Journal of the Korean Statistical Society (2007), **36***: 3, pp 435–445*

PERFORMANCE EVALUATION OF INFORMATION CRITERIA FOR THE NAIVE-BAYES MODEL IN THE CASE OF LATENT CLASS ANALYSIS: A MONTE CARLO STUDY

José G. Dias[1]

Abstract

This paper addresses for the first time the use of complete data information criteria in unsupervised learning of the Naive-Bayes model. A Monte Carlo study sets a large experimental design to assess these criteria, unusual in the Bayesian network literature. The simulation results show that complete data information criteria underperforms the Bayesian information criterion (BIC) for these Bayesian networks.

AMS 2000 subject classifications. Primary 62F07; Secondary 94A17.
Keywords. Bayesian information criterion (BIC), complete data information criteria, latent class model, Monte Carlo simulation, Naive-Bayes model.

1. Introduction

In recent years the Naive-Bayes model has become a popular alternative to more complex classifiers (Duda *et al.*, 2001). This model can be casted as a Bayesian network (Pearl, 1988), where the structure of the network has to be learned (Friedman *et al.*, 1997). Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ represent a sample/training data set of size n; J represents the number of manifest or observed variables; and datum y_{ij} indicates the observed value for variable j in observation i, with $i = 1, \ldots, n$, $j = 1, \ldots, J$. The finite mixture model (a type of Bayesian network for unsupervised learning) with S latent classes for $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ is defined by the density $f(\mathbf{y}_i; \varphi) = \sum_{s=1}^{S} \pi_s f_s(\mathbf{y}_i; \theta_s)$, where the latent class proportions π_s are positive and sum to one; θ_s denotes the parameters of the conditional

Received May 2006; accepted April 2007.

[1] Department of Quantitative Methods and GIESTA-UNIDE, Higher Institute of Social Sciences and Business Studies-ISCTE, Av. das Forças Armadas, 1649-026 Lisboa, Portugal (e-mail: jose.dias@iscte.pt)

distribution of $\mathbf{y}_i$ for the latent class s, defined by $f_s(\mathbf{y}_i;\theta_s)$; $\pi = (\pi_1,\ldots,\pi_{S-1})$, $\theta = (\theta_1,\ldots,\theta_S)$ and $\varphi = (\pi,\theta)$. For nominal data, Y_j has L_j categories, $y_{ij} \in \{1,\ldots,L_j\}$. From the local independence assumption underlying the Naive-Bayes model – the J manifest variables are independent given the latent variable – $f_s(\mathbf{y}_i;\theta_s) = \prod_{j=1}^{J}\prod_{l=1}^{L_j}\theta_{sjl}^{I(y_{ij}=l)}$, where θ_{sjl} is the probability that observation i belonging to latent class s falls in category l of variable j. Category l is associated with the binary variable defined by the indicator function $I(y_{ij}=l)=1$ and 0 otherwise. Note that $\sum_{l=1}^{L_j}\theta_{sjl}=1$. Finally, the Naive-Bayes model has density

$$f(\mathbf{y}_i;\varphi) = \sum_{s=1}^{S}\pi_s\prod_{j=1}^{J}\prod_{l=1}^{L_j}\theta_{sjl}^{I(y_{ij}=l)}, \tag{1.1}$$

where the number of free parameters to be estimated in vectors π and θ are $d_\pi = S-1$ and $d_\theta = S\sum_{j=1}^{J}(L_j-1)$, respectively. The total number of free parameters is $d_\varphi = d_\pi + d_\theta$. The Naive-Bayes model may present problems of identifiability (Goodman, 1974). However, Naive-Bayes models discussed here are identified. The likelihood and log-likelihood functions are $L(\varphi;\mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_i;\varphi)$ and $\ell(\varphi;\mathbf{y}) = \log L(\varphi;\mathbf{y})$, respectively. It is straightforward to obtain the maximum likelihood estimates (MLE) of mixture's parameters by the EM algorithm (Dias and Wedel, 2004).

