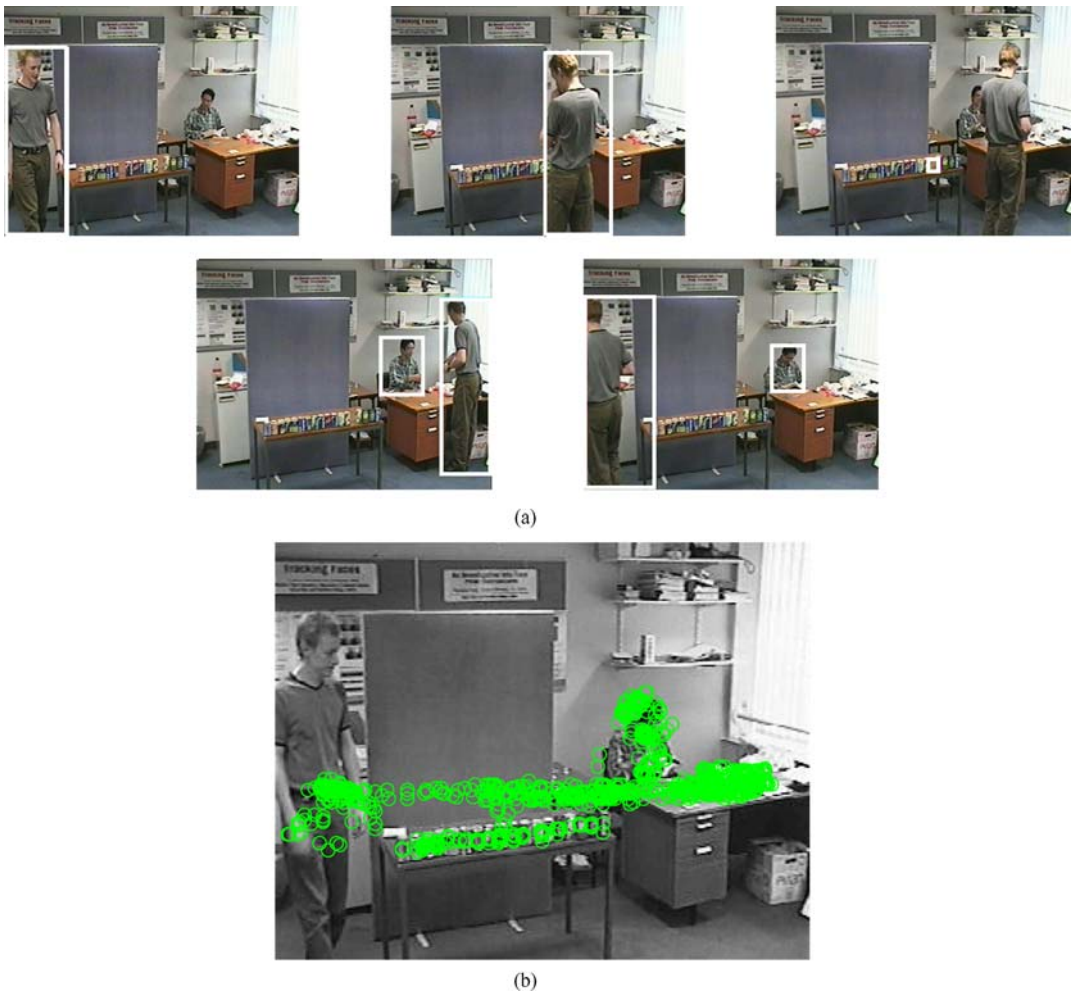


Figure 8. Automatic event detection in an image sequence of a simulated shopping scenario. Example image frames are shown in (a) with automatically detected scene events indicated using bounding boxes. The centroids of the 1642 scene events detected over the 19 minutes of video are shown in (b).

