

# A grade of membership model for rank data

Isobel Claire Gormley\* and Thomas Brendan Murphy†

**Abstract.** A grade of membership (GoM) model is an individual level mixture model which allows individuals have partial membership of the groups that characterize a population. A GoM model for rank data is developed to model the particular case when the response data is ranked in nature. A Metropolis-within-Gibbs sampler provides the framework for model fitting, but the intricate nature of the rank data models makes the selection of suitable proposal distributions difficult. ‘Surrogate’ proposal distributions are constructed using ideas from optimization transfer algorithms. Model fitting issues such as label switching and model selection are also addressed.

The GoM model for rank data is illustrated through an analysis of Irish election data where voters rank some or all of the candidates in order of preference. Interest lies in highlighting distinct groups of voters with similar preferences (i.e. ‘voting blocs’) within the electorate, taking into account the rank nature of the response data, and in examining individuals’ voting bloc memberships. The GoM model for rank data is fitted to data from an opinion poll conducted during the Irish presidential election campaign in 1997.

**Keywords:** Grade of membership models; Plackett-Luce model; surrogate proposal distributions; rank data; voting blocs.

## 1 Introduction

The use of mixture models as a flexible model-based clustering tool is well established both in theory and in practice (Fraley and Raftery 2002). Mixture models describe a population as a finite collection of homogeneous groups, each of which is characterized by a specific probability density. While based on a similar concept, grade of membership (GoM) models allow every individual have partial membership of each of the groups that characterize the population. Thus GoM models have the capability of providing a soft clustering of the population members.

Typically mixture models are fitted via the EM algorithm (eg. McLachlan and Peel 2000; Fraley and Raftery 2002); one advantage of this approach is that, at convergence, the algorithm provides estimates of the posterior conditional probability of the group membership for each individual. The posterior conditional probability estimates can be used to cluster individuals into groups, thereby achieving a model-based clustering. The GoM model (Erosheva 2003) provides similar group membership probabilities but in this case the probabilities are direct parameters of the model.

---

\*School of Mathematical Sciences, University College Dublin, Ireland, <mailto:Claire.Gormley@ucd.ie>

†School of Mathematical Sciences, University College Dublin, Ireland, <mailto:brendan.murphy@ucd.ie>

In this article the GoM model for rank response data is developed; Erosheva (2002) develops the GoM model for multivariate categorical data. Rank data arise when a set of judges rank some (or all) of a set of objects. Rank data emerge in many areas of society; the final ordering of horses in a race, the ranking of relevant web pages by internet search engines and consumer preference data provide examples of such data. Irish society generates a wealth of rank data as under the Irish electoral system (the Single Transferable Vote) voters rank candidates in order of preference. When drawing inferences from such data, the information contained in the different preference levels must be exploited by the use of appropriate modeling tools.

An illustration of the GoM model for rank data methodology is provided through an examination of voting data from the last Irish presidential election in 1997. Interest lies in highlighting distinct groups of voters with similar preferences (i.e. ‘voting blocs’) within the electorate, taking into account the rank nature of the response data. The preferences that the voters have within these voting blocs are also of interest. Additionally the GoM model provides the scope to examine the voting bloc memberships of each individual voter, which this model allows to be mixed across voting blocs.

A latent class representation of the GoM model for rank data is used for model fitting within the Bayesian paradigm. A Metropolis-within-Gibbs sampler is necessary to provide samples from the posterior distribution. Difficulties arise in the Metropolis step of the algorithm however, as the specification of an appropriate proposal distribution is challenging due to the intricate nature of the rank data model. Surrogate proposal distributions, which are updated at each iteration, are constructed via ideas which underpin optimization transfer algorithms (Lange et al. 2000); the method provides suitable and tractable proposal distributions.

The article proceeds as follows: in Section 2 the Irish voting system and details surrounding the 1997 Irish presidential election are addressed. The Plackett-Luce model for rank data is employed in this application as the rank data model; this model and other rank data models are discussed in Section 3.1. The specification of the GoM model for rank data follows in Section 3.2. Estimation of the GoM model for rank data, and details of the construction of surrogate proposal distributions, are detailed in Section 4. Model fitting issues such as the label switching phenomenon and the question of model dimensionality are addressed in Section 5. An illustrative application of the GoM model for rank data to Irish election data is given in Section 3.2. The article concludes in Section 7 with a discussion of the methodology and some proposals for future directions.

## **2 The 1997 Irish presidential election.**

Irish presidential elections employ an electoral system known as the Single Transferable Vote (STV) system; a similar system known as Proportional Representation by means of a Single Transferable Vote (PR-STV) is used in Irish governmental elections. Under these electoral systems a voter ranks, in order of his/her preference, some or all of the electoral candidates on a ballot form. The votes are totalled through a series of

Table 1: The five candidates who ran for the Irish presidential seat in 1997 and their endorsing political parties. Mary McAleese was subsequently elected.

Candidate	Endorsing Party
Mary Banotti	Fine Gael
Mary McAleese	Fianna Fáil
Derek Nally	Independent
Adi Roche	Labour
Rosemary Scallon	Independent

counts, where candidates are eliminated, their votes are distributed, and surplus votes are transferred between candidates. An illustrative example of the manner in which votes were counted and transferred in the 1997 Irish presidential election is provided in [Gormley and Murphy \(2008b\)](#).

The Republic of Ireland has a semi-presidential system in that the head of state (the President) is not the same person as the head of government (An Taoiseach). The eighth (and current) President of Ireland, Mary McAleese, was originally elected in 1997. The number of candidates in the 1997 presidential election was larger than in previous campaigns. There were five candidates that year: Mary Banotti, Mary McAleese, Derek Nally, Adi Roche, and Rosemary Scallon. As the President is not the head of government presidential candidates are not necessarily members of political parties ([van der Brug et al. 2000](#)). However, in 1997, some candidates were endorsed by political parties and others were independent candidates (see Table 1). In 1997 Fianna Fáil were the governing political party, with Fine Gael the main opposition party. The presidential candidates backed by these parties typically had high public profiles. The Labour party was another strong opposition party in 1997 and thus their supported candidate, Adi Roche, also had a large share of the media attention. The candidates who ran on independent tickets, Scallon and Nally, had somewhat lower public profiles. Mary Banotti, Derek Nally and Adi Roche were considered to be liberal candidates where Mary McAleese and Rosemary Scallon were deemed the more conservative candidates. It is also worth noting that Derek Nally entered the election race at a later stage than the other four candidates. A detailed description of the entire presidential election campaign, including the nomination and selection of candidates, is given by [Marsh \(1999\)](#). Good introductions to the Irish political system are given by [Coakley and Gallagher \(1999\)](#), [Sinnott \(1995\)](#) and [Sinnott \(1999\)](#).

Seven opinion polls and an exit poll, conducted on polling day, were completed during the election campaign. This article focusses on the exit poll, conducted by Lansdowne Market Research, in which 2498 voters were asked how they voted at 150 polling stations in all 41 Irish constituencies. Data from these polls have been previously analysed by [Gormley and Murphy \(2006a, 2008a,b\)](#). The sources of all the poll data are given in Appendix A.

### 3 Model specification

Irish voting data possess some unique properties which require careful statistical modeling. The Grade of Membership (GoM) model is used to model the heterogeneity within the electorate; the Plackett-Luce model for rank data is incorporated with the GoM model to account for the ranked nature of the preferences expressed by the voters. Both aspects of the model are developed in this section.

#### 3.1 The Plackett-Luce model for rank data

Under the STV electoral system a voter ranks some or all of the candidates in order of preference. Table 2 illustrates three typical votes from the 1997 Irish presidential election exit poll. Each vote is a rank data point which can be viewed as reflecting the support that the voter has for each candidate.

Table 2: Three sample votes from the 1997 Irish presidential election exit poll. Voter A chose to express only one preference, voter B expressed all five preferences and Voter C expressed two preferences.

	First preference	Second preference	Third preference	Fourth preference	Fifth preference
Voter A	McAleese				
Voter B	Banotti	McAleese	Scallon	Roche	Nally
Voter C	Banotti	Scallon			

In order to appropriately model such election data, a model for rank data is required. A suite of rank data models are available; [Marden \(1995\)](#) provides an excellent review. The Bradley-Terry model ([Bradley and Terry 1952](#)) examines competition between a set of individuals as a set of pairwise comparisons from which an “ability parameter” can be inferred and thus a ranking of the competitors can be formed. [Mallows \(1957\)](#) provides a model where the probability of a ranking decreases as the distance from a central ranking increases; the extension of Mallow’s model to partial rankings is described in detail by [Critchlow \(1985\)](#). [Thurstone \(1927\)](#) uses an order statistic model to describe the ranking procedure, where each object is assigned a random score and the ranking of the item scores is the observed ranking; such models are also called random utility models ([Train 2003](#)). [Chapman and Staelin \(1982\)](#) detail a random utility model with doubly exponentiated errors termed a stochastic utility model. This model gives rise to an ‘exploding’ likelihood in which each term of the explosion is a multinomial choice probability. [Bradlow and Fader \(2001\)](#) develop a times series version of the stochastic utility model, and detail model estimation within the Bayesian paradigm. More recently, [Graves et al. \(2003\)](#) use a combination of the Bradley-Terry, Luce and Stern models for rank data to estimate driver ability in auto car racing.

In this application the Plackett-Luce model ([Plackett 1975](#)) is utilized to model the

rank nature of the data. The Plackett-Luce model was originally developed within the context of horse racing — the model was used to model the probability of the final permutation of horses at the end of a race. A STV ballot form can be thought of in a similar manner to that of the final permutation of horses in a race. For example, not all horses listed to partake in a race necessarily finish; similarly not all candidates must be ranked by a voter. Also, once a candidate has been chosen he or she cannot be selected again. The similarities between the final permutation of horses in a race and an STV ballot form suggest similar models may be useful in both contexts.

