NOVEL BAYESIAN CLUSTER ENUMERATION CRITERION FOR CLUSTER ANALYSIS WITH FINITE SAMPLE PENALTY TERM

Freweyni K. Teklehaymanot[⋆,†]*, Michael Muma*[⋆]*, Abdelhak M. Zoubir*[⋆,†]

[⋆] Signal Processing Group
Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt, Germany
{muma, zoubir}@spg.tu-darmstadt.de

[†] Graduate School CE
Technische Universität Darmstadt
Dolivostraße 15, D-64293 Darmstadt, Germany
teklehaymanot@gsc.tu-darmstadt.de

ABSTRACT

The Bayesian information criterion is generic in the sense that it does not include information about the specific model selection problem at hand. Nevertheless, it has been widely used to estimate the number of data clusters in cluster analysis. We have recently derived a Bayesian cluster enumeration criterion from first principles which maximizes the posterior probability of the candidate models given observations. But, in the finite sample regime, the asymptotic assumptions made by the criterion, to arrive at a computationally simple penalty term, are violated. Hence, we propose a Bayesian cluster enumeration criterion whose penalty term is derived by removing the asymptotic assumptions. The proposed algorithm is a two-step approach which uses a model-based clustering algorithm such as the EM algorithm before applying the derived criterion. Simulation results demonstrate the superiority of our criterion over existing Bayesian cluster enumeration criteria.

Index Terms— Cluster Enumeration, Bayesian Information Criterion, Cluster Analysis, Small Sample Sizes

1. INTRODUCTION

Model selection is concerned with selecting a parsimonious statistical model, that adequately explains the observed data, from a family of candidate models using a predefined criterion. Over the years, many model selection criteria have been proposed in the literature [1–10]. Most model selection criteria contain data fidelity and penalty terms. One of the prominent fields of study where statistical model selection criteria are extensively used is cluster analysis [11–18]. In cluster analysis, estimating the number of data clusters, known as cluster enumeration, poses a major challenge.

Despite the fact that the original Bayesian Information Criterion (BIC) [8,10] is generic, it has been widely used in the literature, without questioning its validity, to estimate the number of clusters in an observed data set [11–15,17,18]. To mitigate this short coming, we have recently proposed a Bayesian cluster enumeration criterion, BIC_{N}, which is specifically derived for cluster analysis problems [19]. The original BIC (BIC_{O}) and BIC_{N} have the same data fidelity terms. But, their penalty terms are significantly different. Incorporating the cluster analysis problem in the derivation of the BIC has shown to be useful in estimating the number of clusters in data sets with overlapping and unbalanced clusters [19].

Like many model selection criteria in the literature, BIC_{N} is derived under asymptotic assumptions on the size of the observed data. However, in the finite sample regime, the asymptotic assumptions made by BIC_{N}, to arrive at a computationally simple penalty term, are violated. Hence, we propose an extension of BIC_{N} for the finite sample regime by removing the asymptotic assumptions. The proposed cluster enumeration criterion, BIC_{NF}, contains the data fidelity and penalty terms of BIC_{N} plus additional penalty terms. BIC_{NF} is able to satisfactorily estimate the number of data clusters in data sets with small sample sizes and in the asymptotic regime it behaves similar to BIC_{N}. The proposed cluster enumeration algorithm is a two-step approach which uses the Expectation Maximization (EM) algorithm to cluster the observed data set according to the specifications of a candidate model prior to the calculation of BIC_{NF} for that particular model.

The paper is organized as follows. Section 2 formulates the cluster enumeration problem. Section 3 discusses the proposed cluster enumeration algorithm. Performance evaluation of the proposed criterion and comparisons to existing cluster enumeration criteria using synthetic data examples is provided in Section 4. Section 5 concludes the paper.

