# Model selection criteria in Classification contexts

Gilles Celeux

INRIA Futurs (orsay)

# Cluster analysis

Exploratory data analysis tools which aim is to find clusters in a large set of data (many observations and often many variables).

# Supervised Classification

Statistical decision methods which aim is to design a classifier to assign in the future unlabelled observations to one of the classes defined a priori.

# The mixture model

Data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{nd}$ are assumed to arise from a mixture

$$p(\mathbf{x}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k)$$

- the $p_k$'s are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$)
- $\phi(. \mid \mathbf{a}_k)$ denotes a parameterized density (usually the $d$-dimensional Gaussian density) with parameter $a_k$,
- $\theta_K = (p_1, \ldots, p_{K-1}, a_1, \ldots, a_K)$.

A mixture model involves label data $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ which are binary vectors with $z_{ik} = 1$ if and only if $\mathbf{x}_i$ arises from component $k$. Those indicator vectors define a partition $P = (P_1, \ldots, P_K)$ of data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \mid z_{ik} = 1\}$.

# Model selection

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.

# AIC vs. BIC

$$\text{AIC}(m) = -2 \log \mathbf{p}(\mathbf{x}|\hat{\theta}_m) + 2\nu_m,$$

$$\text{BIC}(m) = -2 \log \mathbf{p}(\mathbf{x}|\hat{\theta}_m) + \nu_m \log(n),$$

where

- $\mathbf{x} = (x_1, \ldots, x_n)$ denote the data of pdf $\mathbf{p}(\mathbf{x})$
- A model $m$ is characterized with the pdf $\mathbf{p}(\mathbf{x}|\theta_m)$.
- $\hat{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}|\theta_m)$
- $\nu_m =$ is the number of parameters of model $m$.

# Rationale for AIC (1)

Find $(m*, \theta^0_{m*}) \in \mathcal{M}$ minimizing

$$KL(\mathbf{p}, \mathbf{p}(.|\theta^0_m)) = E[\log \mathbf{p}(\mathbf{x})] - E[\log \mathbf{p}(\mathbf{x}|\theta_m)]$$

It is equivalent to find $(m*, \theta^0_{m*})$ such that

$$\theta^0_{m*} = \arg\max E[\log \mathbf{p}(\mathbf{x}|\theta_m)].$$

For fixed $m$, the maximum likelihood estimate $\hat{\theta}_m$ is a consistent estimator of $\theta^0_m$ from SLLN. But, $KL(\mathbf{p}, \mathbf{p}(.|\hat{\theta}_m))$ is an optimistic estimate of $KL(\mathbf{p}, \mathbf{p}(.|\theta^0_m))$.
This optimistic bias is

$$D_m = E_{\mathbf{x}x'}[\log p(x') - \log p(x'|\hat{\theta}_m)]$$

where $x'$ is an observation independent of data $\mathbf{x}$.

# Rationale for AIC (2)

Denoting

$$K_m = \mathsf{Var}[\frac{\partial \log \mathbf{p}(\mathbf{x}|\theta_m^0)}{\partial \theta}],$$

$$J_m = E[\frac{\partial^2 \log \mathbf{p}(\mathbf{x}|\theta_m^0)}{\partial \theta \partial \theta^t}]$$

and

$$q_m = \mathsf{trace}(K_m J_m^{-1}),$$

we have

$$2nD_m* = 2KL(\mathbf{p}, \mathbf{p}(.|\hat{\theta}_m*)) + 2q_m* + O(n^{-1/2}).$$

If there is $m* \in \mathcal{M}$ such that $\mathbf{p} = \mathbf{p}(.|\theta_{m*}^0)$ then $K_m* = J_m*$ and $q_m* = \nu_m*$.

This assumption is made to get the AIC approximation of the deviance $2D_m$.

# AIC asymptotic properties

- If $\mathbf{p} \in \mathcal{M}$, AIC and cross validation estimation of the expected deviance provide asymptotically the same model selection.
- In a regresssion framework: $y_i = f(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$
  - If the number of models with the same dimension does not grow too fast, then the mean squared error of the selected model with AIC is asymptotically equivalent to the smallest error which can be get with $\mathcal{M}$.
  - AIC is minimax optimal.
- AIC is not consistent.

# Rationale for BIC (1)

BIC is a Bayesian criterion. It is approximating asymptotically the integrated likelihood of the model $m$

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|\theta_m)\pi(\theta_m)d\theta_m,$$

$\pi(\theta_m)$ being a prior distribution for parameter $\theta_m$.

This integrated likelihood is a predictive score which allows to compare two models with the Bayes factor $B_{21}$:

$$\frac{\mathbf{p}(m_2|\mathbf{x})}{\mathbf{p}(m_1|\mathbf{x})} = \frac{\mathbf{p}(\mathbf{x}|m_2)}{\mathbf{p}(\mathbf{x}|m_1)}\frac{p(m_2)}{p(m_1)}$$

# Rationale for BIC (2)

BIC is making use of the Laplace approximation for the integrated likelihood of model $m$

$$\int_{R^d} \exp(nL(u))du = \exp\left(nL(u*)\right) \frac{2\pi}{n}^{d/2} |-L''(u*)|^{-1/2} + O(n^{-1})$$

with $L : R^d \longrightarrow R$ is a $C^2$ function with unique maximum $u*$. Here

$$L(\theta_m) = \frac{1}{n}[\log \mathbf{p}(\mathbf{x}|\theta_m) + \log \pi(\theta_m)].$$

Moreover, the posterior mode $\theta_m^*$ is approximated with $\hat{\theta}_m$ and the Hessian of $L$ is approximated with the Fisher information evaluated at $\hat{\theta}_m$. Thence

$$\log \mathbf{p}(\mathbf{x}|m) = BIC(m) + O(1).$$

If the prior pdf $\pi$ is a normal distribution $N(\hat{\theta}_m, I_{\theta_m}^{-1})$ then

$$\log \mathbf{p}(\mathbf{x}|m) = BIC(m) + O(n^{-1/2})$$

# BIC asymptotic properties

- BIC is consistent: If there exists $m*$ such that $\mathbf{p} = \mathbf{p}(.|m*)$ then for n large enough, BIC selects $m*$.
- The existence of $m*$ is not necessary to design BIC and a good behavior of BIC can be expected if $\mathbf{p} \approx \mathbf{p}(.|m*)$.
- In a regresssion framework: $y_i = f(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$, BIC is not minimax optimal.
- BIC does not lead to a prediction asymptotically optimal for non parametric regression problems.