The Plackett-Luce model accounts for the construction of a ranking as a sequential process where the next most preferred candidate is selected from the current choice set. [Fligner and Verducci \(1988\)](#) refer to models of this form as multistage models. Specifically, the Plackett-Luce model formulates the probability of a voter’s ranked preferences as the product of the conditional probabilities of each choice; it models the ranking of candidates by a voter as a set of independent choices by the voter, conditional on the fact that the cardinality of the choice set is reduced by one after each choice. For example, consider a voter in the current application who chooses to rank three candidates on a ballot form, McAleese, then Banotti and then Roche. The Plackett-Luce model determines the probability of this rank data point as the probability of choosing McAleese from the full set of candidates {Banotti, McAleese, Nally, Roche, Scallon}, times the probability of choosing Banotti from the reduced set of candidates {Banotti, Nally, Roche, Scallon}, times the probability of choosing Roche from the set of candidates {Nally, Roche, Scallon}. Such a model is referred to as *exploded logit* models in the discrete choice modeling literature ([Train 2003](#)) since each rank data point is exploded into several (conditionally) independent choices or ‘psuedo-observations’.

The Plackett-Luce model is parameterized by a ‘support’ parameter

$$\underline{p} = (p_1, p_2, \dots, p_N)$$

where  $N$  denotes the total number of electoral candidates. Note that  $0 \leq p_j \leq 1$  and  $\sum_{j=1}^N p_j = 1$ . The parameter  $p_j$  is interpreted as the probability of candidate  $j$  being ranked first by a voter. The probability of candidate  $j$  being given a lower than first preference is proportional to their support parameter  $p_j$ . At preference levels lower than the first the probabilities are normalized to provide valid probability values. Given the notation

- $c_{(i,t)}$  = candidate in preference level  $t$  in vote  $i$
- $n_i$  = number of preferences expressed by voter  $i$

the Plackett-Luce model states that the probability of vote  $\underline{x}_i$ , is:

$$\begin{aligned} \mathbf{P}\{\underline{x}_i | \underline{p}\} &= \prod_{t=1}^{n_i} \frac{p_{c(i,t)}}{p_{c(i,t)} + p_{c(i,t+1)} + \dots + p_{c(i,N)}} \\ &= \prod_{t=1}^{n_i} \frac{p_{c(i,t)}}{\sum_{s=t}^N p_{c(i,s)}} = \prod_{t=1}^{n_i} q_{c(i,t)}, \end{aligned} \tag{1}$$

where  $c(i, n_i + 1), \dots, c(i, N)$  is any permutation of the unchosen candidates. It can be shown that (1) sums to 1 over all  $n_i!$  possible permutations of  $\underline{x}_i = \{c(i, 1), \dots, c(i, n_i)\}$ .

### 3.2 The grade of membership model

GoM models allow every individual in a population have partial membership of each of the homogeneous groups that characterize the population. A soft clustering of the population members is therefore achievable. The GoM model originally appears in the context of medical diagnosis problems where it is employed to characterize sub-patterns of disease (Woodbury et al. 1978). Early parameter estimation methods for the GoM model are maximum likelihood based. The GoM model is reformulated as a hierarchical Bayesian model by Erosheva (2002); a similar hierarchical model, latent Dirichlet allocation, is developed by Blei et al. (2003). Joutard et al. (2008) discuss model choice within the context of hierarchical Bayesian mixed-membership models, of which the GoM model is a special case. In this article a GoM model for rank data is developed within a Bayesian framework; Erosheva (2003) estimates the GoM model for multivariate categorical data in a similar manner.

Under the GoM model each individual  $i = 1, \dots, M$  has an associated *GoM score* or *mixed-membership parameter*  $\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$  which is a direct parameter of the model. The mixed-membership parameter  $\underline{\pi}_i$  describes the degree of membership of individual  $i$  in each of the  $K$  groups which characterize the electorate. Note that  $0 \leq \pi_{ik} \leq 1$  and  $\sum_{k=1}^K \pi_{ik} = 1$  for  $i = 1, \dots, M$ .

The Plackett-Luce model is combined with the GoM model to model the rank nature of the response data. Within the GoM model framework, the Plackett-Luce support parameter  $p_{kj}$  is the conditional probability of candidate  $j$  being ranked first, in the extreme case when the voter's mixed-membership parameter for group  $k$  is equal to 1 (i.e.  $\underline{\pi}_i = (0, \dots, 1, \dots, 0)$ ). The main assumption of the GoM model for rank data is the convexity of these conditional support parameters at each preference level. That is, the probability of voter  $i$  choosing candidate  $j$  at preference level  $t$ , conditional on voter  $i$ 's mixed-membership parameter, is:

$$\mathbf{P}\{c(i, t) = j | \underline{\pi}_i\} = \sum_{k=1}^K \pi_{ik} \frac{p_{kj}}{\sum_{s=t}^K p_{kc(i, s)}} = \sum_{k=1}^K \pi_{ik} q_{kc(i, t)} \quad (2)$$

where  $q_{kc(i, t)}$  is given by (1). Additionally, local independence is then assumed between each preference level  $t$ , given the mixed-membership parameters. The likelihood function based on the data  $\mathbf{x}$  therefore is

$$\mathbf{P}\{\mathbf{x} | \boldsymbol{\pi}, \mathbf{p}\} = \prod_{i=1}^M \prod_{t=1}^{n_i} \left\{ \sum_{k=1}^K \pi_{ik} q_{kc(i, t)} \right\}.$$

Note that under the GoM framework each voter has partial membership of each group and that mixing takes place at each preference level  $t$  rather than at the vote level as

would be typical of a rank data mixture model (Gormley and Murphy 2006b, 2008a,b). Modeling rank data in this manner provides a deeper insight to the structure within the electorate by allowing mixing to occur at a finer level. This is a desirable characteristic as it may be restrictive to assume a voter expresses all preferences in their vote as dictated by a single group; it is likely that a voter may express some preferences in line with the support parameters of one group, and other preferences in line with the support parameters of other groups.

For example, suppose that the electorate is characterized by two groups — a ‘pro-McAleese’ group which has low support for all other candidates and a ‘pro-Banotti’ group which has good support for Banotti and some support for all other candidates except McAleese. A frequent vote recorded in the exit poll consisted of three preferences — McAleese ranked first, Banotti ranked second and Roche ranked third. It seems intuitive to allow such voters have partial membership of both groups and to model their choices within their vote according to the support parameters of both groups. It seems less intuitive to force such voters to belong to one group alone and model their entire vote according to the support parameters of that group.

### The GoM model and latent class models

To provide further insight to the interpretation of the GoM model for rank data it is worthwhile considering the relationship between the well known latent class model (LCM) for discrete data (Lazarsfeld and Henry 1968; Bartholomew and Knott 1999) and the GoM model. Both models are latent structure models.

Latent class models have discrete latent variables, reflecting the assumption that each individual is a full member of one of the latent classes. Hence the conditional support parameter analogous to (2) would be independent of preference level  $t$  under the latent class model i.e. if voter  $i$  is a member of latent class  $k$  they express their entire set of preferences according to the support parameter  $\underline{p}_k$ .

GoM models on the other hand have continuous latent variables, reflecting the assertion that individuals may have partial membership of more than one group. The set of candidates a voter ranks on his/her ballot is determined by the support parameters of the groups of which the voter has partial membership. The mixed-membership parameter  $\pi_{ik}$  can be interpreted as voter  $i$  expresses an average proportion of  $\pi_{ik}$  preferences according to the support parameters of the  $k$ th group.

Given this description of the LCM and the GoM model, the LCM with  $K$  classes is a special case of the  $K$  group GoM model, in that the discrete LCM latent membership vector is a constrained version of the continuous GoM mixed-membership parameter. Haberman (1995) however suggests a latent class representation of the GoM model. This representation provides a view of the GoM model as a special case of the LCM in which the equality constraint on the number of classes and groups is relaxed. Haberman (1995) illustrates that the marginal distribution of the observed data is exactly the same under the standard form of the GoM model and under the latent class representation of the GoM model. Erosheva (2006) deals thoroughly with the relations between the GoM

model and the LCM for the specific case of polytomous categorical response variables.

### A latent class representation of the GoM model

As in Erosheva (2003), a latent class representation of the GoM model for rank data is considered here to provide insight to unobservable underlying phenomena. The latent class representation of the GoM model for rank data involves augmenting the data with categorical latent variables. The discrete distribution on the latent classes is then given by a functional form of the continuous distribution of the GoM mixed-membership parameters (Erosheva 2006).

For each individual  $i$ , binary latent vectors  $\mathbf{z}_{it} = (z_{it1}, \dots, z_{itK})$  are imputed for  $t = 1, \dots, n_i$  where  $\mathbf{z}_{it} \sim \text{Multinomial}(1, \underline{\pi}_i)$ . These imputed binary latent variables define the latent classes. For a  $K$  group GoM model, there are  $K^N$  latent classes in the latent class representation of the GoM model. Specific combinations of preferences expressed according to each GoM group correspond to particular latent classes in the latent class representation of the GoM model. While each individual is considered to have partial membership of the GoM groups, under the latent class representation, they are considered to be a complete member of one of the  $K^N$  latent classes. The realization of the latent variable  $\mathbf{z}_{it}$  determines the group whose support parameters determine the choice made at preference level  $t$  within voter  $i$ 's ballot.

It follows that under the GoM model the 'augmented' data likelihood function based on the data  $\mathbf{x}$  and the binary latent variables  $\mathbf{z}$  is therefore:

$$\mathbf{P}\{\mathbf{x}, \mathbf{z} | \underline{\pi}, \mathbf{p}\} = \prod_{i=1}^M \prod_{k=1}^K \prod_{t=1}^{n_i} \{\pi_{ik} q_{kc(i,t)}\}^{z_{itk}} \quad (3)$$

Employing the GoM model, while modeling the rank data via the Plackett-Luce model, not only allows estimation of the characteristic parameters of each group but also direct estimation of the mixed-membership parameter for each individual. Hence a soft clustering of the population can be achieved.