2. PROBLEM FORMULATION

Let $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, for $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$, denote independent and identically distributed multivariate Gaussian random variables, where K is the number of clusters and $\boldsymbol{\mu}_k \in \mathbb{R}^{r\times 1}$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{r\times r}$ represent the centroid and covariance matrix of the kth cluster, respectively. The realizations of $\boldsymbol{x}_k$, denoted by $\boldsymbol{x}_n \in \mathbb{R}^{r\times 1}$, for $n = 1, \ldots, N_k$,

The work of F. K. Teklehaymanot is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt. The work of M. Muma is supported by the 'Athene Young Investigator Programme' of Technische Universität Darmstadt.

create a cluster $\mathcal{X}_k$ with parameters $\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]^\top$. Consider that the observed data set is a collection of the clusters $\mathcal{X}_k, k \in \mathcal{K}$, such that $\mathcal{X} \triangleq \{\mathcal{X}_1, \ldots, \mathcal{X}_K\} \subset \mathbb{R}^{r \times N}$, where $N \gg r$ and $N = \sum_{k=1}^{K} N_k$. The clusters $\mathcal{X}_k, k \in \mathcal{K}$, are independent, mutually exclusive, and non-empty. Let $\mathcal{M} \triangleq \{M_{L_{\min}}, \ldots, M_{L_{\max}}\}$ be a family of candidate models. Each candidate model M_l, for $l = L_{\min}, \ldots, L_{\max}$, represents a partition of $\mathcal{X}$ into l clusters with associated cluster parameters $\boldsymbol{\Theta}_l = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l]$ which lies in a parameter space $\Omega_l \subset \mathbb{R}^{q \times l}$. Our research goal is to estimate the number of clusters in $\mathcal{X}$ given a family of candidate models $\mathcal{M}$ assuming that the true number of clusters in $\mathcal{X}$ satisfies the constraint $L_{\min} \leq K \leq L_{\max}$.

We have recently derived the BIC from first principles by formulating the cluster enumeration problem as a maximization of the posterior probability of candidate models given data [19]. For a data set $\mathcal{X}$ with multivariate Gaussian distributed random variables

$$\begin{aligned}\mathrm{BIC}_{\mathrm{N}}(M_l) = &\sum_{m=1}^{l} N_m \ln N_m - \sum_{m=1}^{l} \frac{N_m}{2} \ln \left|\hat{\boldsymbol{\Sigma}}_m\right| \\ &- \frac{q}{2} \sum_{m=1}^{l} \ln N_m, \end{aligned} \tag{1}$$

where N_m represents the number of data vectors in $\mathcal{X}_m$, $\hat{\boldsymbol{\Sigma}}_m$ is the covariance matrix estimate of the mth cluster, and $q = \frac{1}{2}r(r+3)$ denotes the number of parameters estimated per cluster. The first two terms on the right hand side of Eq. (1) are data fidelity terms and the last term is the penalty term. In [19], the penalty term was derived from the expression

$$\frac{1}{2} \sum_{m=1}^{l} \ln \left|\hat{\boldsymbol{J}}_m\right| \tag{2}$$

by making asymptotic assumptions on the size of $\mathcal{X}$ to simplify the computation of $\ln \left|\hat{\boldsymbol{J}}_m\right|$, where $\hat{\boldsymbol{J}}_m$ is the observed Fisher Information Matrix (FIM) of the mth cluster. For small sample sizes, $\mathrm{BIC}_{\mathrm{N}}$ tends to select incorrect models. Hence, we derive the penalty term of $\mathrm{BIC}_{\mathrm{N}}$ for the finite sample regime by removing the asymptotic assumptions made to simplify Eq. (2).

3. PROPOSED BAYESIAN CLUSTER ENUMERATION CRITERION WITH FINITE SAMPLE PENALTY TERM

The observed FIM of the mth cluster is defined as

$$\hat{\boldsymbol{J}}_m \triangleq -\frac{d^2 \ln \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)}{d\boldsymbol{\theta}_m d\boldsymbol{\theta}_m^\top}\bigg|_{\boldsymbol{\theta}_m = \hat{\boldsymbol{\theta}}_m} \in \mathbb{R}^{q \times q}, \tag{3}$$

