

# Testing for Differences Among Discrete Distributions: An Application of Model-Based Clustering

Fernando A. Quintana and Andrés Silva \*

March, 2006

## Abstract

We consider the problem of testing for differences among a number of discrete distributions. Our approach is based on viewing the different samples as drawn from a mixture distribution where each mixture component represents a different group or cluster. We adapt the methodology of Dasgupta and Raftery (1998) to this problem, giving explicit details in the case of multinomial and first-order Markov chain distributions. The resulting method is computationally efficient and easy to implement. Applications to two datasets are discussed.

*Key words:* BIC, EM algorithm, mixture models.

---

\*Fernando Quintana is Associate Professor, Departamento de Estadística, Pontificia Universidad Católica de Chile (e-mail: quintana@mat.puc.cl). Andrés Silva is M. S. student, Departamento de Estadística, Pontificia Universidad Católica de Chile (e-mail: asilva@mat.puc.cl). This work was partially funded by Grant FONDECYT 1020712.

# 1 Introduction

Suppose we have collected samples from  $R$  independent populations and that the responses we record are categorical, with  $C \geq 2$  levels each. Let  $\mathbf{m}_i = (m_{i1}, \dots, m_{iC})$  denote the *counts* for the  $i$ th sample, i.e,  $m_{ij}$  represents the number of times the  $j$ th category was observed in the  $i$ th sample,  $i = 1, \dots, R$ ,  $j = 1, \dots, C$ . The data can be arranged in a single  $R \times C$  contingency table, where the  $R$  row margins  $n_i = \sum_{j=1}^C m_{ij}$  are fixed by design.

The above data structure is very common. This explains the abundant number of available references in the statistical literature. See, for instance, Bishop et al. (1975), Agresti (2002) and references therein.

Associated to each row, there is a discrete joint probability distribution concentrated on a subset of

$$\mathcal{M}_i = \{(s_{i1}, \dots, s_{iC}) : s_{ij} \in \{0, \dots, n_i\}, \sum_{j=1}^C s_{ij} = n_i\},$$

which may depend on some parameter vector  $\boldsymbol{\theta}_i \in \Theta$ . Although much of the later development is general, for clarity of presentation the discussion centers around two important cases:

- (i) *Multinomial sampling*: this is the usual setting where the  $n_i$  subjects in the  $i$ th sample are independently classified into one of the  $C$  categories. The likelihood is

$$f(\mathbf{m}_i | \boldsymbol{\theta}_i) = \binom{n_i}{m_{i1} \dots m_{iC}} \theta_{i1}^{m_{i1}} \dots \theta_{iC}^{m_{iC}}, \quad \mathbf{m}_i \in \mathcal{M}_i, \quad (1)$$

and where  $\boldsymbol{\theta}_i \in \Theta_i = \{(\theta_1, \dots, \theta_C) : 0 \leq \theta_i \leq 1, \sum_{i=1}^C \theta_i = 1\}$ .

- (ii) *First-order Markov chains*: assume a binary sequence of length  $l_i$  is available for each of the  $R$  subjects. The simplest model that incorporates serial correlation within sequences is a first-order Markov chain with transition matrix

$$\begin{pmatrix} \theta_{i1} & 1 - \theta_{i1} \\ \theta_{i2} & 1 - \theta_{i2} \end{pmatrix},$$

which can be represented as  $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})$ . Here,  $\theta_{i1}$  and  $\theta_{i2}$  represent the probabilities of a one-step transition from state 0 to 0, and from state 1 to 0, respectively. A sufficient statistic for first-order Markov chains is the vector of *transition counts*  $(t_{00}^i, t_{01}^i, t_{10}^i, t_{11}^i)$ , where  $t_{rs}^i$  is the number of times  $r$  was followed by  $s$  in the  $i$ th sequence, with  $r, s \in \{0, 1\}$ . Therefore, we assume the data  $\mathbf{m}_1, \dots, \mathbf{m}_R$  represent the transition counts for the  $R$  sequences, so that the likelihood is given by

