How to Take into Account the Discrete Parameters in the BIC Criterion?

V. Vandewalle

University Lille 2, IUT STID

COMPSTAT 2010
Paris
August 23th, 2010

Intorduction

Issue

- Some models involve discrete parameters.
- The discrete parameters play a part in the **likelihood overfitting**.
- **But,** they cannot be penalized using standard BIC approximation.

Study

- Study the influence of the discrete parameters in the BIC approximation
- Focus on a simple model : the modal modality model
- Study the accuracy of differents approximations

Outline

The modal modality model

Model

- $\mathbf{X} \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_m)$ ($\sum_{h=1}^m \alpha_h = 1$, $\alpha_h > 0$).
- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ an n i.i.d. sample coming from $\mathbf{X}$
- Constraint proposed by Biernacki et al. (2006) :

$$\alpha_h = \begin{cases} 1 - \varepsilon & \text{if } h = h^* \\ \frac{\varepsilon}{m-1} & \text{otherwise,} \end{cases}$$

  h^* the location of the modal modality and $0 \leq \varepsilon \leq \frac{m-1}{m}$.
- Two parameters must be estimated : ε which is continuous and h^* which is discrete.

Comments

- Intuitive interpretation.
- Useful to get parsimonious models in clustering.

The modal modality model

- Both continuous and discrete parameters, in a simple case.
- In a Bayesian setting integration over both continuous and discrete parameters.

Integrated likelihood

- Prior on (ε, h^*) :

$$p(\varepsilon, h^*) = \frac{1}{m} p(\varepsilon).$$

- Integrated likelihood :

$$p(\mathbf{x}) = \frac{1}{m} \sum_{h^*=1}^{m} \int_0^{\frac{m-1}{m}} p(\mathbf{x}|\varepsilon, h^*) p(\varepsilon) d\varepsilon.$$

- Truncated Dirichlet prior for $p(\varepsilon)$

$$p(\varepsilon) = C\varepsilon^{-\frac{1}{2}}(1-\varepsilon)^{-\frac{1}{2}} \mathbf{1}_{[0,\frac{m-1}{m}]}(\varepsilon),$$

with C some normalization constant.

Integrated likelihood

Let $n_h = \sum_{i=1}^{n} x_{ih}$, the logarithm of the integrated likelihood (IL) is

$$\text{IL} = \log\left(\frac{1}{m}\sum_{h=1}^{m}\int_0^{\frac{m-1}{m}}(1-\varepsilon)^{n_h}\left(\frac{\varepsilon}{m-1}\right)^{n-n_h} C\varepsilon^{-\frac{1}{2}}(1-\varepsilon)^{-\frac{1}{2}}d\varepsilon\right)$$

How can we approximate this integral ?

- Neglect discrete parameters.
- Make Laplace approximation for each term of the sum.
- Take into account the number of states of the discrete variable into account in the penalization.

Standard BIC approximation

- Maximum likelihood estimator of the parameters

$$(\hat{\varepsilon}, \widehat{h^*}) = \arg\max_{\varepsilon,h} (1-\varepsilon)^{n_h} \left(\frac{\varepsilon}{m-1}\right)^{n-n_h},$$

  which gives $\widehat{h^*} = \arg\max_h n_h$ and $\hat{\varepsilon} = 1 - \frac{n_{\widehat{h^*}}}{n}$.
- If the discrete parameters are not taken into account, the BIC criterion is :

$$\text{BIC}_1 = \log\left((1-\hat{\varepsilon})^{n_{\widehat{h^*}}} \left(\frac{\hat{\varepsilon}}{m-1}\right)^{n-n_{\widehat{h^*}}}\right) - \frac{1}{2}\log n,$$

- However this approximation is not justified when considering discrete parameters.

Taking the discrete parameters into account

For the sum into IL, there are terms for which the maximum in reached on the border for which we need the following proposition.

Proposition

Let $L : [a, b] \mapsto \mathbb{R}$, such that L be one time differentiable on $[a, b]$ and that it reaches its maximum at b with $L'(b) > 0$. Then

$$\log\left(\int_a^b e^{nL(u)}du\right) = nL(b) - \log n + O(1).$$

For a comparison note that

$$\log\left(\int_a^b e^{nL(u)}du\right) = nL(c) - \frac{1}{2}\log n + O(1),$$

if L would reach its maximum for $c \in]a, b[$.

Taking the discrete parameters into account

- Applying the previous proposition

$$\log\left(\int_0^{\frac{m-1}{m}}(1-\varepsilon)^{n_h}\left(\frac{\varepsilon}{m-1}\right)^{n-n_h}C\varepsilon^{-\frac{1}{2}}(1-\varepsilon)^{-\frac{1}{2}}d\varepsilon\right)=$$
$$\log p(\mathbf{x}|\hat{\varepsilon},h)-\frac{1+s_h}{2}\log n+O(1)$$

where $s_h=1$ if the constraint is saturated (i.e. $\hat{\varepsilon}=\frac{m-1}{m}$) and 0 otherwise.

- Then replacing these approximations in IL we get

$$\mathrm{BIC}_2=\log\left(\frac{1}{m}\sum_{h=1}^{m}(1-\hat{\varepsilon}_h)^{n_h}\left(\frac{\hat{\varepsilon}_h}{m-1}\right)^{n-n_h}n^{-\frac{1+s_h}{2}}\right)$$

where $\hat{\varepsilon}_h$ is the maximum likelihood estimator of ε when h is constrained to be the modal modality.

