Bayesian Clustering with Priors on Partitions

Tim B. Swartz [*]

Abstract

Traditional clustering algorithms are deterministic in the sense that a given dataset always leads to the same output partition. This paper modifies traditional clustering algorithms whereby data is associated with a probability model, and clustering is carried out on the stochastic model parameters rather than the data. This is done in a principled way using a Bayesian approach which allows the assignment of posterior probabilities to output partitions. In addition, the approach incorporates prior knowledge of the output partitions using Bayesian melding. The methodology is applied to two substantive problems; (1) a question of stylometry involving a simulated dataset and (2) the assessment of potential champions of the 2010 FIFA World Cup.

Keywords: Bayesian Melding, Clustering, Empirical Bayes, FIFA World Cup 2010, Stylometry.

[*] Tim Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6 (email: tim@stat.sfu.ca). Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. This work was initiated during a visit to the Mathematics Department at the University of Bristol. Swartz thanks the Statistics group at Bristol for their hospitality. Swartz also thanks two anonymous reviewers whose comments helped to improve the manuscript.

1 INTRODUCTION

Modern approaches to clustering/prediction such as neural networks (Bishop 1995) and tree methodologies (Breiman et al. 1984, Ripley 1996) rely on a response variable and covariates. These computationally intensive approaches are sometimes described as data mining techniques where large datasets are often encountered.

In the absence of covariates, traditional clustering algorithms (Kaufman and Rousseeuw 1990) continue to be widely used and are available in most statistical software packages. Traditional clustering algorithms consider data $y_1, \ldots, y_n$ with the intention of partitioning the n data points into clusters based on "similarity" where similarity is determined by a distance measure $d(y_i, y_j)$ appropriately defined between data points. Traditional clustering algorithms are either *hierarchical* or *partitional*. In *agglomerative* hierarchical clustering, algorithms begin with each data point assigned to its own cluster, and in subsequent iterations, clusters are merged based on similarity. In *divisive* hierarchical clustering, algorithms begin with a single cluster consisting of all data points, and in subsequent iterations, clusters are divided into smaller clusters based on similarity. Partitional algorithms determine all clusters at once where subsequent iterations rearrange the cluster membership based on similarity.

A drawback of the traditional clustering algorithms and all deterministic clustering algorithms is that they do not take the variability of data into account. For example, suppose that there is some (but insufficient) similarity for the clustering of the two data points y_i and y_j. Whereas running a deterministic clustering algorithm again and again on the same dataset always leads to different clusters for y_i and y_j, we prefer to express cluster uncertainty. Our approach is Bayesian where we view the clustering problem somewhat differently. We are not primarily interested in clustering the data, but rather, in clustering the populations from which the data arise. And we argue that although the difference is subtle, the analysis of populations is often the problem of interest in many applications. For example, suppose that we have product samples from various factories. Our interest is in the assessment and categorization of the factories rather than the quality of the particular samples produced by the factories. If we assume that datum y_i arises from a population characterized by a parameter θ_i, then traditional clustering algorithms may be applied to $\theta_1, \ldots, \theta_n$ where the variability of the θ's is expressed via the posterior distribution. As the θ's are stochastic, it then follows that the clustering results are stochastic.

Related to this general philosophy of clustering are approaches based on latent class clustering (Vermunt and Magidson 2002). In latent class clustering, a statistical model is also imposed on the data. However, the model typically takes the form of a mixture distribution where the underlying component distributions define the cluster groups. Unknowns in the model formulation include the parameters of the component distributions, the mixture weights and possibly the number of mixture components. As with our approach, latent class mixture modelling is probabilistic in the sense that there is a probability associated with group membership. In other words, there is a latent variable structure where membership in a particular group is unknown. There are many variations of latent class clustering models and we note that the methodology has been developed in both classical and Bayesian settings.

We describe our proposed procedures in full detail in Section 2 where we note that the approach is generally applicable to any of the traditional clustering algorithms. In addition, a by-product of the approach is that missing distance measurements $d(y_i, y_j)$ and repeated distance measurements cause no additional problems. The idea of clustering on parameters was initially proposed by Gill, Swartz and Treschow (2007). This paper extends the original approach in a number of directions. Later in Section 2, we suggest that the proposed default priors for the θ's may not utilize full prior knowledge. In many applications, prior knowledge concerning cluster membership is available. We demonstrate how the Bayesian melding technique of Poole and Raftery (2000) can be applied to impose a prior on the θ's based on our knowledge of the output partitions. The implementation of Bayesian melding is simplified as we take advantage of the fact that the space of output partitions is a finite discrete set.

In Section 3, we apply the proposed methodology to a problem of stylometry involving a simulated dataset. We are interested whether texts cluster according to their respective authors. In actual applications, authorship is typically unknown. With frequency data on non-contextual keywords, the underlying data model is multinomial and this provides an improvement upon standard stylometric analyses that rely on the multivariate normal distribution. Ward's method (Ward 1963) is used as the clustering technique where algorithmic stopping rules are designed for the given application.

In Section 4, we apply the proposed methodology to a topical problem involving the 2010 FIFA World Cup. Whereas many football analyses consider the ranking of teams, we are interested in determining whether participant teams are amongst the elite teams that stand a realistic chance of winning the World Cup. Posterior probabilities of membership to this elite class (cluster) are provided. The underlying data model is an extension (Davidson 1970) of the Bradley-Terry model (Bradley and Terry 1952). The clustering algorithm is the unweighted paired group method of averaging algorithm (UPGMA) proposed by Lance and Williams (1966). Some surprising results are obtained. We conclude with a short discussion in Section 5.

