With the underlying probability model, the problem of choosing the number of components can be reformulated as a statistical model choice problem. The purpose of model selection is to select a candidate model among a finite collection of models *{, . . . , }*, in order to estimate function from which the data are drawn. Each model is characterized by a density whose parameters are of the particular dimension. Testing for the number of components in a mixture appears to be difficult since the classical likelihood ratio test does not hold for mixtures. On the contrary, criteria based on penalized likelihood, such as the Bayesian Information Criterion (BIC) have been successfully applied to mixture models (Picard 2007). Nevertheless, it appears that those criteria do not consider the specific objective of mixture models in the clustering context (taking into account the modelling purpose when selecting a model would lead to use data-driven penalizations favoring useful and parsimonious models). To overcome this limitation, it can be advantageous to choose number of clusters in order to get the mixture giving rise to partitioning data with the greatest evidence. This has lead to the construction of classification-based criteria (Picard 2007).

When choosing a parsimonious model in a collection of models, the problem is to solve the bias-variance dilemma:

* a too simple model leads to a large approximation error;
* a too complex model leads to a large estimation error.

Standard criteria of model selection are AIC and BIC criteria. Both  
criteria are penalized likelihood criteria:

(1.10)

(1.11)

Here denote the data of probability density function *p(x).*  
A model *m* is characterized with the probability density function *p(x|).*

; is the number of parameters of model *m .*

In the Bayesian context, model density parameters and are viewed as random variables with prior distributions when model is fixed. This formulation is flexible since additional information can be modelled through prior distributions, and if no information is available a non-informative prior can be used. The BIC aims at selecting the model which maximizes the posterior probability *P(|X)* (Picard 2007). Picard (2007) states that regularity conditions for BIC do not hold for mixture models, since the estimates of some mixing proportions can be on the boundary of the parameter space.

Other criteria that have been defined for the special case of mixture models can be based on Bayesian methods, on the entropy function of the mixture, or on information theory (Picard 2007). The use of the BIC can lead to an overestimation of the number of clusters regardless the clusters separation. Moreover estimating the “true” numbers of clusters, which is the objective of the BIC, is not necessarily suitable in a practical context. For these reasons, a new criterion, the Integrated Classification Criterion (ICL) that considers the clustering objective of mixture models, was proposed (Picard 2007). It leads to consider the integrated likelihood of the complete data (or integrated completed likelihood). To approximate this integrated complete likelihood, a BIC-like ap-  
proximation is possible. It leads to the criterion , where the missing data have been replaced by their most probable value for parameter estimate ˆθ. Roughly speaking criterion ICL is the criterion BIC penalized by the estimated mean entropy. Because of this additional entropy term, ICL favors values giving rise to partitioning the data with the greatest evidence. ICL appears to provide a stable and reliable estimate of for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping. But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components.

The performance of ICL have been tested based on real and simulated data sets. Picard (2007) emphasizes that compared with BIC, ICL tends to select a lower number of clusters which provides good clustering results in real situations, compared with BIC which tends to select a too overly high number of clusters. When the data are simulated, ICL tends to select a lower number of clusters if the groups are not well separated, contrary to BIC which finds the true number of classes (Picard 2007).