Bayesian versus data driven model selection for microarray data

Raffaele Giancarlo · Giosué Lo Bosco · Filippo Utro

Abstract Clustering is one of the most well known activities in scientific investigation and the object of research in many disciplines, ranging from Statistics to Computer Science. In this beautiful area, one of the most difficult challenges is a particular instance of the *model selection* problem, i.e., the identification of the correct number of clusters in a dataset. In what follows, for ease of reference, we refer to that instance still as model selection. It is an important part of any statistical analysis. The techniques used for solving it are mainly either Bayesian or data-driven, and are both based on *internal knowledge*. That is, they use information obtained by processing the input data. Although both techniques have been evaluated in the realm of microarray data analysis, their merits (relative to each other) has not been assessed. Here we will fill this gap in the literature by comparing three Bayesians versus several state of the art data-driven model selection methods. Our results show that, although in some cases Bayesian methods guarantee good results, they are not able to compete in terms of ability to predict the correct number of clusters in a dataset with the data-driven methods.

Keywords Clustering · Model selection · Bayesian information criterion · Akaike information criterion · Minimum message length · Bioinformatics

Electronic supplementary material The online version of this article (doi:10.1007/s11047-014-9446-5) contains supplementary material, which is available to authorized users.

R. Giancarlo · G. Lo Bosco (✉)
Dipartimento di Matematica e Informatica, University of Palermo, Via Archirafi 34, 90123 Palermo, Italy
e-mail: giosue.lobosco@unipa.it

R. Giancarlo
e-mail: raffaele@math.unipa.it

G. Lo Bosco
Istituto Euro Mediterraneo di Scienza e Tecnologia, Via Emerico Amari 123, 90139 Palermo, Italy

F. Utro
Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
e-mail: futro@us.ibm.com

1 Introduction

The advent of high throughput technologies, in particular microarrays, for biological research has revived interest in clustering, resulting in a plethora of new clustering algorithms. Indeed, experiments based on them are common practice in biological and medical research to address a wide range of problems, including the classification of tumors (Alizadeh et al. 2000; Alon et al. 1999; Dudoit and Fridlyand 2002; Golub et al. 1999; Perou et al. 1999; Pollack et al. 1999; Ross et al. 2000), where a reliable and precise classification is essential for successful diagnosis and treatment.

In the classic statistics and data analysis literature, there are two essential aspects of clustering: finding a "good" partition of the datasets and estimating the number of clusters, if any, in a dataset. The former problem is usually solved by the use of a clustering algorithm. For recent reviews on clustering algorithms, in particular for biomedical research, the reader is referred to Andreopoulos et al. (2009) and D'haeseleer (2006). However, the most fundamental issue is the latter problem, here referred to as model selection.

In general, it is usually solved with the use of internal/relative measures (defined in Sect. 2). The excellent survey by Handl et al. (2005) makes the study of those techniques

a central part of both research and practice in bioinformatics. It is also worth mentioning that a recent systematic presentation of statistical measures for clustering, with particular attention to microarray data, is given in Giancarlo et al. (2009). The two most prominent categories in which measures for model selection fall are as follows:

- Bayesian: Methods belonging to this category use Bayes rule and perform model selection by suitably choosing among a finite set of a priori fixed models. The Bayesian information criterion (BIC) (Schwarz 1978), the Akaike information criterion (AIC) (Akaike 1978) and the minimum message length (MML) (Wallace and Boulton 1968) are three Bayesian methods that have been used to efficiently estimate the number of clusters in a dataset (Wallace and Boulton 1968; Pelleg and Moore 2000; Wallace and Dowe 2000; Figuereido and Jain 2002; Bouguila and Ziou 2007). It is worthy of mention that the well known minimum description length (MDL) (Rissanen 1978) also falls in this category but, being equivalent to the BIC (Rissanen 1978), is implicitly accounted for here.
- Data-driven: Nothing is assumed about the structure of the dataset, which is inferred directly from it. We recall that, in the data analysis literature and for the special case of microarray data, Giancarlo et al. (2008a) and Giancarlo and Utro (2011) have recently proposed an extensive comparative analysis of data-driven validation measures taken from the most relevant paradigms in the area: (a) compactness e.g., Tibshirani et al. (2001) and Krzanowski and Lai (1985); (b) hypothesis testing in statistics, e.g., Tibshirani et al. (2001); (c) stability-based techniques, e.g., Dudoit and Fridlyand (2002), Ben-Hur et al. (2002), and Monti et al. (2003); and (d) jackknife techniques, e.g., Yeung et al. (2001). These benchmarks consider both the ability of a measure to predict the correct number of clusters in a dataset and, departing from the current state of the art in that area, the computer time it takes for a measure to complete its task.

Here, we compare the ability of AIC, BIC and MML to predict the correct number of clusters against the best data-driven methods identified by the study in Giancarlo et al. (2008a) and Giancarlo and Utro (2011), and for the specific research field of microarray data analysis.

The reminder of this paper is organized as follows: Sect. 2 presents a formal statement of the problems we are interested in. Section 3 describes the details about all the methods for model selection that have been used here. Section 4 is devoted to the description of the datasets and clustering algorithms used for the experiments, while Sect. 5 reports the results of the experiments. Finally, the last section offers some conclusions.

2 Basic notions and definitions

The aim of cluster analysis is to determine a partition of n items according to a *similarity/distance* function. Intuitively, the partition should be such that items in the same cluster are "similar", while items in different clusters are "dissimilar". Following Handl et al. (2005), cluster analysis here is seen as a three step process: (1) pre-processing, (2) clustering and (3) cluster validation. For the first step, the state of the art is given in Quackenbush (2002) for normalization, in Liu and Motoda (1998) for feature selection and in Giancarlo et al. (2010, 2011, 2013), and Priness et al. (2007) for the choice of similarity/distance functions. Regarding the other two steps, in what follows, we highlight the essential aspects of them, with some emphasis on cluster validation since it is central for this paper. To this end, we need to introduce some notation.