2 METHODOLOGY

Consider the problem of clustering where we have data $y_1, \ldots, y_n$ which are typically multivariate. We require a statistical model

$$f(y \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta_i) \tag{1}$$

where $y = (y_1, \ldots, y_n)'$, $\theta = (\theta_1, \ldots, \theta_n)'$, and θ_i is a continuous parameter that characterizes the population from which y_i arose. Note that the statistical model (1) is an ingredient that is not required for traditional clustering methodologies (hierarchical and partitional). However, its inclusion is often straightforward when we consider the stochastic manner in which the data were generated, and its inclusion allows us to take the variability of the data into account. Note also that (1) presumes conditional independence amongst the y_i's; this is an unnecessary assumption which we relax in our second example.

In many situations, and in the examples that we consider, direct prior knowledge concerning θ is unavailable, and therefore we suggest a vague default prior which we denote by $\pi_\theta(\theta)$. This leads to the posterior density given by

$$\pi(\theta \mid y) \propto f(y \mid \theta)\ \pi_\theta(\theta). \tag{2}$$

At this stage, we have a standard Bayesian application. However, our interest is in clustering, and as previously argued, we cluster on the θ's rather than cluster on the y's. The reason is that the θ's represent the underlying populations of interest whereas the y's are simply observations from the populations. We therefore need to define a distance measure $d(\theta_i, \theta_j)$ on the parameters, and there often exist sensible metrics as demonstrated in the two examples. Since the θ's are typically of lower dimension than the y's, this also aids in the specification of distance measures. With the distance measure defined, we describe the clustering procedure via

$$\theta \to C(\theta) \to \phi \tag{3}$$

where $C = C(\theta)$ denotes the clustering algorithm which may be any of the traditional clustering algorithms (hierarchical and partitional) and ϕ denotes the output partition from the clustering algorithm. For clarification, given a realization $\theta_1, \ldots, \theta_n$, the clustering algorithm C produces a partition ϕ which is a division of the set $\{\theta_1, \ldots, \theta_n\}$ into non-overlapping components.

Therefore our clustering methodology readily presents itself via a sampling-based approach. We generate θ from the posterior distribution (2). We then provide θ as input to the clustering procedure (3) which produces an output partition ϕ. Repeating the above steps and averaging over a large number of simulations yields posterior probabilities with respect to the output partitions. For example, the clustering of θ_i and θ_j may be of interest. In this case, it is a simple matter to calculate the proportion of simulated output partitions ϕ where θ_i and θ_j cluster together.

There are four comments which are relevant to the proposed clustering methodology. First, there is flexibility in the simulation procedure. If possible, one may sample θ directly from the posterior. Alternatively, schemes such as Markov chain methods or importance sampling may be utilized. Second, we note that traditional clustering methodologies which cluster data are unable to handle missing distances $d(y_i, y_j)$. In our sampling-based approach, a complete vector $(\theta_1, \ldots, \theta_n)$ is generated in every iteration of sampling. Hence, we never encounter missing distances $d(\theta_i, \theta_j)$. Third, traditional clustering methodologies struggle with repeated measurements on the y's. This causes no problem for our proposed methodology since extra data are simply incorporated into the model (1). Fourth, we may think of the clustering algorithm C in (3) as a black-box procedure. An input θ yields an output ϕ after some complex operations. This is a useful way to think about things as our focus now shifts to the more interpretable parameter ϕ.

2.1 Implementing Priors on Partitions

In clustering as described by (3), the parameter ϕ has physical meaning since the output partition directly addresses the inferential question of interest. It follows that more is generally known about the output parameter ϕ than the input parameter θ. It is therefore conceivable that subjective prior opinion concerning ϕ may be available and expressed via $\pi_\phi(\phi)$. A difficulty however is that $\pi_\theta(\theta)$ is typically incompatible with $\pi_\phi(\phi)$ and it is not clear how a principled Bayesian approach ought to proceed. An immediate thought is to ignore $\pi_\theta(\theta)$ and obtain an induced distribution on θ via $\pi_\phi(\phi)$. However, this is not possible in our context as the space of output partitions corresponding to ϕ is a finite discrete set and θ is a continuous and possibly multivariate parameter. There are a potentially infinite number of θ's that map to a particular ϕ.

The general problem of reconciling prior opinion on inputs and outputs is an important topic which is fundamental to the discipline of computer experiments. Two prominent approaches have been considered. Raftery, Givens and Zeh (1995) proposed the Bayesian synthesis technique which introduced a marginal "postmodel" distribution for θ that takes into account the relationship between the input parameter θ and the output parameter ϕ. The distribution was later observed to suffer from the Borel paradox (Wolpert 1995). The Borel paradox concerns an ill-defined conditional probability with respect to the choice of parametrization and occurs when conditioning on an event of probability zero.

