

# Chapter 3

## Mixture Models and Clustering

In this thesis, we focus on the use of mixture models to perform clustering. By clustering we mean finding groups (or clusters) of observations in a finite data set, such that each group represents observations sharing characteristics, which are distinct from the overall data set. Mixture models, being based on statistics, tackle this problem in a formal and principled way. The applications of mixture models for clustering [79] has a number of advantages in contrast to classical clustering methods such as  $k$ -means and hierarchical clustering: it quantifies the uncertainty of a given cluster assignment; the estimated models are descriptors of the groups found; and it is possible to answer questions such as the number of clusters in a purely statistical way [145].

The characteristics of mixture models make this approach of great value in the analysis of biological data. In particular, for gene expression analysis [11], the main interest is on finding groups of genes that have similar expression patterns through a set of experimental conditions, and possibly are part of a biological functional module [71, 135]. However, a single gene (and its products) can participate simultaneously in more than one functional module, for example by taking part in distinct protein complexes, each with its particular function [122]. Furthermore, data arising from large-scale experiments, such as microarray measurements of gene expression, contain large amount of noise [135]. In this context, overlapping clusterings, such as the ones given by a mixture models, represent the results of gene expression clustering analysis in a more natural way than “hard” clusterings. Also, the uncertainty of a given cluster assignment returned by the mixture model is a valuable information in the distinction of assignments derived from relevant and noisy observations. Furthermore, the mixture components can model particular assumptions about the data, such as temporal dependencies, therefore producing more reliable estimates. As an evidence, there is a vast list of publications that successfully applied mixture models in finding potentially overlapping groups in gene expression analysis [14, 138, 143, 147, 155, 185–187, 232, 234].

This chapter is organized as follows: Section 3.1 gives a definition of the use of mixture models to perform clustering, and introduces how clusters can be obtained from mixtures. Later, we propose a novel external index to perform validation of mixture models in Section 3.2, which is evaluated with simulated data in Section 3.3.

### 3.1 Clustering with Mixture Models

Clustering is the task of partitioning a data set of  $N$  objects (or observations) from  $\mathbf{X}$  into  $K$  disjoint groups (or clusters). We represent a data set by a  $N \times L$  matrix  $\mathbf{X}$ , where entry  $x_{ij}$  denotes the values of the  $j$ th variable of the  $i$ th object. The clustering (or partition) can be represented by  $\mathbf{Y}$ , where  $y_i \in \{1, \dots, K\}$  indicates the group to which a given object  $x_i$  belongs to [142]. In mixture model based clustering, we assume that each component in the mixture represents a group of objects. In other words, the density of the  $k$ th component can be interpreted as the conditional of  $x_i$  on  $y_i$ , i.e.,  $p_k(x_i|\theta_k) = p(x_i|y_i = k, \theta_k)$ , and the mixing coefficient as the prior probability of the component, i.e.,  $\alpha_k = p(y_i = k)$ .

For a given data set  $\mathbf{X}$  with  $N$  observations and a variable  $Y$  defining the component assignments, the likelihood of the complete data assuming that the  $x_i$  are independently distributed is given by

$$p(\mathbf{X}, \mathbf{Y}|\Theta) = \mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y}) = \prod_{k=1}^K \prod_{i=1}^N (\alpha_k \cdot p_k(x_i|\theta_k))^{\mathbf{1}(y_i=k)} \quad (3.1)$$

Thus, the problem of clustering observations from  $\mathbf{X}$  can be formulated as finding the maximum likelihood estimate (MLE)

$$\Theta^* = \arg \max_{\Theta} \mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y}). \quad (3.2)$$

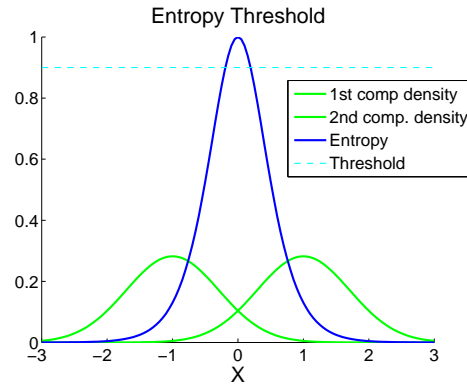
This problem, as described in Section 2.3, can be solved by maximizing the complete likelihood using the EM algorithm [61].