Consider a set of n items $X = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^m$, $1 \leq i \leq n$, and m is referred to as the number of features or conditions. The set X is represented in one of two different ways: (1) a data matrix D, of size $n \times m$, in which the rows represent the items and the columns represent the condition values; (2) a similarity/dissimilarity matrix S, of size $n \times n$, in which each entry $S_{i,j}$, $1 \leq i \neq j \leq n$, quantifies the similarity/dissimilarity of the pair of items (i, j). Specifically, the value of $S_{i,j}$ can be computed using rows i and j of D.

In what follows, let $C_k = \{c_1, c_2, \ldots, c_k\}$ be a *partition* of X. Each subset $c_i \subseteq X$, $1 \leq i \leq k$, is referred to as a *cluster*, and C_k is referred to as a *clustering solution*. Let $\overline{C}_j$ be a reference classification for X consisting of j classes. That is, $\overline{C}_j$ may either be a partition of X into j groups or a division of the universe generating X into j categories, usually referred to as *class labels*. Such a reference classification is referred to as *gold solution*. Intuitively, the partition of the dataset in classes is based on external knowledge that leaves no ambiguity on the actual number of classes and on the membership of elements to classes.

It is worth pointing out that although there exist real microarray datasets for which such an *a priori* division is known, in a few previous studies of relevance here, a more relaxed criterion has been adopted. Indeed, datasets with high quality partitions that have been inferred by the use of internal knowledge via data analysis tools such as clustering algorithms. In strict technical terms, there is a difference between the two types of gold solutions. For their datasets, Dudoit and Fridlyand (2002) elegantly make clear that difference in a related study and we closely follow their approach here.

2.1 Clustering algorithms

Usually, the partition of the items in X is accomplished by means of a clustering algorithm $\mathcal{A}$. The literature on clustering algorithms is very rich and a recent survey of

classic as well as more innovative methods, specifically designed for microarray data, is given in Andreopoulos et al. (2009), Shamir and Sharan (2003) and a more in depth treatment can be found for instance in Handl et al. (2005), Everitt (1993), Hartigan (1975), Jain and Dubes (1988), and Kaufman and Rousseeuw (1990). For the convenience of the reader, we recall that clustering algorithms are classified into: *partitional* and *hierarchical*.

The first type of clustering algorithms takes as input X and most of them also an integer k that in some cases can be estimated automatically. The final output is a partition C_k. It is worth pointing out that a partitional clustering algorithm can take as input a partition of the data and use it as an initial clustering solution that the algorithm refines. In this paper, we refer to this input option as *external initialization*. The second type of clustering algorithm produces a nested sequence of partitions, i.e. a tree. However, it can be easily adapted to generate a partition of a dataset into k clusters. The details are left to the reader.

2.2 Cluster validation

2.2.1 External validation measures

An *external measure E* is a function that takes as input two partitions C_j and C_k and returns a value assessing how close C_k is to C_j. It is external because the quality assessment of the partition is established via criteria external to the data. The three most prominent external measures known in the literature are: the Adjusted Rand Index (Hubert and Arabie 1985), the F-index (Rijsbergen 1979) and the Fowlkes and Mallows Index (FM-Index for short) (Fowlkes and Mallows 1983). For definitions and examples of their uses, the interested reader is referred to Jain and Dubes (1988) for a general presentation and to Giancarlo et al. (2008b) for one specific to microarray data.

2.2.2 Internal validation measures

An *internal measure I* is a function defined on the set of all possible partitions of X and with values in $\mathbb{R}$. It should measure the quality of a partition according to some suitable criteria, based on information contained in the dataset without resorting to external knowledge, i.e. a partition of the data known a priori (e.g. a gold solution).

It is worth pointing out that, in the specialistic literature, it is usual to refer to the internal measures also with the term *relative* when they are used to establish the optimal number of clusters k^*. For the state of the art on internal measures, the reader is referred to Handl et al. (2005), Giancarlo et al. (2008a, b), and Giancarlo and Utro (2012a).

3 Model selection methods

Some of the most prominent internal measures are based on: (a) compactness (b) hypothesis testing in statistics; (c) stability-based techniques; (d) jackknife techniques and (e) Bayesian scores. In particular, for each of the mentioned classes we consider the following:

(a) within clusters sum of square (`WCSS` for short) (Hastie et al. 2003) and Krzanowski and Lai Index (`KL` for short) (Krzanowski and Lai 1985).
(b) Gap statistics (`Gap` for short) (Tibshirani et al. 2001).
(c) CLEST (Dudoit and Fridlyand 2002), model explorer (`ME` for short) (Ben-Hur et al. 2002), consensus clustering (`Consensus` for short) (Monti et al. 2003) and fast consensus (`FC` for short) (Giancarlo and Utro 2011).
(d) Figure of merit (`FOM` for short) (Yeung et al. 2001).
(e) Schwarz's `BIC` (Schwarz 1978), `AIC` (Akaike 1978), `MML` (Wallace and Boulton (1968)).

Those measures have been selected among the many that have been proposed in the relevant literature for their prominence and based on an accurate and robust comparative experimental analysis Giancarlo et al. (2008a) that indicates them as the most reliable in this class. We highlight here a few key facts about them, relevant for our experiments. We refer the interested reader to Schwarz (1978), Giancarlo et al. (2008a), and Giancarlo and Utro (2011) for a more in-depth presentation of each of them, as well as additional references to textbooks and papers covering additional aspects of them.