Despite the increasing widespread application of this Naive-Bayes model, estimating the number of latent classes to retain remains an important topic of research. Information criteria have become popular as a useful approach to model selection. The basic principle under these criteria is parsimony that results from the trade-off between model fit and model complexity. A number of model selection criteria has been suggested, the most prominent and widely used being the Akaike information criteria (AIC) of Akaike (Akaike, 1974) and the Bayesian information criteria (BIC) of Schwarz (Schwarz, 1978). Recently, new criteria have been introduced such as the classification likelihood criterion (Biernacki and Govaert, 1997) and the integrated classification likelihood criterion (Biernacki *et al.*, 2000).

Despite extensive study of the performance of information criteria in the statistical literature, little is known about the performance of these criteria for the Naive-Bayes model. Therefore, a Monte Carlo experiment is designed to assess the ability of the different information criteria to learn the true model and to measure the effect of the design factors.

This paper is organized as follows. Section 2 reviews the literature on model selection criteria from a Bayesian viewpoint. Section 3 discusses model selection based on data augmentation. Section 4 describes the design of the Monte Carlo study. Section 5 presents and discusses the results. The paper concludes with a summary of main findings, implications and suggestions for further research.

2. Bayesian Model Selection

Bayesian estimation focuses on the posterior distribution of parameters $p(\varphi|\mathbf{y})$, which is proportional to $L(\varphi;\mathbf{y})p(\varphi)$. The prior distribution $p(\varphi)$ represents how likely different values of φ are, before seeing the data. It is assumed here that parameters are *a priori* independent, $p(\varphi) = p(\pi)\prod_{s=1}^{S}\prod_{j=1}^{J} p(\theta_{sj})$. The Dirichlet distribution is a natural prior for these parameters. For $\omega = (\omega_1, \omega_2, \ldots, \omega_k)$, it is denoted by $\mathcal{D}(\xi_1, \ldots, \xi_k)$ with parameters $(\xi_1, \ldots, \xi_k)$ and density function $p(\omega_1, \omega_2, \ldots, \omega_k) = \left(\Gamma(\xi_0)/\prod_{j=1}^{k}\Gamma(\xi_j)\right)\prod_{j=1}^{k}\omega_j^{\xi_j - 1}$, where $\omega_j \geq 0$ for $j = 1, \ldots, k$, $\sum_{j=1}^{k}\omega_j = 1$, $\Gamma(.)$ is the gamma function and $\xi_0 = \sum_{j=1}^{k}\xi_j$. The expected value and variance of ω_j are $\mathrm{E}(\omega_j) = \xi_j/\xi_0$ and $\mathrm{Var}(\omega_j) = \xi_j(\xi_0 - \xi_j)/\left[\xi_0^2(\xi_0 + 1)\right]$, respectively. In the analyses below, we consider two special types of priors:

1. The uniform prior (U) corresponding to a Dirichlet distributions with $\pi \sim \mathcal{D}(1, \ldots, 1)$ and $\theta_{sj} \sim \mathcal{D}(1, \ldots, 1)$ is given by

$$\log p(\varphi) = \log\left[(S-1)!\right] + S\sum_{j=1}^{J}\log\left[(L_j - 1)!\right]. \tag{2.1}$$

2. The Jeffreys' prior (J) corresponding to a Dirichlet distributions with $\pi \sim \mathcal{D}(1/2, \ldots, 1/2)$ and $\theta_{sj} \sim \mathcal{D}(1/2, \ldots, 1/2)$ is

$$\begin{aligned}\log p(\varphi) = {} & S\sum_{j=1}^{J}\log\Gamma\left(\frac{L_j}{2}\right) - S\log\Gamma\left(\frac{1}{2}\right)\sum_{j=1}^{J}L_j + \frac{1}{2}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{l=1}^{L_j}\log\theta_{sjl} \\ & + \log\Gamma\left(\frac{S}{2}\right) - S\log\Gamma\left(\frac{1}{2}\right) + \frac{1}{2}\sum_{s=1}^{S}\log\pi_s. \end{aligned} \tag{2.2}$$