### 3.3 Prior and posterior distributions

A Bayesian approach is taken when estimating the GoM model for rank data and thus the specification of prior distributions for the parameters of the model is required. It is assumed that the mixed-membership parameters follow a Dirichlet( $\underline{\alpha}$ ) distribution and that the support parameters follow a Dirichlet( $\underline{\beta}$ ) distribution i.e.

$$\begin{aligned} \underline{\pi}_i &\sim \text{Dirichlet}\{\underline{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)\} \\ \underline{p}_k &\sim \text{Dirichlet}\{\underline{\beta} = (\beta_1, \beta_2, \dots, \beta_N)\}. \end{aligned}$$

The conjugacy of the Dirichlet distribution with the multinomial distribution means the use of a Dirichlet prior is naturally attractive. The use of a Dirichlet prior does

however induce a negative correlation structure between parameters. The sensitivity of inferences drawn under the GoM model for rank data to this prior specification is considered in Section 6.3. In practice the prior parameters are fixed as  $\underline{\alpha} = (0.5, \dots, 0.5)$  and  $\underline{\beta} = (0.5, \dots, 0.5)$  which is the Jeffreys prior for the multinomial distribution (O’Hagan and Forster 2004). Pritchard et al. (2000) employ a Dirichlet prior in a similar context but place hyperpriors on the Dirichlet parameters. Erosheva et al. (2007) discuss the pitfalls of employing fixed values of the hyperparameters without sufficient prior knowledge and suggest a reparametrisation of the hyperparameters as a solution in such a case. In any case, we explored the use of alternative values of  $\underline{\alpha}$  and  $\underline{\beta}$  and the results were robust to the choice for a reasonable range of values (Section 6.3).

Given these prior distributions and the augmented data likelihood function (3) from the GoM model for rank data, the posterior distribution based on the data is:

$$\mathbf{P}\{\boldsymbol{\pi}, \mathbf{p} | \mathbf{x}, \mathbf{z}\} \propto \left[ \prod_{i=1}^M \prod_{k=1}^K \prod_{t=1}^{n_i} \{\pi_{ik} q_{kc(i,t)}\}^{z_{itk}} \right] \left[ \prod_{i=1}^M \prod_{k=1}^K \pi_{ik}^{\alpha_k - 1} \right] \left[ \prod_{k=1}^K \prod_{j=1}^N p_{kj}^{\beta_j - 1} \right].$$

This posterior distribution differs from the posterior distribution in the case of the original GoM model (Erosheva 2002, 2003) in the form of the likelihood function. In the original GoM model, discrete response variables are treated as independent given the mixed-membership parameters. The likelihood function is therefore the product of independent Bernoulli distributions. In the GoM model for rank data however, the dependence of choices within a rank response leads to a more complex likelihood function that is the product of terms that share parameter values.

### 4 Parameter estimation

Due to the intricate nature of the posterior distribution, Markov chain Monte Carlo methods are necessary to produce realizations of the model parameters. In particular, the Gibbs sampler algorithm can be employed if the full conditional distributions for all model parameters are available for sampling.

The full conditional distributions of the latent variables  $z_{it}$  and the mixed-membership parameters  $\underline{\pi}_i$  are readily available i.e.

$$z_{it} \sim \text{Multinomial} \left\{ 1, \left( \frac{\pi_{i1} q_{1c(i,t)}}{\sum_{k'=1}^K \pi_{ik'} q_{k'c(i,t)}}, \frac{\pi_{i2} q_{2c(i,t)}}{\sum_{k'=1}^K \pi_{ik'} q_{k'c(i,t)}}, \dots, \frac{\pi_{iK} q_{Kc(i,t)}}{\sum_{k'=1}^K \pi_{ik'} q_{k'c(i,t)}} \right) \right\}$$

for  $i = 1, \dots, M, \quad t = 1, \dots, n_i$  and

$$\underline{\pi}_i \sim \text{Dirichlet}(\alpha_1 + \sum_{t=1}^{n_i} z_{it1}, \dots, \alpha_K + \sum_{t=1}^{n_i} z_{itK}) \quad \text{for } i = 1, \dots, M.$$

In the case of the support parameters, the full conditional distributions are

$$\mathbf{P}\{\underline{p}_k | \boldsymbol{\pi}, \mathbf{x}, \mathbf{z}\} \propto \left[ \prod_{i=1}^M \prod_{t=1}^{n_i} \left\{ \frac{\pi_{ik} p_{kc(i,t)}}{\sum_{s=t}^N p_{kc(i,s)}} \right\}^{z_{itk}} \right] \left[ \prod_{j=1}^N p_{kj}^{\beta_j - 1} \right]. \quad (4)$$

Due to the form of the likelihood function based on the rank data, the complete conditional distribution of the support parameters is not readily available for sampling and a straight forward Gibbs sampler algorithm cannot be fully implemented.

#### 4.1 The Metropolis-within-Gibbs sampler

Different MCMC algorithms may be combined to draw on and accumulate their individual strengths. The Metropolis-within-Gibbs (or the univariate Metropolis) algorithm imbeds  $T$  Metropolis steps within an outer Gibbs sampler algorithm. Generally  $T = 1$  is used which in effect simply substitutes a Metropolis step for a Gibbs step. [Carlin and Louis \(2000\)](#) detail the conditions necessary for the convergence of such a hybrid algorithm.

In any Metropolis-based algorithm, the rate of convergence of the chain depends on the relationship between the proposal and target distributions. The use of a proposal distribution which is closely related to the shape and orientation of the target distribution provides an improved rate of convergence and good mixing. Additionally, a proposal distribution which is easy to sample from is preferable. Choosing a suitable proposal distribution for a complex target distribution, such as (4), is therefore difficult and is often done in an ad hoc manner.

To implement a Metropolis step to sample support parameter values a suitable proposal distribution is required. A satisfactory proposal distribution for the target distribution (4) is not initially apparent. One possibility examined was to approximate the full conditional density (4) using Rosén's approximation ([Rosén 1972](#)) to the Plackett-Luce model. Rosén derived an approximation which states that when  $n_i \ll N$  and the values in  $\underline{p}$  are not too variable,

$$\mathbf{P}\{\underline{x}_i | \underline{p}\} \approx p_{c(i,1)} p_{c(i,2)} \cdots p_{c(i,n_i)},$$

that is, under these conditions the Plackett-Luce probability of a vote  $\underline{x}_i$  may be approximated by the product of the support parameters for the ranked candidates. Under this approximation the distribution of the support parameter  $\underline{p}_k$  is a Dirichlet distribution which would provide the basis for a suitable and tractable proposal distribution. This approximation approach provided a poor proposal distribution however, demonstrated by a lack of mixing in the chain. The poor performance of this approximation approach can be attributed to the fact that the conditions required for Rosén's approximation are not satisfied in this case i.e. in this application  $n_i$  is often close to  $N$  and the values in  $\underline{p}$  are very variable. A more satisfactory proposal distribution is therefore required. One solution is to employ a 'surrogate proposal distribution'.

### Surrogate proposal distributions

Optimization transfer via surrogate objective functions (or the MM algorithm) (Lange et al. 2000) is an optimization tool which operates by creating a surrogate function for a problematic objective function which requires optimization. Different approaches are taken to construct the necessary surrogate function depending on the form of the problematic objective function. Iteratively maximizing a minorizing surrogate function, for example, produces a sequence of new parameter estimates which converges to a local maximum of the objective function. Thus in a maximization problem the initials MM stand for minorize/maximize. (In a minimization problem MM stands for majorize/minimize.) It emerges that the well known EM algorithm (Dempster et al. 1977) is in fact a special case of the MM algorithm. Good practical examples of the MM algorithm and details of the relationship between the EM and MM algorithms are provided by Lange et al. (2000) and Hunter and Lange (2004).

In a similar vein, tractable proposal distributions for use in a Metropolis step may be formed by the construction of a surrogate function for a complex full conditional distribution. The approach taken to construct a surrogate proposal distribution depends on the mathematical form of (a part of) the complete conditional distribution. For example, a typical technique used in the MM algorithm literature is to exploit the supporting hyperplane property of a convex function: for convex function  $f(\theta)$  with differential  $df(\theta)$ :

$$f(\theta) \geq f(\bar{\theta}) + df(\bar{\theta})(\theta - \bar{\theta}) \quad (5)$$

where  $\bar{\theta}$  denotes a constant value of the parameter  $\theta$ . This inequality provides a linear minorizing surrogate function of  $f(\theta)$  which may be more tractable than the original function. Alternative approaches can be taken to provide quadratic or higher order surrogate functions (see Hunter and Lange (2004)).

Here the technique of constructing a surrogate function, and iteratively updating it, is borrowed to form a suitable and tractable proposal distribution when sampling the Plackett-Luce support parameters. Taking logs of the full conditional of the support parameters (4) gives

$$\log \mathbf{P}\{\underline{p}_k | \boldsymbol{\pi}, \mathbf{x}, \mathbf{z}\} \propto \sum_{i=1}^M \sum_{t=1}^{n_i} z_{itk} \left\{ \log p_{kc(i,t)} - \log \sum_{s=t}^N p_{kc(i,s)} \right\} + \sum_{j=1}^N (\beta_j - 1) \log p_{kj}. \quad (6)$$

The function  $-\log(\cdot)$  is a convex function and thus the supporting hyperplane property (5) can be applied to the complex term  $-\log \sum_{s=t}^N p_{kc(i,s)}$  in (6). The resulting function, a minorizing surrogate function for the log of the full conditional, can then be used as a proposal distribution. Full details are provided in Appendix B. The surrogate proposal distribution emerges as a Normal density with mean and variance dependent on the previously sampled values of the model parameters. Each time the Metropolis step occurs within the Metropolis-within-Gibbs sampler the surrogate proposal distribution is updated to depend on the previously sampled values of the model parameters. This methodology thus provides both a suitable and tractable proposal distribution selected in a theoretically sound manner.