WCSS (Hastie et al. 2003) measures the "goodness" of a cluster via its compactness, one of the most fundamental indicators of cluster quality. Indeed, for each k in $[2, k_{max}]$, the method consists of computing the sum of the square distance between each element in a cluster and the centroid of that cluster. The "correct" number of clusters k^* is predicted according to the following rule of thumb. For values of $k < k^*$, the value of `WCSS` should be substantially decreasing, as a function of the number of clusters k. On the other hand, for values of $k^* \leq k$, the compactness of the clusters will not increase as much, causing the value of `WCSS` not to decrease as much. The following heuristic approach comes out Tibshirani et al. (2001): plot the values of `WCSS`, computed on the given clustering solutions, in the range $[1, k_{max}]$; choose as k^* the abscissa closest to the "knee" in the `WCSS` curve.

KL (Klie et al. 2010) is an internal measure based on `WCSS`, but it is automatic, i.e., a numeric value for k^* is returned. Let

$$DIFF(k) = (k-1)^{\frac{2}{m}}\texttt{WCSS}(k-1) - k^{\frac{2}{m}}\texttt{WCSS}(k) \quad (1)$$

whose expected behavior is:

(i) for $k < k^*$, both $DIFF(k)$ and $DIFF(k+1)$ should be large positive values.
(ii) for $k > k^*$, both $DIFF(k)$ and $DIFF(k+1)$ should be small values, and one or both might be negative.
(iii) for $k = k^*$, $DIFF(k)$ should be large positive, but $DIFF(k+1)$ should be relatively small (might be negative).

Based on these considerations, Krzanowski and Lai proposed to choose, as prediction of k^*, the k maximizing:

$$KL(k) = \left|\frac{DIFF(k)}{DIFF(k+1)}\right| \quad (2)$$

That is,

$$k^* = \arg\max_{2 \le k \le k_{max}} KL(k) \quad (3)$$

Gap (Tibshirani et al. 2001) as KL, is an automatic internal measure based on WCSS. The method computes the gap between the WCSS curve computed on datasets produced by a null model and the one computed on a real dataset. Since WCSS is expected to decrease sharply up to k^*, on the real dataset, while it is expected to have a nearly constant slope on the null model datasets, the size of the gap is expected to increase up to k^* and then to decrease. Moreover, the WCSS curves are normalized via logs and a simulation error is also considered. Finally, it is worth pointing out that a more accurate prediction of k^* is based on a Monte Carlo simulation, i.e., the method is executed several times and the most frequent outcome is taken as the prediction.

G-Gap (Giancarlo et al. 2008a) is a geometric approximation of the Gap Statistics (Tibshirani et al. 2001). It allows for the identification of the "knee" in the WCSS curve via a geometric approach, rather than a Monte Carlo simulation as in the Gap Statistics.

CLEST (Dudoit and Fridlyand 2002) generalizes in many aspects an approach proposed by Breckenridge (1989) and can be regarded as a clever combination of hypothesis testing and stability-based techniques. It estimates the number of clusters in a dataset by iterating the following: for each $k \in [k_{min}, k_{max}]$ randomly partition, H times, the original dataset in a *learning* set and a *training* set. The learning set is used to build a classifier [e.g. diagonal linear discriminant analysis, see Dudoit and Fridlyand (2002) for details] for the data. That is, the classifier is assumed to be a reliable model for the data. Therefore, the classifier is used to derive a "gold solution" partition of the training set, which is then used to assess the quality of the partitions of the training set obtained by a given clustering algorithm. Then, method computea a p value for each k. Finally, it estimates k^* as the maximum value of k that satisfies a given threshold criteria on the computed p values.

FOM (Yeung et al. 2001) is a family of internal validation measures specifically designed for microarray data. It is based on the jackknife approach and has been designed for use as a relative measure assessing the predictive power of a clustering algorithm, i.e., its ability to predict the correct number of clusters in a dataset. We use the adjusted aggregate FOM for our experiments and, for brevity, we refer to it simply as FOM. The FOM computation is based on a root mean square deviation over all conditions. Formally, for each k, it is computed as:

$$\texttt{FOM}_k = \sum_1^m \sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{x \in c_i}(D(x,e) - m_i(e))^2}$$

where $D(x,e)$ is the be the feature level of x without the column e and $m_i(e)$ is the average feature level of condition e for items in cluster c_i (note that the clustering is also computed to the matrix D without the feature e). FOM uses the same heuristic methodology outlined for WCSS, i.e., one tries to identify the "knee" in the FOM plot as a function of the number of clusters.

Diff-FOM (Giancarlo et al. 2008a) is an extension of KL (Krzanowski and Lai 1985) to FOM. It is based on the computation of the following formula:

$$DFOM(k) = (k-1)^{2/m}\texttt{FOM}(k-1) - k^{2/m}\texttt{FOM}(k).$$

The "rule of thumb" that one uses to predict k^*, via Diff-FOM, is the same as for KL, being based on the same intuition. Therefore, as k increases towards k^*, $DFOM(k)$ increases to decrease sharply and then assumes nearly constant values as it moves away from k^*. So, one can take as k^* the abscissa corresponding to the maximum of $DFOM(k)$ in the interval $[3, k_{max}]$.