The Bayesian information criterion (BIC), proposed by Schwarz (Schwarz, 1978), utilizes the integrated likelihood $p(\mathbf{y}) = \int L(\varphi;\mathbf{y})p(\varphi)d\varphi$, which is the weighted average of the likelihood values. Using the Laplace approximation about the posterior mode $\tilde{\varphi}$ where $L(\varphi;\mathbf{y})p(\varphi)$ is maximized, it results (Tierney and Kadane, 1986)

$$\log p(\mathbf{y}) \approx \ell(\tilde{\varphi};\mathbf{y}) + \log p(\tilde{\varphi}) - \frac{1}{2}\log\left|\mathbf{H}(\tilde{\varphi};\mathbf{y})\right| + \frac{d_\varphi}{2}\log(2\pi), \tag{2.3}$$

where $\mathbf{H}(\tilde{\varphi};\mathbf{y})$ is the negative of the Hessian matrix of the log-posterior function, $\log L(\varphi;\mathbf{y})p(\varphi)$, evaluated at the modal value $\varphi = \tilde{\varphi}$. BIC assumes a proper prior, which assigns positive probability to lower dimensional subspaces of the parameter vector. For a very diffuse (almost non-informative and consequently ignorable) prior distribution, $\mathbf{H}(\tilde{\varphi};\mathbf{y})$ can be replaced by the observed information matrix $\mathbf{I}(\tilde{\varphi};\mathbf{y})$. Replacing the posterior mode by the MLE $\widehat{\varphi}$, the approximation becomes

$$\log p(\mathbf{y}) \approx \ell(\widehat{\varphi};\mathbf{y}) + \log p(\widehat{\varphi}) - \frac{1}{2}\log|\mathbf{I}(\widehat{\varphi};\mathbf{y})| + \frac{d_\varphi}{2}\log(2\pi). \tag{2.4}$$

From the asymptotic behavior of (2.4), the Bayesian information criterion (BIC) chooses S that minimizes

$$\mathrm{BIC} = -2\ell(\widehat{\varphi};\mathbf{y}) + d_\varphi \log n. \tag{2.5}$$

Therefore, BIC selects the model with the greatest asymptotic posterior probability and does not depend on the prior. From the notion of stochastic complexity, a criterion identical in form to BIC is derived, usually known as minimum descriptive length (MDL), but with a broader justification (Rissanen, 1987).

3. COMPLETE DATA INFORMATION CRITERIA

Complete data information criteria are based on data augmentation, where the observed data $(\mathbf{y})$ is expanded to a new space $(\mathbf{y},\mathbf{z})$, which includes the missing data $(\mathbf{z})$. The missing datum (z_{is}) indicates whether latent class s has generated observation i. The expected value of z_{is} is given by

$$\alpha_{is} = \frac{\pi_s f_s(\mathbf{y}_i;\theta_s)}{\sum\limits_{v=1}^{S} \pi_v f_v(\mathbf{y}_i;\theta_v)} \tag{3.1}$$

and corresponds to the posterior probability that $\mathbf{y}_i$ was generated by latent class s. Note that α is function of φ and $\mathbf{y}$. The entropy of the matrix $\alpha = (\alpha_{is})$, $i = 1,\ldots,n$ and $s = 1,\ldots,S$ is defined by $\mathrm{EN}(\alpha) = -\sum_{i=1}^{n}\sum_{s=1}^{S}\alpha_{is}\log\alpha_{is}$. For $\mathrm{EN}(\alpha) \simeq 0$, latent classes are well separated.

Celeux and Soromenho (1996) introduced an entropic measure for the selection of S. As $\mathrm{EN}(\alpha)$ has no upperbound, they introduced the normalized entropy criterion (NEC). NEC chooses S that minimizes

$$\mathrm{NEC} = \frac{\mathrm{EN}(\widehat{\alpha})}{\ell(\widehat{\varphi};\mathbf{y}) - \ell_1(\widehat{\varphi};\mathbf{y})}, \tag{3.2}$$

where $\ell_1(\widehat{\varphi}; \mathbf{y})$ is the log-likelihood value for the one-latent-class model and $\widehat{\alpha}$ comes from (3.1) at the MLE. To overcome the impossibility of deciding between $S = 1$ and $S > 1$, Biernacki *et al.* (1999) proposed the following rule: if there is no S such that NEC < 1, then $S = 1$ has to be preferred.