As the Normal distribution extends beyond the  $[0, 1]$  interval in which the support parameters lie, proposed values from this surrogate proposal must be suitably normalized. This adjustment is dealt with within the Metropolis step as detailed in Appendix B.

## 5 Model features

When estimating parameters via MCMC algorithms some special features of the GoM model require attention. A fundamental issue in the fitting of any mixture based model within a Bayesian framework is that of label switching i.e. the invariance of posterior distributions to permutations in the labeling of the homogeneous groups. Another obvious issue is inferring the correct number of groups present in the population. Both features of the GoM model for rank data are dealt with in this section.

### 5.1 Label switching

The likelihood function based on the GoM model for rank data is invariant under re-labeling of the homogeneous groups of the population. If the prior distribution does not discriminate between the homogeneous groups then the posterior distribution will be symmetric and thus estimating model parameters by their posterior mean is inappropriate. [Jasra et al. \(2005\)](#) provide an overview of identifiability issues within Bayesian mixture models and of currently popular solutions to the problem. One approach ([Richardson and Green 1997](#)) minimizes label switching by imposing artificial identifiability constraints such as ordering the mixing proportions or other model parameters. Which parameters or combination thereof on which to base the ordering, and indeed selecting the ordering itself, is somewhat ad hoc however. [Celeux \(1998\)](#) details a clustering approach — the points in the MCMC sampler are permuted via a clustering algorithm and the permutation closest to a chosen standard is selected. Relabeling strategies using a decision theoretic approach as proposed by [Celeux et al. \(2000\)](#) and [Stephens \(2000\)](#) are implemented here.

A decision theoretic approach involves defining a loss function  $\mathcal{L}(\hat{\mathbf{p}}; \mathbf{p})$  which quantifies the loss incurred by choosing  $\hat{\mathbf{p}}$  when the true parameter value is  $\mathbf{p}$ . The aim is thus to minimize the posterior expected loss  $\mathbf{E}\{\mathcal{L}(\hat{\mathbf{p}}; \mathbf{p})|\mathbf{x}\}$ . In this case the support parameters  $\mathbf{p}$  of the Plackett-Luce model are the parameters used to rectify the label switching issue. Their reference value is set to be the *maximum a posteriori* (MAP) estimate  $\tilde{\mathbf{p}}$  obtained after a number of initial uphill only moves in the Metropolis step, subsequent to a burn-in period of the Markov chain. This MAP value is used as the template to which each estimate at the  $t^{th}$  iteration  $\hat{\mathbf{p}}^t$  will be ‘matched’ to correct for any label switching that may occur during estimation. A sum of squares function is employed as the loss function to be minimized:

$$\mathcal{L}(\hat{\mathbf{p}}; \tilde{\mathbf{p}}) = \sum_{k=1}^K \sum_{j=1}^N (\hat{p}_{kj}^t - \tilde{p}_{kj})^2.$$

Thus, once the MAP estimate  $\tilde{\mathbf{p}}$  has been obtained, following each Metropolis step the rows of the matrix  $\hat{\mathbf{p}}^t$  are permuted until the loss function is minimized. For large values of  $K$ , searching for the optimal permutation becomes computationally expensive. However in cases where interest lies in many label dependent quantities (as is true here) relabelling approaches are deemed to be the preferred approach from currently available solutions (see [Jasra et al. \(2005\)](#)). Alternative approaches such as those based on the introduction of artificial identifiability constraints suffer from similar problems when  $K$  is large. An online algorithm ([Stephens 2000](#)) is utilized here and is detailed in Appendix C. It follows that label switching will be minimized thus ensuring the validity of posterior estimates.

## 5.2 Model selection

Another feature of the GoM model is the need to infer the model dimensionality i.e. the value  $K$ , the number of groups present in the population. Within the Bayesian paradigm the natural approach would appear to be to base inference on the posterior distribution of  $K$  given the data  $\mathbf{x}$ ,  $\mathbf{P}\{K|\mathbf{x}\}$ . However this posterior can be strangely dependent on the model definition and is typically computationally challenging to construct. [Joutard et al. \(2008\)](#) provide a comprehensive overview and comparison of model selection criteria within the context of GoM models.

In this application of the GoM model for rank data, the Deviance Information Criterion (DIC) introduced by [Spiegelhalter et al. \(2002\)](#) is used to provide a measure of model fit. It penalizes the posterior mean deviance of a model by the ‘effective number of parameters’. The effective number of parameters is derived to be the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. Explicitly for data  $\mathbf{x}$  and parameters  $\theta$  the DIC is

$$DIC = \overline{D(\theta)} + p_D$$

where  $D(\theta) = -2\log(\mathbf{P}\{\mathbf{x}|\theta\}) + 2\log\{h(\mathbf{x})\}$  is the Bayesian deviance and  $h(\mathbf{x})$  is a function of the data only. The effective number of parameters is defined as  $p_D = \overline{D(\theta)} - D(\bar{\theta})$ . The criterion has an approximate decision theoretic justification. Models with smaller DIC values are preferable.

## 6 Application to the Irish electorate

The GoM model for rank data was applied to the preferences expressed in an exit poll conducted on the day of the 1997 presidential election. The Metropolis-within-Gibbs sampler was run over 50000 iterations, with a burn-in period of 10000 iterations, over the range  $K = 1, \dots, 5$  voting blocs. Dirichlet priors with  $\underline{\alpha} = (0.5, \dots, 0.5)$  and  $\underline{\beta} = (0.5, \dots, 0.5)$  were imposed on the mixed-membership variables and the support parameters respectively.

Figure 1 illustrates the model selection criterion values obtained when fitting the GoM model for rank data to the 1997 presidential exit poll data. The DIC suggests an

electorate composed of four voting blocs. A previous analysis of this data ([Gormley and Murphy 2008a](#)) suggests a mixture model with four Plackett-Luce models is appropriate here.

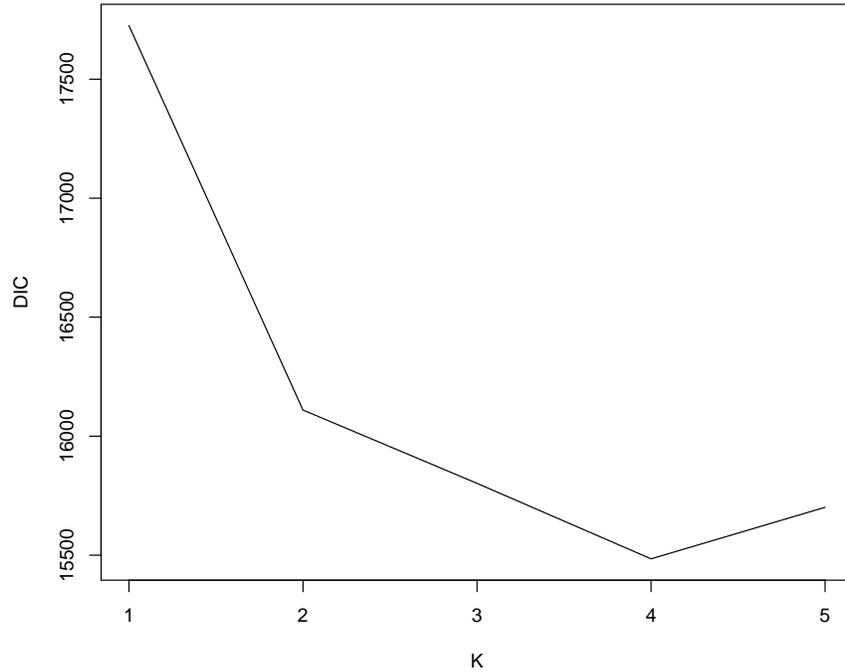
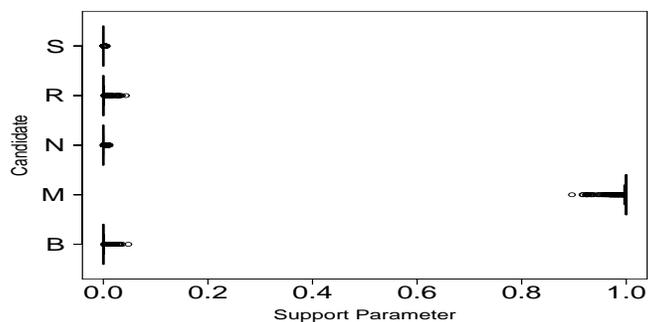


Figure 1: Values of the DIC for the GoM model for rank data fitted to the 1997 exit poll data over different values of the number of voting blocs  $K$ .

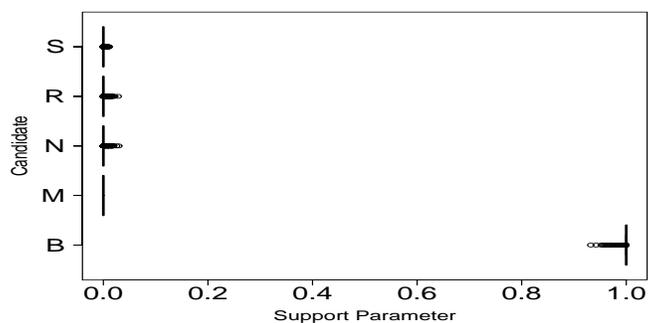
## 6.1 Support for the presidential candidates

The posterior mean support parameters and their associated uncertainty for each electoral candidate within the four voting blocs are illustrated in Figure 2. The five candidates were Banotti, McAleese, Nally, Roche and Scallon with McAleese winning the presidential seat. The four voting blocs have distinct and intuitive interpretations within the context of the 1997 Irish presidential election. The uncertainty associated with the posterior means is relatively small throughout.