Poole and Raftery (2000) subsequently proposed the Bayesian melding approach which combines $\pi_\theta(\theta)$ and $\pi_\phi(\phi)$ using logarithmic pooling. In logarithmic pooling applied to our clustering context, a *combined* prior on the output parameter ϕ is proportional to

$$\pi^*_\theta(\phi)^\alpha \pi_\phi(\phi)^{1-\alpha} \tag{4}$$

where $\pi_\theta^*(\phi)$ is the induced probability mass function corresponding to $\pi_\theta(\theta)$ on the ϕ-space and $0 \leq \alpha \leq 1$ is the pooling weight where larger values of α assign more emphasis to $\pi_\theta^*(\phi)$ relative to $\pi_\phi(\phi)$. The term "logarithmic" is understood in the sense that the logarithm of (4) is additive in its components. However, for our application, we require a density defined on the continuous input parameter θ rather than the discrete output parameter ϕ. Accordingly, what does (4) say about the probability distribution associated with θ? Since the mapping from $\phi \rightarrow \theta$ is not 1:1, the standard change-of-variable formula (which requires a jacobian) cannot be applied to (4); for a given ϕ_0, there are an infinite number of θ's that map to ϕ_0. Referring to (4), Poole and Raftery (2000) proposed a density which is proportional to

$$\pi_\theta^*(C(\theta))^\alpha \pi_\phi(C(\theta))^{1-\alpha} \left(\frac{\pi_\theta(\theta)}{\pi_\theta^*(C(\theta))} \right) \tag{5}$$

and is based on the rationale that all θ's mapped to the same $\phi = C(\theta)$ ought to have densities proportional to $\pi_\theta(\theta)$.

We utilize the Bayesian melding approach where features of our clustering problem lead to simplifications. In particular, our implementation does not require the use of the sampling-importance-resampling algorithm (Rubin 1988) nor the use of nonparametric density estimation techniques which often face difficulties in high dimensional problems (Liu, Lafferty and Wasserman 2007).

Our first attempt at an alternative prior for θ corresponds to $\alpha = 0$ in (5) using logarithmic pooling. The prior is given by

$$p(\theta) = \left(\frac{\pi_\theta(\theta)}{\pi_\theta^*(C(\theta))} \right) \pi_\phi(C(\theta)). \tag{6}$$

From (6), we observe that the output prior π_ϕ is the prominent component since all θ's that map to $C(\theta)$ have the common factor $\pi_\phi(C(\theta))$. For values of θ that share a common $C(\theta)$, their relative contributions are distributed according to the first factor in (6). Since $\pi_\theta(\theta)$ is often a vague default prior, this suggests that the output prior π_ϕ is of overriding importance in determining $p(\theta)$.

Is (6) a proper density? The answer is yes provided that π_θ is proper. To see that this is the case, consider a partition of the θ-space $\{A_1, \ldots, A_m\}$ where $C(\theta) = C_i$ if $\theta \in A_i$. Note that the partition is finite since the number of output partitions is necessarily finite. Then

$$\int p(\theta) d\theta \;=\; \sum_{i=1}^{m} \frac{\pi_\phi(C_i)}{\pi_\theta^*(C_i)} \int_{A_i} \pi_\theta(\theta) d\theta \;=\; \sum_{i=1}^{m} \frac{\pi_\phi(C_i)}{\pi_\theta^*(C_i)}\, \pi_\theta^*(C_i) \;=\; \sum_{i=1}^{m} \pi_\phi(C_i) \;=\; 1.$$

A difficulty with (6) lies in the specification of the subjective prior $\pi_\phi(\phi)$. If the clustering problem has input parameters $\theta_1, \ldots, \theta_n$ corresponding to n underlying populations, then the number of possible output partitions m is the Bell number B_n (Rota 1964). The Bell number B_n is the number of ways that a set of n elements can be partitioned into nonempty subsets. For example, $B_1 = 1$, $B_2 = 2$ and $B_3 = 5$ where B_3 corresponds to the five partitions $\{\{\theta_1\}, \{\theta_2\}, \{\theta_3\}\}$, $\{\{\theta_1, \theta_2\}, \{\theta_3\}\}$, $\{\{\theta_1, \theta_3\}, \{\theta_2\}\}$, $\{\{\theta_2, \theta_3\}, \{\theta_1\}\}$ and $\{\{\theta_1, \theta_2, \theta_3\}\}$. In even a small application with $n = 10$ populations, the specification of a subjective prior becomes unwieldy as $B_{10} = 115975$.

Our solution to the combinatorial problem is to provide a non-negative score $S(C(\theta))$ to each output partition $C(\theta)$. The idea is that an output partition $C(\theta)$ can be assessed by various criteria and we can easily assign a numerical score which describes our relative belief in $C(\theta)$. We keep the number of possible scores modest, much smaller than the associated Bell number. In the two substantive examples, we propose S-scores that are driven by prior knowledge. With a non-negative score $S(C(\theta))$ defined on output partitions, this suggests an alternative prior on the input parameter θ given by

$$p(\theta) \propto \left(\frac{\pi_\theta(\theta)}{\pi_\theta^*(S(C(\theta)))} \right) S(C(\theta)) \tag{7}$$

where $\pi_\theta^*(S(C(\theta)))$ is the induced probability mass function corresponding to $\pi_\theta(\theta)$ on the S-space.

To help clarify the proposed methodology, we consider a toy example where (7) can be obtained analytically. Consider a data model (1) where there are $n = 2$ underlying population parameters θ_1 and θ_2. The inferential problem concerns the potential clustering of these input parameters. A uniform prior $\pi_\theta(\theta) = \pi_\theta(\theta_1, \theta_2) = 1$ for $\theta_1, \theta_2 \in (0, 1)$ is chosen and we note that a clustering algorithm C applied to a realization of θ leads to one of the $m = 2$ output partitions $\phi_1 = \{\{\theta_1\}, \{\theta_2\}\}$ and $\phi_2 = \{\theta_1, \theta_2\}$. Suppose that the clustering algorithm $C(\theta)$ merges θ_1 and θ_2 (i.e. yields partition ϕ_2) when $\mid \theta_1 - \theta_2 \mid < 0.5$. Assume further that we have prior knowledge on the output partitions expressed via an S-score where $S(\phi_1) = 1.0$ and $S(\phi_2) = 4.0$. Equivalently, the S-score implies $\pi_\phi(\phi_1) = 0.2$ and $\pi_\phi(\phi_2) = 0.8$. Then referring to (7), $\pi_{\theta^*}(S(\phi_2)) = \int \int I(\mid \theta_1 - \theta_2 \mid < 0.5)\ d\theta_1 d\theta_2 = 0.75$ and its complement probability $\pi_{\theta^*}(S(\phi_1)) = 0.25$. One then obtains