where $\boldsymbol{\theta}_m$ denotes the parameters of the mth cluster and $\mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)$ is the likelihood function. Since we are assuming that the data vectors in $\mathcal{X}$ are multivariate Gaussian distributed, $\ln \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)$ can be written as

$$\begin{aligned}\ln \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m) = N_m \ln \frac{N_m}{N} - \frac{rN_m}{2} \ln 2\pi - \frac{N_m}{2} \ln |\boldsymbol{\Sigma}_m| \\ - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Sigma}_m^{-1} \boldsymbol{\Delta}_m\right), \end{aligned} \tag{4}$$

where $\boldsymbol{\Delta}_m \triangleq \sum_{\boldsymbol{x}_n \in \mathcal{X}_m} (\boldsymbol{x}_n - \boldsymbol{\mu}_m)(\boldsymbol{x}_n - \boldsymbol{\mu}_m)^\top$. Since the covariance matrix of the mth cluster, $\boldsymbol{\Sigma}_m$, is a symmetric matrix, the relation $\mathrm{vec}(\boldsymbol{\Sigma}_m) = \boldsymbol{D}\boldsymbol{u}_m$ holds [20, pp. 56–57]. $\mathrm{vec}(\boldsymbol{\Sigma}_m) \in \mathbb{R}^{r^2 \times 1}$ denotes the stacking of the elements of $\boldsymbol{\Sigma}_m$ into one long column vector. The unique elements of $\boldsymbol{\Sigma}_m$ are stored in $\boldsymbol{u}_m \in \mathbb{R}^{\frac{1}{2}r(r+1) \times 1}$ and $\boldsymbol{D} \in \mathbb{R}^{r^2 \times \frac{1}{2}r(r+1)}$ represents the duplication matrix of $\boldsymbol{\Sigma}_m$. The duplication matrix, $\boldsymbol{D}$, is calculated as [21]

$$\boldsymbol{D}^\top = \sum_{i \geq j} \boldsymbol{v}_{ij} \mathrm{vec}\left(\boldsymbol{Y}_{ij}\right)^\top, \tag{5}$$

where $1 \leq j \leq i \leq r$, $\boldsymbol{v}_{ij} \in \mathbb{R}^{\frac{1}{2}r(r+1) \times 1}$ is a unit vector with one at its $(j-1)r + i - \frac{1}{2}j(j-1)$ entry and zero elsewhere and $\boldsymbol{Y}_{ij} \in \mathbb{R}^{r \times r}$ is given by

$$\boldsymbol{Y}_{ij} = \begin{cases} \boldsymbol{E}_{ii}, & i = j \\ \boldsymbol{E}_{ij} + \boldsymbol{E}_{ji}, & i \neq j \end{cases}, \tag{6}$$

where $\boldsymbol{E}_{ij}$ contains one at its i, jth entry and zero elsewhere. Taking the symmetry of $\boldsymbol{\Sigma}_m$ into account the parameter vector $\boldsymbol{\theta}_m = [\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m]^\top$ is replaced by $\check{\boldsymbol{\theta}}_m = [\boldsymbol{\mu}_m, \boldsymbol{u}_m]^\top$. A straight forward calculation of the second derivative of $\ln \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)$ with respect to $\check{\boldsymbol{\theta}}_m$, see [19] for details, results in

$$\begin{aligned}\frac{d^2 \ln \mathcal{L}(\boldsymbol{\theta}_m|\mathcal{X}_m)}{d\check{\boldsymbol{\theta}}_m d\check{\boldsymbol{\theta}}_m^\top} = &\frac{N_m}{2} \left(\frac{d\boldsymbol{u}_m}{d\boldsymbol{u}_m}\right)^\top \boldsymbol{D}^\top \boldsymbol{V}_m \boldsymbol{D} \frac{d\boldsymbol{u}_m}{d\boldsymbol{u}_m^\top} \\ &- \left(\frac{d\boldsymbol{u}_m}{d\boldsymbol{u}_m^\top}\right)^\top \boldsymbol{D}^\top \boldsymbol{W}_m \boldsymbol{D} \frac{d\boldsymbol{u}_m}{d\boldsymbol{u}_m} \\ &- N_m \left(\frac{d\boldsymbol{u}_m}{d\boldsymbol{u}_m}\right)^\top \boldsymbol{D}^\top \boldsymbol{Z}_m \mathrm{vec}\left(\frac{d\boldsymbol{\mu}_m^\top}{d\boldsymbol{\mu}_m^\top}\right) \\ &- N_m \frac{d\boldsymbol{\mu}_m^\top}{d\boldsymbol{\mu}_m} \boldsymbol{\Sigma}_m^{-1} \frac{d\boldsymbol{\mu}_m}{d\boldsymbol{\mu}_m^\top}, \end{aligned} \tag{7}$$