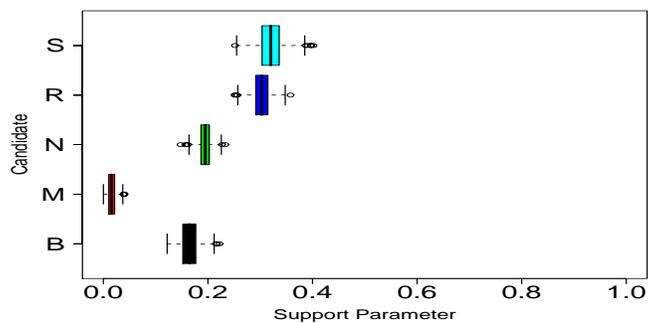
Trace plots for the support parameters estimated by the Markov chain are illustrated in Figure 5 in Appendix D. Convergence of the Markov chain was assessed using the multivariate version of Gelman and Rubin's convergence diagnostic ([Gelman and Rubin 1992](#); [Brooks and Gelman 1998](#)) — multiple chains were run from overdispersed starting values and a multivariate potential scale reduction factor of 1.08 was obtained. Approximate convergence is diagnosed when the factor is close to 1.



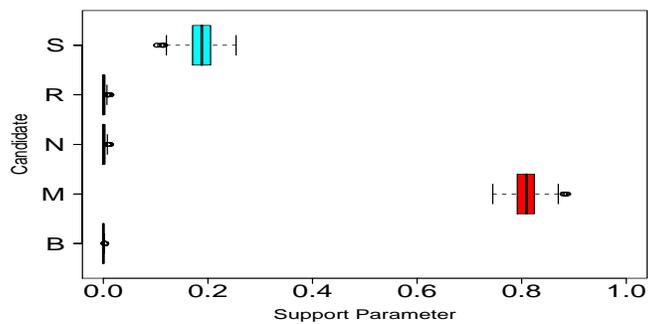
(a) Voting bloc 1.



(b) Voting bloc 2.



(c) Voting bloc 3.



(d) Voting bloc 4.

Figure 2: Box and whisker plots of the posterior mean support parameter estimates, with their associated uncertainty, for each of the five electoral candidates within the four voting blocs highlighted. Each candidate is denoted by their initial.

**Voting bloc one: pro-McAleese voters.**

(Figure 2(a)) The posterior mean support parameter estimate for candidate McAleese within this voting bloc is 0.99 with low associated uncertainty. It follows therefore that within voting bloc one there is little or no support for the other candidates. This voting bloc models voters who strongly favor McAleese; Mary McAleese was elected as President of Ireland in the 1997 election. Banotti and Roche have the largest associated uncertainty of the other candidates and thus may have support parameters slightly larger than zero. Banotti was McAleese's closest challenger and although Roche was not a major challenger on polling day, she had maintained a large public profile throughout the campaign.

**Voting bloc two: pro-Banotti voters.**

(Figure 2(b)) Banotti has high support in this voting bloc. While there is essentially zero support for McAleese the other candidates have some uncertainty around zero. Banotti supporters appear to dislike McAleese strongly, where McAleese supporters (voting bloc one) tend to be less extreme in their views of the other candidates.

**Voting bloc three: anti-McAleese voters.**

(Figure 2(c)) With the exception of McAleese, each candidate has some level of support in this voting bloc. The candidates with larger support parameters had smaller public profiles during the presidential campaign and were backed by smaller, if any, political parties. This voting bloc models voters who are generally in favor of any candidate except McAleese.

**Voting bloc four: conservative voters.**

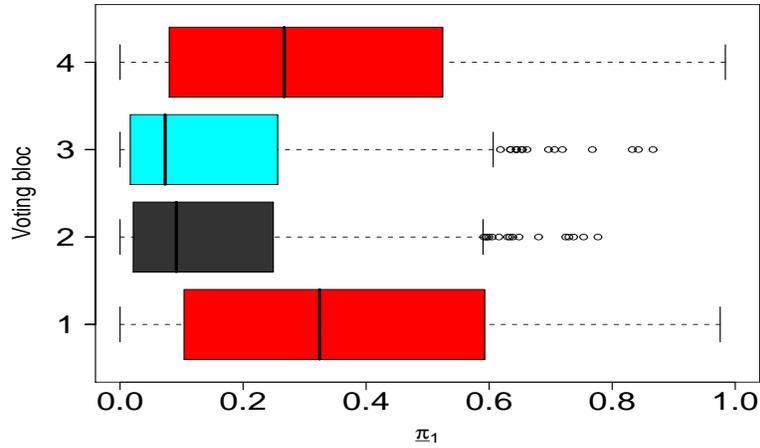
(Figure 2(d)) The final voting bloc encapsulates a conservative group of voters; McAleese and Scallon emerged as the more conservative candidates during the campaign and have the larger support parameters.

## 6.2 Mixed-membership parameters for the electorate

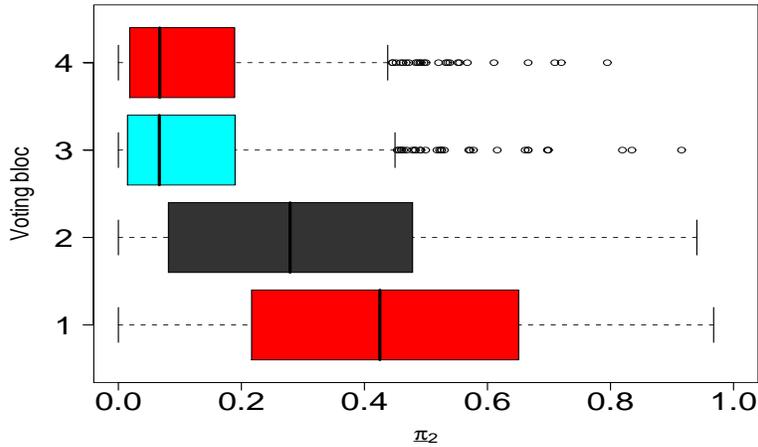
The unique feature of the GoM model is that the partial memberships of the voting blocs for each voter are inferred directly when estimating the model. Figure 3 illustrates the estimates of the mixed-membership realizations sampled during the Metropolis-within-Gibbs algorithm (subsequent to burn-in) for three randomly selected voters. All have mixed-membership parameters which are interpretable within the context of the 1997 Irish presidential election. The preferences expressed by each voter are detailed under each figure.

**Voter one.**

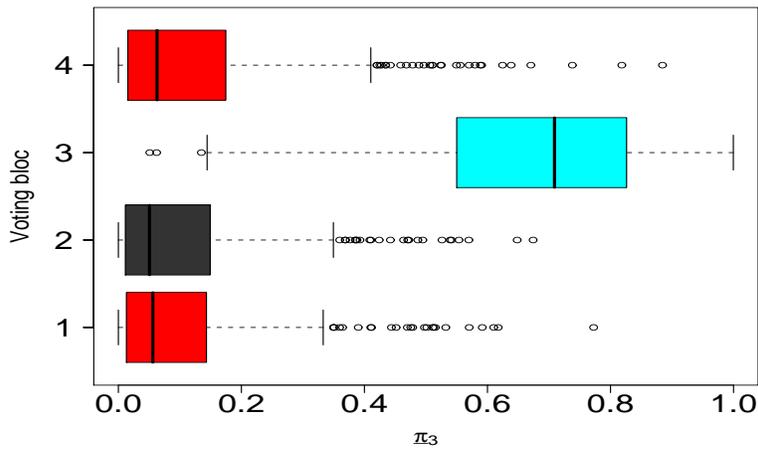
(Figure 3(a)) This voter, who only ranked McAleese, has a larger degree of membership in voting blocs one (the pro-McAleese voting bloc) and four (the conservative voting bloc). The degree of membership of voting bloc three (the anti-



(a) Voter 1: McAleese - - -



(b) Voter 2: McAleese Banotti Nally Roche Scallon



(c) Voter 3: Scallon Nally Roche - -

Figure 3: Box plots of thinned realizations (subsequent to burn-in) of the mixed membership parameter  $\pi_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$  for three randomly selected voters. The preferences expressed by each voter are detailed under each figure. The symbol - denotes the case where a voter chose not to express any further preferences. The four voting blocs referred to are as reported in Figure 2.

McAleese voting bloc) is distributed close to zero. The posterior mean mixed-membership parameter for voter one was  $\pi_1 = (0.36, 0.16, 0.16, 0.32)$ . Thus 36% of this voter's voting behavior can be characterized by voting bloc one and 32% of it by voting bloc four.

The uncertainty in this voter's group membership arises because, given their ballot, the voter clearly belongs to a group which has large support for McAleese. There exist two groups which have large support parameters for McAleese, voting bloc 1 and voting bloc 4. Under the GoM model for rank data voter 1 is not constrained to be assigned to only one of these voting blocs, but has the flexibility of being partially characterized by both.

#### **Voter two.**

(Figure 3(b)) This voter chose to express all five preferences and has larger degree of membership in voting blocs one (the pro-McAleese voting bloc) and two (the pro-Banotti voting bloc). This is an intuitive assignment as the first two preferences expressed were McAleese and then Banotti. The degree of membership in either voting bloc three or four is small. This again makes intuitive sense as voting bloc three encapsulates the anti-McAleese voters, which clearly voter two is not. Also voting bloc four models the conservative voters who favor McAleese and Scallon. Since Scallon was ranked last by this voter it follows that their degree of membership in the conservative voting bloc should be small.