The posterior probability (Eq. 2.16) of a mixture model reveals the probability of a cluster assignment. The simplest way of decoding a mixture, that is, to infer clusters in the data, is to interpret the mixture components as descriptive models of non-overlapping clusters and assign each object  $x_i$  to the cluster  $k$  of maximal posterior,

$$y_i = \arg \max_{1 \leq k \leq K} (r_{ik}). \quad (3.3)$$

In model-based clustering as well as  $k$ -means, these hard assignments are performed after each E-Step, while for the mixtures this is only necessary after estimation is finished. Indeed, a mixture of Gaussians with the identity covariance described in Section 2.3.3, where  $\sigma^2 \rightarrow 0$  and hard assignments are performed, is equivalent to the  $k$ -means algorithm [88]. In the next chapters, we will refer to model-based clustering whenever such hard assignments are performed during the EM, and to mixture estimation otherwise.

An inspection of the distribution of the posterior probability of component assignments given an object  $x_i$ , i.e.,  $r_i = (r_{i1}, \dots, r_{iK})$ , reveals the level of ambiguity in making the cluster assignments. Therefore, we propose here a novel decoding method, entropy thresholding, which takes the ambiguity of assignments into account. As depicted in Figure 3.1,



**Figure 3.1:** Entropy of posterior assignments for the bimodal density from Figure 2.1. Values in between the two density functions have high entropy, and would be discarded by the entropy threshold method. If we select  $\varphi = 0.9$ , objects in the range  $[-0.2, 0.2]$  will be assigned to the cluster  $K + 1$ .

this ambiguity can be quantified by computing the Shannon entropy [54]

$$H(r_i) = - \sum_{k=1}^K r_{ik} \log \frac{1}{r_{ik}}. \quad (3.4)$$

Choosing a threshold  $\varphi$  for the entropy yields a grouping of the data into at most  $K + 1$  groups. If  $H(r_i) < \varphi$ , we assign  $x_i$  to the component with maximal posterior as in Eq. 3.3. Otherwise,  $x_i$  is assigned to the  $(K + 1)$ -st group, which contains all objects which cannot be assigned unambiguously, that is

$$y_i = \begin{cases} \arg \max_{1 \leq k \leq K} (r_{ik}), & H(r_i) < \varphi \\ K + 1, & \text{otherwise.} \end{cases} \quad (3.5)$$

## 3.2 Validation of Mixture Models

The task of obtaining a mixture model does not end with the parameter estimation. Questions on the number of components and the quality of the representation of the data often arise after this step. In classical clustering, there are several methodologies, under the name of cluster validation, proposed for answering these questions. These methodologies, mainly based on re-sampling techniques and fit indices, have been proposed in the vast cluster validation literature [28, 56, 66, 124, 141, 233] and reviewed for example in [89, 109]. Nevertheless, the mixture model framework embraces challenges and characteristics not explored by “classical” cluster validation techniques.

One often over-looked aspect is the use of external indices, which are used to compare the similarity of a cluster solution to a gold standard or to another clustering solution.

Most external indices proposed so far are only able to measure the agreement between two non-overlapping clusterings [109]. Mixtures, however, can be interpreted as a partition with overlap, and encode more information than non-overlapping partitions. Therefore, the overlap, encoded by the posterior distributions of the mixture, should be taken into consideration when for example two mixtures are compared. Additionally, there are cases, where even if the clustering results are non-overlapping partitions, the a priori labels are based on overlapping partitions. This is the case, for example, of functional annotation of genes [9].

Motivated by the previous problem, in the next Section, as one of the contribution of this thesis, we propose a novel external index that can be used for the comparison of mixture models and overlapping partitions [51]. Such an index is an extension of a widely employed external index for comparing hard partitions — the corrected Rand index [103].

In Section 3.2.1, we introduce the basics of external indices of non-overlapping partitions, and we define the extension for the overlapping case. Finally, in Section 3.3, we employ simulated data for assessing the characteristics of the external indices in data with overlap.