Taking the discrete parameters into account

- Simplify BIC_2 to avoid the integration on the states of the discrete variable, which gives the alternative criterion

$$\text{BIC}_3 = \log\left((1-\hat{\varepsilon}_{\widehat{h^*}})^{n_{\widehat{h^*}}}\left(\frac{\hat{\varepsilon}_{\widehat{h^*}}}{m-1}\right)^{n-n_{\widehat{h^*}}}\right) - \frac{1}{2}\log n - \log m.$$

- It is the standard BIC criterion penalized by the logarithm of the number of possible states of the discrete variable.

Numerical experiments

- Study the accuracy of the approximation in a simple case.
- Study the accuray for parsimonious models on binary data.

$\mathbf{X} \sim \mathcal{M}(1, 0.40, 0.35, 0.25)$

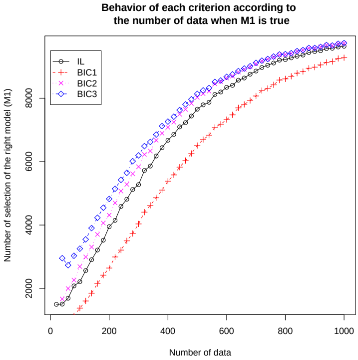

FIG.: Number of times where the true model is selected.

$\mathbf{X} \sim \mathcal{M}(1, 0.40, 0.30, 0.30)$

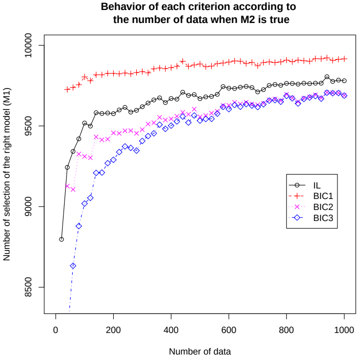

FIG.: Number of times where the parsimonious model is selected.

Binary simulated data

Model

- Binary data in the mulvariate case (in dimension d).
- $\mathbf{x}_i$ ($i \in \{1, \ldots, n\}$) with $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^d)$.
- $\mathbf{x}_i^j$ drawn from a Bernoulli distribution.
- Equality of ε for each variable (Celeux and Govaert (1991)).

Experimental setting

- If d is large it is not possible to perform the integration over all the states of the discrete variable.
- Importance sampling (IS) to compute the sum.
- Compare the different approximations of the integrated likelihood without considering the model choice issue.
- $d = 5$, $d = 10$ and $d = 20$ variables.
- $\varepsilon = 0.45$ for each variable.
- 100 datasets, simulate $10,000$ modal positions for IS.

Binary simulated data

Crit \ n	20	50	100	1000
		$d = 5$ dimensions		
IL	−70.91 (0.9)	−174.96 (1.2)	−347.77 (1.7)	−3448.18 (7.2)
BIC_1	−68.77 (1.4)	−172.59 (1.7)	−345.03 (2.2)	−3444.38 (7.2)
BIC_2	−**70.50** (0.8)	−**174.59** (1.2)	−**347.41** (1.6)	−3447.84 (7.2)
BIC_3	−72.23 (1.4)	−176.05 (1.7)	−348.49 (2.2)	−**3447.85** (7.2)
		$d = 10$ dimensions		
IL	−140.24 (1.0)	−348.15 (1.2)	−693.71 (2.4)	−6891.66 (10)
BIC_1	−135.98 (2.1)	−343.32 (2.1)	−688.22 (3.3)	−6884.02 (10)
BIC_2	−**139.49** (1.0)	−**347.44** (1.2)	−**693.01** (2.3)	−**6890.97** (10)
BIC_3	−142.91 (2.1)	−350.25 (2.1)	−695.15 (3.3)	−6890.95 (10)
		$d = 20$ dimensions		
IL	−279.01 (0.8)	−694.51 (1.4)	−1385.87 (2.4)	−13795.88 (14)
BIC_1	−271.06 (2.6)	−685.31 (3.2)	−1374.98 (3.5)	−13765.95 (11)
BIC_2	−**277.93** (0.8)	−**693.46** (1.4)	−**1384.84** (2.4)	−**13794.85** (14)
BIC_3	−284.93 (2.6)	−699.18 (3.2)	−1388.85 (3.5)	−13779.81 (11)

TAB.: Mean value of the criterion according the values of n and d, the standard deviation is given into parenthesis.

Binary real data

- Binary data from the UCI database repository and the Statlog database.
- Parsimonious product of binary distributions model.
- Comparison the integrated likelihood without considering the model choice issue.
- If the initial data are continuous they are discretized using the Fisher algorithm (Fisher (1958)).

Binary real data

Dataset	n	d	IL	BIC_1	BIC_2	BIC_3
SPECT Heart (Test)	187	23	−2759.1	−2742.5	−2758.0	−**2758.5**
SPECT Heart (Train)	80	23	−1015.5	−999.0	−1014.5	−**1014.9**
Acute Inflammations	120	7	−572.7	−568.1	−572.2	−**572.9**
Abalone	34	7	−164.1	−159.6	−163.6	−**164.4**
Breast Cancer Diagnostic	569	30	−9978.9	−9958.5	−9977.6	−**9979.3**
Crab	200	5	−695.9	−693.6	−**695.5**	−697.1
Cushings	27	2	−23.7	−22.5	−**23.8**	−23.9
Fglass	214	9	−947.6	−940.7	−**947.0**	−946.9

TAB.: Comparison of the approximations of the log-likelihood value for binary data of the UCI and Statlog databases.

Conclusion and perspectives

Conclusion

- The number of possible states of the discrete variable should be taken into account.
- At least in the penalty of the BIC criterion.

Perspectives

- Estimate the integrated likelihood via posterior simulation using the harmonic mean identity.
- Study the setting where the number of possible states of the discrete variable grows to infinity.