In the case of voter two the posterior mean mixed-membership parameter was  $\pi_2 = (0.44, 0.31, 0.13, 0.12)$ ; 44% of voter two's behavior can be characterized by the pro-McAleese voting bloc with 31% characterized by the pro-Banotti voting bloc.

#### **Voter three.**

(Figure 3(c)) Voter three chose not to rank either of the high profile candidates, McAleese or Banotti. Voter three has a very high degree of membership of the voting bloc of anti-McAleese voters and very small degree of membership of any of the alternative voting blocs, all of which have support for McAleese and/or Banotti. The posterior mean mixed-membership parameter of  $\pi_3 = (0.10, 0.10, 0.68, 0.12)$  further highlights how voter three is mostly characterized by the anti-McAleese voting bloc.

### **6.3 Model assessment**

The use of the Plackett-Luce model and the Dirichlet prior distributions as the probability model for the election data may yield misleading inferences if the model fit is poor. In this section, the sensitivity of the inferences drawn to the choice of the Dirichlet priors and posterior predictive model checks are outlined.

### Sensitivity analysis

The use of Dirichlet priors for the support parameters  $\mathbf{p}$  and the mixed-membership parameters  $\boldsymbol{\pi}$  induces a potentially undesirable negative correlation structure between parameters. The sensitivity of inferences drawn under the GoM model to the prior specification therefore needs to be addressed. Importance sampling is employed to assess the sensitivity of the support parameter and mixed-membership parameter estimates to changes in the prior. Two types of alternative prior are examined — the logistic-normal approximation to the Dirichlet (Aitchison and Shen 1980) and Dirichlet distributions with parameter values different to the employed (0.5, . . . , 0.5) values. In particular, the logistic-normal distribution that is closest to the Dirichlet in terms of Kullback-Leibler divergence (see Aitchison (1986)) was considered, as well as other logistic-normal distributions with other correlation structures.

Table 3: Posterior means of the support parameters resulting from the use of three different priors for both the support parameters and the mixed-membership parameters in the GoM model for rank data. The Dirichlet(0.5, . . . , 0.5) was originally employed when modeling the Irish electorate and the posterior means reported in this table correspond to Figure 2. The consistent posterior mean estimates indicate a lack of sensitivity to the form of the prior.

Prior	Voting bloc	Banotti	McAleese	Nally	Roche	Scallon
Dirichlet(0.5, . . . , 0.5)	1	0.02	0.95	0.01	0.02	0.00
	2	0.99	0.00	0.01	0.00	0.00
	3	0.14	0.00	0.20	0.30	0.36
	4	0.00	0.81	0.01	0.01	0.17
Logistic-normal	1	0.03	0.93	0.01	0.03	0.00
	2	1.00	0.00	0.00	0.00	0.00
	3	0.14	0.00	0.19	0.30	0.37
	4	0.00	0.82	0.00	0.00	0.18
Dirichlet(0.1, . . . , 0.1)	1	0.02	0.95	0.01	0.02	0.00
	2	0.99	0.00	0.00	0.01	0.00
	3	0.13	0.00	0.19	0.33	0.35
	4	0.00	0.78	0.01	0.0	0.21

Support parameter estimates were very insensitive to any change in prior specification. The mixed-membership parameters were less insensitive but given the conjugacy of the Dirichlet with the multinomial distribution, the Dirichlet prior is still an attractive option.

### Posterior predictive model checks

Posterior predictive simulation (Gilks et al. 1996) was employed to assess model fit. Subsequent to a burn-in period of 10000 iterations, 40000 samples thinned every 100th

iteration were drawn from the posterior distribution  $\mathbf{P}\{\boldsymbol{\pi}, \mathbf{p}|\mathbf{x}, \mathbf{z}\}$ , giving  $R = 400$  sets of simulated parameters. A predictive election data set  $\mathbf{x}^r$  was then simulated from the GoM model for rank data, given each of the  $r = 1, \dots, R$  draws of the parameters from the posterior distribution. Due to the discrete nature of the simulated votes, the number of first preference votes obtained by the five candidates in each simulated data set was recorded. Figure 4 illustrates the number of first preferences received by each candidate in each simulated posterior predictive data set, and in the exit poll data. The posited model appears to fit well, as indicated by the lack of discrepancy between the observed number of first preferences and the simulated values. Similar results were obtained from three additional parallel runs of the Metropolis-within-Gibbs sampler.

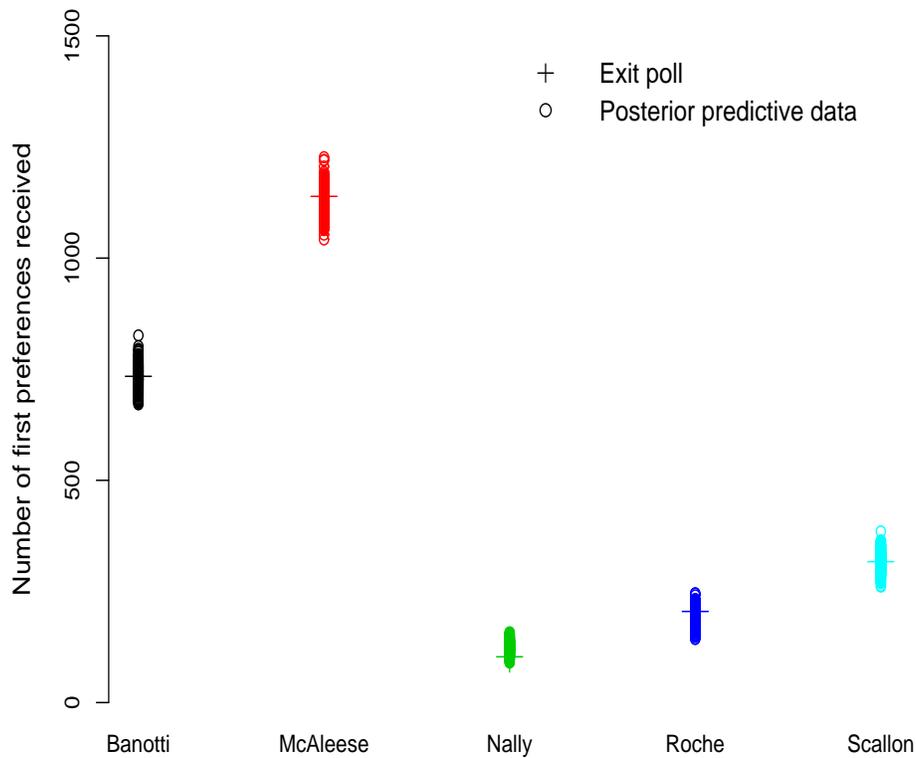


Figure 4: Each circle indicates the number of first preference votes received by the five candidates in each of 400 simulated posterior predictive data sets. The crosses indicate the number of first preferences received by each candidate in the observed exit poll data set.

## 7 Conclusions

A particular case of the GoM model, the GoM model for rank data, has been developed. In the context of rank response data, the model provides scope to examine a population

for the presence of homogeneous groups, to estimate the common characteristics that members of these groups share and to investigate the group membership of population members on a case by case basis. The loss of information which results from a hard clustering is avoided by providing a soft clustering of a heterogeneous population; the uncertainty associated with the group membership of each individual is directly inferable within the model structure.

A central part of the methodology involved the construction of surrogate proposal distributions for use in a Metropolis style step. This method, which approximates a complex conditional distribution by a surrogate distribution, provides a good and statistically sound proposal distribution for any Metropolis algorithm, in contrast to the often ad hoc selection of such proposals. Additionally the evolving nature of the surrogate proposal provides a good approximation to the complex conditional distribution at each Metropolis step in the algorithm. The construction of surrogate proposals is applicable in a much wider area than the GoM model for rank data — a linear minorizing surrogate function is utilized here but as detailed by [Hunter and Lange \(2004\)](#) many other higher order approximations may be employed if necessary.

The developed methodology provides a suitable and necessary framework in which the structure of the Irish electorate in particular may be examined. The mixed-membership parameters of the voters provide a deep insight to the mechanisms and opinions that drive each voter individually. In [Gormley and Murphy \(2008a\)](#) a mixture of Plackett-Luce models was fitted to the exit poll electorate; a four component model was deemed the optimal model. The four voting blocs highlighted in this article through the GoM methodology differ slightly from those under the mixture of Plackett-Luce models. Both solutions suggest the presence of a pro-Banotti voting bloc and the presence of a conservative voting bloc. However the remaining two voting blocs under the mixture model are a ‘pro-McAleese and Banotti’ bloc and a pro-Scallan bloc — these are quite different to the remaining two voting blocs under the GoM model (a pro-McAleese bloc and an anti-McAleese bloc). As voters under the GoM model may have partial membership of each voting bloc, it appears this extra model flexibility has allowed the discovery of the more detailed structure within the electorate. For example, the voting blocs in the GoM solution are more refined — the pro-McAleese and pro-Banotti voting blocs are deemed to be distinct whereas the mixture model finds it difficult to separate them. Also the GoM anti-McAleese voting bloc is a more refined version of the mixture model pro-Scallan voting bloc, in that the McAleese support parameter is reduced. In [Gormley \(2006\)](#) a latent space model is also used to explore the electorate from the 1997 presidential exit poll — similar characteristics as highlighted in this article are uncovered. In particular the ‘McAleese versus the rest’ election theme and the link between the conservative candidates are highlighted.

The Plackett-Luce model for rank data provided a good model for the rank nature of the voting data and allowed estimation of the common voting preferences within each voting bloc. Alternative choice models are available. In particular the Benter model for rank data ([Benter 1994](#)) is one suitable alternative. The Benter model is similar to the Plackett-Luce model but has an additional parameter (the dampening parameter) which models the way in which some preferences may be chosen less carefully than

others. However as noted by [Gormley and Murphy \(2008b\)](#), when the cardinality of the choice set is small (as in the current application) the dampening parameters are often irrelevant and the Benter model reduces to the Plackett-Luce model.