### 3.2.1 External Indices

External indices assess the agreement between two partitions defined over the same set of objects, where one partition  $\mathbf{Y}$  represents the result of a clustering method, and the other partition  $\mathbf{Y}'$  represents class labels<sup>1</sup>. While a number of external indices have been introduced in the literature, the use of corrected Rand (CR) is recommended [103]. CR has its value corrected for chance agreement, it is not dependent on the cluster size distributions and can compare partitions with distinct number of clusters [149]. See [110] for a comprehensive review of external indices.

Let  $\mathbf{Y}$  and  $\mathbf{Y}'$  be discrete vectors representing the partitions yielded by a clustering method and the class labels. Let  $y_i \in \{1, \dots, K\}$  and  $y'_i \in \{1, \dots, L\}$  be, respectively, observations from  $\mathbf{Y}$  and  $\mathbf{Y}'$ , where  $y_i = k$  indicates that object  $i$  belongs to cluster  $k$ . Note,  $K$  and  $L$  can be distinct. Thus, the following indicator functions can be defined

$$\mathbf{1}(y_i = y_j) = \begin{cases} 1, & \text{if } y_i = y_j \\ 0, & \text{otherwise} \end{cases} \quad \text{and,} \quad (3.6)$$

$$\mathbf{1}(y'_i = y'_j) = \begin{cases} 1, & \text{if } y'_i = y'_j \\ 0, & \text{otherwise.} \end{cases} \quad (3.7)$$

From these, we can define the following terms

---

<sup>1</sup> $\mathbf{Y}$  and  $\mathbf{Y}'$  can also be partitions from two distinct clustering methods applied to the same data set.

$$a = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathbf{1}(y_i = y_j) \mathbf{1}(y'_i = y'_j), \quad (3.8)$$

$$b = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (1 - \mathbf{1}(y_i = y_j)) \mathbf{1}(y'_i = y'_j), \quad (3.9)$$

$$c = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathbf{1}(y_i = y_j) (1 - \mathbf{1}(y'_i = y'_j)), \text{ and} \quad (3.10)$$

$$d = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (1 - \mathbf{1}(y_i = y_j)) (1 - \mathbf{1}(y'_i = y'_j)). \quad (3.11)$$

The term  $a$  measures the number of object pairs that are found in the same cluster in both  $\mathbf{Y}$  and  $\mathbf{Y}'$ . It is the equivalent of the number of true positives commonly used in the machine learning literature. Analogously,  $b$ ,  $c$  and  $d$  correspond respectively to the number of false positives, false negatives and true negatives. The total number of object pairs  $p$  is equal to  $p = a + b + c + d$ . From these terms, the corrected Rand is defined as [103],

$$\text{CR} = \frac{(a + d) - ((a + b)(a + c) + (c + d)(b + d))p^{-1}}{p - ((a + b)(a + c) + (c + d)(b + d))p^{-1}}. \quad (3.12)$$

CR takes values from -1 to 1, where 1 represents perfect agreement while values of CR near or below 0 represent agreements occurring by chance. The correction of Rand index, proposed in [103], estimates the expected Rand index value by assuming that the baseline distributions of the partitions are fixed. This is equivalent to calculating the expected Rand index value for random permutations of the objects labels in one partition, while the other is fixed.

Two other interesting external indices, which can be defined by the terms in Eq. 3.8, 3.9, 3.10 and 3.11, are the sensitivity and specificity [199],

$$\text{Sens} = \frac{a}{a + c} \quad (3.13)$$

$$\text{Spec} = \frac{a}{a + b} \quad (3.14)$$

They both take values from 0 to 1, where 1 indicates perfect agreement. The use of these indices is complementary to CR, as they indicate for example a tendency to make more false positives or false negative errors — CR treats both errors equally. In practice, a lower sensitivity (more false positives) is an indicator of joining real clusters; while a lower specificity (more false negatives) indicates a tendency to split real clusters.