# The practice

- Bias of Monte Carlo numerical experiments.
- Different practical behavior according to the modelling setting.
  - Assessing the number of components $K$ in a Gaussian mixture model:

    AIC has a high tendency to overestimate $K$.

    BIC has a more satisfactory behavior.
  - Model selection in Regression: AIC is almost equivalent to the Mallows Cp criterion.
  - AIC is a reference criterion to assess the order of ARMA models, BIC is not.

# Motivation of the present talk

- Assuming that the data arose from one of the models in competition is unrealistic and can be misleading when using AIC or BIC.
- A common feature of standard penalized likelihood criteria is to not take into account the modelling purpose.
- Our opinion is that taking account of the modelling purpose when selecting a model would lead to use data-driven penalisations favoring useful and parsimonious models.
- This view point is exploited in a classification context.

# Model-based cluster analysis

- Model-based clustering (MBC) consists of assuming that the data come from a source with several subpopulations.
- Each subpopulation is modeled separetaly.
- The overall population is a mixture of these subpopulations.
- The resulting model is a finite mixture model.

# The mixture model

Data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{nd}$ are assumed to arise from a mixture

$$p(\mathbf{x}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k)$$

- the $p_k$'s are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$)
- $\phi(. \mid \mathbf{a}_k)$ denotes a parameterized density (usually the $d$-dimensional Gaussian density) with parameter $a_k$,
- $\theta_K = (p_1, \ldots, p_{K-1}, a_1, \ldots, a_K)$.

A mixture model involves missing data $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ which are binary vectors with $z_{ik} = 1$ if and only if $\mathbf{x}_i$ arises from component $k$.

# An hidden structure model

The mixture model is an incomplete data structure model:
The complete data are

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$$

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ are binary vectors such that $z_{ik} = 1$ iff $\mathbf{x}_i$ arises from group $k$.

The $\mathbf{z}$'s define a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \mid z_{ik} = 1\}$.

# Multivariate Gaussian Mixture (MGM)

Multidimensional observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^d$ are assumed to be a sample from a probability distribution with density

$$f(\mathbf{x}_i \mid K, \theta) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i \mid \mathbf{m}_k, \Sigma_k)$$

where the $p_k$'s are the mixing proportions and $\phi(. \mid \mathbf{m}_k, \Sigma_k)$ denotes a Gaussian density with mean $\mathbf{m}_k$ and variance matrix $\Sigma_k$.

This is the most popular model for clustering of quantitative data.

# Discrete Data

Observations to be classified are described with $d$ discrete variables.
Each variable $j$ has $m_j$ response levels.

Data are represented in the following way:

$(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ with

$$\begin{cases} x_i^{jh} = 1 & \text{if } i \text{ has response level } h \text{ for variable } j \\ x_i^{jh} = 0 & \text{otherwise.} \end{cases}$$

# The standard latent class model (LCM)

Data are supposed to arise from a mixture of $g$ multivariate multinomial distributions with pdf

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} p_k m_k(\mathbf{x}_i;\boldsymbol{\alpha}_k) = \sum_{k} p_k \prod_{j,h} (\alpha_k^{jh})^{x_i^{jh}}$$

where

- $\alpha_k^{jh}$ is denoting the probability that variable $\mathbf{x}^j$ has level $h$ if object $i$ in cluster $k$, and $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, p; h = 1, \ldots, m_j)$,
- $\mathbf{p} = (p_1, \ldots, p_g)$ is denoting the vector of mixing proportions of the $g$ latent clusters,
- $\boldsymbol{\theta} = (p_k, \boldsymbol{\alpha}_k, k = 1, \ldots, g)$ denoting the vector parameter of the latent class model to be estimated.

Latent class model is assuming that the variables are conditionnally independent knowing the latent clusters.

# First Interest of MBC

Many versatile or parsimonious models available

## MGM: The variance matrix eigenvalue decomposition

$$\Sigma_k = V_k D_k^t A_k D_k$$

where

- $V_k = |\Sigma_k)|^{1/d}$ defines the component volume ($d$ is the dimension of the observation space),
- $D_k$ the matrix of eigenvectors of $\Sigma$ defines the component orientation
- $A_k$ the diagonal matrix of normalised eigenvalues defines the component shape.

By allowing some of these quantities to vary between components, we get different and easily interpreted models.

## 28 different models

Following Banfield & Raftery (1993) or Celeux & Govaert (1995), a large range of 28 versatile (from the most complex to the simplest one) models derived from this eigenvalue decomposition can be considered.

- The general family: Assuming equal or free proportions, volumes orientations and shapes leads to 16 different models.
- The diagonal family: Assuming in addition that the component variances matrices are diagonal leads to 8 models.
- The spherical family: Assuming in addition that the component variance matrices are proportional to the identity matrix leads to 4 models.

# LMC: a Reparameterization

$$\boldsymbol{\alpha} \longleftrightarrow (\mathbf{a}, \boldsymbol{\varepsilon})$$

where binary vector $\mathbf{a}_k = (\mathbf{a}_k^1, \ldots, \mathbf{a}_k^d)$ provides the mode levels in cluster $k$ for variable $j$

$$(a^{jh}) = \begin{cases} 1 & \text{if } h = \arg\max_h \alpha^{jh} \\ 0 & \text{otherwise.} \end{cases}$$

and the $\varepsilon_k^j$ can be regarded as scattering values.

$$(\varepsilon^{jh}) = \begin{cases} 1 - \alpha^{jh} & \text{if } a^{jh} = 1 \\ \alpha^{jh} & \text{if } a^{jh} = 0. \end{cases}$$

For instance, if $\boldsymbol{\alpha}^j = (0.7, 0.2, 0.1)$, the new parameters are $\mathbf{a}^j = (1, 0, 0)$ and $\boldsymbol{\varepsilon}^j = (0.3, 0.2, 0.1)$.