Hathaway (1986) observed that the complete log-likelihood can be decomposed into

$$\ell_C(\varphi; \mathbf{y}, \mathbf{z}) = \ell(\varphi; \mathbf{y}) + \sum_{i=1}^{n} \sum_{s=1}^{S} z_{is} \log \alpha_{is}. \qquad (3.3)$$

The classification likelihood criterion (CL) proposed by Biernacki and Govaert (1997) chooses S that minimizes

$$\mathrm{CL} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{ENC}(\tilde{\mathbf{z}}, \widehat{\alpha}), \qquad (3.4)$$

where the level of penalization depends on the entropy of the classification given by $\mathrm{ENC}(\tilde{\mathbf{z}}, \widehat{\alpha}) = -\sum_{i=1}^{n} \sum_{s=1}^{S} \tilde{z}_{is} \log \widehat{\alpha}_{is}$, with $\tilde{z}_{is} = 1$, if $\arg\max_s \widehat{\alpha}_{is} = s$ and 0 otherwise. Biernacki *et al.* (2000) named this function as MAP (Maximum A Posteriori): $\tilde{\mathbf{z}} = \mathrm{MAP}(\widehat{\alpha})$. McLachlan and Peel (2000) suggested instead replacing $\mathbf{z}$ by its estimated expected value $\widehat{\alpha}$ given the observed data $(\mathbf{y})$. The complete likelihood criterion (CLC) chooses S that minimizes

$$\mathrm{CLC} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{EN}(\widehat{\alpha}). \qquad (3.5)$$

The integrated or marginal likelihood of the complete data, given by $p(\mathbf{y}, \mathbf{z}) = \int L_C(\varphi; \mathbf{y}, \mathbf{z}) p(\varphi) d\varphi$, can be used as a model selection criterion. Given that $\log p(\mathbf{y}, \mathbf{z}) = \log p(\mathbf{y}|\mathbf{z}) + \log p(\mathbf{z})$, $\log p(\mathbf{y}|\mathbf{z})$ can be approximated by $\log p(\mathbf{y}|\mathbf{z}, \widehat{\theta}) - (d_\theta/2) \log n$, where $\widehat{\theta}$ is the MLE of θ, with $\log p(\mathbf{y}|\mathbf{z}, \theta) = \ell_C(\varphi; \mathbf{y}, \mathbf{z}) - \sum_{i=1}^{n} \sum_{s=1}^{S} z_{is} \log \pi_s$. For multinomial data $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_S)$ and adopting the Dirichlet distribution for π with parameters $\epsilon = (\epsilon, \ldots, \epsilon)$, one has $\log p(\mathbf{z}) = \sum_{s=1}^{S} \log \Gamma(\epsilon + n\pi_s) - \log \Gamma(S\epsilon + n) + \log \Gamma(S\epsilon) - S \log \Gamma(\epsilon)$. The integrated classification likelihood criterion (ICL), proposed by Biernacki *et al.* (2000), chooses S that minimizes

$$\mathrm{ICL} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{ENC}(\tilde{\mathbf{z}}, \widehat{\alpha}) + 2n \sum_{s=1}^{S} \tilde{\pi}_s \log \tilde{\pi}_s + d_\theta \log n - 2 \log p(\tilde{\mathbf{z}}), \quad (3.6)$$

with $\tilde{\pi}_s = (1/n) \sum_{i=1}^{n} \tilde{z}_{is}$. The integrated complete likelihood criterion (ICOMPL) chooses S that minimizes

$$\mathrm{ICOMPL} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{EN}(\widehat{\alpha}) + 2n \sum_{s=1}^{S} \widehat{\pi}_s \log \widehat{\pi}_s + d_\theta \log n - 2 \log p(\widehat{\alpha}), \quad (3.7)$$

where $\widehat{\pi}_s = (1/n)\sum_{i=1}^{n} \widehat{\alpha}_{is}$. The robustness of these criteria was analyzed using two different priors introduced before. We set uniform ($\epsilon = 1$) and Jeffreys' priors ($\epsilon = 1/2$) represented by U (2.1) and J (2.2), respectively.