where

$$\boldsymbol{V}_m \triangleq \boldsymbol{\Sigma}_m^{-1} \otimes \boldsymbol{\Sigma}_m^{-1} \in \mathbb{R}^{r^2 \times r^2} \tag{8}$$

$$\boldsymbol{W}_m \triangleq \boldsymbol{\Sigma}_m^{-1} \otimes \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\Delta}_m \boldsymbol{\Sigma}_m^{-1} \in \mathbb{R}^{r^2 \times r^2} \tag{9}$$

$$\boldsymbol{Z}_m \triangleq \boldsymbol{\Sigma}_m^{-1} (\bar{\boldsymbol{x}}_m - \boldsymbol{\mu}_m) \otimes \boldsymbol{\Sigma}_m^{-1} \in \mathbb{R}^{r^2 \times r}. \tag{10}$$

Here, $\bar{\boldsymbol{x}}_m \triangleq \frac{1}{N_m} \sum_{\boldsymbol{x}_n \in \mathcal{X}_m} \boldsymbol{x}_n$ is the sample mean of the data vectors that belong to the mth cluster. A compact matrix representation of $\hat{\boldsymbol{J}}_m$ is given by

$$\hat{\boldsymbol{J}}_m = \begin{bmatrix} -\frac{\partial^2 \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_m|\mathcal{X}_m)}{\partial \boldsymbol{\mu}_m \partial \boldsymbol{\mu}_m^\top} & -\frac{\partial^2 \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_m|\mathcal{X}_m)}{\partial \boldsymbol{\mu}_m \partial \boldsymbol{u}_m^\top} \\ -\frac{\partial^2 \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_m|\mathcal{X}_m)}{\partial \boldsymbol{u}_m \partial \boldsymbol{\mu}_m^\top} & -\frac{\partial^2 \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_m|\mathcal{X}_m)}{\partial \boldsymbol{u}_m \partial \boldsymbol{u}_m^\top} \end{bmatrix} \tag{11}$$

The maximum likelihood estimator of the mean and covariance matrix of the mth Gaussian cluster are given by

$$\hat{\boldsymbol{\mu}}_m = \frac{1}{N_m} \sum_{\boldsymbol{x} \in \mathcal{X}_m} \boldsymbol{x}_n \tag{12}$$

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{N_m} \sum_{\boldsymbol{x}_n \in \mathcal{X}_m} (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_m)^\top . \tag{13}$$

Hence, simplifying Eq. (7) using Eqs. (12) and (13) results in

$$\hat{\boldsymbol{J}}_m = \begin{bmatrix} N_m \hat{\boldsymbol{\Sigma}}_m^{-1} & \boldsymbol{0}_{r \times \frac{1}{2}r(r+1)} \\ \boldsymbol{0}_{\frac{1}{2}r(r+1) \times r} & \frac{N_m}{2} \boldsymbol{D}^\top \hat{\boldsymbol{F}}_m \boldsymbol{D} \end{bmatrix}, \tag{14}$$