# Five latent class models

Denoting $h(ij)$ the level of object $i$ for the variable $j$, the model can be written

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_k p_k \prod_j \left( (1-\varepsilon_k^{jh(jk)})^{x_i^{jh(jk)}} (\varepsilon_k^{jh(ij)})^{x_i^{jh(ij)} - x_k^{jh(jk)}} \right).$$

Using this form, it is possible to impose various constraints to the scattering parameters $\varepsilon_k^{jh}$. The models we consider are the following

- the standard latent class model $[\varepsilon_k^{jh}]$: The scattering is depending upon clusters, variables and levels.
- $[\varepsilon_k^{j}]$: The scattering is depending upon clusters and variables but not upon levels.
- $[\varepsilon_k]$: The scattering is depending upon clusters, but not upon variables.
- $[\varepsilon^{j}]$: The scattering is depending upon variables, but not upon clusters.
- $[\varepsilon]$: The scattering is constant upon variables and clusters.

# Maximum likelihood estimation

The EM algorithm is the reference tool to derive the ML estimates in a mixture model.

- **E step** Compute the conditional probabilities $t_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, K$ that $x_i$ arises from the $k$th component for the current value of the mixture parameters.
- **M step** Update the mixture parameter estimates maximising the expected value of the completed likelihood. It leads to weight the observation $i$ for group $k$ with the conditional probability $t_{ik}$.

# Assessing a mixture model

- In an unsupervised setting, the question is: Choose a sensible mixture model with an adequate number of components.
- But, a good answer to this question depends of the focus of the mixture model.

# Choosing a mixture model in density estimation context

In a Bayesian perspective, a classical way for choosing a model is to select the model maximizing the integrated likelihood,

$$\mathbf{f}(\mathbf{x} \mid K) = \int \mathbf{f}(\mathbf{x} \mid K, \theta)\pi(\theta \mid K)d\theta,$$

$$\mathbf{f}(\mathbf{x} \mid K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid K, \theta),$$

$\pi(\theta \mid K)$ being a non or weakly informative prior distribution on $\theta$. It can be approximated with the BIC criterion

$$\log \mathbf{f}(\mathbf{x} \mid K) \approx \log \mathbf{f}(\mathbf{x} \mid K, \hat{\theta}) - \frac{\nu_K}{2} \log(n),$$

where $\hat{\theta}$ is the m.l. estimate of $\theta$ and $\nu_K$ is the number of free parameters of the model. Simulation experiments (see Roeder and Wasserman 1997) show that BIC works well at a practical level.

# Assessing $K$ in a clustering context

Assessing the number of components $K$ is an important but difficult problem.

Mixture modelling can be regarded as a semi parametric tool for density estimation purpose or as a model for cluster analysis.

- In the density estimation context, BIC is doing the job quite well.
- In the cluster analysis context, since BIC does not take into account the clustering purpose for assessing $K$, BIC has a tendency to overestimate $K$ regardless of the separation of the clusters.

To overcome this limitation, it can be advantageous to choose $K$ in order to get the mixture giving rise to partitioning data with the greatest evidence.

# The ICL criterion: definition

It leads to consider the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ (or integrated completed likelihood),

$$\mathbf{p}(\mathbf{x}, \mathbf{z} \mid K) = \int_{\Theta_K} \mathbf{p}(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta,$$

where

$$\mathbf{p}(\mathbf{x}, \mathbf{z} \mid K, \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta)$$

with

$$p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta) = \prod_{k=1}^{K} p_k^{z_{ik}} \left[\phi(\mathbf{x}_i \mid \mathbf{a}_k)\right]^{z_{ik}}.$$

To approximate this integrated complete likelihood, a BIC-like approximation is possible. It leads to the criterion

$$\text{ICL}(K) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} \mid K, \hat{\theta}) - \frac{\nu_K}{2} \log n,$$

where the missing data have been replaced by their most probable value for parameter estimate $\hat{\theta}$.

# Behavior of the ICL criterion

Roughly speaking criterion ICL is the criterion BIC penalized by the estimated mean entropy

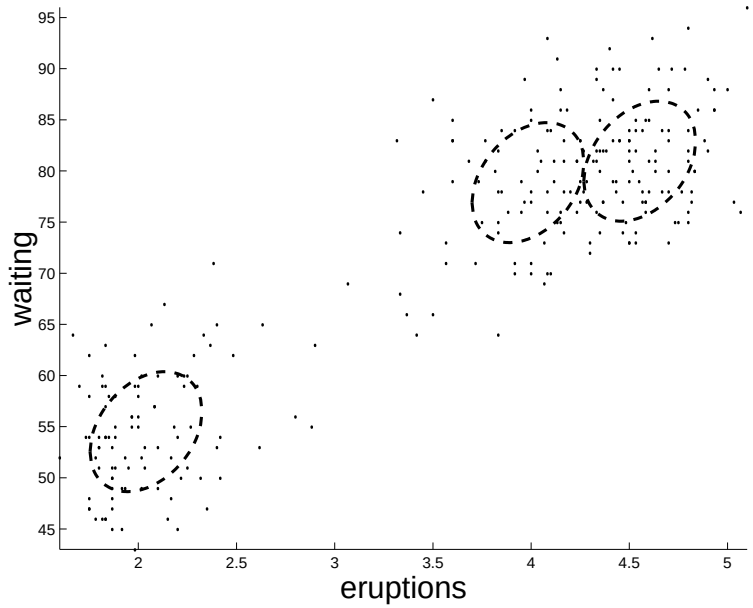
$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik} \log t_{ik} \geq 0,$$

$t_{ik}$ denoting the conditional probability that $\mathbf{x}_i$ arises from the $k$th mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$).
Because of this additional entropy term, ICL favors $K$ values giving rise to partitioning the data with the greatest evidence.