Alternatively, using the BIC-like approximation, we have $\log p(\mathbf{y}, \mathbf{z}) \approx \log p(\mathbf{y}, \mathbf{z}; \tilde{\varphi}) - (d_\varphi/2)\log n$, where $\tilde{\varphi}$ is the maximum of $p(\mathbf{y}, \mathbf{z}; \varphi)$. For large n, $\tilde{\varphi}$ can be replaced by the MLE $\widehat{\varphi}$. The integrated classification likelihood criterion with BIC approximation (ICL-BIC) proposed by Biernacki *et al.* (2000) chooses S that minimizes $-2\log p(\mathbf{y}, \tilde{\mathbf{z}}; \widehat{\varphi}) + d_\varphi \log n$, which is the same as

$$\mathrm{ICL} - \mathrm{BIC} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{ENC}(\tilde{\mathbf{z}}, \widehat{\alpha}) + d_\varphi \log n. \tag{3.8}$$

Similarly, using the expected value of $\mathbf{z}$ instead of $\tilde{\mathbf{z}}$, one obtains the integrated complete likelihood criterion with BIC approximation (ICOMPL-BIC) that chooses S that minimizes

$$\mathrm{ICOMPL} - \mathrm{BIC} = -2\ell(\widehat{\varphi}; \mathbf{y}) + 2\mathrm{EN}(\widehat{\alpha}) + d_\varphi \log n. \tag{3.9}$$

4. Experimental Design

A Monte Carlo (MC) study was conducted for the assessment of the performance of these criteria and robustness across experimental conditions. Its experimental design controls the number of variables and categories, the sample size, the balance of latent class sizes and the level of separation of the latent classes. The number of variables (J) was set at levels 5 and 8; and the number of categories (L_j) at levels 2 and 3. From preliminary analyses with $L_j = 2$, $J = 5$ and $S = 3$, we concluded that data sets with a non-singular estimated information matrix for the three-latent-class model with sample sizes smaller than 300 are difficult to generate. Therefore, the number of latent classes is set to two ($S = 2$) and the factor sample size (n) assumes the levels: 300, 600, 1200 and 2400. The latent class sizes were generated using the expression $\pi_s = a^{s-1}(\sum_{v=1}^{S} a^{v-1})^{-1}$, with $s = 1, \ldots, S$ and $a \geq 1$. With $a = 1$, equal proportions are yield; for larger values of a, latent class sizes become more unbalanced. For example, for $S = 2$ and $a = 3$, latent class sizes are $\pi = (1/4,\ 3/4)$. In our MC study, we set three levels for a: 1, 2 and 3.

Controlling the level of separation of the latent classes is more challenging. Because other factors influencing the level of separation are already controlled, latent class separation is based exclusively on the separation of the parameters θ. In this paper, we apply a sampling procedure proposed by Dias (2004). The vector θ is generated as:

1. Draw θ_{1j} from the Dirichlet distribution with parameters $(\phi_1, \dots, \phi_{L_j})$, $j = 1, \dots, J$,
2. Draw θ_{sj} from the Dirichlet distribution with parameters $(\delta\theta_{1j1}, \dots, \delta\theta_{1jL_j})$, $j = 1, \dots, J$, $s = 2, \dots, S$.

This procedure assumes that parameters θ of the Naive-Bayes model are sampled from a superpopulation defined by the hyperparameters δ and $(\phi_1, \dots, \phi_{L_j})$, $j = 1, \dots, J$, and defines a hierarchical (Bayesian) structure. We set $(\phi_1, \dots, \phi_{L_j}) = (1, \dots, 1)$, which corresponds to the uniform distribution. For $s = 2, \dots, S$, we have $E(\theta_{sjl}) = \theta_{1jl}$ and $\text{Var}(\theta_{sjl}) = \theta_{1jl}\,(1 - \theta_{1jl})\,/\,(\delta + 1)$. With this procedure, on average, all latent classes are centered at the same parameter value generated from a uniform distribution (first latent class). The constant $\delta > 0$ controls the level of separation of the latent classes. As δ increases, the latent class separation decreases as a consequence of the decreasing of the variance. As $\delta \to \infty$, all latent classes tend to share the same parameters. Based on results reported in Dias (2004), three levels of δ give a good coverage of the level of separation of the latent classes for this model: 0.1 (well-separated latent classes), 1 (moderately-separated latent classes) and 10 (weakly-separated latent classes). These values of δ were set in this study.