$$f(\mathbf{m}_i | \boldsymbol{\theta}_i) = \binom{c_0^i - 1}{c_0^i - t_{00}^i - 1} \binom{c_1^i - 1}{c_1^i - t_{11}^i - 1} \theta_{i1}^{t_{00}^i} (1 - \theta_{i1})^{t_{01}^i} \theta_{i2}^{t_{10}^i} (1 - \theta_{i2})^{t_{11}^i}, \quad (2)$$

where  $C = 4$ ,  $c_s^i$  is the number of times  $s \in \{0, 1\}$  was observed in the sequence,  $\boldsymbol{\theta}_i \in \Theta_i = (0, 1)^2$ , and  $\mathbf{m}_i \in \mathcal{M}_i$  is further restricted to be a valid set of transition counts. See details in Quintana and Newton (1998).

Other examples, including Markov chains of higher orders, fit into the general structure represented by the  $R \times C$  contingency table.

In this work we concentrate on the problem of comparing the  $R$  discrete distributions, which includes as a particular sub-problem, assessing whether the distributions are the same or not. We do so by viewing the data for each row as independently sampled from a mixture model with an unknown number of components. The problem is then solved by carrying out a cluster analysis of the  $R$  samples, adapting the work by Dasgupta and Raftery (1998) to our framework. If the selected cluster consists of a single group, then the distributions are indeed identical. Otherwise, there are significant differences, and the selected cluster gives us much more information than obtained when using routine procedures like Pearson's chi-square tests. The idea is analogous in spirit to the method proposed by Quintana (1998), but applied now to a wider context and under a different methodological view. An alternative computational strategy based on Gibbs sampling for the multinomial case can be found in Kuo and Yang (2006). An advantage of the approach we develop now is that it is computationally efficient (we do not require posterior simulation schemes) and easy to implement, and works well when  $R$  is large and the number of subsets to be found is not necessarily small or structured.

The rest of this article is organized as follows. Section 2 presents the mixture model and the proposed methodology, giving specific details in the multinomial and first-order

Markov chain cases. Section 3 illustrates the methods with two examples, and Section 4 provides a summary and further discussion.

## 2 The Model

Let  $f(\mathbf{m}|\boldsymbol{\theta})$  be a probability distribution for the vector of counts, parametrized by  $\boldsymbol{\theta} \in \Theta$ . We consider finite mixtures of such distributions. Concretely, we assume that the vectors  $\mathbf{m}_1, \dots, \mathbf{m}_R$  are independently drawn from a  $K$ -component mixture:

$$p(\mathbf{m}_1, \dots, \mathbf{m}_R | \boldsymbol{\theta}, \boldsymbol{\pi}_K) = \prod_{i=1}^R \sum_{j=1}^K \pi_{K,j} f(\mathbf{m}_i | \boldsymbol{\theta}_j), \quad (3)$$

where  $\boldsymbol{\pi}_K = (\pi_{K,1}, \dots, \pi_{K,K})$  with  $0 \leq \pi_{K,j} \leq 1$  are constrained by  $\sum_{j=1}^K \pi_{K,j} = 1$  for all  $K$ , with  $K \leq R$ , and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  are distinct points in the appropriate parameter space  $\Theta$ . Mixture models like (3) are very popular and have therefore been extensively studied in the statistical literature. See McLachlan and Basford (1988), McLachlan and Peel (2000), Fraley and Raftery (2002) and references therein.

The main purpose of using a mixture model is that rows in the contingency table form groups or clusters, according to the value of their corresponding parameter vector. Thus, the  $R$  distributions are effectively only  $K$ , and the probability that the  $i$ th row was sampled from the  $j$ th cluster is  $\pi_{K,j}$ . Under this view, equality of the  $R$  distributions is equivalent to the existence of only a single cluster and so the comparison that motivates this work reduces to carrying out a cluster analysis. To do so, we follow the model-based clustering (MBC) approach discussed in Dasgupta and Raftery (1998), which was initially proposed in Banfield and Raftery (1993). See also Murtagh and Raftery (1984). The procedure involves fitting the mixture model (3) and then comparing such fits for different values of  $K$  using the Bayesian Information Criterion (BIC) proposed in Schwarz (1978). We describe these steps next.