### Extended Corrected Rand

The main idea of the extended corrected Rand (ECR) is to redefine the indicator functions, as defined in Eq. 3.6 and Eq. 3.7, giving them a probabilistic interpretation [51]. The posterior distribution defines the probability that a given object  $x_i$  from  $\mathbf{X}$  belongs to the component  $k$ , i.e.,  $y_i = k$ , in a mixture model parameterized by  $\Theta$ , i.e.  $p(y_i = k|x_i, \Theta)$ . This is exactly the Eq. 2.16, which we refer to as  $r_{ik}$  for simplicity. Likewise, we have  $r'_{il}$  for indicating the posterior that  $x_i$  belongs to component  $l$  in  $\mathbf{Y}'$ . We denote the event that a pair of objects has been generated by the same component in  $\mathbf{Y}$ , the co-occurrence event, as  $x_i \equiv x_j$  given  $\mathbf{Y}$ . Assuming independence of the clusters from  $\mathbf{Y}$ , the probability of the co-occurrence of  $x_i$  and  $x_j$  given  $\mathbf{Y}$  for  $1 \leq i \leq j \leq N$  can be estimated as

$$p(y_i \equiv y_j \text{ given } \mathbf{Y}) = \sum_{k=1}^K r_{ik}r_{jk}. \quad (3.15)$$

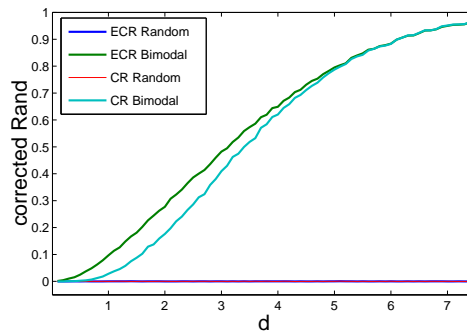
We use the previous equation to redefine the variables  $a$ ,  $b$ ,  $c$  and  $d$ , used in the definition of CR

$$\begin{aligned} a &= \sum_{i=1}^{N-1} \sum_{j=i+1}^N p(y_i \equiv y_j \text{ given } \mathbf{Y})p(y_i \equiv y_j \text{ given } \mathbf{Y}'), \\ b &= \sum_{i=1}^{N-1} \sum_{j=i+1}^N (1 - p(y_i \equiv y_j \text{ given } \mathbf{Y}))p(y_i \equiv y_j \text{ given } \mathbf{Y}'), \\ c &= \sum_{i=1}^{N-1} \sum_{j=i+1}^N p(y_i \equiv y_j \text{ given } \mathbf{Y})(1 - p(y_i \equiv y_j \text{ given } \mathbf{Y}')), \text{ and} \\ d &= \sum_{i=1}^{N-1} \sum_{j=i+1}^N (1 - p(y_i \equiv y_j \text{ given } \mathbf{Y}))(1 - p(y_i \equiv y_j \text{ given } \mathbf{Y}')). \end{aligned} \quad (3.16)$$

From these, the extended corrected Rand (ECR) can be computed by the original formula Eq. 3.12. ECR also takes values from -1 to 1, where 1 represents perfect agreement while values of ECR near or below zero represent agreements occurred by chance. By definition, it works exactly as the corrected Rand when “hard” partitions are given.

## 3.3 Experiments

To evaluate the extended corrected Rand, we make use of simulated data from mixtures of Gaussians. In the first experiment, we define a very simple scenario with a mixture of two Gaussians components in an univariate space. Hence, we can compare the characteristics



**Figure 3.2:** We depict the mean CR and ECR for the results of the mixture estimation with the normal bimodal density. The larger  $d$ , the lower is the overlap between the two components.

of ECR and CR when distinct degrees of overlap between the components are present. In the second experiment, we sample data from a mixture, where components show a large degree of overlap. From an initial mixture, we vary the number of components and distributions of objects in the components.

The EM algorithm is used to fit the multivariate Gaussian mixtures with full covariance matrices as described in Section 2.3.3. The EM method is initialized as described in Section 2.3.2. In the simulated data experiments, 50 data sets are generated for each proposed mixture.

### 3.3.1 Simulated Data 1

We perform experiments with a normal mixture with two equiprobable components to evaluate the proposed index characteristics in the presence of distinct degrees of overlap. The components have means  $\mu_1 = [0, 0]^T$ ,  $\mu_2 = [d, 0]^T$ , and covariance matrices  $\Sigma_1 = \Sigma_2 = I$ , as suggested in [77]. For obtaining mixtures with distinct degrees of overlap (bimodal data), we vary  $d$  in the range  $[0.0, 7.5]$ . The lower the value  $d$ , the higher is the overlap between the two components. For each component we draw 200 objects. The density function given the original mixture parameterization is used to obtain the posterior  $r'_{il}$ . We also calculate the values of CR after performing hard assignments of the solutions (Eq. 3.3).