Consensus (Monti et al. 2003) is a stability-based technique (Giancarlo and Utro 2012a, b). Therefore, a large number of clustering solutions, each obtained via a sample of the original dataset, are used in order to identify the correct number of clusters. Intuitively, it computes a consensus matrix that indicates the level of agreement of clustering solutions that have been obtained via independent sampling of the dataset. Based on experimental observations and sound arguments, Monti et al. (2003) derive a "rule of thumb" in order to estimate the real number k^* of clusters present in D. For brevity, in what follows, only the key points are presented. The interested reader can find a full discussion in Monti et al. (2003), Giancarlo and Utro (2012a, b). The empirical cumulative distribution of the entries of the consensus matrix is computed. In an ideal situation in which there are k clusters and the clustering algorithm is so good to provide a perfect

classification, such a curve is bimodal, with peaks at zero and one. Monti et al. observe and validate experimentally that the area under the CDF curve is an increasing function of k. That result has also been confirmed by the experiments in Giancarlo et al. (2008a). In particular, for values of $k = k^*$, that area has a significant increase, while its growth flattens out for $k > k^*$. However, Monti et al. propose a closely associated method, described next. For a given k, the area of the corresponding CDF curve is estimated and an increasing function Δ is computed for successive values of k. Again, Monti et al. observe experimentally that: (i) For each $k \leq k^*$, there is a pronounced decrease of the Δ curve. That is, the incremental growth of the area under the CDF decreases sharply. (ii) For $k > k^*$, there is a stable plot of the Δ curve. That is, for $k > k^*$, the growth of the area flattens out. From this behaviour, the "rule of thumb" to identify k^* with the use of the Δ curve is: take as k^* the abscissa corresponding to the smallest non-negative value where the curve starts to stabilize; that is, no big variation in the curve takes place from that point on.

FC (Giancarlo and Utro 2011) is a fast approximation of `Consensus`, based on the observation that costly computational duplications can be avoided when the clustering algorithm is hierarchical.

BIC, AIC MML (Schwarz 1978; Akaike 1978; Wallace and Boulton 1968) They are useful in model selection since the evidence that a dataset could be generated by a particular model can be quantified by particular scores, defined as follows:

$$\texttt{BIC}(M_r) = -ln\ p(X|\Theta, M_r) + \frac{r}{2} ln\ n, \tag{4}$$

$$\texttt{AIC}(M_r) = -ln\ p(X|\Theta, M_r) + r, \tag{5}$$

$$\begin{aligned}\texttt{MML}(M_r) = &-ln\ p(X|\Theta, M_r) + \frac{1}{2} ln\ |I(\Theta)| \\ &-\frac{r}{2}(1 - ln\ 12),\end{aligned} \tag{6}$$

where X is the dataset, M_r indicates a model with r parameters $\Theta = \{\theta_1, .., \theta_r\}$, $p(X|\Theta, M_r)$ is the likelihood, n is the number of observations in X, $I(\theta)$ denotes the Fisher information matrix, i.e. the expectation of the Hessian of $p(X|\Theta, M_r)$ with respect to Θ, and $|.|$ indicates the determinant of a matrix.

Once the model M_r is assumed, all the three above mentioned scores are computed by performing a maximum likelihood estimation on $p(X|\Theta, M_r)$.

The general idea for selecting the best model over a finite set M_r for $r = 1, \ldots, r_{max}$, is to compute the score for each r, and finally choose as best model the one which minimizes such a score.

The justification behind this criterion is that all the scores have in common the log likelihood, that is always minimized due to the maximum likelihood estimation phase. Since this term decreases while the number of parameter increases, each score formulation introduces a different penalty term for the number of parameters r in order to avoid overfitting.

In particular, `BIC` introduces a penalty that results only from Bayes factors, *AIC* and *MML* take into account the entropy of the model also in terms of compressibility of a message containing the data.

One of the possible uses of the three mentioned scores is the estimation of the correct number of clusters in a dataset. The idea is to use a clustering algorithm compatible to a particular model M_{r_k}, where k denotes the number of clusters, and use that clustering algorithm to compute k_{max} partitions, each for $k = 1, \ldots, k_{max}$. Finally, the number of clusters k^* can be estimated as:

$$k^* = \underset{k=1,\ldots,k_{max}}{\arg\min}\ \texttt{S}(M_{r_k}).$$

where S can indicate `BIC`, `AIC` or `MML`.

In the case of Gaussian model assumption, the log likelihood of each each element $x_i \in c_j$ assumes the form

$$\begin{aligned}ln\ p(x_i|\mu_j, \Sigma_j) = &ln\ n_j - ln\ n \\ &-\frac{1}{2}(x_i - \mu_j)\Sigma_j^{-1}(x_i - \mu_j)^T \\ &-\frac{m}{2}(ln 2\pi + ln(|.|\Sigma_j))\end{aligned} \tag{7}$$

and can be estimated by computing the estimation of the centroids $\widehat{\mu_j}$ and covariance matrices $\widehat{\Sigma_j}$ on the cluster c_j that contains n_j elements.

Finally, the three score values can be computed by imposing the number of parameters $r_k = 2 \cdot k \cdot m + k - 1$, where $2 \cdot k \cdot m$ represents the number of centroids and covariances to estimate, and $k - 1$ is the number of class probabilities.

4 Experimental setup

We now detail the experimental methodology used to obtain the results presented in this manuscript. As it will be evident, it is a de facto standard in this area.

4.1 Clustering algorithms

In our experiments, we have chosen K-means among the *Partitional Methods*, and Average Link (Hier-A), Complete Link (Hier-C) and Single Link (Hier-S) among the

Hierarchical Methods. Of course, each of the above mentioned algorithms has already been used for data analysis of microarray data, e.g. Giancarlo et al. (2008b); Giancarlo and Utro (2012a); Gesú et al. (2005). The interested reader is referred to Fig. S1 to see the performance of these algorithms, on microarray data, evaluated via with the Adjusted Rand Index. We use K-means both in the version that starts the clustering from a random partition of the data and in the version where it takes as part of the input an initial partition produced by one of the chosen hierarchical methods. The acronyms of those versions are K-means-R, K-means-A, K-means-C and K-means-S, respectively.