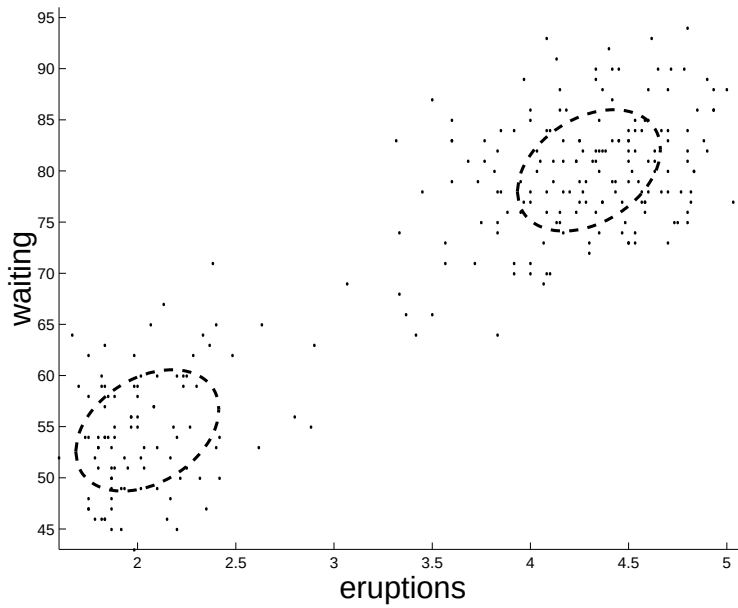
- ICL appears to provide a stable and reliable estimate of $K$ for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping.
- But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components.

# BIC vs. ICL: an illustration

BIC




ICL



# Supervised classification

- The problem of supervised classification is to assign a $d$-dimensional vector $\mathbf{x}$ to one class from $g$ classes $C_1, \ldots, C_g$.
- A decision function, called a classifier, $\delta(\mathbf{x}) : \mathbf{R}^d \rightarrow \{1, \ldots, g\}$ is to be designed from a learning sample $(\mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, n$,

  $\mathbf{x}_i$ being the description vector of observation $i$ on $d$ variables and $\mathbf{z}_i$ denoting the label indicator vector of observation $i$:

$$z_{ik} = \begin{cases} 1 \text{ if } i \in C_k \\ 0 \text{ otherwise} \end{cases}$$

# Generative and Discriminative approaches

- The generative approach is representing the class conditional densities with a parametric model $\mathbf{p}(\mathbf{x}|m, z_k = 1, \theta_m)$ for $k = 1, \ldots, g$.
- Examples of generative classifiers are Linear Discriminant Analysis, Quadratic Discriminant Analysis.
- The discriminative approach is representing the conditional probability of a class with a semi parametric or a non parametric model.
- Examples of discriminative classifiers are Logistic Regression, Classification trees (CART), Support Vector Machines (SVM), $k$ neearest neighbors, ...

# The generative approach

- The parameter $\theta_m$ of the parametric model $\mathbf{p}(\mathbf{x}|m, z_k = 1, \theta_m)$ for $k = 1, \ldots, g$ is estimated from the learning sample $(\mathbf{x}_i, z_i), i = 1, \ldots, n$.
- Then the classifier consists of assigning an observation $\mathbf{x}$ to the class $k$ maximizing the estimated conditional probability of a class $p(z_k = 1|m, \mathbf{x}, \hat{\theta}_m)$.
- It leads to set $\delta(\mathbf{x}) = j$ if and only if

$$j = \arg\max_k p_k \mathbf{p}(\mathbf{x}|m, z_k = 1, \hat{\theta}_m),$$

$\hat{\theta}_m$ being the ml estimate of the class conditional parameters $\theta$ and $p_k$ being the prior probability of class $k$.

# Model selection in supervised classification

It is often of interest to consider a large collection of models with different numbers of parameters and to select the model expected to provide the lowest actual error rate.

- Minimizing the $v$-fold cross-validated error rate can be regarded as a nearly optimal solution. But it is highly CPU time consuming and the choice of $v$ can be sensitive.
- An alternative is BIC which takes the form

$$\text{BIC}(m) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \frac{\nu_m}{2} \log(n),$$

  $\nu_m$ being the dimension of $\theta_m$. But, BIC measures the fit of the model $m$ to the data $(\mathbf{x}, \mathbf{z})$ rather than its ability to produce a reliable classifier. . .

# The Bayesian Entropy Criterion

The classifier related to model $m$ is designed from the conditional likelihood $\mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta_m)$.

Thus, instead of choosing the model maximizing the integrated likelihood, we propose to select a relevant model by maximizing the integrated conditional likelihood

$$\mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \int \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta_m)\pi(\theta_m|\mathbf{x})d\theta_m,$$

$\pi(\theta_m|\mathbf{x})$ being the posterior distribution of $\theta_m$ knowing $\mathbf{x}$. Acting in such a way, we are measuring the ability of model $m$ to answer the classification task rather than its fit to the data $(\mathbf{x}, \mathbf{z})$.

The BEC criterion is approximating $\log \mathbf{p}(\mathbf{z}|m, \mathbf{x})$.

# Computing BEC (1)

We have

$$\mathbf{p}(\mathbf{z}|m,\mathbf{x}) = \frac{\mathbf{p}(\mathbf{x},\mathbf{z}|m)}{\mathbf{p}(\mathbf{x}|m)}$$

with

$$\mathbf{p}(\mathbf{x},\mathbf{z}|m) = \int \mathbf{p}(\mathbf{x},\mathbf{z}|m,\theta_m)\pi(\theta_m)d\theta_m,$$

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|m,\theta_m)\pi(\theta_m)d\theta_m.$$

# Computing BEC (2)

Both log-integrals can be approximated with the BIC criterion:

$$\log \mathbf{p}(\mathbf{x}, \mathbf{z}|m) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \frac{\nu_m}{2} \log n + \mathsf{O}(1)$$

$$\log \mathbf{p}(\mathbf{x}|m) = \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m) - \frac{\nu_m}{2} \log n + \mathsf{O}(1),$$

with

$$\hat{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \theta_m),$$

$$\tilde{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}|m, \theta_m)$$

$\nu_m$ being the dimension of the vector parameter $\theta_m$. Thus

$$\log \mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m) + \mathsf{O}(1).$$

And, the BEC criterion is

$$\mathsf{BEC} = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m).$$

# Why the name BEC?

Denoting $t_{ik}(m, \hat{\theta}_m)$ the conditional probability that $\mathbf{x}_i$ arises from class $k$ in model $m$ with ml parameter estimate $\hat{\theta}_m$, we have

$$\log \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \hat{\theta}) = \sum_{i=1}^{n} \log t_{iz_i}(m, \hat{\theta}_m)$$

which can be regarded as the entropy of the classification $\mathbf{z}$.