Additionally, we generate random data to function as a null case. This consists of data generated from a single normal component with  $\mu = [d/2, 0]^T$  and  $\Sigma = I$ . A random solution ( $Y'$ ) with the same number of components and object distributions as the corresponding bimodal data is calculated. For each particular  $d$ , we carried out a non parametric equal-means hypothesis test based on bootstrap [70] to compare the mean ECR (or CR) obtained with the bimodal and random data.

**Results.** As displayed in Figure 3.2, for data with high overlap, ECR has higher values than CR, while for data with low overlap both indices have similar values. With random data, the indices take on mean values near zero and low variance ( $< 0.001$ ), which indicate that ECR is successful in the correction for randomness.

With respect to the hypothesis test, the equal means hypothesis is rejected with the use of ECR in all  $d > 0.0$  with  $p$ -value  $< 0.001$ . On the other hand, with the use of CR, the null hypothesis (equal means) is only rejected ( $p$ -value  $< 0.001$ ) when low overlap is presented ( $d > 0.4$ ). We can conclude that ECR is able to detect the distinction between the agreement of the random and bimodal data in all cases, while CR fails when a high degree of overlap is present. Furthermore, when overlap is low, both indices behave similarly.

### 3.3.2 Simulated Data 2

We use a more extensive set of simulated data to evaluate ECR. Based on a mixture defined in [77], which will be called “base mixture”, we change and extend its definition to generate data with distinct components densities and number of components. The “base mixture” has four components, three of them with a large overlap and two of them with same mean vectors

$$\begin{aligned}\mu_1 &= [-4, -4], \Sigma_1 = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix}, \\ \mu_2 &= [-4, -4], \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \\ \mu_3 &= [-1, -6], \Sigma_3 = \begin{bmatrix} 0.125 & 0.0 \\ 0.0 & 0.125 \end{bmatrix}, \\ \mu_4 &= [2, 2], \Sigma_4 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.\end{aligned}$$

An example of a data set sampled from this mixture can be seen in Figure 3.3.

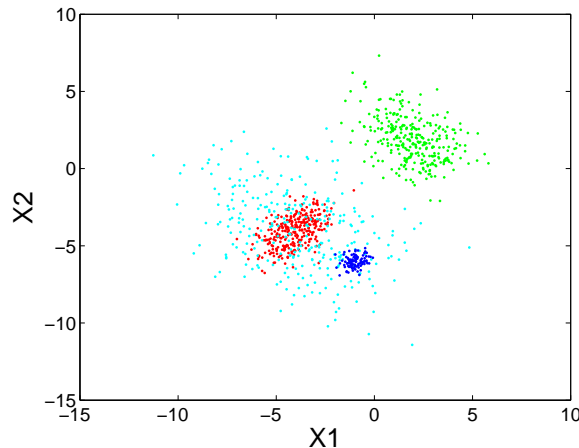
As with the bimodal data, we also generate random data from the normal

$$\mu = [0, 0], C = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

For each data set the EM is performed with 2 to 10 components. As in [149] it is expected that ECR should obtain mean values near zero for random data and low standard error. Additionally, ECR will be maximum at the correct number of components. For comparison, we also compute BIC and CR.

**Components Distribution.** We use three types of component distributions for the “base mixture”: equal density (ED), ( $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$ ), 10% density (10%) ( $\alpha_1 =$





**Figure 3.3:** We depict the points of data sampled from the base mixture in the two dimensional space. Objects in light blue corresponds to component 1, in red to component 2, in dark blue to component 3, and in green to component 4. As can be seen, components 1 and 2 have the same mean, but distinct orientations and sizes. Furthermore, component 3 is also inside component 1.

$\alpha_2 = \alpha_3 = 0.3$  and  $\alpha_4 = 0.1$ ) and 60% density (60%) ( $\alpha_1 = 0.6$  and  $\alpha_2 = \alpha_3 = \alpha_4 = 0.16$ ).