This MC study sets a $2^2 \times 3^2 \times 4$ factorial design with 144 cells. The main performance measure used is the frequency with which each criterion picks the correct model. For each data set, each criterion is classified as *under-fitting*, *fitting* or *over-fitting*, based on the relation between S and the estimated S by those criteria.

Special care needs to be taken before arriving at conclusions based on MC results. In this study, we performed 100 replications within each cell to obtain the frequency distribution of selecting the true model, resulting in a total of 14400 data sets. To avoid local optima, for each number of latent classes (2 and 3) the EM algorithm was repeated 5 times with random starting centers, and the best solution (maximum likelihood value out of those 5 runs) and model selection results were kept. The EM algorithm ran for 1500 iterations, which was enough to ensure the convergence for all cells of the design. The programs were written in MATLAB.

5. RESULTS

Although good results have been reported in the model learning with complete data information criteria, especially for Gaussian models, the key feature of our

TABLE 5.1 *Results of the Monte Carlo study – percentages of under-fit, fit and over-fit*

Factors		*Criteria*									
		BIC	*NEC*	*CL*	*CLC*	*ICL*			*ICOMPL*		
						BIC	*U*	*J*	*BIC*	*U*	*J*
Sample size (n)											
	Under-fit	48.67	65.78	46.89	65.78	71.06	70.94	71.31	82.36	82.36	82.42
300	*Fit*	51.33	27.83	33.75	25.75	28.89	28.97	28.64	17.50	17.39	17.44
	Over-fit	0.00	6.39	19.36	8.47	0.06	0.08	0.06	0.14	0.25	0.14
	Under-fit	40.39	71.17	56.25	71.17	66.75	66.75	66.97	77.86	77.86	77.86
600	*Fit*	59.61	25.86	36.36	25.33	33.14	33.14	32.92	21.94	21.94	21.94
	Over-fit	0.00	2.97	7.39	3.50	0.11	0.11	0.11	0.19	0.19	0.19
	Under-fit	35.83	76.44	61.14	76.44	65.86	65.86	65.89	79.61	79.61	79.67
1200	*Fit*	64.17	23.11	36.28	22.81	34.08	34.08	34.06	20.36	20.36	20.31
	Over-fit	0.00	0.44	2.58	0.75	0.06	0.06	0.06	0.03	0.03	0.03
	Under-fit	31.53	76.72	61.94	76.72	64.44	64.39	64.50	78.39	78.39	78.39
2400	*Fit*	68.47	23.03	36.97	22.94	35.44	35.50	35.39	21.58	21.58	21.58
	Over-fit	0.00	0.25	1.08	0.33	0.11	0.11	0.11	0.03	0.03	0.03
Number of variables (J)											
	Under-fit	42.17	82.47	66.75	82.47	76.82	76.82	76.94	87.76	87.76	87.76
5	*Fit*	57.83	15.94	28.63	15.78	23.07	23.06	22.94	12.07	12.01	12.07
	Over-fit	0.00	1.58	4.63	1.75	0.11	0.13	0.11	0.17	0.22	0.17
	Under-fit	36.04	62.58	46.36	62.58	57.24	57.15	57.39	71.35	71.35	71.40
8	*Fit*	63.96	33.97	43.06	32.64	42.71	42.79	42.56	28.63	28.63	28.57
	Over-fit	0.00	3.44	10.58	4.78	0.06	0.06	0.06	0.03	0.03	0.03
Number of categories (L)											
	Under-fit	39.58	72.44	58.75	72.44	67.14	67.14	67.17	77.89	77.89	77.93
2	*Fit*	60.42	24.61	34.19	24.15	32.69	32.68	32.67	21.92	21.86	21.88
	Over-fit	0.00	2.94	7.06	3.40	0.17	0.18	0.17	0.19	0.25	0.19
	Under-fit	38.63	72.61	54.36	72.61	66.92	66.83	67.17	81.22	81.22	81.24
3	*Fit*	61.38	25.31	37.49	24.26	33.08	33.17	32.83	18.78	18.78	18.76
	Over-fit	0.00	2.08	8.15	3.12	0.00	0.00	0.00	0.00	0.00	0.00
Proportions (a)											
	Under-fit	36.10	73.29	56.94	73.29	66.06	65.92	66.29	79.52	79.52	79.54
1	*Fit*	63.90	24.12	35.21	23.33	33.81	33.96	33.58	20.29	20.21	20.27
	Over-fit	0.00	2.58	7.85	3.38	0.13	0.13	0.13	0.19	0.27	0.19
	Under-fit	38.96	71.29	55.46	71.29	65.90	65.90	66.02	78.56	78.56	78.60
2	*Fit*	61.04	26.15	36.71	25.33	34.08	34.08	33.96	21.38	21.38	21.33
	Over-fit	0.00	2.56	7.83	3.38	0.02	0.02	0.02	0.06	0.06	0.06
	Under-fit	42.25	73.00	57.27	73.00	69.13	69.15	69.19	80.58	80.58	80.60
3	Fit	57.75	24.60	35.60	23.96	30.77	30.73	30.71	19.38	19.38	19.35
	Over-fit	0.00	2.40	7.12	3.04	0.10	0.13	0.10	0.04	0.04	0.04