## 2.1 Fitting the Mixture Model

Usage of the EM algorithm (Dempster et al., 1977) has become a standard tool for fitting mixture models like (3) via maximum likelihood estimation. The simplest way to proceed is to introduce latent cluster indicators  $Z_{ij}$ , defined as 1 if the  $i$ th row was sampled from the  $j$ th distribution (cluster), and 0 otherwise, with  $j = 1, \dots, K$  and  $i = 1, \dots, R$ . Note that, by definition,  $\sum_{j=1}^K Z_{ij} = 1$  for all  $i$ . Thus the complete-data likelihood is

$$p(\mathbf{m}_1, \dots, \mathbf{m}_R, \{Z_{ij}\} \mid \boldsymbol{\theta}, \boldsymbol{\pi}_K) = \prod_{i=1}^R \prod_{j=1}^K (\pi_{K,j} f(\mathbf{m}_i \mid \boldsymbol{\theta}_j))^{Z_{ij}}, \quad (4)$$

which transforms (3) to an expression that is much easier to work with in practice.

The E step of the algorithm consists of computing the expected complete-data log-likelihood given the observed data and currently imputed parameters. This reduces to evaluating the posterior probability of cluster assignment:

$$\hat{Z}_{ij} = \frac{\pi_{K,j} f(\mathbf{m}_i \mid \boldsymbol{\theta}_j)}{\sum_{\ell=1}^K \pi_{K,\ell} f(\mathbf{m}_i \mid \boldsymbol{\theta}_\ell)}, \quad j = 1, \dots, K; \quad i = 1, \dots, R. \quad (5)$$

In the M step, the expected complete-data log-likelihood

$$l^*(\boldsymbol{\pi}_K, \boldsymbol{\theta}) = \sum_{i=1}^R \sum_{j=1}^K \hat{Z}_{ij} [\log(\pi_{K,j}) + \log(f(\mathbf{m}_i \mid \boldsymbol{\theta}_j))] \quad (6)$$

is maximized with respect to  $\boldsymbol{\pi}_K$  and  $\boldsymbol{\theta}$ . We get

$$\hat{\pi}_{K,j} = \frac{\sum_{i=1}^R \hat{Z}_{ij}}{\sum_{i=1}^R \sum_{\ell=1}^K \hat{Z}_{i\ell}}, \quad j = 1, \dots, K, \quad (7)$$

but the estimation of  $\boldsymbol{\theta}$  depends on the specific assumptions for  $f(\mathbf{m}_i \mid \boldsymbol{\theta}_j)$ . For the multinomial model (1) we have

$$\hat{\theta}_{jg} = \frac{\sum_{i=1}^R \hat{Z}_{ij} m_{ig}}{\sum_{i=1}^R \sum_{\ell=1}^C \hat{Z}_{i\ell} m_{i\ell}}, \quad j = 1, \dots, K; \quad g = 1, \dots, C, \quad (8)$$

while for the first-order Markov chain (2) we get

$$\hat{\theta}_{j1} = \frac{\sum_{i=1}^R \hat{Z}_{ij} t_{00}^i}{\sum_{i=1}^R \hat{Z}_{ij} (t_{00}^i + t_{01}^i)}, \quad \hat{\theta}_{j2} = \frac{\sum_{i=1}^R \hat{Z}_{ij} t_{10}^i}{\sum_{i=1}^R \hat{Z}_{ij} (t_{10}^i + t_{11}^i)}, \quad j = 1, \dots, K. \quad (9)$$

The two above steps are iterated until convergence is reached.

## 2.2 Choosing the Number of Mixture Components

Selecting the number of clusters is an essential task for any clustering algorithm. This is also valid in the context of our MBC approach. We adopt here the method by Dasgupta and Raftery (1998) who fit the mixture (3) using values of  $K$  ranging from 1 to a certain default maximum, and choosing the highest BIC value among these alternatives. For a given  $K$ , this is defined as

$$\text{BIC} = 2l^*(\hat{\boldsymbol{\pi}}_K, \hat{\boldsymbol{\theta}}) - d_K \log(n), \quad (10)$$

where  $d_K$  is the number of independent parameters in the corresponding model. For (1) we get  $d_K = KC - 1$ , while for (2) we find  $d_K = 3K - 1$ .