4.2 Datasets

The eleven datasets, together with the acronyms used in this paper, are reported next. For conciseness, we mention only some relevant facts about them. The interested reader can find additional information in Dudoit and Fridlyand (2002) for the Lymphoma and NCI60 datasets, Di Gesú et al. (2005) for the CNS Rat, Leukemia and Yeast datasets and in Monti et al. (2003), for the remaining ones.

It is worth pointing out that, although microarray technology is capable of producing very large datasets, i.e., elements and features in the millions each, those datasets go through substantial size reduction steps for two main reasons. The first is to identify, and therefore focus, on part of the dataset giving the "most informative" variations, the second is computational. Indeed, as clearly pointed out by many studies, e.g., Giancarlo and Utro (2011), clustering algorithms and internal validation measures have serious limitations for large datasets, being able to process reliably and efficiently those with a number of elements and features in the hundreds. The choice of our datasets reflects such a state of the art. However, for each of them, we point out, or give reference to, the process that has been followed to obtain them from the original microarray experiment.

CNS Rat It is a 112×17 data matrix, obtained from the expression levels of 112 genes during a rat's central nervous system development. The dataset is studied by Wen et al. (1998), where they suggest a partition of the genes into six classes, four of which are composed of biologically, functionally-related genes. This partition is taken as the gold solution, which is the same one used for the validation of FOM (Yeung et al. 2001).

Leukemia It is a 38×100 data matrix, where each row corresponds to a patient with acute leukemia and each column to a gene. The original microarray experiment consists of a 72×6817 matrix, due to Golub et al. (1999). In order to obtain the current dataset, Handl et al. (2005) extracted from it a 38×6817 matrix, corresponding to the learning set in the study of Golub et al. and, via preprocessing steps, they reduced it to the current dimension by excluding genes that exhibited no significant variation across samples. The interested reader can find details of the extraction process in Handl et al.. For this dataset, there is a partition into three classes and that is taken as gold solution.

NCI60 It is a 57×200 data matrix, where each row corresponds to a cell line and each column to a gene. This dataset originates from a microarray study in gene expression variation among the sixty cell lines of the National Cancer Institute anti-cancer drug screen (NCI 2008), which consists of a 61×5244 data matrix. There is a partition of the dataset into eight classes, for a total of 57 cell lines, and it is taken as the gold solution. The dataset has been obtained from the original microarray experiments as described by Dudoit and Fridlyand (2002).

Lymphoma It is a 80×100 data matrix, where each row corresponds to a tissue sample and each column to a gene. The dataset comes from the study of Alizadeh et al. (2000) on the three most common adult lymphoma tumors. There is a partition into three classes and it is taken as the gold solution. The dataset has been obtained from the original microarray experiments, consisting of an 80×4682 data matrix, following the same preprocessing steps detailed in Dudoit and Fridlyand (2002).

Yeast It is a 698×72 data matrix, studied by Spellman et al. (1998) whose analysis suggests a partition of the genes into five functionally-related classes, which is taken as the gold solution and which has been used by Shamir and Sharan for a case study on the performance of clustering algorithms (Shamir and Sharan 2003).

St. Jude It is a 248×985 data matrix, where each row corresponds to a tissue sample and each column to a gene. The dataset comes from the study of Yeoh et al. (2002) on diagnostic bone marrow samples from pediatric acute leukemia patients corresponding to 6 prognostically important leukemia sub-types. There is a partition into 6 classes and we take that as the gold solution.

Novartis It is a 103×1000 data matrix, where each row corresponds to a tissue sample and each column to a gene. The dataset comes from the study of Su et al. (2002) on four distinct cancer types. There is a partition into four classes and we take that as the gold solution.

Normal tissue It is a 90×1277 data matrix, where each row corresponds to a tissue sample and each column to a gene. The dataset comes from the study of Su et al. (2002) on the four distinct cancer types. There is a partition into four classes and we take that as the gold solution.

Gaussian3 It is a 60×600 data matrix. It is generated by having 200 distinctive features out of the 600 assigned to each cluster. There is a partition into three classes and that is taken as the gold solution. The data simulates a pattern whereby a distinct set of 200 genes is up-regulated in one of the three clusters, and down-regulated in the remaining two.

Gaussian5 It is a 500×2 data matrix. It represents the union of observations from 5 bivariate Gaussians, 4 of which are centered at the corners of the square of side length λ, with the 5th Gaussian centered at $(\lambda/2, \lambda/2)$. A total of 250 samples, 50 per class, were generated, where two values of λ are used, namely, $\lambda = 2$ and $\lambda = 3$, to investigate different levels of overlapping between clusters. There is a partition into five classes and that is taken as the gold solution.

Simulated6 It is a 60×600 data matrix. It consists of a 600-gene by 60-sample dataset. It can be partitioned into 6 classes with 8, 12, 10, 15, 5, and 10 samples respectively, each marked by 50 distinct genes uniquely up-regulated for that class. In addition, a list of 300 noise genes (i.e., genes having the same distribution within all clusters) are included. In particular, such genes are generated with decreasing differential expression and increasing variation, following the same distribution. Finally, the first block of 50 genes of the list is assigned to cluster 1, the second block to cluster 2 and so on. This partition into 6 classes is taken as the gold solution.

5 Results and discussion

The plots of the Adjusted Rand Index values, shown in Fig. S1, indicate the K-means algorithm with its different initializations as one of the best performers on all the considered datasets, since it exhibits a maximum for a k very close to the gold solution. Since K-means works very well when the structure of the clusters is spherical, we have taken the outcome of the mentioned experiment as an indication that a spherical Gaussian model would be appropriate for our datasets. Therefore, we have used it for all the Bayesian model selection methods considered here. This means that the likelihood is computed as in formula 7. However, in order to have robustness in our findings, we have also evaluated the Bayesians methods in conjunction with all clustering algorithms considered in this study.