$$p(\theta) \propto \begin{cases} \left(\frac{1}{0.25}\right)(1.0) & \mid \theta_1 - \theta_2 \mid \geq 0.5 \\ \left(\frac{1}{0.75}\right)(4.0) & \mid \theta_1 - \theta_2 \mid < 0.5 \end{cases}$$

which leads to

$$p(\theta) = \begin{cases} 0.800 & \mid \theta_1 - \theta_2 \mid \geq 0.5 \\ 1.067 & \mid \theta_1 - \theta_2 \mid < 0.5 \end{cases} . \tag{8}$$

We therefore observe the effect of the S-scores in (8) as the initial uniform probability on (θ_1, θ_2) corresponding to π_θ is modified to give more probability to the region $\mid \theta_1 - \theta_2 \mid < 0.5$. If we had chosen a more severe S-score, say $S(\phi_1) = 1.0$ and $S(\phi_2) = 9.0$, then $p(\theta)$ in (8) is even more greatly affected resulting in $p(\theta) = 0.4$ for $\mid \theta_1 - \theta_2 \mid \geq 0.5$ and $p(\theta) = 1.2$ for $\mid \theta_1 - \theta_2 \mid < 0.5$.

Now suppose that we are interested in the posterior probability of an event A defined on output partitions. For example, A could be the event that θ_i and θ_j cluster together. As another example, A could be the event that θ_i is a singleton such that it clusters into a group of its own. Using the preferred prior (7), the posterior probability of the event A is expressed by the ratio of integrals

$$I(A) = \frac{\int I(C(\theta) \in A) f(y \mid \theta) p(\theta) d\theta}{\int f(y \mid \theta) p(\theta) d\theta}. \tag{9}$$

As (9) is generally intractable, we consider an approximation based on importance sampling. Let $\theta^{(1)}, \ldots, \theta^{(N)}$ be a sample from an importance sampler with density $h(\theta)$. Then (9) can be approximated by

$$\hat{I}(A) = \frac{\sum_{i=1}^{N} I(C(\theta^{(i)}) \in A) f(y \mid \theta^{(i)}) \pi_\theta(\theta^{(i)}) S(C(\theta^{(i)}))/(h(\theta^{(i)}) \pi_\theta^*(S(C(\theta^{(i)}))))}{\sum_{i=1}^{N} f(y \mid \theta^{(i)}) \pi_\theta(\theta^{(i)}) S(C(\theta^{(i)}))/(h(\theta^{(i)}) \pi_\theta^*(S(C(\theta^{(i)}))))}. \tag{10}$$

where the choice of the importance sampler is based on the particular application. In the example of Section 3, we note that (10) simplifies according to $h(\theta) \propto f(y \mid \theta)$. Simplifications to (10) also occur if $h(\theta) = \pi_\theta(\theta)$ or if $\pi_\theta(\theta) \propto 1$.

A potential challenge in the approximation of (9) via (10) is the term $\pi_\theta^*(S(C(\theta^{(i)})))$. However, we are reminded that the number of S-scores is not large, say s, and it may therefore be possible to estimate $\pi_\theta^*(S_1), \ldots, \pi_\theta^*(S_s)$ using the same set of importance sampling variates $\theta^{(1)}, \ldots, \theta^{(N)}$ where

$$\hat{\pi}_\theta^*(S_j) = \frac{1}{N} \sum_{i=1}^{N} I(S(C(\theta^{(i)})) = S_j) \pi_\theta(\theta^{(i)})/h(\theta^{(i)}). \tag{11}$$

3 EXAMPLE: STYLOMETRY

Given texts of unknown origin, the essential problem of stylometry concerns inference relating to authorship where linguistic style is taken into account. Stylometry may therefore be viewed as a clustering problem where texts are partitioned into groups sharing a common author. Whereas many approaches to stylometry have been proposed (Holmes 1999), we specifically refer to two Bayesian approaches; (1) Mosteller and Wallace (1963, 1984) in the study of the Federalist papers, and (2) Gill, Treschow and Swartz (2007) in the study of texts associated with the reign of King Alfred.

To assess our methodology in a general stylometric setting, we considered a dataset where authorship is "known". Accordingly, we generated data

$$y_i = (y_{i1}, \ldots, y_{iK})' \sim \text{Multinomial}(m_i, \theta_{i1}, \ldots, \theta_{iK}) \tag{12}$$

where y_{ik} is the frequency count corresponding to the k-th *keyword* of the i-th text, $k = 1, \ldots, K-1$, y_{iK} is the frequency count corresponding to non-keywords and m_i is the number of words in the i-th text, $i = 1, \ldots, n$. The intention is that keywords are non-contextual and are used unreflectively by authors such that keyword frequency is constant over texts for a given author. Hence, differential rates of usage of keywords form *word-prints* for authors. Keywords are typically prepositions, conjunctions, articles and common verbs. Clearly, the choice of keywords ought to be determined by a subject matter expert.