Usage of BIC as a selection procedure in this context has been justified in Dasgupta and Raftery (1998) as computing approximate Bayes factors for a model selection scheme with a prior distribution that assigns equal probability to all the models. From a frequentist viewpoint, it has been shown (Jeffreys, 1961) that choosing between two nested models on the basis of Bayes factors minimizes the sum of type I and II error rates. But more generally, such methodology can be applied to the case of more than two not necessarily nested models. For a thorough review of Bayes factors, see Kass and Raftery (1995). Thus, we view the mixtures (3) as models for different values of  $K$ , so that the actual Bayes factor is approximated in terms of the BIC values. Generally speaking, this approximation can be close (Kass and Wasserman, 1995), but the usual proofs use conditions not satisfied by finite mixture models. Nevertheless, Keribin (2000) showed that likelihood-penalized methods such as BIC, generate consistent estimates of the number of clusters. See further discussion about these and other related issues in Leroux (1992) and in Fraley and Raftery (2002).

Therefore, after computing the BIC values for  $K = 1, \dots, K_{max}$ , where  $K_{max}$  is a reasonable upper limit, the selected number of components  $K^*$  is the one attaining the highest BIC, which is consistent with the interpretation in terms of approximate Bayes factors. As a reasonable default choice, and considering that in practice we never encountered a problem where more than 5 clusters were detected, we recommend setting  $K_{max} = \min\{R, 10\}$  or even  $K_{max} = \min\{R, 8\}$ . See additional details about this method

in Dasgupta and Raftery (1998).

## 2.3 Selecting the Partition

After the value of  $K^*$  has been determined, we need to actually select a single partition. We do so by computing, for each  $i = 1, \dots, R$ , the value  $j^*$  such that for the model with  $K^*$  mixture components

$$\hat{Z}_{ij^*} = \max\{\hat{Z}_{ij} : 1 \leq j \leq K^*\}, \quad (11)$$

i.e., the component that maximizes the posterior probability of cluster assignment (5), with  $\theta_j$  and  $\pi_{K,j}$  replaced by  $\hat{\theta}_j$  and  $\hat{\pi}_{K,j}$  for all  $1 \leq j \leq K^*$ . We take the  $j^*$  values as indicators of cluster membership, thus completely defining the selected partition.

Table 1 gives a summarized version of the algorithm just described in the form of pseudo-code.

## 3 Data Illustrations

We consider here two datasets. The first one concerns performance of a group of students in an elementary Mathematics class. The second example involves sequences of hits and outs for 127 baseball players in the 1990 season of both American and National League Baseball.

### 3.1 Comparing Student's Performance

Our first dataset was introduced in Quintana (1998). It consists of Pass/Fail grades for a group of 203 students that enrolled in an “Algebra and Introductory Calculus” class during the 1996 Fall Semester in the Pontificia Universidad Católica de Chile. This particular class (we refer to it as MAT 1492) groups students from 4 different programs:

```

1. for ( $K$  in 1 to  $K_{max}$ ) {

    Iterate E and M steps until convergence: {

        E step:
            Compute  $\hat{Z}_{ij}, i = 1, \dots, R, j = 1, \dots, K$  using (5)

        M step:
            Compute  $\hat{\pi}_{K,j}, j = 1, \dots, K$  using (7)
            and  $\hat{\theta}_i, i = 1, \dots, R$  using either (8) or (9)
    }

    Evaluate  $BIC(K)$  using (10) and the appropriate definition of  $d_K$ 

}

2. Select the number of clusters as  $K^* = \operatorname{argmax}_{K=1, \dots, K_{max}} BIC(K)$ 

3. The cluster index for the  $i$ th subject is  $j^*$ , obtained according to (11)

```

Table 1: *Pseudo-code version of the clustering algorithm.*

(I) a general purpose Program in Chemistry; (II) B. S. in Chemistry and Pharmacy; (III) a general purpose Program in Biological Sciences; and (IV) B. S. in Biochemistry.

This class consisted entirely of students in their first semester. The students were randomly divided into two groups and assigned separate instructors (A and B). However, the material covered in the lectures, and all class activities were common for both groups. Even exams were graded by a common staff of Teaching Assistants. See additional background and analysis concerning these data in Quintana (1998) and in Kuo and Yang (2006).