We report here the comparison between the predictive power of the three Bayesian methods `BIC`, `AIC` and `MML` and that of the data-driven methods. To this end, we have computed the "optimal" number k^* of clusters in the case of the Bayesian methods, and used the benchmarking results reported in Giancarlo et al. (2008a) and Giancarlo and Utro (2011) in regard to the data-driven methods. Moreover, we also discuss the assessment of the Bayesian methods in absolute terms, i.e, by themselves. As discussed in Sect. 3, for most measures, the prediction of k^* is based on the visual inspection of curves and histograms. Here we limit ourselves to produce summary tables, based on our analysis of the relevant curves and experiments.

5.1 The performance of Bayesian methods

Each of the Bayesian methods has been computed for each dataset and each clustering algorithm. Each such a computation involves the estimation of a covariance matrix, which suffers from the dimensionality of the dataset. Indeed, in cases where a dataset has a number of features m much greater than the number of items n, the following problem comes up. Since the rank of the dataset matrix X is $r \leq min(m, n)$, there exist at least $m - n$ columns that are linear combination of at most n columns of X. The presence of linearly dependent columns in X implies the presence of linearly dependent rows in the $m \times m$ covariance matrix of X, making impossible the computation of its inverse and thus making impossible the computation of the multivariate normal distribution. In this case, the reduction of the data dimension by principal component analysis (Jain et al. 1999) can help because, if we consider the list of principal components ordered in decreasing order with respect to their corresponding variances, and we select the first ones of them which explain the 100 % of the total variance, the eigenvectors associated to these selected principal components will define a projection Y of X into an (at least) n-dimensional space where the computed covariance will be invertible. Since most of the datasets used here have the mentioned dimensionality problem, we have used the data reduction technique just outlined in order to circumvent it.

Tables 1, 2, and 3 summarize the results obtained by the Bayesian methods. Columns indicate the datasets, rows indicate the used measure in conjunction with a specific clustering algorithm. Each cell in a table displays a precision result, a number in a circle with a black background indicates a prediction in agreement with the number of classes in the dataset, while a number in a circle or a square with a white background indicates a prediction that differs, in absolute value, by 1 or 2 from the number of classes in the dataset respectively; a number not in a circle indicates the remaining predictions. Based on these results, Bayesian methods are able to provide the correct number of clusters only on very few datasets. In particular, `AIC` seems to be the best between the three Bayesian methods considered in this paper. `BIC` (see Table 1) is able to correctly estimate k^* for $7/8$ of the clustering algorithms on Gaussian5 and a close estimation is obtained for Normal Tissue. However, only in conjunction with Hier-A, it is able to estimate k^* for Gaussian3 and with K-means-R on NCI60. It is worth pointing out that Hier-A, K-means-A and K-means-C provide the "best" performance across the several datasets. Table 2 shows the results for `AIC`, the perfomance for Gaussian3 are the same as `BIC`, while in conjunction with K-means-R `AIC` is able to correctly estimate for CNS Rat. Based on the results on Table 2, it is worth pointing out

Table 1 A summary of the results for BIC on all algorithms and on all datasets

	CNS Rat	Leukemia	NCI60	Lymphoma	Yeast	Novartis	St.Jude	Normal	Gaussian3	Gaussian5	Simulated6
Hier-A	4	4	7	6	3	2	2	5	3	5	7
Hier-C	2	4	11	14	20	6	5	11	4	5	10
Hier-S	1	1	1	1	4	1	1	3	1	1	7
K-means-R	2	7	8	12	20	6	10	11	6	5	6
K-means-A	3	7	7	14	20	11	7	11	5	5	7
K-means-C	4	9	7	14	20	11	7	11	4	5	7
K-means-S	4	7	7	14	20	11	7	11	5	5	7
Gold solution	**6**	**3**	**8**	**3**	**5**	**4**	**6**	**13**	**3**	**5**	**6**

Each cell in a table displays a precision result,a number in a circle with a black background indicates a prediction in agreement with the number of classes in the dataset,while a number in a circle or a square with a white background indicates a prediction that differs, in absolute value, by 1 or 2 from the number of classes in the dataset respectively; a number not in a circle indicates the remaining predictions

Table 2 A summary of the results for AIC on all algorithms and on all datasets

	CNS Rat	Leukemia	NCI60	Lymphoma	Yeast	Novartis	St.Jude	Normal	Gaussian3	Gaussian5	Simulated6
Hier-A	4	4	7	6	4	2	2	5	3	5	7
Hier-C	5	4	11	14	20	6	5	11	4	5	10
Hier-S	1	1	7	1	1	1	1	3	1	1	7
K-means-R	6	7	10	14	20	6	10	11	6	5	6
K-means-A	7	7	7	14	20	11	7	11	5	5	7
K-means-C	8	9	7	14	20	11	7	11	4	5	7
K-means-S	7	7	7	14	20	11	7	11	5	5	7
Gold solution	**6**	**3**	**8**	**3**	**5**	**4**	**6**	**13**	**3**	**5**	**6**

Each cell in a table displays a precision result,a number in a circle with a black background indicates a prediction in agreement with the number of classes in the dataset,while a number in a circle or a square with a white background indicates a prediction that differs, in absolute value, by 1 or 2 from the number of classes in the dataset respectively; a number not in a circle indicates the remaining predictions