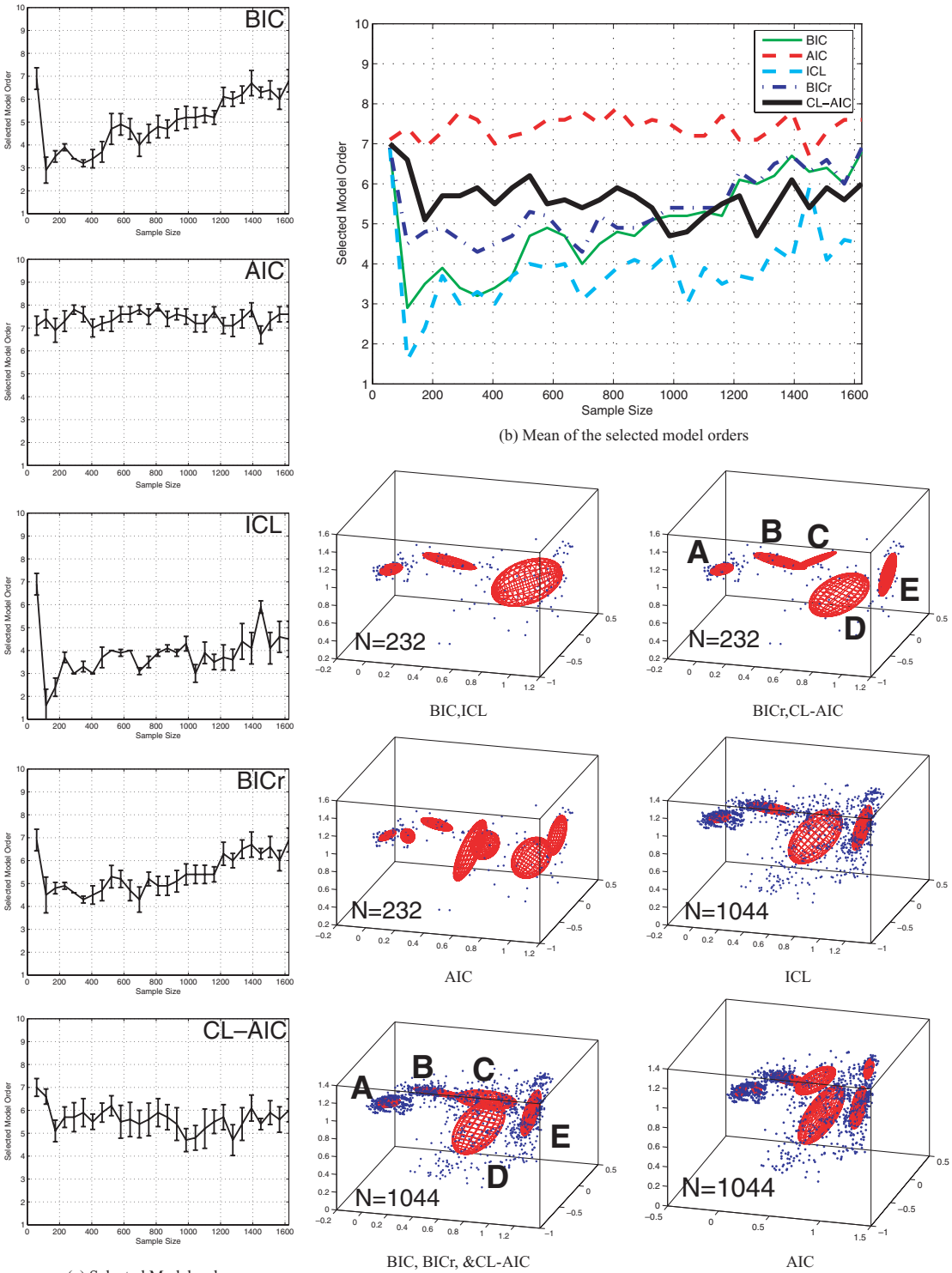
6.3. *Learning Scene Event Context*

A simulated shopping scenario was captured at 25 Hz, 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 Fig. 8(a). The scene consists of a shopkeeper sitting 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 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

Table 6. Percentage of correct model order selection (over 50 trials) by different criteria for learning scene event context with 232 and 1044 samples respectively.

	BIC	AIC	ICL	BICr	CL-AIC
232	4	2	2	**84**	32
1044	54	2	6	48	**56**



(a) Selected Model orders

(b) Mean of the selected model orders

(c) Typical examples of the selected models

Figure 9. 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.

by the location, shape and direction of the changes. These significant scene changes, referred to as scene events, are detected and clustered with the number of clusters being determined using a model selection criterion. 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 Fig. 8) and defined as a 7D feature vector (see Xiang et al. (2002) for details). A total of 1642 scene events were detected from the 19 minutes of video data (see Fig. 8(b)).

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 Figs. 9 and Table 6. Note that in Figs. 9(c) 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 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 before turning into slightly over-fitting, with CL-AIC performing slightly better than the other two. In the meantime, ICL remained under-fitting and AIC remained over-fitting. Examples of estimated models shown in Fig. 9(c) demonstrate that each estimated cluster corresponded to one scene event class when the model order was selected correctly.

6.4. Discussions

Our experiment results demonstrate the effectiveness of our BICr and CL-AIC model selection criteria on unsupervised learning of visual context. More specifically, given sparse data, BICr rectifies the under-fitting tendency of BIC and also outperforms ICL, AIC and CL-AIC. Given moderate to large data sample sizes, CL-AIC appears to be the best choice among the 5 criteria considered. By both direct observation of the data and comparing Figs. 5, 7 and 9 with Figs. 2 and 3, it appears that the true kernel functions of typical visual data are clearly non-Gaussian and severely overlapped. It is worth pointing out that the noise that is inevitably contained in the visual data can distribute in a very complex manner. For instance, we notice that the noise can form an additional cluster in the spatial context learning case, whereas the noise appeared to distribute randomly over the whole feature space in the scene event context learning case. Due to the existence of noise, most model selection criteria are more likely to over-fit even when sufficiently large data samples are available.

It is interesting to note that in our real data experiments, there were considerable amount of variations in the selected model orders by all the criteria even when the sample size was large (see the error bars in Figs. 5(a), 7(a) and 9(a)). Given large sample size, it is unlikely that these variations were caused by variations in the data distribution for different trials because the data sets for different trials had most data samples in common. These variations can only be explained by the sensitivity of the model selection criteria to the initialization of the EM algorithm for model parameter estimation. Again it is the noise contained in the visual data that should be blamed.