(continued)

Factors		*Criteria*									
		BIC	*NEC*	*CL*	*CLC*	*ICL*			*ICOMPL*		
						BIC	*U*	*J*	*BIC*	*U*	*J*
Level of separation (δ)											
0.1	*Under-fit*	3.71	39.19	18.81	39.19	29.02	29.04	29.23	49.69	49.69	49.71
	Fit	96.29	56.33	69.31	54.79	70.83	70.81	70.63	50.04	50.04	50.02
	Over-fit	0.00	4.48	11.88	6.02	0.15	0.15	0.15	0.27	0.27	0.27
1	*Under-fit*	16.56	80.83	55.58	80.83	72.06	71.92	72.27	88.98	88.98	89.04
	Fit	83.44	16.54	34.73	15.88	27.83	27.96	27.63	11.00	10.92	10.94
	Over-fit	0.00	2.63	9.69	3.29	0.10	0.13	0.10	0.02	0.10	0.02
10	*Under-fit*	97.04	97.56	95.27	97.56	100.00	100.00	100.00	100.00	100.00	100.00
	Fit	2.96	2.00	3.48	1.96	0.00	0.00	0.00	0.00	0.00	0.00
	Over-fit	0.00	0.44	1.25	0.48	0.00	0.00	0.00	0.00	0.00	0.00
Overall											
	Under-fit	39.10	72.53	56.56	72.53	67.03	66.99	67.17	79.56	79.56	79.58
	Fit	60.90	24.96	35.84	24.21	32.89	32.92	32.75	20.35	20.32	20.32
	Over-fit	0.00	2.51	7.60	3.26	0.08	0.09	0.08	0.10	0.12	0.10

results is that those results may not apply to Naive-Bayes models (Table 5.1). They consistently underperform with success rate no larger than 35.8% (CL) and they tend to under-fit the correct number of latent classes. By comparing different approximations to ICL and ICOMPL (based on BIC approximation, uniform prior and Jeffreys' prior), we conclude that ICL and ICOMPL are very robust to the prior setting. It is also observed that classification criteria (MAP($\widehat{\alpha}$)) perform better than complete criteria based on the expected value of $\mathbf{z}$ ($\widehat{\alpha}$).

A second objective of the study was the comparison of these criteria across the factors in the design. The classification likelihood criterion (CL) dominates other complete data criteria across experimental conditions with exception of well-separated latent classes, where ICL performs better. Therefore, we will focus on the performance of BIC and CL across the design factors. Increasing the sample size almost always improves the performance of BIC and CL. However, these criteria showed a tendency to underestimate the true number of latent classes when the sample size decreases. Increasing the number of variables (J) and categories (L_j) mostly reduces the under-fitting, and improves the performance of BIC and CL. For BIC, more balanced latent classes sizes are associated with an improved performance. The level of separation of the latent classes plays an important role in the performance of these criteria. BIC finds the correct model in 96.3% of the cases for the well-separated situation, but just in 3.0% for ill-separated latent classes. The same effect can be found for the CL criterion.