**Number of Components.** In addition to the components in the base mixture, we also included the components ( $\mu_5 = [-6, -1]^T$ ,  $\mu_6 = [-12, -12]^T$  and  $\Sigma_4 = \Sigma_5$ ,  $\Sigma_3 = \Sigma_6$ ). We generated data sets with two to six components with 700 observations. For each number of component  $K$ , we select the first  $K$  components from the mixture. The component distribution used is  $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.21$  and  $\alpha_5 = \alpha_6 = 0.07$ .

**Results** As can be observed in Table 3.1, CR and ECR indicate the right number of clusters (four) in all three scenarios of component distributions. Nevertheless, in the setting 10%, the equal means hypothesis test was not rejected, when comparing the mean of the CR with 4 and 5 components. In relation to BIC, it overestimates the number of cluster as five, in the distribution setting 10%, and as 8 in the distribution setting 60%.

In relation to the data with distinct number of components, ECR indicates the right number in all data sets, as shown in Table 3.2. CR overestimates the number of components of the data set with 5 and 6 components indicating 8 components in both cases. BIC can only correctly predict the number of clusters in the data with 2 and 3 components. Note that the degree of overlap varies in the models with distinct number of components, in special for the highly overlapping  $c = 2$  and  $c = 3$ . This makes the mixture estimation task harder, as a result, lower ECR (and CR) values are obtained in those data sets. Additionally, the estimated mixtures obtain mean ECR values near zero and low standard errors ( $< 0.001$ ) in all situations with the random data (not shown).

**Table 3.1:** We depict the mean values for data with distributions ED (top), 10% (middle) and 60% (bottom) against the number of components in the mixtures for corrected Rand, extended corrected Rand and BIC. For all indices, the maximum values (in bold) indicate the predicted number of components.

| no clusters | 2     | 3     | <b>4</b>    | 5            | 6     | 7     | 8            | 9     | 10    |
|-------------|-------|-------|-------------|--------------|-------|-------|--------------|-------|-------|
| ED CR       | 0.32  | 0.68  | <b>0.76</b> | 0.71         | 0.71  | 0.69  | 0.66         | 0.64  | 0.65  |
| ECR         | 0.31  | 0.64  | <b>0.68</b> | 0.63         | 0.63  | 0.60  | 0.57         | 0.54  | 0.55  |
| -BIC        | -8.95 | -8.20 | -8.08       | <b>-7.94</b> | -8.12 | -8.14 | -8.00        | -8.02 | -8.21 |
| 10% CR      | 0.34  | 0.50  | <b>0.66</b> | 0.66         | 0.64  | 0.59  | 0.57         | 0.58  | 0.58  |
| ECR         | 0.43  | 0.45  | <b>0.58</b> | 0.55         | 0.52  | 0.48  | 0.45         | 0.44  | 0.44  |
| -BIC        | -9.03 | -8.97 | -8.77       | <b>-8.72</b> | -8.74 | -8.76 | -8.78        | -8.80 | -8.80 |
| 60% CR      | 0.36  | 0.75  | <b>0.81</b> | 0.76         | 0.74  | 0.72  | 0.67         | 0.66  | 0.62  |
| ECR         | 0.35  | 0.64  | <b>0.71</b> | 0.65         | 0.61  | 0.55  | 0.50         | 0.44  | 0.41  |
| -BIC        | -8.16 | -7.70 | -7.45       | -7.31        | -7.48 | -7.35 | <b>-6.92</b> | -7.09 | -7.11 |

We analyzed how distinct criteria measure the agreement of two mixtures when data is generated from highly overlapping mixtures. The extended Corrected Rand displays better results than the corrected Rand in discriminating the right solutions in all scenarios. Furthermore, ECR behaves similarly to CR when no great overlap is present in the data, and in the correction for randomness. It is important to stress that CR and ECR do not substitute BIC for finding the right number of components, because they require the true labels (or true posteriors). These labels are often not present, and despite the sub-optimal results in this analysis, BIC works reasonably in practice. Nevertheless, if true labels and overlapping assignments are present, ECR is more precise.

In summary, this chapter covers the basic aspects of the use of mixture models to perform clustering. All results discussed here are based on the use of multivariate Gaussians as the components of the mixture. Nevertheless, for specific applications, one can take advantage of the characteristics of the data at hand, and choose the component models accordingly. The EM algorithm offers a flexible framework for such extensions. In practice, for a given model choice, one only needs to redefine the M-Step accordingly.