7. Conclusion

In conclusion, two novel probabilistic model selection criteria BICr and CL-AIC were proposed to improve existing model selection criteria for variable data sample sizes. The effectiveness of BICr and CL-AIC were demonstrated on discovering visual context information for dynamic scene modelling. Their performance is superior to that of a number of existing popular model selection criteria including BIC, AIC and ICL. Our study suggests that for modelling visual data using a mixture model, BICr is the most appropriate criterion given sparse visual data. When moderate or large data samples are available, CL-AIC should be chosen.

Our experiment results demonstrate that given a sparse visual data set, there are alway considerable variations in the selection model orders even when the proposed BICr criterion is used. Under this situation, one must be careful in selecting the optimal model order and estimating model parameters. Our results seem to suggest that selecting model structure over multiple trials using BICr followed by model averaging (Hoeting et al., 1995) could be a suitable strategy for unsupervised sparse visual data modelling. Our ongoing work

is along this line. Finally, it is worth pointing out that BICr and 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.

Acknowledgments

The authors thank Lukasz Zalewski for valuable discussions and helping with the facial expression data collection.

Notes

1. MDL formally coincides with BIC, although they are conceptually different.
2. The Dirichlet prior then becomes a noninformative prior (Bernardo and Smith, 1994).

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pp. 267–28.

Bernardo, J. and Smith, A. 1994. *Bayesian Theory*. Wiley and Sons.

Biernacki, C., Celeux, G., and Govaert, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Bishop, C. 1995. *Neural Networks for Pattern Recognition*. Cambridge University Press.

Brand, M. and Kettnaker, V. 2000. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):844–851.

Brand, M., Oliver, N., and Pentland, A. 1996. Coupled hidden markov models for complex action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, pp. 994–999.

Celeux, G. and Soromenho, G. 1996. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification*, 13:195–212.

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

Cherkassky, V. and Ma, Y. 2003. Comparison of model selection for regression. *Neural Computation*, 15(2):1691–1714.

Cohen, I., Sebe, N., Chen, L., Garg, A., and Huang, T. 2003. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187.

Cootes, T.F., Edwards, G.J., and Taylor, C.J. 1998. Active appearance models. In *European Conference on Computer Vision*, Freiburg, Germany, pp. 484–498.

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

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

Figueiredo, M. and Jain, A.K. 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Fitzgerald, W. 1996. *Numerical Bayesian Methods Applied to Signal Processing*. Springer Verlag.

Gath, I. and Geva, B. 1989. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–781.

Gong, S. and Xiang, T. 2003. Recognition of group activities using dynamic probabilistic networks. In *IEEE International Conference on Computer Vision*, pp. 742–749.

Haritaoglu, I., Harwood, D., and Davis, L.S. 2000. w^4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The elements of statistical learning: Data mining, inference and prediction*. Springer.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. 1995. Bayesian model averaging, a tutorial. *Statistical Science*, 14:382–417.

Hongeng, S. and Nevatia, R. 2001. multi-agent event recognition. In *IEEE International Conference on Computer Vision*, pp. 80–86.

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

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

Johnson, N., Galata, A., and Hogg, D. 1998. The acquisition and use of interaction behaviour models. In *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, USA, pp. 866–871.

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

Kullback, S. 1968. *Information Theory and Statistics*. Dover: New York.

Lange, T., Roth, V., Braun, M.L., and Buhmann, J.M. 2004. Stability-based validation of clustering solutions. *Neural Computation*, 16:1299–1323.

McKenna, S., Jabri, S., Duric, Z., Rosenfeld, A., and Wechsler, H. 2000. Tracking group of people. *Computer Vision and Image Understanding*, 80:42–56.

McKenna, S. and Nait-Charif, H. 2004. Learning spatial context from tracking using penalised likelihoods. In *International Conference on Pattern Recognition*, pp. 138–141.

Mclachlan, G. and Peel, D. 1997. *Finite Mixture Models*. John Wiley & Sons.

Oliver, N., Rosario, B., and Pentland, A. 2000. A bayesian computer vision system for modelling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843.

Raftery, A. 1995. Bayes model selection in social research. *Sociological Methodology*, 90:181–196.

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

Roberts, S. 1997. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272.

Roberts, S., Husmeier, D., Rezek, I., and Penny, W. 1998. Bayesian approaches to Gaussian mixture modelling. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 20(11):1133–1142.

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

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

Stauffer, C. and Grimson, W. 2000. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–758.

Tian, Y., Kanade, T., and Cohn, J. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:97–115.

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

Wada, T. and Matsuyama, T. 2000. Multiobject behavior recognition by event driven selective attention method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):873–887.

Xiang, T., Gong, S., and Parkinson, D. 2002. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. In *British Machine Vision Conference*, pp. 233–242.

Zalewski, L. and Gong, S. 2004. Modelling facial expression as probabilistic hierarchical dynamical states. Technical Report 0043, Vision Lab, Queen Mary, University of London.