where $\hat{\boldsymbol{F}}_m \triangleq \hat{\boldsymbol{\Sigma}}_m^{-1} \otimes \hat{\boldsymbol{\Sigma}}_m^{-1} \in \mathbb{R}^{r^2 \times r^2}$ and $\boldsymbol{0}_{\frac{1}{2}r(r+1) \times r} \in \mathbb{R}^{\frac{1}{2}r(r+1) \times r}$ is a zero matrix. Using this result, the BIC of the candidate model M_l with finite sample penalty term, referred to as $\text{BIC}_{\text{NF}}(M_l)$, can be written as

$$\begin{aligned} \text{BIC}_{\text{NF}}(M_l) &= \sum_{m=1}^{l} N_m \ln N_m - \sum_{m=1}^{l} \frac{N_m}{2} \ln \left| \hat{\boldsymbol{\Sigma}}_m \right| \\ &\quad - \frac{1}{2} \sum_{m=1}^{l} \ln \left| \hat{\boldsymbol{J}}_m \right| \\ &= \sum_{m=1}^{l} N_m \ln N_m - \sum_{m=1}^{l} \frac{N_m}{2} \ln \left| \hat{\boldsymbol{\Sigma}}_m \right| \\ &\quad - \frac{1}{4} r(r+3) \sum_{m=1}^{l} \ln N_m + \frac{1}{4} r(r+1) l \ln 2 \\ &\quad + \frac{1}{2} \sum_{m=1}^{l} \ln \left| \hat{\boldsymbol{\Sigma}}_m \right| - \frac{1}{2} \sum_{m=1}^{l} \ln \left| \boldsymbol{D}^\top \hat{\boldsymbol{F}}_m \boldsymbol{D} \right| . \end{aligned} \tag{15}$$

Comparing Eqs. (1) and (15) we notice that

$$\begin{aligned} \text{BIC}_{\text{NF}}(M_l) &= \text{BIC}_{\text{N}}(M_l) + \frac{1}{4} r(r+1) l \ln 2 \\ &\quad + \frac{1}{2} \sum_{m=1}^{l} \ln \left| \hat{\boldsymbol{\Sigma}}_m \right| - \frac{1}{2} \sum_{m=1}^{l} \ln \left| \boldsymbol{D}^\top \hat{\boldsymbol{F}}_m \boldsymbol{D} \right| . \end{aligned} \tag{16}$$

Unlike BIC_{N} and BIC_{O}, the penalty term of BIC_{NF} depends on the covariance matrix of the individual clusters in $M_l \in \mathcal{M}$. This allows the proposed criterion, BIC_{NF}, to lower the penalty term when the determinant of the covariance matrices are high and penalize more severely when they are low. The calculation of $\text{BIC}_{\text{NF}}(M_l)$ using Eq. (16) requires the estimation of the covariance matrix, $\boldsymbol{\Sigma}_m$, and the number of data vectors per cluster, N_m, for $m = 1, \ldots, l$. Algorithm 1 shows the proposed two-step approach which uses the EM algorithm to estimate cluster parameters prior to the calculation of BIC_{NF}.

The additional complexity of $\text{BIC}_{\text{NF}}(M_l)$ compared to $\text{BIC}_{\text{N}}(M_l)$ comes from the term $\frac{1}{2} \sum_{m=1}^{l} \ln \left| \boldsymbol{D}^\top \hat{\boldsymbol{F}}_m \boldsymbol{D} \right|$. The duplication matrix $\boldsymbol{D}$ is computed only once, and thus it can be ignored in the complexity analysis. Hence, the excess computational cost is $\mathcal{O}(lr^6)$.

Note that the penalty term of BIC_{NF} contains covariance matrix estimate of each cluster in $M_l \in \mathcal{M}$. If the observations span a large range of values, then the covariances of individual clusters are very large and their inverses become close to zero. As a result, the penalty term of the proposed criterion, BIC_{NF}, might go to infinity. Hence, in such cases, we recommend normalizing the data prior to the estimation of cluster parameters.