This thesis focuses on two types of components models for analyzing gene expression profiles. The use of HMMs to analyze gene expression time-courses will be the focus of Chapter 4. While in Chapter 5, we propose a new type of probabilistic model, dependence trees, to model gene expression profiles during a developmental process. Furthermore, once the M-Step for a given model is defined, one can straight-forwardly apply any other extensions of the EM algorithm. We explore, in Chapter 6, the use of semi-supervised extension of the EM to integrate additional data biological and improve clusterings of gene expression time-courses.

**Table 3.2:** We present the mean values for data with 2, 3, 4, 5 and 6 components (top to bottom) against the number of components of the estimated mixture for corrected Rand, extended corrected Rand and BIC . For all indices, the maximum values (in bold) indicate the predicted number of components and the line preceding the indices values states the correct number of components.

|              |              |              |             |              |             |       |              |       |              |
|--------------|--------------|--------------|-------------|--------------|-------------|-------|--------------|-------|--------------|
| no clusters  | <b>2</b>     | 3            | 4           | 5            | 6           | 7     | 8            | 9     | 10           |
| CR           | <b>0.37</b>  | 0.30         | 0.24        | 0.20         | 0.17        | 0.15  | 0.14         | 0.13  | 0.12         |
| ECR          | <b>0.21</b>  | 0.15         | 0.12        | 0.09         | 0.07        | 0.06  | 0.05         | 0.04  | 0.04         |
| -BIC         | <b>-4.74</b> | -4.75        | -4.77       | -4.79        | -4.81       | -4.84 | -4.86        | -4.88 | -4.91        |
| no clusters  | 2            | <b>3</b>     | 4           | 5            | 6           | 7     | 8            | 9     | 10           |
| CR           | 0.33         | <b>0.52</b>  | 0.48        | 0.43         | 0.40        | 0.36  | 0.33         | 0.30  | 0.30         |
| ECR          | 0.31         | <b>0.39</b>  | 0.35        | 0.30         | 0.27        | 0.23  | 0.21         | 0.19  | 0.18         |
| -BIC         | -5.74        | <b>-5.49</b> | -5.50       | -5.52        | -5.54       | -5.56 | -5.58        | -5.60 | -5.63        |
| no. clusters | 2            | 3            | <b>4</b>    | 5            | 6           | 7     | 8            | 9     | 10           |
| CR           | 0.34         | 0.50         | <b>0.66</b> | 0.66         | 0.64        | 0.59  | 0.57         | 0.58  | 0.58         |
| ECR          | 0.43         | 0.45         | <b>0.58</b> | 0.55         | 0.52        | 0.48  | 0.45         | 0.44  | 0.44         |
| -BIC         | -9.03        | -8.97        | -8.77       | <b>-8.72</b> | -8.74       | -8.76 | -8.78        | -8.80 | -8.80        |
| no. clusters | 2            | 3            | 4           | <b>5</b>     | 6           | 7     | 8            | 9     | 10           |
| CR           | 0.33         | 0.46         | 0.48        | 0.53         | 0.54        | 0.55  | <b>0.56</b>  | 0.55  | 0.56         |
| ECR          | 0.33         | 0.42         | 0.44        | <b>0.47</b>  | 0.478       | 0.46  | 0.46         | 0.44  | 0.44         |
| -BIC         | -1.00        | -0.99        | -0.98       | -0.97        | -0.97       | -0.97 | <b>-0.96</b> | -0.97 | -0.96        |
| no. clusters | 2            | 3            | 4           | 5            | <b>6</b>    | 7     | 8            | 9     | 10           |
| CR           | 0.21         | 0.44         | 0.48        | 0.54         | 0.56        | 0.56  | <b>0.57</b>  | 0.55  | 0.56         |
| ECR          | 0.18         | 0.44         | 0.48        | 0.51         | <b>0.52</b> | 0.52  | 0.52         | 0.50  | 0.50         |
| -BIC         | -7.34        | -6.79        | -6.74       | -6.71        | -6.71       | -6.70 | -6.70        | -6.68 | <b>-6.58</b> |