And, our criterion is related to this term.

This is the reason why we called it Bayesian Entropy Criterion (BEC).

# Deriving $\tilde{\theta}$

The criterion BEC needs to compute $\tilde{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}|m, \theta_m)$. Since, for $i = 1, \ldots, n$,

$$p(\mathbf{x}_i|m, \theta_m) = \sum_{k=1}^{g} p(z_{ik} = 1|m, \theta_m) p(\mathbf{x}_i|z_{ik} = 1, m, \theta_m),$$

$\tilde{\theta}$ is the ml estimate of a finite mixture distribution.

- It can be derived from the EM algorithm initiated with $\hat{\theta}$.
- Moreover, when the learning data set has been obtained through the diagnosis paradigm, the proportions in the mixture distribution are fixed: $p_k = \text{card}\{i \text{ such that } z_{ik} = 1\}/n$ for $k = 1, \ldots, g$.
- Thus $\tilde{\theta}$ would be estimated in a stable and reliable way.

# Alternative criteria?

A BIC-like approximation of $\log \mathbf{p}(\mathbf{y}|\mathbf{x})$

$$\log \mathbf{p}(\mathbf{y}|\mathbf{x}) \approx \log \mathbf{p}(\mathbf{y}|\mathbf{x}, \theta_m^\star) - \frac{\nu_m}{2} \log n,$$

where

$$\theta_m^\star = \arg\max_{\theta_m} \mathbf{p}(\mathbf{y}|\mathbf{x}, \theta_m),$$

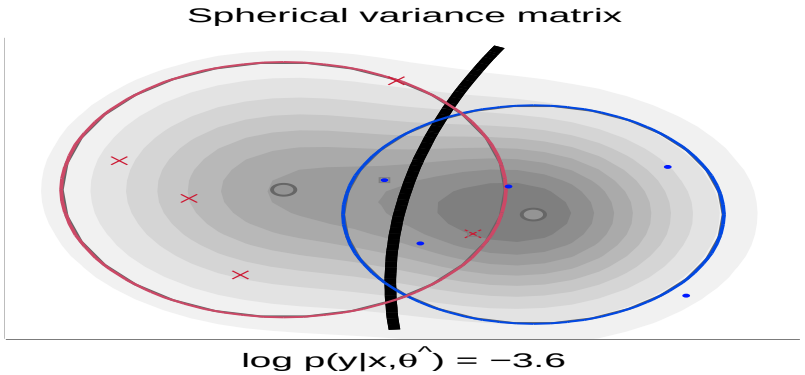
is not valid since the posterior distribution $\pi(\theta_m|\mathbf{x})$ depends on $n$.

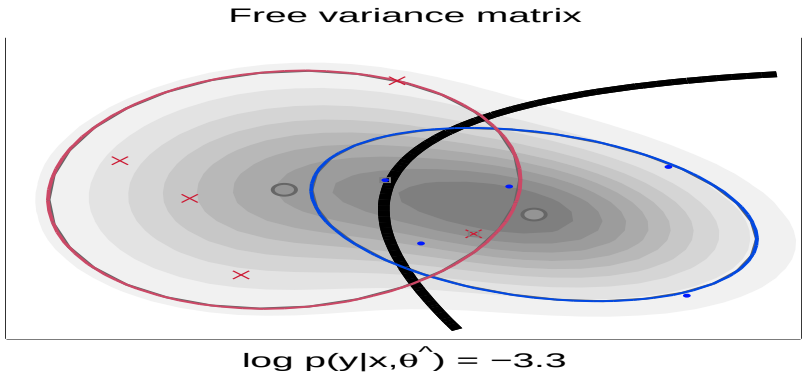
In a discriminative approach of classification for which $\mathbf{x}$ is assumed to be not dependent on $\theta$ this BIC-like approximation would be valid.

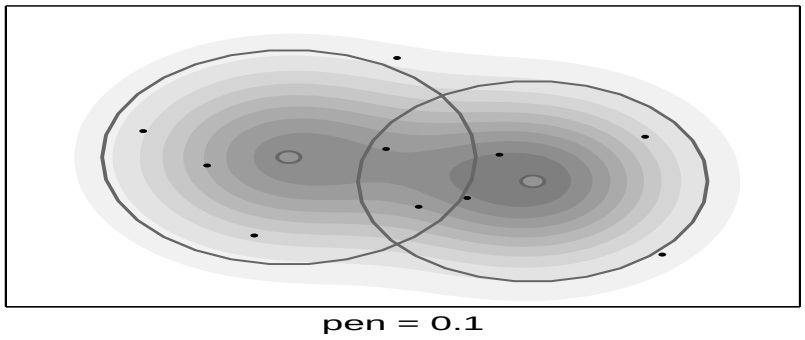
# BEC as a penalized likelihood criterion