We observe that BIC tends to be extremely conservative for ill-separated latent classes.

6. Conclusion

This paper discussed model selection for Naive-Bayes models under unsupervised learning. Because most of the information criteria are based on asymptotics or other type of approximations (*e.g.*, Laplace approximation), most of these model selection heuristics have to be dealt with care. The extensive Monte Carlo study allowed their assessment for realistic sample sizes and under different design conditions such as: number of variables, number of categories, relative latent class sizes and the level of separation of the latent classes.

It was shown that Bayesian information criterion (BIC/MDL) outperforms the complete data information criteria for all the conditions under study, and the CL criterion outperforms the remaining complete data information criteria. Indeed, BIC picks the correct model in 60.9% of the simulated data sets, against just 35.84% for the CL. Therefore, we can conclude that the complete information criteria tend to overpenalize. We conclude that BIC is the best model selection criterion for the Naive-Bayes model with a discrete latent variable. However, it was also shown that the level of separation of components plays a pivotal role in model selection (number of latent classes) using the BIC.

Our findings are valid only for the Naive-Bayes model with a single discrete latent variable. Therefore, the conclusions of this study point out the need for detailed replications of these results for more complex Bayesian networks. On the other hand, this research shows that the approximations and assumptions under which these criteria are derived may play an important role that may invalidate complete information criteria. Moreover, as it has been shown these results are not sensitive to the prior setting (prior distribution and hyper-parameters), as it can be seen comparing BIC-, J- and U-type specifications. Future research should pay more attention to the improvement of the approximations underlying these criteria for the Naive-Bayes model, in particular from the likelihood side, where the Laplace approximation may not perform well under multimodality.

References

Akaike, H. (1974). "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, **19**, 716–723.

Biernacki, C. and Govaert, G. (1997). "Using the classification likelihood to choose the number of clusters", *Computing Science and Statistics*, **29**, 451–457.

BIERNACKI, C., CELEUX, G. AND GOVAERT, G. (1999). "An improvement of the NEC criterion for assessing the number of clusters in a mixture model", *Pattern Recognition Letters*, **20**, 267–272.

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**, 719–725.

CELEUX, G. AND SOROMENHO, G. (1996). "An entropy criterion for assessing the number of clusters in a mixture model", *Journal of Classification*, **13**, 195–212.

DIAS, J. G. (2004). "Controlling the level of separation of components in Monte Carlo studies of latent class models", In *Classification, Clustering, and Data Mining Applications* (Banks, D., House, L., McMorris, F. R., Arabie, P. and Gaul, W., eds.), 77–84, Springer, Berlin.

DIAS, J. G. AND WEDEL, M. (2004). "An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods", *Statistics and Computing*, **14**, 323–332.

DUDA, R. O., HART, P. E. AND STORK, D. G. (2001). *Pattern Classification*, 2nd ed., Wiley-Interscience, New York.

FRIEDMAN, N., GEIGER, D. AND GOLDSZMIDT, M. (1997). "Bayesian network classifiers", *Machine Learning*, **29**, 131–163.

GOODMAN, L. A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models", *Biometrika*, **61**, 215–231.

HATHAWAY, R. J. (1986). "Another interpretation of the EM algorithm for mixture distributions", *Statistics & Probability Letters*, **4**, 53–56.

MCLACHLAN, G. AND PEEL, D. (2000). *Finite Mixture Models*, Wiley-Interscience, New York.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo.

RISSANEN, J. (1987). "Stochastic complexity", *Journal of the Royal Statistical Society*, Ser. B, **49**, 223–239.

SCHWARZ, G. (1978). "Estimating the dimension of a model", *The Annals of Statistics*, **6**, 461–464.

TIERNEY, L. AND KADANE, J. B. (1986). "Accurate approximations for posterior moments and marginal densities", *Journal of the American Statistical Association*, **81**, 82–86.