Several questions about these data are of interest. In particular, we wish to know

whether the students' performance is affected by the program (i.e., a "program effect") or the instructor (i.e., an "instructor effect"). To answer the first question we initially consider the data as a  $4 \times 4$  contingency table, with rows representing Programs (I) through (IV) as described earlier, and columns corresponding to the four combinations of instructor/result, where instructor is either "A" or "B" and result is either "pass" or "fail". The data are reported in Table 2.

Table 2: *Student performance in an elementary "Algebra and Introductory Calculus" class.*

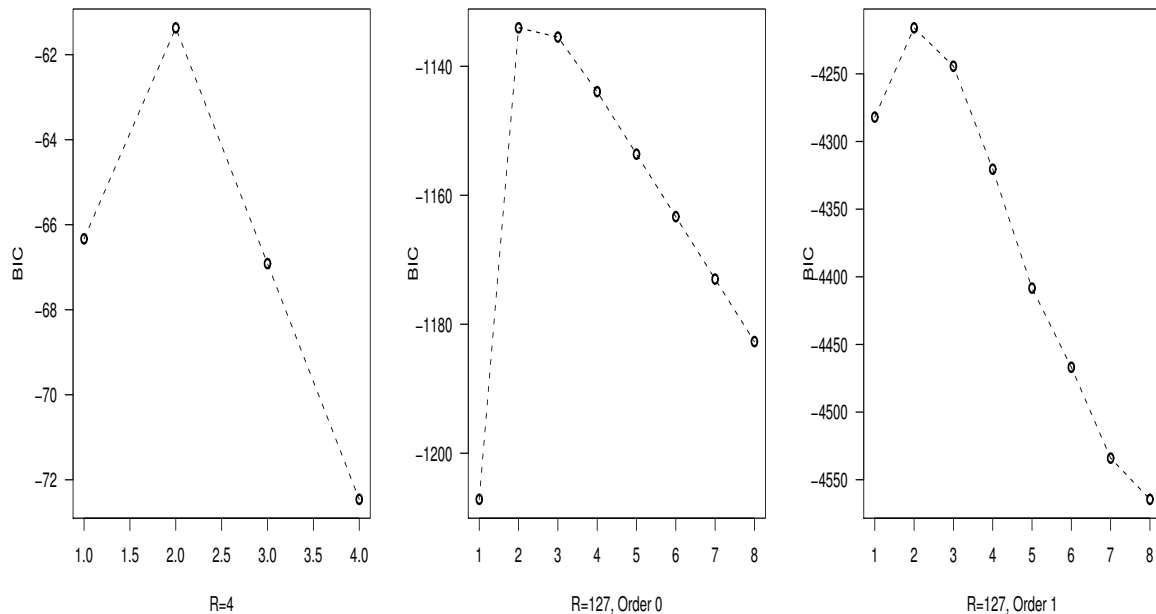
	Instructor A		Instructor B		
Program	Pass	Fail	Pass	Fail	Total
I	8	11	11	13	43
II	10	14	13	9	46
III	19	25	20	18	82
IV	14	2	12	4	32
Total	51	52	56	44	203

The BIC values for this example are plotted in Figure 1, from which it is concluded that the appropriate number of components is 2. Applying the procedure in Section 2.3, the optimal partition is (1, 1, 1, 2) which means that students from Program (IV) have different performance patterns than the other 3 programs, which in turn behave alike. Indeed, the proportion of students from Program (IV) that passed the class is much higher than the rest, which explains the results. This is also consistent with the findings in Quintana (1998). It is worth noting that the standard Pearson's chi-square statistic is 15.650 with 9 degrees of freedom, yielding a  $p$ -value of 0.0748, which under a standard decision rule, would not detect differences in the four distributions.

To answer the second question, i.e., assessing the "instructor effect", we consider an exploratory initial analysis where programs are assessed separately. Thus, we consider four  $2 \times 2$  contingency tables where, for each program, rows are the instructors and columns represent the pass/fail counts. We found that the BIC criterion favored  $K = 1$

in all cases (data not shown). This suggests that we can collapse the four programs and view the data as a single  $2 \times 2$  table, with instructors as rows and pass/fail counts in the columns. The BIC values are  $-11.688$  ( $K = 1$ ) and  $-13.075$  ( $K = 2$ ) so we do not actually find differences in the student's performance across instructors. Reinforcing this finding, the chi-square test statistic is 0.61657 with 1 degree of freedom and  $p$ -value 0.4326 so we reach the same conclusion.