We considered $n = 7$ texts each consisting of $m_i = 1000$ words where the words were classified according to $K - 1 = 5$ keywords. Furthermore, the data were generated according to the usage rate parameters $\theta_1 = \theta_2 = \theta_3 = (0.02, 0.02, 0.02, 0.02, 0.02, 0.9)'$, $\theta_4 = \theta_5 = (0.09, 0.03, 0.01, 0.01, 0.01, 0.85)'$ and $\theta_6 = \theta_7 = (0.01, 0.01, 0.01, 0.03, 0.09, 0.85)'$. Therefore, the seven texts correspond to three authors, and we emphasize that the number of clusters (authors) is unknown in advance. In this problem, a given stylometric procedure takes the keyword frequency data y corresponding to the 7 texts, and clusters the texts according to their authorship parameters $\theta_1, \ldots, \theta_7$. Therefore, the goal of our stylometric procedure is to obtain the "true" partition $\{\{\theta_1, \theta_2, \theta_3\}, \{\theta_4, \theta_5\}, \{\theta_6, \theta_7\}\}$ with high posterior probability. Of course, the data are generated according to (12) and not every simulated dataset provides strong evidence in favour of the true partition.

The multinomial model (12) used for data generation also serves as the statistical model from which the likelihood (1) is formed. The multinomial model represents an improvement over more tractable normal models that are typically used in stylometric problems (Holmes and Forsyth 1995). A difficulty with normal models is that they fail to account for the negative correlations between frequency counts when diagonal variance-covariance matrices are assumed.

In this application, we demonstrate the use of Ward's method (Ward 1963) as the traditional clustering algorithm. Needing a distance measure to differentiate the θ's, we defined

$$d(\theta_i, \theta_j) = \sum_{k=1}^{K} (\theta_{ik} - \theta_{jk}) \log(\theta_{ik}/\theta_{jk}) \tag{13}$$

which is sometimes referred to as Jeffreys divergence (Jeffreys 1946). We note that the distance measure (13) is a natural choice since it takes into account the fact that θ_i forms a probability distribution corresponding to a categorical random variable. Ward's method uses (13) to calculate the increase in the "sum of squares" when merging two clusters. We also note that clustering on the θ's rather than data better addresses the inferential question of interest. The θ's represent the characteristics of authors, and it is the authors whom we attempt to distinguish.

We invoked a stopping rule in the clustering algorithm which considers the within-author variability. This is done by first dividing the texts into blocks of roughly the same size, and then initializing Ward's algorithm by placing each of the blocks of the i-th text into the i-th cluster, $i = 1, \ldots, 7$. For our dataset, we divided each text of 1000 words into 10 blocks of 100 words. The idea is that the variation between blocks by an author is similar to the variation between texts by the same author. We therefore stopped the clustering algorithm when a proposed merge between two clusters is such that the cluster distance exceeds the α-quantiles of the two within-cluster distances. We have chosen $\alpha = 0.80$ but have found that results do not differ greatly for small changes in α.

To complete the model specification, we require the prior $p(\theta)$ in (7) which incorporates knowledge on inputs and outputs. For this, it is tempting to define π_θ according to a flat

Dirichlet prior. This is consistent with our claim that we don't know a great deal about the inputs (i.e. the keyword frequencies). The flat Dirichlet also has the advantage that it is proper and is conjugate to the multinomial distribution (12). However, it is important to remember that a flat "non-informative" distribution for a parameter does not imply a flat distribution on a transformation of the parameter. This is the case in this application as a flat distribution π_θ induces an informative distribution on ϕ whereby Ward's algorithm yields a single cluster with high probability with respect to π_θ. Alternatively, we consider a prior π_θ that is more in line with the observed data. For this, we set π_θ as the product of Dirichlet$(a_{i1}, \ldots, a_{iK})$ densities where $a_{ik} = \lambda y_{ik} + 1$. The Dirichlet provides the conjugacy which is desirable for importance sampling and we chose $\lambda = 1.0$. Our procedure may therefore be described as empirical Bayes.

As for the S-scores, we note that with $n = 7$ texts, there are $B_7 = 877$ possible output partitions. This number appears too large to elicit a subjective prior on $C(\theta)$. Instead, we assigned points to partitions $C(\theta)$ based on various desirable characteristics. Specifically, we began with $S(C(\theta)) = 0$. If the total number of output clusters in the final partition is T, we modified S according to $S(C(\theta)) \leftarrow S(C(\theta)) + 8 - T$. This has the effect of assigning greater scores when the resultant numbers of clusters is small. This is in keeping with a subjective belief that there are few authors who had the capability of producing the texts in question. We increased S by an additional point for each pair of the texts 1, 2 and 3 clustered in the output partition, and by another point for each of texts 4, 5 and 6 when they were not clustered with any of texts 1, 2 or 3. Finally, an extra point was given if text 4 was clustered with text 5. These additional points are simply illustrative of potential prior beliefs that one might hold concerning the texts. Under the given point scheme, a little analysis shows that $S(C(\theta))$ takes on the values $S = 4, \ldots, 11$ which is more manageable yet still reflects subjective beliefs on the output partitions.

For the importance sampler $h(\theta)$ in (10), we took advantage of the conjugacy mentioned earlier whereby $h(\theta) \sim f(y \mid \theta)\pi_\theta(\theta)$ gives a Dirichlet importance sampler. In the case of (11), we used the flat Dirichlet $h(\theta) = \pi_\theta(\theta)$ as the importance sampler. Both of these choices lead to simplications in the estimators (10) and (11).