There are several models detailed in the literature which deal with modeling rank data generated by a heterogeneous population. However most of these models are mixture based models which have discrete distributions on the latent variables and thus have inherent differences to the GoM model for rank data (which have continuous distributions on the latent variables). In [Gormley and Murphy \(2006b\)](#) for example a mixture of Plackett-Luce models is fitted to rank college applications data within a maximum likelihood framework. An EM algorithm is employed to facilitate parameter estimation. When fitting this mixture model of Plackett-Luce models via the EM algorithm the posterior group membership probabilities of each observation emerge as a by-product; in the GoM model for rank data these individual group membership probabilities are direct parameters of the model. Additionally, due to the discrete nature of the distribution on the latent variables in the mixture model, the mixture of Plackett-Luce models approach requires that each college applicant belongs to a single group rather than having the flexibility of mixed group membership. In [Gormley and Murphy \(2008a\)](#) another mixture model is used to model rank voting data but an alternative rank data model, the Benter model ([Benter 1994](#)), is employed. This model differs from the GoM model for rank data in both the type of rank data model and in the inherent latent structural differences between mixture models and GoM models. The use of mixtures of rank data models for voting data is extended by [Gormley and Murphy \(2008b\)](#) where voter covariates (such as gender and age) are also incorporated. A deeper insight is provided to the heterogeneous structure of the electorate through the inclusion of the covariates. The extension of the GoM model for rank data to incorporate such covariates is an area of ongoing research. In [Gormley and Murphy \(2006a\)](#) a quite different approach to modeling rank voting data is taken where inferences are drawn on the locations of both voters and electoral candidates in a latent space. The probability of a voter choosing a candidate is modeled as a function of the distance between them in the latent space. The probability of an entire vote is then modeled via a single Plackett-Luce model. The application of this latent space model for rank data to Irish general election data highlighted interesting structures in the set of candidates. Irish general elections are quite different in nature to the presidential election examined here as typically the number of electoral candidates is quite large and party politics play a large role.

The GoM model for rank data has modeling applications in addition to Irish election data. Rank data appears in many areas of society — within the Irish context again third level college applications involve applicants ranking courses in order of preference. Such data has been analysed using the Plackett-Luce mixture model ([Gormley and Murphy 2006b](#)) but the GoM model for rank data would also be applicable. In terms of voting data, STV elections are widely employed. The system is also known as the alternative vote (or instant run-off voting) system and it is used in elections in Australia and Fiji and in some regional elections in North America ([Farrell 2001](#)). The Proportional Representation by means of a Single Transferable Vote (PR-STV) system is perhaps more widely used than STV — PR-STV is essentially the STV system in the case where

more than one seat is to be elected. Irish governmental elections employ this system and can be modeled using rank data models (Gormley and Murphy 2006a, 2008a,b). Within our own field of Statistics, the PR-STV system is employed to elect the councils of the Royal Statistical Society and the Institute of Mathematical Statistics. Several other European Union member states use a PR system (Regenwetter et al. 2006) but this is usually achieved using a list voting system. The GoM model for rank data would be a suitable tool for exploring the populations in these contexts.

The GoM model for rank data could be developed in several directions. In terms of the application in this article further model accuracy could be attained by imposing a hierarchical framework — a hyperprior could be introduced for the Dirichlet parameters  $\underline{\alpha}$  and  $\underline{\beta}$  of the mixed membership and support parameter priors respectively. Pritchard et al. (2000) and Erosheva (2003) employ such hierarchical priors. However, as the results obtained in this application were insensitive to prior specification and were intuitive further complication of the model was deemed unnecessary.

No consensus on the topic of model choice within the GoM framework has been achieved in the literature (Joutard et al. 2008); given prior investigations of the 1997 Irish presidential exit poll, the model dimensionality selected in this application appeared intuitive. However other model comparison tools such as the AICM or BICM (Raftery et al. 2007) could perhaps be employed to aid the model choice procedure.

## Appendix A: Data sources

The various 1997 Irish presidential election opinion poll data sets were collected by the three companies: Lansdowne Market Research, Irish Marketing Surveys, and the Market Research Bureau of Ireland. These data sets are available through the Irish Elections Data Archive

<http://www.tcd.ie/PoliticalScience/elections/elections.html>

and the Irish Opinion Poll Archive

<http://www.tcd.ie/PoliticalScience/cgi/>

which are maintained by Professor Michael Marsh in the Department of Political Science, Trinity College Dublin, Ireland.

## Appendix B: Construction of a surrogate proposal distribution.

The conditional distribution of the Plackett-Luce support parameters required for the Gibbs sampler is not in standard form. One approach is to impute a Metropolis style step within the Gibbs sampler for the remaining model parameters. To implement a Metropolis step to sample Plackett-Luce support parameter values a tractable proposal

distribution which approximates the full conditional is required. Taking logs of (4) gives

$$\log \mathbf{P}\{\underline{p}_k | \boldsymbol{\pi}, \mathbf{x}, \mathbf{z}\} + C = \sum_{i=1}^M \sum_{t=1}^{n_i} z_{itk} \left\{ \log p_{kc(i,t)} - \log \sum_{s=t}^N p_{kc(i,s)} \right\} + \sum_{j=1}^N (\beta_j - 1) \log p_{kj} \quad (7)$$

where  $C$  is a constant. The function  $-\log(\cdot)$  is a convex function and thus the supporting hyperplane property (5) can be applied to the term  $-\log \sum_{s=t}^N p_{kc(i,s)}$  in (7). This provides a minorizing surrogate function for the log of the full conditional of  $\underline{p}_k$ . By (5):

$$-\log \sum_{s=t}^N p_{kc(i,s)} \geq -\log \sum_{s=t}^N \bar{p}_{kc(i,s)} - \frac{\sum_{s=t}^N p_{kc(i,s)}}{\sum_{s=t}^N \bar{p}_{kc(i,s)}} + 1$$

where  $\bar{p}_{kc(i,s)}$  is the previously sampled value of the respective support parameter. Denoting

$$\begin{aligned} \delta_{kj} &= \sum_{i=1}^M \sum_{t=1}^{n_i} z_{itk} \mathbf{1}_{\{c(i,t)=j\}} & \text{and} \\ \psi_{ijt} &= \begin{cases} 1 & \text{if } t = 1 \\ 1 & \text{if } t > 1 \text{ and } c(i, 1), \dots, c(i, t-1) \neq j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

then (7) becomes:

$$\log \mathbf{P}\{\underline{p}_k | \boldsymbol{\pi}, \mathbf{x}, \mathbf{z}\} + C \geq \sum_{j=1}^N \{\beta_j + \delta_{kj} - 1\} \log p_{kj} - \sum_{i=1}^M \sum_{t=1}^{n_i} \left\{ \frac{\sum_{j=1}^N z_{itk} p_{kj} \psi_{ijt}}{\sum_{s=t}^N \bar{p}_{kc(i,s)}} \right\}.$$

This expression is in the form of the log of a Gamma distribution and hence the support parameters are approximately Gamma distributed i.e.

$$p_{kj} \sim \text{Gamma} \left( \beta_j + \delta_{kj}, \left[ \sum_{i=1}^M \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^N \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt} \right]^{-1} \right).$$

This Gamma function becomes computationally unstable in that the shape parameter  $\beta_j + \delta_{kj}$  is generally large — the definition of  $\delta_{kj}$  involves  $M$ , the number of voters, which in an electoral data set is typically large. However, since

$$\text{Gamma}(r, \lambda) \rightarrow \text{Normal}(r\lambda, r\lambda^2) \quad \text{as } r \rightarrow \infty$$

a Normal( $\mu_{kj}, \sigma_{kj}^2$ ) distribution is employed as a tractable proposal for the support parameter  $p_{kj}$  where:

$$\begin{aligned} \mu_{kj} &= \frac{\beta_j + \delta_{kj}}{\sum_{i=1}^M \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^N \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt}} \\ \sigma_{kj}^2 &= \frac{\beta_j + \delta_{kj}}{\left[ \sum_{i=1}^M \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^N \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt} \right]^2}. \end{aligned}$$

Note that the parameters of the Normal proposal distribution are functions of  $\bar{p}_{kj}$ , the sampled values of the support parameters from the previous iteration of the Metropolis-within-Gibbs algorithm. Hence the surrogate proposal distribution is not static but is updated in an online fashion and thus provides a good approximation of the full conditional distribution.

Since the parameters to be sampled are constrained such that  $0 \leq p_{kj} \leq 1$  and  $\sum_{j=1}^N p_{kj} = 1$  normalization must be performed during the Metropolis step within the Gibbs sampler. Thus, subsequent to choosing suitable starting values for the support parameters, the Metropolis step within the Gibbs sampler proceeds as follows for each  $k = 1, \dots, K$ :

1. For  $j = 1, \dots, N$  generate  $\check{p}_{kj}$  where  $\check{p}_{kj} \sim N(\mu_{kj}, \sigma_{kj}^2)$ .
2. Set  $\underline{\check{p}}_k = (\check{p}_{k1}/S, \dots, \check{p}_{kN}/S)$  where  $S = \sum_{j=1}^N \check{p}_{kj}$ .
3. Let  $\underline{\bar{p}}_k$  denote the value of  $\underline{p}_k$  from the previous iteration of the Metropolis-within-Gibbs algorithm. Calculate the acceptance probability  $\alpha$  where:

$$\begin{aligned} \alpha &= \min \left[ \log \left\{ \frac{\mathbf{P}(\underline{\check{p}}_k | \dots) q(\underline{\bar{p}}_k | \dots)}{\mathbf{P}(\underline{\bar{p}}_k | \dots) q(\underline{\check{p}}_k | \dots)}, 1 \right\} \right] \\ &= \min \left[ \sum_{i=1}^M \sum_{t=1}^{n_i} \left[ z_{itk} \left\{ \log(\check{p}_{kc(i,t)}) - \log \left( \sum_{s=t}^N \check{p}_{kc(i,s)} \right) - \log(\bar{p}_{kc(i,t)}) \right. \right. \right. \\ &\quad \left. \left. + \log \left( \sum_{s=t}^N \bar{p}_{kc(i,s)} \right) \right\} \right] + \sum_{j=1}^N [(\beta_j - 1) \{ \log(\check{p}_{kj}) - \log(\bar{p}_{kj}) \}] \\ &\quad \left. + \frac{(\check{p}_{kj} - \mu_{kj}/S)^2 - (\bar{p}_{kj} - \mu_{kj}/S)^2}{2\sigma_{kj}^2/S^2} \right], 0 \end{aligned}$$

where ... represents all other parameters and  $q(\cdot)$  the Normal surrogate proposal distribution.