Figure 1: Values of the BIC criterion for the data on student's performance (left panel) and for the baseball data (middle and right panels). Horizontal axes represent potential values for the number of mixture components  $K$  in (3).



We can also summarize both analyses by viewing the data as an  $8 \times 2$  contingency table, having the 8 combinations of program/instructor in the rows and the pass/fail counts in the columns. To this effect, we ignore the random assignment of students to instructors. This time the model selected has 2 components (data not shown), and the corresponding 2 subsets in the best partition are of sizes 2 (the two groups from Program

IV) and 6 (all the rest). This suggests that Program IV is different from the other ones, but no instructor effect is found.

It is interesting to compare these results with those reported in Quintana (1998), using a clustering structure induced by a Dirichlet process prior on the  $\theta_i$  vectors of classification probabilities. It was found there that for the  $8 \times 2$  version of the data, the maximum a posteriori (MAP) estimate of the number of clusters is 3, but the MAP partition contains only 2 subsets, one formed by students from Program IV that were taught by Instructor A, and the other containing all the rest. The posterior probability for the second most likely partition is very close to the MAP, and coincides with what results using our proposed method. Using the same model, but with a different computational approach, Kuo and Yang (2006) found analogous results.

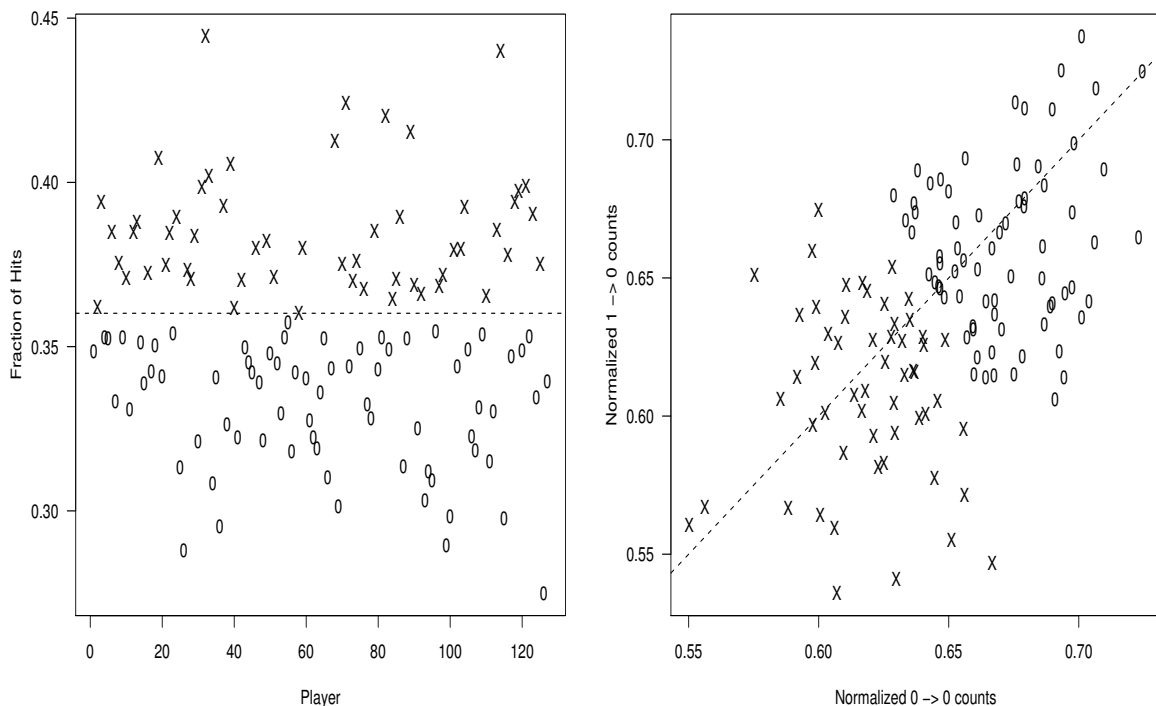
### 3.2 Comparing Hits and Outs in Baseball