Algorithm 1 Proposed two-step cluster enumeration algorithm

Inputs: data set $\mathcal{X}$; set of candidate models $\mathcal{M} \triangleq \{M_{L_{\min}}, \ldots, M_{L_{\max}}\}$
Calculate the duplication matrix $\boldsymbol{D}$ via Eq. (5)
for $l = L_{\min}, \ldots, L_{\max}$ **do**
  for $m = 1, \ldots, l$ **do**
    Estimate $\boldsymbol{\Sigma}_m$ using the EM algorithm
    Estimate N_m via hard clustering [19]
  end for
  calculate $\text{BIC}_{\text{NF}}(M_l)$ using Eq. (16)
end for
Estimate the number of clusters in $\mathcal{X}$:
$\hat{K} = \underset{l=L_{\min},\ldots,L_{\max}}{\arg\max} \text{BIC}_{\text{NF}}(M_l)$

4. RESULTS

4.1. Performance Measures

The two main performance measures used to compare different Bayesian cluster enumeration criteria are the empirical probability of detection

$$p_{\text{det}} = \frac{1}{\text{MC}} \sum_{s=1}^{\text{MC}} \mathbb{1}_{\{\hat{K}_s = K\}}, \tag{17}$$

where MC is the number of Monte-Carlo experiments and $\mathbb{1}_{\{\hat{K}_s = K\}}$ is the indicator function given by

$$\mathbb{1}_{\{\hat{K}_s = K\}} \triangleq \begin{cases} 1, & \text{if} \quad \hat{K}_s = K \\ 0, & \text{otherwise} \end{cases}, \tag{18}$$

and the Mean Absolute Error (MAE), which is computed as

$$\text{MAE} = \frac{1}{\text{MC}} \sum_{s=1}^{\text{MC}} \left| K - \hat{K}_s \right| . \tag{19}$$

We also consider the empirical probability of over estimation p_{over}, which is the probability that $\hat{K} > K$, as an additional performance measure.

4.2. Simulation Setup

In all simulations, we assume that $\mathcal{M} \triangleq \{M_{L_{\min}}, \ldots, M_{L_{\max}}\}$ is given with $L_{\min} = 1$ and $L_{\max} = 2K$, where K is the true number of data clusters in $\mathcal{X}$. All simulation results are an average of 1000 Monte-Carlo experiments. We compare our proposed criterion, BIC_{NF}, with BIC_{N} and BIC_{O}. The compared criteria use the same implementation of EM algorithm. For Data-1, we generate realizations of the random variable $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, for $k = 1, \ldots, 5$, with cluster centroids $\boldsymbol{\mu}_1 = [-2, 0]^\top$, $\boldsymbol{\mu}_2 = [5, 0]^\top$, $\boldsymbol{\mu}_3 = [0, 7]^\top$, $\boldsymbol{\mu}_4 = [8, 4]^\top$, $\boldsymbol{\mu}_5 = [3, 10]^\top$, and covariance matrices $\boldsymbol{\Sigma}_1 = \text{diag}\,(0.2, 0.2)$, $\boldsymbol{\Sigma}_2 = \text{diag}\,(0.6, 0.6)$, $\boldsymbol{\Sigma}_3 = \text{diag}\,(0.4, 0.4)$, $\boldsymbol{\Sigma}_4 = \text{diag}\,(0.2, 0.2)$, $\boldsymbol{\Sigma}_5 = \text{diag}\,(0.3, 0.3)$, where $\text{diag}(a, b)$ puts a and b in the main diagonal of a 2×2 matrix and sets the off-diagonal elements to zero. Data-2 contains realizations of the random variable $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, for $k = 1, \ldots, 6$, with $\boldsymbol{\mu}_1 = [-1, 0, 7]^\top$, $\boldsymbol{\mu}_2 = [3, 0, 8]^\top$, $\boldsymbol{\mu}_3 = [0, 5, 1]^\top$, $\boldsymbol{\mu}_4 = [9, 4, 4]^\top$, $\boldsymbol{\mu}_5 = [3, 9, 5]^\top$, $\boldsymbol{\mu}_6 = [5, 5, 1.5]^\top$, and $\boldsymbol{\Sigma}_1 = \text{diag}\,(0.6, 1.2, 0.6)$, $\boldsymbol{\Sigma}_2 = \text{diag}\,(1.8, 0.9, 1.5)$, $\boldsymbol{\Sigma}_3 = \text{diag}\,(1.2, 0.6, 0.3)$, $\boldsymbol{\Sigma}_4 = \text{diag}\,(0.9, 0.9, 0.9)$, $\boldsymbol{\Sigma}_5 = \text{diag}\,(0.9, 1.5, 0.9)$, $\boldsymbol{\Sigma}_6 = \text{diag}\,(1.2, 1.2, 1.2)$.