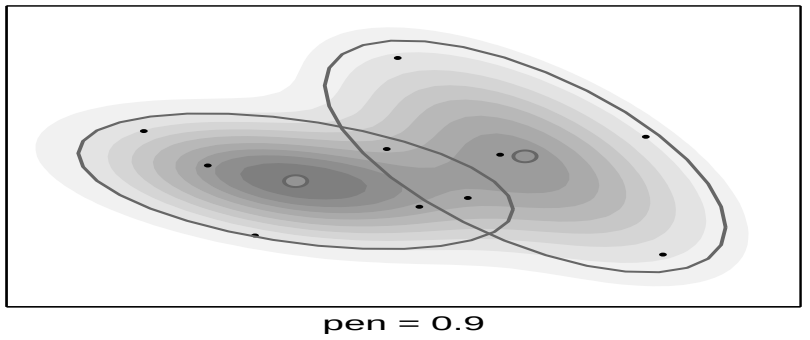
$$\text{BEC} = \log \mathbf{p}(\mathbf{y}|\mathbf{x}, \hat{\theta}_m) - \Big(\log \mathbf{p}(\mathbf{x}|\tilde{\theta}_m) - \log \mathbf{p}(\mathbf{x}|\hat{\theta}_m)\Big).$$

$\log \mathbf{p}(\mathbf{x}|\tilde{\theta}_m) - \log \mathbf{p}(\mathbf{x}|\hat{\theta}_m)$ is positive because $\tilde{\theta}$ maximizes the marginal likelihood $\mathbf{p}(\mathbf{x}|\theta_m)$. This penalty is minimum when $\hat{\theta} = \tilde{\theta}$. Its implicit dependency on the model complexity is now illustrated



# Asymptotic behavior of BEC (1)

Proposition 1

If the sample joint distribution belongs to exactly one model $m^*$ in a finite family of candidate models $\{m_1, \ldots, m_M\}$,
and under standard regularity conditions on the model family,
Then BEC criterion selects $m^*$ with probability one as the sample size $n$ of the training set tends to infinity.

# Asymptotic behavior of BEC (2)

Proposition 2 Assuming that the true distribution $\mathbf{p}(\mathbf{x}, \mathbf{y})$ belongs to two nested models $m$ and $m'$, with $\nu$ and $\nu'$ parameters, for any $\varepsilon > 0$, we have for $n$ large enough

$$E(\text{BEC}(m)) - E(\text{BEC}(m')) < \varepsilon.$$

Actually, 2[BEC($m$)-BEC($m'$)] tends in distribution to $\chi'^2_{\delta_\nu} - \chi^2_{\delta_\nu}$ with $\delta_\nu = \nu' - \nu$

$\rightarrow$ It induces a plateau rule.

# Computational cost of BEC

Roughly speaking, computing BEC is equivalent to Half Sampling.

And Half Sampling is the crudest and fastest version of cross validation.

# An Illustrative Monte Carlo experiment (1)

- 500 samples of $n = 120$ points from two classes with equal prior probabilities have been generated with the following class conditional densities:

$$X|Z_1 = 1 \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

and

$$X|Z_2 = 1 \sim \mathcal{N}\left(\begin{bmatrix} \Delta \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}\right).$$

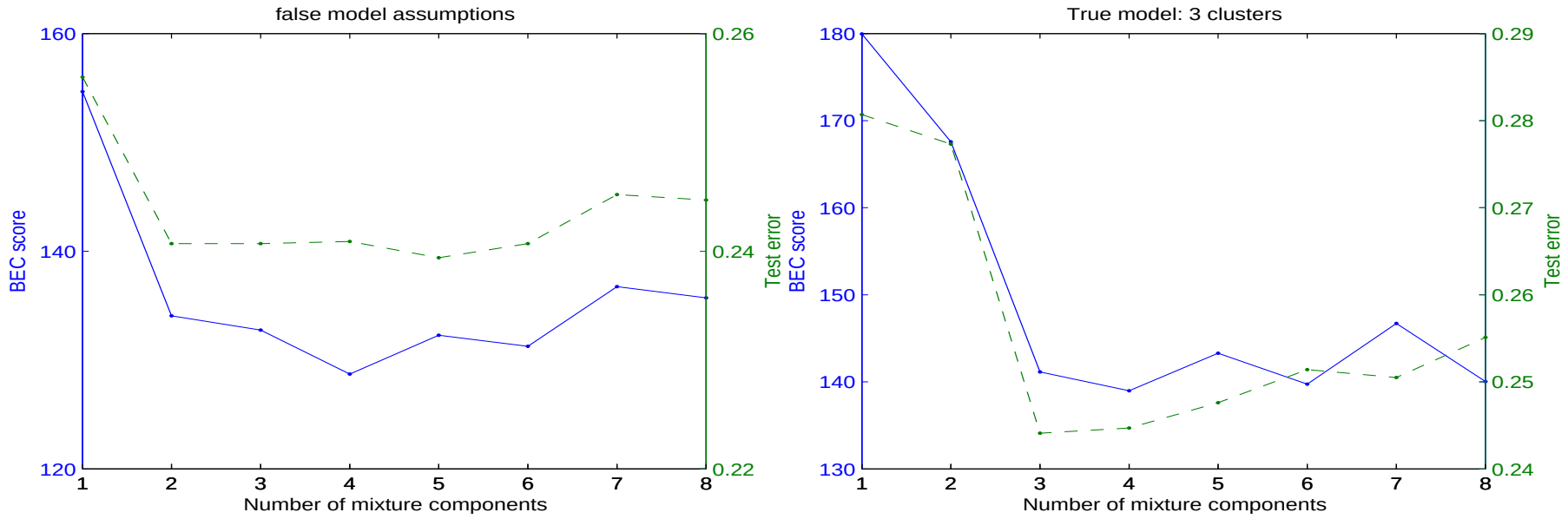
- Two models different from the true one are considered.
  - The first model DIAG is considering a Gaussian class conditional distribution with a diagonal variance matrix,
  - the second model BALL is considering a Gaussian class conditional distributions with a spherical variance matrix.
- The performances of criteria BEC and BIC are compared.