Table 3 A summary of the results for MML on all algorithms and on all datasets

	CNS Rat	Leukemia	NCI60	Lymphoma	Yeast	Novartis	St.Jude	Normal	Gaussian3	Gaussian5	Simulated6
Hier-A	10	17	15	8	15	17	11	14	13	15	17
Hier-C	3	11	13	14	14	5	15	16	13	3	14
Hier-S	8	14	17	8	12	14	14	14	14	2	15
K-means-R	5	4	12	10	14	2	14	7	8	4	4
K-means-A	5	16	16	15	16	6	5	6	10	3	5
K-means-C	6	5	16	15	16	6	5	6	4	1	5
K-means-S	6	17	16	15	16	6	5	6	3	3	5
Gold solution	**6**	**3**	**8**	**3**	**5**	**4**	**6**	**13**	**3**	**5**	**6**

Each cell in a table displays a precision result,a number in a circle with a black background indicates a prediction in agreement with the number of classes in the dataset,while a number in a circle or a square with a white background indicates a prediction that differs, in absolute value, by 1 or 2 from the number of classes in the dataset respectively; a number not in a circle indicates the remaining predictions

that Hier-A, Hier-C and K-means-R provide the "best" performance across several datasets. Finally, Table 2 shows the results for MML, which is able to estimate correctly Gaussian3 and CNS Rat only in conjunction with Kmeans-S.

5.2 Bayesian versus data-driven methods

In order to compare Bayesian with the data-driven methods, we take as reference the benchmarking results reported in Giancarlo et al. (2008a) and Giancarlo and Utro (2011) in regard to the data-driven methods. The results are summarized in Table 4. Due to its lack of precision, we do not consider MML in this comparison. It is evident that even the remaining Bayesian methods (i.e. BIC and AIC) are not able to compete in terms of precision with the data-driven ones.

6 Conclusion

We have presented a study about model selection methodologies in the specific contest of microarray data clustering. In particular, our intention was to compare several new state of the art data-driven model selection methodologies with respect to the Bayesian inspired ones, since the latter represent a general and classical way of performing model selection. Among the Bayesian methods, we have used BIC, AIC and MML because of their wide usage in the field of clustering. Experiments have been

Table 4 Summary of results for the best and on all datasets

	CNS Rat	Leukemia	NCI60	Lymphoma	Yeast	Novartis	St.Jude	Normal	Gaussian3	Gaussian5	Simulated6
WCSS-K-means-C	⑤	❸	❽	8	④	⑤	❻	9	❸	④	⑦
WCSS-R-R0	⑤	④	❽	❸	④	⑤	❻	9	-	❸	⑦
G-Gap-K-means-R	⑦	❸	4	④	⑥	③	4	7	❸	8	3
G-Gap-R-R5	⑤	④	2	②	④	⑤	⑦	7	❸	7	3
FOM-K-means-C	⑦	8	❽	④	④	❹	⑤	6	❸	-	⑤
FOM-K-means-S	❻	❸	❽	8	④	❹	⑦	4	-	-	4
FOM-R-R5	❻	❸	⑦	5	❺	7	-	10	❸	-	⑦
FOM-Hier-A	⑦	❸	⑦	6	⑥	❹	8	6	❸	-	⑤
DIFF-FOM-K-means-C	⑦	❸	⑦	④	3	③	3	7	❸	29	3
FC-Hier-A	⑦	❸	❽	❸	❺	⑤ - 6	❻	10	❸	❺	⑤
FC-Hier-C	❻	④	❽	5	⑥	❹ - ⑤	⑤ - ❻	10	❸	❺	⑤
Consensus-Hier-A	⑦	❸	❽	❸	❺	⑤ - 6	❻	10	❸	❺	⑤
Consensus-Hier-C	❻	④	❽	5	⑥	❹ - ⑤	⑤ - ❻	10	❸	❺	⑤
BIC-Hier-A	4	④	⑦	6	3	2	2	5	❸	❺	⑦
BIC-K-means-C	4	9	⑦	14	20	11	⑦	11	④	❺	⑦
BIC-K-means-S	4	7	⑦	14	20	11	⑦	11	5	❺	⑦
AIC-Hier-A	4	④	⑦	6	④	2	2	5	❸	❺	⑦
AIC-Hier-C	⑤	④	11	14	20	6	⑤	11	④	❺	10
AIC-K-means-R	❻	7	10	14	20	6	10	11	6	❺	❻
AIC-K-means-A	⑦	7	⑦	14	20	11	⑦	11	5	❺	⑦
AIC-K-means-C	8	9	⑦	14	20	11	⑦	11	④	❺	⑦
AIC-K-means-S	⑦	7	⑦	14	20	11	⑦	11	5	❺	⑦
Gold solution	**6**	**3**	**8**	**3**	**5**	**4**	**6**	**13**	**3**	**5**	**6**

Each cell in a table displays the ability to estimate the correct number of clusters in a dataset. An entry containing a dash only indicates that either the experiment was stopped because ofits high computational demand or that no useful indication was given by the method

carried out on 11 benchmark microarray datasets, and the Bayesian methods have been compared with 9 state of the art data-driven model selection methods. Results show the merit of Bayesian methods only in some cases, suggesting that they do not seem to be able to compete in terms of precision with the data-driven methods.

Funding Giosué Lo Bosco and Raffaele Giancarlo were supported by Progetto di Ateneo dell'Universitá degli Studi di Palermo 2012-ATE-0298 *Metodi Formali e Algoritmici per la Bioinformatica su Scala Genomica*.