There may be various inferential questions that are of interest to us, and the complexity of the questions poses no additional difficulties. For example, we may be interested in the probability that texts 1, 2 and 3 have a common author and that texts 4, 5 and 6 each have authors that are different from the authors of texts 1, 2 and 3. This event corresponds to the way in which the data were generated, and we hope that the posterior probability of the event is appreciable. The posterior probability that texts clustered in this manner is 0.47. To check the sensitivity of the empirical Bayes prior π_θ, we obtained the posterior probability 0.49 when changing $\lambda = 1.0$ to $\lambda = 0.9$. This suggests a lack of sensitivity with respect to the prior on the input parameter θ for which we typically do not have interpretable beliefs.

The probabilities reported above were based on $N = 50000$ simulations which required approximately 12.0 minutes of computation on a Sun Workstation. The probabilities can be regarded as accurate to the second decimal place. To investigate the effect of the output prior on ϕ described by S-scores, we modified the S-score to a prior given by $S(C(\theta)) = 20.0$ if the true partition is obtained and $S(C(\theta)) = 1.0$ otherwise. With the stronger prior, one would expect the posterior probability of the event described above to increase. In this case, the posterior probability increased to 0.98 when $\lambda = 1.0$. This suggests a sensitivity with respect to the output parameter ϕ for which we typically have interpretable beliefs.

4 EXAMPLE: 2010 FIFA WORLD CUP

The FIFA World Cup is a month-long international football (soccer) competition that is held once every four years. With over 200 international football nations governed by FIFA (Fédération Internationale de Football Association), only the host country South Africa and 31 qualifying nations competed in the prestigious 2010 edition of the FIFA World Cup. In the final match, Spain defeated the Netherlands 1-0 in extra time, giving Spain their first world title.

With the eyes of the world on South Africa, a lot of attention was given to the participant teams. In particular, teams were ranked, betting odds for matches were established, etc. Our focus was slightly different. Prior to the World Cup tournament, we wanted to know which of the 32 teams had a realistic chance of becoming World Cup champions. There is considerable uncertainty in the outcome of high-level football matches as the "best" teams do not always win. Reasons for the uncertainty include the dearth of good scoring opportunities, and the unpredictability of influential events such as offsides, red cards and hand balls. We wanted to take the uncertainty into account and restrict the field to only those teams that were capable of becoming champions.

The dataset which we considered involved all of the matches played between 41 teams of interest during the years 2008 and 2009 prior to the World Cup. This was an important time frame as as the teams were often competing in World Cup qualifying matches and were participating in important tournaments. Consequently, of the 216 matches, only 98 games were *friendlies* where teams may not have played to their upmost capabilities. Also, we chose not to go back further in time than two years since we wanted the dataset to reflect the current abilities of teams. The 41 teams under consideration consisted of the 32 World Cup qualifying nations plus the following 9 additional teams: Bahrain, Bosnia-Herzegovina, Costa Rica, Egypt, Gabon, Republic of Ireland, Russia, Tunisia and Ukraine. The 9 extra teams were added to augment the dataset. The additional teams were all strong sides that were involved in the qualification process right until the very end, and therefore these teams were involved in meaningful matches. In addition, the extra teams provided additional crossover information. For example, since Asian teams rarely play South American teams, it was useful to include teams that play both Asian and South American teams.

The statistical model which we implemented is due to Davidson (1970) and is an extension of the popular Bradley-Terry model (Bradley and Terry 1952). The model takes into account three outcomes; wins, losses and draws. We decided against models that take goal differential into account as results are sometimes distorted. For example, when a team is losing by a goal late in a match, the team often presses and ends up losing by two goals. Also, a bad loss by three goals may not be qualitatively much different than a loss by five goals. In the Bradley-Terry setup, the parameter θ_i is used to represent the strength of the i-th team. Therefore the θ's are population parameters, and this fits in line with our clustering philosophy. Following Davidson (1970), we have

$$\begin{aligned}
\text{Prob}(i \text{ defeats } j) &= \frac{\exp\{\theta_i + \beta\}}{\exp\{\theta_i + \beta\} + \exp\{\theta_j\} + \exp\{\alpha + (\theta_i + \theta_j + \beta)/2\}} \\
\text{Prob}(j \text{ defeats } i) &= \frac{\exp\{\theta_j\}}{\exp\{\theta_i + \beta\} + \exp\{\theta_j\} + \exp\{\alpha + (\theta_i + \theta_j + \beta)/2\}} \qquad (14) \\
\text{Prob}(\text{draw}) &= 1 - \text{Prob}(i \text{ defeats } j) - \text{Prob}(j \text{ defeats } i).
\end{aligned}$$

In (14), β is the home field parameter where $\beta = 0$ if the match was played on a neutral site, $\beta = \gamma$ if the match was played on i's home field and $\beta = -\gamma$ if the match was played on j's home field. Therefore a home field advantage is conferred when $\gamma > 0$. The parameter α affects the probability of drawn matches where larger values of α increase the frequency of draws. It may also be possible to include a covariate that distinguishes between friendly and non-friendly matches. However, we made no distinction as we take the view of the sporting maxim "you are what your record says you are". We also note that our dataset consisted of a majority of non-friendly matches and that it is difficult to assign relative weights to friendlies and non-friendlies.

The likelihood in this application deviates from the structure in (1) as it was formed from the product of the appropriate probabilities in (14) over the 216 matches. The likelihood is similar to that used by Hallinan (2005) in the ranking of football teams. To simplify the analysis, the nuisance parameters α and γ were set to the values obtained through maximization of the likelihood function.