# An Illustrative Monte Carlo experiment (2)

| | model | $\overline{err}$ | BIC | BEC | BIC choice(%) | BEC choic |
|---|---|---|---|---|---|---|
| $\Delta = 1$ | DIAG | **0.250** | 502.331 | **64.108** | 24 | 98 |
| $\Delta = 1$ | BALL | 0.268 | **500.422** | 69.665 | 76 | 2 |
| $\Delta = 3.5$ | DIAG | **0.070** | 502.331 | **22.067** | 24 | 94 |
| $\Delta = 3.5$ | BALL | 0.076 | **500.422** | 26.120 | 76 | 6 |
| $\Delta = 5$ | DIAG | **0.019** | 502.331 | **6.081** | 24 | 84 |
| $\Delta = 5$ | BALL | 0.023 | **500.422** | 8.310 | 76 | 16 |
| $\Delta = 7$ | DIAG | **0.002** | 502.331 | **0.458** | 24 | 80 |
| $\Delta = 7$ | BALL | 0.004 | **500.422** | 1.046 | 76 | 20 |
| $\Delta = 10$ | DIAG | **0.000** | 502.331 | **0.001** | 24 | 60 |
| $\Delta = 10$ | BALL | 0.000 | **500.422** | 0.002 | 76 | 40 |

Column $\overline{err}$ gives the error rate evaluated on a test sample of size 50,000.

# Illustration of the plateau rule




In each plot the full line gives the variations of -BEC whose values appears on the left scale and the dashed line gives the variations of the classification error rate for a test sample of size 50,000. The error rate are given on the right scale.

# Experiments on real data sets

## The EDDA family of models

It is a family of Gaussian classification models using the variance matrix eigenvalue decomposition of class $k$, k=1,. . . ,g

$$\Sigma_k = V_k D_k^t A_k D_k$$

where

- $V_k = |\Sigma_k)|^{1/d}$ defines the component volume ($d$ is the dimension of the observation space),
- $D_k$ the matrix of eigenvectors of $\Sigma$ defines the class orientation
- $A_k$ the diagonal matrix of normalized eigenvalues defines the class shape.

By allowing some of these quantities to vary between classes, 14 different and easily interpreted models are get.

## 14 different models

The models proposed in EDDA are

- The general family: Assuming equal or free volumes orientations and shapes leads to 8 different models.
- The diagonal family: Assuming in addition that the class variance matrices are diagonal leads to 4 models.
- The spherical family: Assuming in addition that the class variance matrices are proportional to the identity matrix leads to 2 models.

In the original version (Bensmail and Celeux, JASA 1996) one of the 14 models is selected by minimizing the cross-validated error rate.

# Conditions of Experiments

- Six benchmark data sets from statlog.
- For avoiding numerical problems, we restricted the data sets to the four first axes of PCA.
- For assessing the performances of the classifiers, a test error rate has been commputed from test data sets selected at random. This operation has been repeated 20 times.
- The criteria in competition are BIC, AIC, BEC, CV3.
  - CV3 is a three-fold cross validation procedure.

# Australian dataset

2 classes, 200 training data, 490 test data (Selected randomly 100 times)

10 variables reduced in 4 dimensions by PCA

| model | $\nu$ | BIC | AIC | BEC | CV3 | test error |
|---|---|---|---|---|---|---|
| $\lambda I$ | 10 | 0 | 0 | 0 | 0 | 0.293 |
| $\lambda_k I$ | 13 | 0 | 0 | 0 | 0 | 0.289 |
| $\lambda B$ | 13 | 0 | 9 | 32 | 28 | 0.23 |
| $\lambda_k B$ | 14 | 0 | 0 | 1 | 1 | 0.264 |
| $\lambda B_k$ | 16 | 0 | 0 | 0 | 0 | 0.287 |
| $\lambda_k B_k$ | 17 | 93 | 0 | 0 | 0 | 0.276 |
| $\lambda D^t A D$ | 19 | 0 | 23 | 38 | 36 | 0.229 |
| $\lambda_k D^t A D$ | 20 | 0 | 0 | 0 | 1 | 0.261 |
| $\lambda D_k^t A D_k$ | 25 | 0 | 68 | 25 | 34 | 0.23 |
| $\lambda_k D_k^t A D_k$ | 26 | 0 | 0 | 3 | 0 | 0.258 |
| $\lambda D_k^t A_k D_k$ | 28 | 0 | 0 | 0 | 0 | 0.291 |
| $\lambda_k D_k^t A_k D_k$ | 29 | 7 | 0 | 1 | 0 | 0.274 |

# A collection of benchmark data sets

| Dataset | $K$ | $N$ | $d$ | BIC | AIC | BEC | CV3 | oracle |
|---|---|---|---|---|---|---|---|---|
| Abalone | 3 | 4177 | 7 | 47.3 | 47.4 | 46.1 | 45.9 | 45.4 |
| Bupa | 2 | 345 | 6 | 37.5 | 38.3 | 33.5 | 34.6 | 31.6 |
| Haberman | 2 | 306 | 3 | 25.0 | 25.0 | 25.1 | 24.9 | 23.7 |
| Pageblocks | 5 | 5473 | 10 | 4.4 | 4.4 | 2.8 | 2.8 | 2.5 |
| Teaching | 3 | 151 | 5 | 63.8 | 63.3 | 63.8 | 61.1 | 56.9 |
| Australian | 2 | 690 | 14 | 26.3 | 26.4 | 22.6 | 22.8 | 21.9 |
| Diabetes | 2 | 768 | 8 | 26.0 | 25.6 | 23.9 | 24.2 | 23.0 |
| German | 2 | 1000 | 20 | 25.3 | 25.4 | 25.1 | 24.9 | 24.0 |
| Heart | 2 | 270 | 10 | 17.5 | 18.3 | 17.6 | 17.3 | 15.6 |

# Model selection in computer vision (1)

- Object categorization problem: Finding images containing a motorbike.
- Training data set of 826 images, Test data set of 900 images.
- Each image is categorized into a 1000-dimensional vector.
- Generative model: each class is modelled with a mixture of diagonal Gaussian distributions.
- The problem is to find a suitable number of components to describe each class.

# Model selection in computer vision (2)

Mixtures were learned with 1 to 5 clusters for the motorbike images and with 1 to 7 clusters for the background images.