References

Akaike H (1978) A new look at the statistical model identification. IEEE Trans Autom Control 9(6):716–723

Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson JJ, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403:503–511

Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine A (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750

Andreopoulos B, An A, Wang X, Schroeder M (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 10(3):297–314

Ben-Hur A, Elisseeff A, Guyon I (2002) A stability based method for discovering structure in clustering data. In: Seventh pacific symposium on biocomputing, ISCB, pp 6–17

Bouguila N, Ziou D (2007) High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Trans Pattern Anal Mach Intell 29(10):1716–1731

Breckenridge J (1989) Replicating cluster analysis: method, consistency, and validity. Multivar Behav Res 24(2):147–161

D'haeseleer P (2006) How does gene expression cluster work? Nat Biotechnol 23:1499–1501

Di Gesú V, Giancarlo R, Lo Bosco G, Raimondi A, Scaturro D (2005) A genetic algorithm for clustering gene expression data. BMC Bioinform 6:289

Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3:1–21

Everitt B (1993) Cluster analysis. Edward Arnold, London

Figuereido MAT, Jain AK (2002) Unsupervised learning of fInite mixture models. IEEE Trans Pattern Anal Mach Intell 24(3):381–396

Fowlkes E, Mallows C (1983) A method for comparing two hierarchical clusterings. J Am Stat Assoc 78:553–584

Giancarlo R, Utro F (2011) Speeding up the consensus clustering methodology for microarray data analysis. Algorithms Mol Biol 6:1

Giancarlo R, Utro F (2012a) Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis. Theor Comput Sci 428:58–79

Giancarlo R, Utro F (2012b) Stability-based model selection for high throughput genomic data: an algorithmic paradigm. In: Artificial immune systems. Lecture notes in computer science, vol 7597, pp 260–270

Giancarlo R, Scaturro D, Utro F (2008a) Computational cluster validation for microarray data analysis: experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer. BMC Bioinform 9:462

Giancarlo R, Scaturro D, Utro F (2008b) A tutorial on computational cluster analysis with applications to pattern discovery in microarray data. Math Comput Sci 1:655–672

Giancarlo R, Scaturro D, Utro F (2009) Statistical indices for computational and data driven class discovery in microarray data. In: Chen JY, Lonardi S (eds) Biological data mining. CRC Press, San Francisco, pp 295–335

Giancarlo R, Lo Bosco G, Pinello L (2010) Distance functions, clustering algorithms and microarray data analysis. In: Learning and intelligent optimization. Lecture notes in computer science, pp 125–138

Giancarlo R, Lo Bosco G, Pinello P, Utro F (2011) The three steps of clustering in the post-genomic Era. In: Computational intelligence methods for bioinformatics and biostatistics. Lecture notes in computer science, pp 13–30

Giancarlo R, Lo Bosco G, Pinello L, Utro F (2013) A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. BMC Bioinform 14:S6

Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeeck M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(531):531–537

Handl J, Knowles J, Kell D (2005) Computational cluster validation in post-genomic data analysis. Bioinformatics 21(15):3201–3212

Hartigan J (1975) Clustering algorithms. Wiley, New York

Hastie T, Tibshirani R, Friedman J (2003) The elements of statistical learning. Springer, Heidelberg

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Jain A, Dubes R (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs

Jain A, Murty M, Flynn P (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Klie S, Nikoloski Z, Selbig J (2010) Biological cluster evaluation for gene function prediction. J Comput Biol 17:1–18

Krzanowski W, Lai Y (1985) A criterion for determining the number of groups in a dataset using sum of squares clustering. Biometrics 44:23–34

Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Kluwer Academic Publishers, Norwell

Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52:91–118

NCI 60 Cancer Microarray Project (2008) http://genome-www.stanford.edu/NCI60

Pelleg D, Moore A (2000) X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning, Morgan Kaufmann, San Francisco, pp 727–734

Perou C, Jeffrey S, van de Rijn M, Rees C, Eisen M, Ross D, Pergamenschikov A, Williams C, Zhu S, Lee J, Lashkari D, Shalon D, Brown P, Botstein D (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 96:9212–9217

Pollack J, Perou C, Alizadeh A, Eisen M, amd CF, Williams AP, Jeffrey S, Botstein D, Brown P (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41–46

Priness I, Maimon O, Ben-Gal I (2007) Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinform 8:111

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32:496–501

Rijsbergen CV (1979) Information retrieval, 2nd edn. Butterworths, London

Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–471

Ross D, Scherf U, Eisen M, Perou C, Spellman P, Iyer V, Jeffrey S, van de Rijn M, Walthama M, Pergamenschikov A, Lee J, Lashkari D, Shalon D, Myers T, Weistein J, Botstein D, Brown P (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227–235

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464. doi:10.2307/2958889

Shamir R, Sharan R (2003) Algorithmic approaches to clustering gene expression data. In: Jiang T, Smith T, Xu Y, Zhang MQ (eds) Current topics in computational biology. MIT Press, Cambridge, pp 120–161

Spellman P, Sherlock G, Zhang M, Iyer VR, Anders K, Eisen M, Brown P, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 9:3273–3297

Su A, Cooke M, Ching K, Hakak Y, Walker J, Wiltshire T, Orth A, Vega R, Sapinoso L, Moqrich A, Patapoutian A, Hampton G, Schultz P, Hogenesch J (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465–4470

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the gap statistics. J R Stat Soc B 2:411–423

Wallace CS, Boulton DM (1968) An information measure for classification. Comput J 11(2):185–194

Wallace CS, Dowe DL (2000) MML clustering of multi-state, poisson, von mises circular and Gaussian distributions. Stat Comput 10(1):73–83

Wen X, Fuhrman S, Michaels G, Carr D, Smith S, Barker J, Somogyi R (1998) Large scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95:334–339

Yeoh EJ, Ross M, Shurtleff S, Williams W, Patel D, Mahfouz R, Behm F, Raimondi S, Relling M, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans W, Naeve C, Wong L, Downing J (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1:133–143

Yeung K, Haynor D, Ruzzo W (2001) Validating clustering for gene expression data. Bioinformatics 17:309–318