4.3. Simulation Results

Comparison of the three Bayesian cluster enumeration criteria as a function of the number of data vectors per cluster, $N_k, k \in \mathcal{K}$, for Data-1 is given in Table 1. The proposed criterion, BIC_{NF}, outperforms the other criteria when the number of data vectors per cluster is small and it exhibits a very small MAE. The empirical probability of over estimation, p_{over}, of BIC_{N} and BIC_{O} is very high especially when the number of data vectors per cluster is small. As expected the cluster number estimates of all compared criteria converge to the correct number of clusters, $K = 5$, when the number of data vectors per cluster increases. A comparison of the penalty terms of the different criteria when the number of data vectors per cluster $N_k = 10$ is shown in Fig. 1. BIC_{NF} penalizes over estimation more severely than the other criteria.

Next, we compare the cluster enumeration performance of different criteria for Data-2 by setting the number of data vectors per cluster to one of the values in $\{50, 100, 250, 1000\}$. BIC_{NF} outperforms the other criteria when N_k is small and it exhibits a small MAE. For small values of N_k, BIC_{N} performs better than BIC_{O}, while BIC_{O} tends to under estimate the number of clusters. Similar to the results of Data-1, asymptotically, all cluster enumeration criteria behave satisfactorily.

Table 1. Comparison of different Bayesian cluster enumeration criteria for Data-1.

		10	50	100	1000
$p_{\text{det}}(\%)$	BIC_{NF}	**77.6**	**100**	**100**	**100**
	BIC_{N}	0	77.8	96.2	**100**
	BIC_{O}	26.4	99.3	99.7	**100**
$p_{\text{over}}(\%)$	BIC_{NF}	0	0	0	0
	BIC_{N}	100	22.2	3.8	0
	BIC_{O}	73.1	0.7	0.3	0
MAE	BIC_{NF}	**0.228**	**0**	**0**	**0**
	BIC_{N}	4.768	0.483	0.043	**0**
	BIC_{O}	2.461	0.007	0.003	**0**

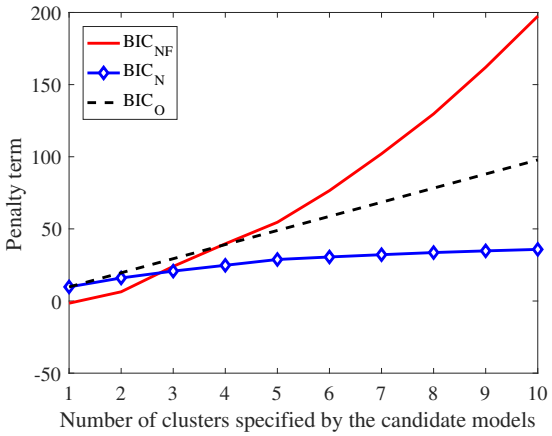

Fig. 1. The penalty terms of different Bayesian cluster enumeration criteria for Data-1 when $N_k = 10$.

Table 2. Comparison of different Bayesian cluster enumeration criteria for Data-2.