−-BIC ($\times 10^5$)

| | $R_1$ | | | | |
|---|---|---|---|---|---|
| $R_2$ | 1 | 2 | 3 | 4 | 5 |
| 1 | -9.111 | -9.227 | -9.255 | -9.263 | -9.264 |
| 2 | -9.260 | -9.257 | -9.126 | -9.243 | -9.271 |
| 3 | -9.279 | **-9.281** | -9.275 | -9.273 | -9.126 |
| 4 | -9.242 | -9.270 | -9.278 | -9.279 | -9.275 |
| 5 | -9.272 | -9.122 | -9.239 | -9.267 | -9.275 |
| 6 | -9.276 | -9.271 | -9.269 | -9.115 | -9.231 |
| 7 | -9.259 | -9.267 | -9.268 | -9.264 | -9.261 |

−BEC ($\times 10^3$)

| | $R_1$ | | | | |
|---|---|---|---|---|---|
| $R_2$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 3.06 | 1.18 | 0.91 | 0.75 | 0.63 |
| 2 | 1.35 | 1.27 | 6.24 | 1.09 | 0.75 |
| 3 | 0.51 | 0.46 | 0.46 | 0.39 | 6.93 |
| 4 | 1.99 | 0.80 | 0.52 | 0.48 | 0.37 |
| 5 | 0.32 | 7.95 | 2.35 | 0.80 | 0.53 |
| 6 | 0.45 | 0.34 | **0.29** | 8.57 | 2.44 |
| 7 | 0.91 | 0.58 | 0.51 | 0.38 | 0.32 |

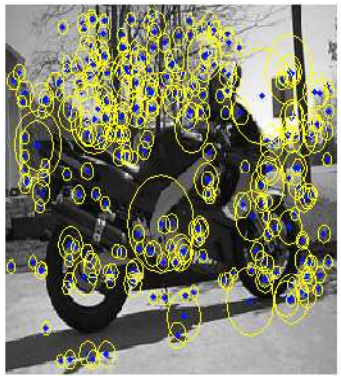
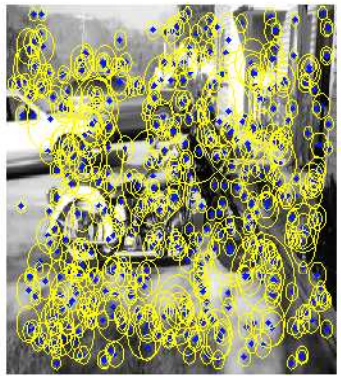
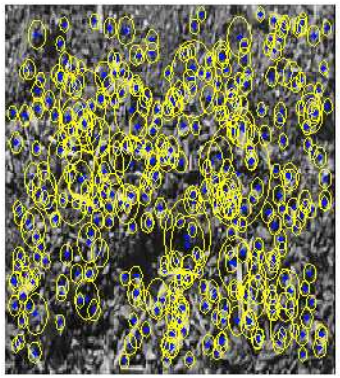
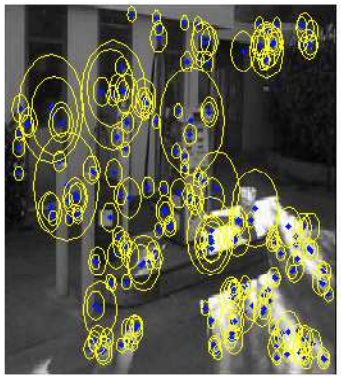
CV10 error rate ($\times 100$)

| | $R_1$ | | | | |
|---|---|---|---|---|---|
| $R_2$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 7.19 | 9.04 | 6.61 | 4.98 | 6.95 |
| 2 | 6.95 | 7.42 | 9.04 | 7.18 | 6.61 |
| 3 | 6.26 | 5.91 | 4.98 | 4.75 | 9.62 |
| 4 | 7.42 | 6.61 | 5.79 | 5.45 | 4.87 |
| 5 | 4.75 | 9.85 | 6.84 | 5.79 | 5.68 |
| 6 | 5.33 | 4.29 | **4.09** | 11.47 | 6.61 |
| 7 | 5.91 | 6.14 | 5.79 | 4.72 | 4.72 |

Test error rate ($\times 100$)

| | $R_1$ | | | | |
|---|---|---|---|---|---|
| $R_2$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 6.26 | 8.34 | 5.56 | 6.49 | 4.85 |
| 2 | 5.56 | 5.10 | 7.76 | 6.72 | 5.91 |
| 3 | 5.91 | 5.33 | 5.56 | 4.87 | 8.69 |
| 4 | 6.95 | 5.68 | 5.56 | 5.21 | 5.21 |
| 5 | 4.98 | 9.50 | 6.84 | 5.45 | 5.91 |
| 6 | 4.87 | 4.52 | **3.84** | 10.08 | 6.84 |
| 7 | 5.33 | 6.03 | 4.75 | 4.85 | 4.59 |

# Examples of misclassified images

# Discussion

- Taking into of the modelling purpose is an interesting view to select a reliable and useful model.
- This point of view lead to criteria with data driven penalties: It is a desirable feature of model selection criteria.
- Taking account of the model purpose is interesting when assessing a model, but it does not seem to be useful when estimating a model.
  - In supervised classification, maximizing the conditional likelihood from a generative model, lead to difficult optimisation problems with unstable solutions.
- In a small sample setting, a full Bayesian approach embedding the derivation of a classification entropy in the predictive approach to model selection is desirable.

# References


- C. Biernacki, G. Celeux and G. Govaert (2000) Assessing a mixture model for clustering with the integrated completed likelihood *IEEE, Trans. on PAMI*, **22**, 719-725.
- G. Bouchard and G. Celeux (2004) Model selection in supervised classification. *IEEE, Trans. on PAMI*, **28**, 544-554.
- E. Lebarbier and T. Mary-Huard (2006) Une introduction au critère BIC. Fondements théoriques et interprétation. *Journal de la SFdS*, prochain numéro.
- Y. Yang (2005) Can the strenghts of AIC and BIC be shared? A confict between model identification and regression estimation. *Biometrika*, **92**, 927-950.