The traditional clustering algorithm that was implemented is the popular UPGMA algorithm proposed by Lance and Williams (1966). For comparing two teams, the distance measure

$$d(\theta_i, \theta_j) \ = | \ \text{Prob}(i \text{ defeats } j) - \text{Prob}(j \text{ defeats } i) \ \ | \tag{15}$$

with home field parameter $\beta = 0$ is a natural choice. When calculating distances between clusters, the distance measure (15) is averaged over all distances between component θ's in one cluster to the other. We invoked a stopping rule in the clustering algorithm whenever the minimum cluster distance exceeded 0.5. The stopping rule stipulates that a team is not of the same quality as teams in another cluster when its probability of winning matches against such teams is less than 0.25. Likewise, a team is not of the same quality as teams in another cluster when its probability of winning matches against such teams exceeds 0.75. For a team in a lower cluster to win four consecutive matches against teams in the highest cluster (as might be needed in the knockout stage to win the World Cup), this probability would be less than $(0.25)^4 = 0.004$.

The model specification requires the prior $p(\theta)$ in (7) which incorporates knowledge on inputs and outputs. For this, we considered an empirical Bayes approach where diffuse normal probabilites are assigned to π_θ. Specifically, let x_i be the points assigned to the i-th team in the November 2009 FIFA Coca-Cola World Rankings (www.fifa.com/worldfootball/ranking). The World Rankings are based on four years of data with decreasing weight assigned to matches over successive years. In addition, the World Rankings are based on points awarded to matches where points depend on the match result, the importance of the match and the strength of the opposition. The mean of the normal distribution for the i-th team is then given by $\log(x_i / \sum_{j=1}^{41} x_j)$. Although the FIFA Coca-Cola points are based on seemingly sensible criteria, there is no interpretation associated with the complex point scheme other than the obvious rank ordering of the teams. It is therefore sensible that the normal variances should be large ($\sigma = 3.0$) to reflect our uncertainty concerning the θ's. With $\sigma = 3.0$, the prior probabilities in (14) cover the range of all realistic possibilities. The sole exception is θ_1 which we fix according to $\theta_1 = \log(x_1 / \sum_{j=1}^{41} x_j)$ since there is a nonidentifiability associated with (14). As for the S-scores, we have a much better prior understanding of the relative rankings of the teams. We assigned $S(C(\theta))$ where a greater S-score, $S = 1, \ldots, 10$, corresponds to a less varied permutation of the output partition $C(\theta)$ when compared to the FIFA Coca-Cola rankings. More specifically, we first obtain the average number of FIFA Coca-Cola points for teams within each cluster corresponding to the output partition $C(\theta)$. We then create a secondary partition which has the same cluster sizes as $C(\theta)$ and whose cluster members are determined according to the FIFA Coca-Cola rankings. For example, if the first cluster in the output partition $C(\theta)$ has five members, then the first cluster in the secondary partition has the five teams with the lowest Coca-Cola rankings. Then, for each team in the secondary partition, we calculate the absolute difference between its points and the average number of points for the corresponding cluster in the output partition. The absolute differences are summed over all teams where large totals indicate a greater deviation from the FIFA Coca-Cola rankings, for which lower S-scores are assigned.

For the importance sampler $h(\theta)$ in (10), we used independent normals for the θ's. The means of the normals were obtained by maximizing the likelihood function and the variances were set according to $\sigma = 1.0$. For importance sampling in the calcuation of (11), we used $h(\theta) = \pi_\theta(\theta)$ as the importance sampler. Various issues associated with the choice of importance samplers are discussed in Evans and Swartz (1995).

In Table 1, we present the results of the clustering methodology. Estimates of the form (10) were based on $N = 20000$ simulations which required approximately one hour of computation on a Sun Workstation. The table entries can be regarded as accurate to within one digit in the second decimal place. For each team, we calculated the posterior probability of inclusion in the highest ranking cluster. For teams with less than 10% probability, we designated these teams as non-contenders. There is no surprise that Brazil and Spain were designated as the most highly probable contenders as their form prior to the World Cup had been exceptional and this is consistent with the prior $p(\theta)$ derived from the FIFA Coca-Cola rankings. Traditional powers Argentina, England, Germany, the Netherlands and Italy also made our contenders list. What is somewhat surprising was the exclusion of 2006 World Cup finalist France and highly-ranked Portugal from the contenders list. In both cases, although the prior $p(\theta)$ supported their inclusion, the recent form of both France and Portugal expressed through the likelihood had not been up to standard. Perhaps the most surprising contender in Table 1 is Algeria. Algeria had the fewest matches (four) amongst the 41 teams in the dataset. In these matches, Algeria won three games, two against Egypt (one home match and one at a neutral site) and one against Uruguay at home. Their lone defeat was an away match versus Egypt. Although Egypt and Uruguay had respectable records, it might be argued that a stronger prior ought to have been placed on Algeria. A similar comment concerns the unexpectedly high probability assigned to North Korea. In their 6 matches, North Korea had one away loss (to South Korea) and five draws (four to South Korea and one to Japan).

In Table 1, we also provide a summary of the 2010 World Cup results. We observe that 10 of our 14 contenders qualified to the knockout stage. Moreover, all four semifinalists were included in our list of contenders.

5 DISCUSSION

The clustering approach presented in this paper has a number of key features. The methodology is based on hierarchical and partitional algorithms and therefore retains many of the appealing features of these traditional clustering algorithms. However, the approach is Bayesian which permits the quantification of the uncertainty in output partitions. Clustering is carried out on population characteristics rather than data, and this typically addresses the inferential question of interest in a more direct manner. As a by-product, one does not need to worry about missing data nor repeated measurements. In addition, the approach proposes a principled approach for combining prior information on inputs and outputs to the clustering algorithm.