Our second example concerns data for 127 players in the 1990 season of both American and National Baseball Leagues. For each of these players, the entire sequence of hits and outs at bat was recorded. The binary outcomes are defined to be 1 if a hit, walk or sacrifice occurred and 0 if an out occurred. All of these players have at least 500 binary measurements each, and several occasion-specific covariates are available as well, but for the purpose of this application, they are not considered. The data were first introduced in Albright (1993), and were later analyzed, among others, in Quintana and Newton (2000), Quintana and Müller (2004) and some references therein. It is important to note that using a Monte Carlo conditional test with exact size, Quintana and Newton (1998) concluded that zero-order Markov chains are appropriate for the analysis of these data. The test statistic they used was defined as the sum of likelihood-ratio test statistics for order zero versus one from individual sequences. However, under an entirely different semi-parametric Bayesian formulation, Quintana and Müller (2004) found that the selected order is one.

We begin the analysis by ignoring the underlying serial dependence, and consider the multinomial model, accommodating the data in a  $127 \times 2$  contingency table, with

rows representing players and columns the counts of hits and outs. The BIC values computed as in (10) chose 2 clusters. See Figure 1, middle panel. Next, we constructed the two clusters as indicated in Section 2.3. To have a better idea of what these clusters represent, the left panel in Figure 2 shows the 127 proportions of hits, defined as the number of hits divided by the total number of at bat occasions. The grouping of players is indicated by the symbols “X” and “O”. The clusters are perfectly correlated with such values, separating the players in two groups, according to whether their batting averages are above or below 0.360. These two groups have 55 and 72 players, respectively.

Figure 2: *Final clustering of 127 baseball players in the 1990 season of American and National Baseball Leagues. The left panel represents the proportion of hits for all players versus player number, from 1 to 127. The right panel plots the MLE of  $\theta_{i1}$  versus the MLE of  $\theta_{i2}$ , as specified in (12). In both plots, the two final clusters found using the proposed algorithm are marked by the “X” and “O” symbols.*



Next we model each sequence as a first-order Markov chain, and repeat the analysis. We found once more 2 clusters. Interestingly, the corresponding partition is identical to that selected for serial independence, suggesting that for these data, the clustering is not related to assumptions on the serial correlation.

To visualize the clusters in this new context, the right panel in Figure 2 shows the normalized 0 to 0 versus 1 to 0 transition counts. These are defined as

$$\frac{t_{0,0}^i}{t_{0,0}^i + t_{0,1}^i} \quad \text{and} \quad \frac{t_{1,0}^i}{t_{1,0}^i + t_{1,1}^i}, \quad (12)$$

respectively, and correspond to the maximum likelihood estimates of  $\theta_{i1}$  and  $\theta_{i2}$ , the parameters representing the first order transition matrix in (2). The type of scatter shown in this plot does support a first-order Markov model, as the points do not appear to strongly line up along the overlaid diagonal line representing equality in the rows of the transition probability matrix and henceforth, serial independence. In fact, the normalized transition counts for the cluster marked with an “X” are more separated from this diagonal than those players in the cluster marked with a “0”. Moreover, players in the “0”-clusters are in general more likely to have an out in the next occasion at bat than those in the “X”-cluster. This also serves as an interpretation of the clusters meaning, also suggesting the presence of streakiness in these players’ batting patterns.

Finally, Quintana and Newton (2000) found that, under a nonparametric Bayesian model based on a Dirichlet process prior, the MAP estimate for the number of clusters is also 2. Finding the MAP partition, however, is much more difficult due to the huge number of possible groupings that can be formed. A brute force method based on tallying partitions directly from MCMC output is likely going to take too many iterations to produce significant output, specially for the case at hand where  $R = 127$ . For instance, the number of possible partitions of 127 objects into 2 subsets is  $8.5 \times 10^{38}$ . An “ad-hoc” procedure based on applying the k-means algorithm (Hartigan and Wong, 1979) to estimates of the posterior means of transition matrix parameters  $\theta_i$  gave essentially the same result as our proposed method (data not shown).