		50	100	250	1000
$p_{\text{det}}(\%)$	BIC_{NF}	**82.1**	**96.7**	**98.7**	**99.3**
	BIC_{N}	64.7	92.9	98.1	**99.3**
	BIC_{O}	51.7	91.1	**98.7**	**99.3**
$p_{\text{over}}(\%)$	BIC_{NF}	0.6	0.6	0.6	0.2
	BIC_{N}	30.9	5.7	1.3	0.2
	BIC_{O}	0	0.2	0.2	0
MAE	BIC_{NF}	**0.19**	**0.033**	**0.013**	**0.007**
	BIC_{N}	0.851	0.079	0.019	**0.007**
	BIC_{O}	0.602	0.089	**0.013**	**0.007**

5. CONCLUSION

We propose a Bayesian cluster enumeration criterion whose penalty term is derived for the finite sample regime. Further, the proposed criterion is integrated into a two-step algorithm to provide an optimal estimate of the number of data clusters. Simulation results confirm the strength of the proposed criterion for estimating the number of clusters in data sets with small sample sizes. Our proposed criterion, BIC_{NF}, achieves good performance results with a small additional computational complexity compared to BIC_{N}.

6. REFERENCES

[1] P. M. Djurić, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[2] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge University Press, New York, 2008.

[3] G. Claeskens and N. L. Hjort, "The focused information criterion," *J. Am. Stat. Assoc.*, vol. 98, no. 464, pp. 900–916, Dec. 2003.

[4] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *2nd Int. Symp. Inf. Theory*, 1973, pp. 267–281.

[5] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. R. Statist. Soc. B*, vol. 41, no. 2, pp. 190–195, 1979.

[6] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *J. R. Statist. Soc. B*, vol. 64, no. 4, pp. 583–639, 2002.

[7] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[8] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[9] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, June 1989.

[10] J. E. Cavanaugh and A. A. Neath, "Generalizing the derivation of the Schwarz information criterion," *Commun. Statist.-Theory Meth.*, vol. 28, no. 1, pp. 49–66, 1999.

[11] D. Pelleg and A. Moore, "X-means: extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. Mach. Learn. (ICML)*, 2000, pp. 727–734.

[12] M. Shahbaba and S. Beheshti, "Improving X-means clustering with MNDL," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. and Appl. (ISSPA)*, Montreal, Canada, 2012, pp. 1298–1302.

[13] T. Ishioka, "An expansion of X-means for automatically determining the optimal number of clusters," in *Proc. 4th IASTED Int. Conf. Comput. Intell.*, Calgary, Canada, 2005, pp. 91–96.

[14] Q. Zhao, V. Hautamaki, and P. Fränti, "Knee point detection in BIC for detecting the number of clusters," in *Proc. 10th Int. Conf. Adv. Concepts Intell. Vis. Syst. (ACIVS)*, Juan-les-Pins, France, 2008, pp. 664–673.

[15] Q. Zhao, M. Xu, and P. Fränti, "Knee point detection on Bayesian information criterion," in *Proc. 20th IEEE Int. Conf. Tools with Artificial Intell.*, Dayton, USA, 2008, pp. 431–438.

[16] T. Huang, H. Peng, and K. Zhang, "Model selection for Gaussian mixture models," *Statistica Sinica*, vol. 27, no. 1, pp. 147–169, 2017.

[17] A. Mehrjou, R. Hosseini, and B. N. Araabi, "Improved Bayesian information criterion for mixture model selection," *Pattern Recognit. Lett.*, vol. 69, pp. 22–27, Jan. 2016.

[18] F. K. Teklehaymanot, M. Muma, J. Liu, and A. M. Zoubir, "In-network adaptive cluster enumeration for distributed classification/labeling," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, 2016, pp. 448–452.

[19] F. K. Teklehaymanot, M. Muma, and A. M. Zoubir, "A novel Bayesian cluster enumeration criterion for unsupervised learning," *IEEE Trans. Signal Process. (under review), [Online-Edition:* `https://arxiv.org/abs/1710.07954v2`*]*, 2018.

[20] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics (3 ed.)*, Wiley Series in Probability and Statistics. John Wiley & Sons Ltd, Baffins Lane, Chichester, West Sussex PO19 1UD, England, 2007.

[21] J. R. Magnus and H. Neudecker, "The elimination matrix: some lemmas and applications," *SIAM J. Algebraic Discrete Meth.*, vol. 1, no. 4, pp. 422–449, Dec. 1980.