As in any clustering algorithm, a user needs to specify a distance measure which describes closeness between populations. In the applications that we have studied, the θ's are primary parameters for which intuition is available and distance measures present themselves. A user does need to think carefully about the prior $p(\theta)$ in (7). Typically knowledge on the input component θ is vague although subjective knowledge exists on the output partitions ϕ. It is a matter of quantifying the subjective knowledge via S-scores. It is also necessary to think carefully about the choice of the importance sampler $h(\theta)$.

6 REFERENCES

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Bradley, R.A. and Terry, M.E. (1952). "Rank analysis of incomplete block designs I. The method of paired comparisons", *Biometrika*, 39, 324-345.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Davidson, R.R. (1970). "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments", *Journal of the American Statistical Association*, 65, 317-328.

Evans, M. and Swartz, T.B. (1995). "Methods for approximating integrals in Statistics with special emphasis on Bayesian integration", *Statistical Science*, 10, 254-272. Discussion in 1996, 11, 54-64.

Gill, P.S., Swartz, T.B. and Treschow, M. (2007). "A stylometric analysis of King Alfred's literary works", *Journal of Applied Statistics*, 34, 1251-1258.

Hallinan, S.E. (2005). "Paired comparison models for ranking national soccer teams", MSc project, Worcester Polytechnic Institute, Worcester, MA.

Holmes, D.I. and Forsyth, R.S. (1995). "The Federalist revised: New directions in author attribution", *Literary and Linguistic Computing*, 10, 111-127.

Holmes, D.I. (1999). "Stylometry", *Encyclopedia of Statistical Sciences: Update Volume 3*, Wiley, New York, 721-727.

Jeffreys, H. (1946). "An invariant form the prior probability in estimation problems", *Proceedings of the Royal Society of London, Series A*, 186, 453-461.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Lance, G.N. and Williams, W.T. (1966). "A general theory of classificatory sorting strategies: 1. Hierarchical systems", *The Computer Journal*, 9, 373-380.

Liu, H., Lafferty, J. and Wasserman, L. (2007). "Sparce nonparametric density estimation in high dimensions using the rodeo", *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Juan, Puerto Rico.

Mosteller, F. and Wallace, D.L. (1963). "Inference in an authorship problem. A comparative study of discrimination methods applied to the authorship of the disputed Federalist papers", *Journal of the American Statistical Association*, 58, 275-309.

Mosteller, F. and Wallace, D.L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Springer-Verlag, New York.

Poole, D. and Raftery, A.E. (2000). "Inference for deterministic simulation models: The Bayesian melding approach", *Journal of the American Statistical Association*, 95, 1244-1255.

Raftery, A.E., Givens, G.H. and Zeh, J.E. (1995). "Inference from a deterministic population dynamics model for bowhead whales", (with discussion), *Journal of the American Statistical Association*, 90, 402-430.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Rota, G-C. (1964). *The number of partitions of a set*, *American Mathematical Monthly*, 71, 498-504.

Rubin, D.B. (1988). "Using the SIR algorithm to simulate posterior distributions", In *Bayesian Statistics 3*, editors J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Clarendon Press, Oxford, 395-402.

Vermunt, J.K. and Magidson, J. (2002). "Latent class cluster analysis", In *Latent Class Analysis*, editors J.A. Hagenaars and A.L. McCutcheon, Cambridge University Press, Cambridge, 89-106.

Ward, J.H. (1963). "Hierarchical grouping to optimise an objective function", *Journal of the American Statistical Association*, 58, 236-244.

Wolpert, R.L. (1995). Comment on "Inference from a deterministic population dynamics model for bowhead whales", by A.E. Raftery, G.H. Givens and J.E. Zeh, *Journal of the American Statistical Association*, 90, 426-427.

Contenders		Non-Contenders		Knockout Stage
Team	Probability	Team	Probability	Teams
Brazil	0.97	Denmark	0.09	Spain$^{(1)}$
Spain	0.87	North Korea	0.09	Netherlands$^{(2)}$
Algeria	0.50	Paraguay	0.06	Germany$^{(4)}$
Argentina	0.48	Cameroon	0.06	Uruguay$^{(4)}$
England	0.36	United States	0.06	Paraguay$^{(8)}$
Chile	0.34	Japan	0.05	Argentina$^{(8)}$
Australia	0.29	Honduras	0.05	Ghana$^{(8)}$
Germany	0.27	Greece	0.03	Brazil$^{(8)}$
Uruguay	0.23	Nigeria	0.02	Portugal$^{(16)}$
Netherlands	0.14	Cote d'Ivoire	0.02	Japan$^{(16)}$
Mexico	0.14	France	0.02	Chile$^{(16)}$
Switzerland	0.11	Portugal	0.01	Slovakia$^{(16)}$
Italy	0.10	Slovakia	0.00	Mexico$^{(16)}$
South Korea	0.10	Serbia	0.00	England$^{(16)}$
		Ghana	0.00	United States$^{(16)}$
		Slovenia	0.00	South Korea$^{(16)}$
		New Zealand	0.00	
		South Africa	0.00	

Table 1: Posterior probabilities of inclusion for teams that were deemed potential 2010 FIFA World Cup champions. For comparison, the 16 teams that advanced to the knockout stage are listed where the superscript (i) denotes that a team advanced to the group of size i.