## 4 Summary and Discussion

In this article we developed a method for comparing discrete distributions that was motivated as a special case of a model-based cluster analysis using mixture models. Each mixture component represents a different cluster, and the number of clusters is estimated using the method of Dasgupta and Raftery (1998). Our proposal has the advantage of being efficient and easy to implement. Compared to standard chi-square tests, the underlying clustering construction provides much more informative results, which can be interpreted in the context of the specific problem. This is certainly a better outcome than a plain “yes/no” answer to the question: “are these distributions different?”

The methodology we have discussed can be applied to any type of discrete distribution. We gave details for two important special cases, but other situations can be handled using the same general clustering procedure. The case of continuous distributions is not specially different, although accommodating the data in a contingency table may be hopeless, or would require loss of information via discretizations.

Finally, the proposed method is computationally efficient, and can very easily produce reasonable results. It also works well when the number of individuals is moderate to large. This was shown to be the case in the baseball example discussed in Section 3.2, where the number of possible partitions is huge. In fact, this is a substantial advantage of our method over alternatives such as finding the MAP partition, which is typically much harder to find than using our MBC method.

## References

- Agresti, A. (2002), *Categorical Data Analysis, Second Edition*, New York: Wiley-Interscience.
- Albright, S. C. (1993), “A statistical analysis of hitting streaks in baseball (Disc: p1184-1196),” *Journal of the American Statistical Association*, 88, 1175–1183.

- Banfield, J. D. and Raftery, A. E. (1993), “Model-based Gaussian and non-Gaussian clustering,” *Biometrics*, 49, 803–821.
- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Massachusetts: M.I.T. Press.
- Dasgupta, A. and Raftery, A. E. (1998), “Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering,” *Journal of the American Statistical Association*, 93, 294–302.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm (with discussion),” *Journal of the Royal Statistical Society Series B*, 39, 1–37.
- Fraley, C. and Raftery, A. E. (2002), “Model-Based Clustering, Discriminant Analysis, and Density Estimation,” *Journal of the American Statistical Association*, 97, 611–631.
- Hartigan, J. A. and Wong, M. A. (1979), “A K-means clustering algorithm,” *Applied Statistics*, 28, 100–108.
- Jeffreys, H. (1961), *Theory of probability*, Third edition, Clarendon Press, Oxford.
- Kass, R. and Raftery, A. (1995), “Bayes Factors,” *Journal of the American Statistical Association*, 90, 773–795.
- Kass, R. and Wasserman, L. (1995), “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *Journal of the American Statistical Association*, 90, 928–934.
- Keribin, C. (2000), “Consistent estimation of the order of mixture models,” *Sankhyā. The Indian Journal of Statistics. Series A*, 62, 49–66.
- Kuo, L. and Yang, T. Y. (2006), “An improved collapsed Gibbs sampler for Dirichlet process mixing models,” *Computational Statistics & Data Analysis*, 50, 659–674.

- Leroux, B. G. (1992), “Consistent estimation of a mixing distribution,” *The Annals of Statistics*, 20, 1350–1360.
- McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, New York: Marcel Dekker.
- McLachlan, G. J. and Peel, D. (2000), *Finite Mixture models*, New York: Wiley.
- Murtagh, F. and Raftery, A. E. (1984), “Fitting straight lines to point patterns,” *Pattern Recognition*, 17, 479–483.
- Quintana, F. and Müller, P. (2004), “Nonparametric Bayesian Assessment of the Order of Dependence for Binary Sequences,” *Journal of Computational and Graphical Statistics*, 13, 213–231.
- Quintana, F. A. (1998), “Nonparametric Bayesian analysis for assessing homogeneity in  $k \times l$  contingency tables with fixed right margin totals,” *Journal of the American Statistical Association*, 93, 1140–1149.
- Quintana, F. A. and Newton, M. A. (1998), “Assessing the order of dependence for partially exchangeable binary data,” *Journal of the American Statistical Association*, 93, 194–202.
- (2000), “Computational aspects of Nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences,” *Journal of Computational and Graphical Statistics*, 9, 711–737.
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.