4. Generate a uniform random variable  $u \sim U(0, 1)$ .
5. If  $\log(u) \leq \alpha$  define  $\underline{p}_k = \tilde{\underline{p}}_k$ .

## Appendix C: online label switching algorithm

The online algorithm to correct for label switching which occurs during the Metropolis-within-Gibbs sampler proceeds as follows:

1. Generate all  $K!$  permutations  $\nu_l$  for  $l = 1, \dots, K!$ . Set  $t = 0$ .
2. After discarding the burn-in Metropolis steps, denote the support parameters estimated at step  $t$  by  $\hat{\mathbf{p}}^t$ .
3. Choose permutation  $\nu_l$  for  $l = 1, \dots, K!$  which minimizes the loss function:

$$\mathcal{L}(\hat{\mathbf{p}}_{\nu_l}^t; \tilde{\mathbf{p}}) = \sum_{k=1}^K \sum_{j=1}^N (\hat{p}_{\nu_l(k)j}^t - \tilde{p}_{kj})^2 \quad (8)$$

where  $\tilde{\mathbf{p}}$  denotes the MAP estimate of the support parameters and  $\nu_l(k)$  denotes permutation  $\nu_l$  applied to integer  $k$ .

4. Given  $\nu_l$ , the permutation which minimizes the loss function (8), update the model parameters:

$$p_{kj} = \frac{t}{t+1} p_{kj} + \frac{1}{t+1} \hat{p}_{\nu_l(k)j}^t.$$

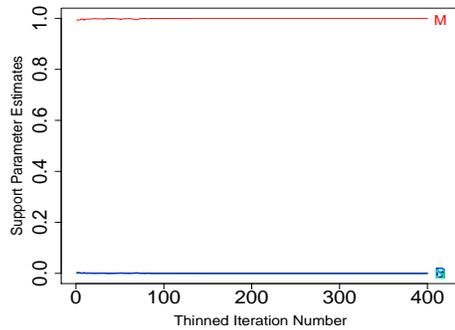
The mixed-membership parameters  $\boldsymbol{\pi}$  are updated via the same permutation. Set  $t = t + 1$  and repeat steps 3 and 4 subsequent to each Metropolis step within the sampler.

It follows that label switching will be minimized thus somewhat ensuring the validity of posterior mean estimates.

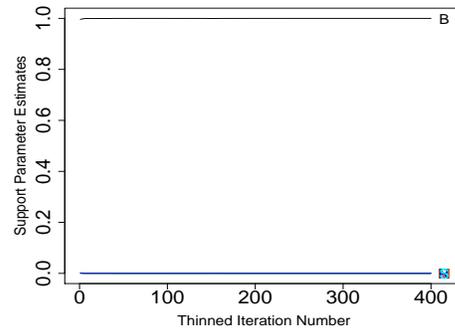
## Appendix D: Metropolis-within-Gibbs sampler trace plots

### References

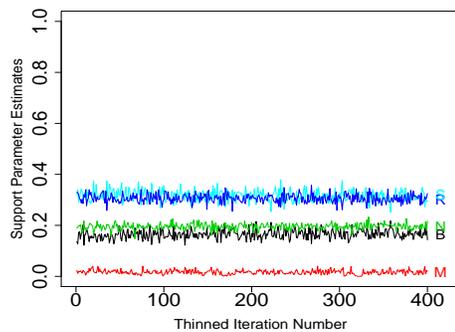
- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall. 283
- Aitchison, J. and Shen, S. (1980). "Logistic-normal distributions: Some properties and uses." *Biometrika*, 67(2): 261–272. 283



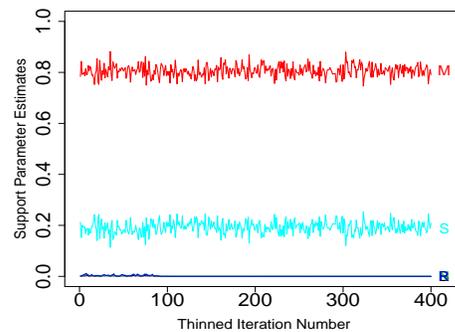
(a) Voting bloc 1.



(b) Voting bloc 2.



(c) Voting bloc 3.



(d) Voting bloc 4.

Figure 5: Trace plots of samples of support parameters for the presidential candidates within each voting bloc obtained after convergence of the Markov chain. The initial at the right hand side of each trace indicates the candidate whose support parameter is traced. Surrogate proposal distributions were employed and random starting values were used to initialize the chain. Each figure illustrates 40000 samples thinned every 100th iteration, subsequent to a burn-in period of 10000 iterations.

Bartholomew, D. J. and Knott, M. (1999). *Latent variable models and factor analysis*. London: Edward Arnold, second edition. 271

Benter, W. (1994). “Computer-based Horse Race Handicapping and Wagering Systems: A Report.” In Ziemba, W. T., Lo, V. S., and Haush, D. B. (eds.), *Efficiency of Racetrack Betting Markets*, 183–198. San Diego and London: Academic Press. 285, 286

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent Dirichlet allocation.” *Journal of Machine Learning Research*, 3: 993–1022. 270

- Bradley, R. A. and Terry, M. E. (1952). “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” *Biometrika*, 39: 324–345. 268
- Bradlow, E. T. and Fader, P. S. (2001). “A Bayesian Lifetime Model for the “Hot 100” Billboard Songs.” *Journal of the American Statistical Association*, 96: 368–381. 268
- Brooks, S. P. and Gelman, A. (1998). “General methods for monitoring convergence of iterative simulations.” *J. Comput. Graph. Statist.*, 7(4): 434–455. 278
- Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis.*. New York: Chapman & Hall, 2nd edition. 274
- Celeux, G. (1998). “Bayesian inference for Mixtures: the label switching problem.” In *Proceedings Compstat*, 227–232. Physica-Verlag. 276
- Celeux, G., Hurn, M., and Robert, C. P. (2000). “Computational and inferential difficulties with mixture posterior distributions.” *Journal of the American Statistical Association*, 95(451): 957–970. 276
- Chapman, R. and Staelin, R. (1982). “Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model.” *Journal of Marketing Research*, 19: 288–301. 268
- Coakley, J. and Gallagher, M. (1999). *Politics in the Republic of Ireland*. London: Routledge in association with PSAI Press, 3rd edition. 267
- Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data*. Lecture Notes in Statistics, 34. Berlin: Springer-Verlag. 268
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38. With discussion. 275
- Erosheva, E. (2002). “Grade of membership and latent structure models with application to disability survey data.” Ph.D. thesis, Department of Statistics, Carnegie Mellon University. 266, 270, 273
- Erosheva, E. A. (2003). “Bayesian Estimation of the Grade of Membership Model.” In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics*, 7, 501 – 510. Oxford: Oxford University Press. 265, 270, 272, 273, 287
- (2006). “Latent class representation of the Grade of Membership Model.” Technical Report 492, Department of Statistics, University of Washington. 271, 272
- Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). “Describing disability through individual-level mixture models for multivariate binary data.” *The Annals of Applied Statistics*, 1(2): 502–537. 273
- Farrell, D. M. (2001). *Electoral Systems: A Comparative Introduction*. New York: St. Martin’s Press. 286

- Fligner, M. A. and Verducci, J. S. (1988). “Multistage Ranking Models.” *Journal of the American Statistical Association*, 83: 892–901. 269
- Fraley, C. and Raftery, A. E. (2002). “Model-Based Clustering, Discriminant Analysis, and Density Estimation.” *Journal of the American Statistical Association*, 97(458): 611–631. 265
- Gelman, A. and Rubin, D. B. (1992). “Inference From Iterative Simulation Using Multiple Sequences (Disc: P483-501, 503-511).” *Statistical Science*, 7: 457–472. 278
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall. 283
- Gormley, I. C. (2006). “Statistical Models for Rank Data.” Ph.D. thesis, Department of Statistics, University of Dublin, Trinity College. 285
- Gormley, I. C. and Murphy, T. B. (2006a). “A Latent Space Model for Rank Data.” In Arioldi, E., Blei, D., Fienberg, S., Goldenberg, A., Xing, E., and Zheng, A. (eds.), *Statistical Network Analysis: Models, Issues, and New Directions*, volume 4503 of *Lecture Notes in Computer Science*, 90–102. Springer Verlag. 267, 286, 287
- (2006b). “Analysis of Irish third-level college applications data.” *Journal of the Royal Statistical Society, Series A*, 169(2): 361–379. 271, 286
- (2008a). “Exploring Voting Blocs Within the Irish Electorate: A Mixture Modeling Approach.” *Journal of the American Statistical Association*, To appear. 267, 271, 278, 285, 286, 287
- (2008b). “A mixture of experts model for rank data with applications in election studies.” *The Annals of Applied Statistics*, To appear. 267, 271, 286, 287
- Graves, T., Reese, C. S., and Fitzgerald, M. (2003). “Hierarchical Models for Permutations: Analysis of Auto Racing Results.” *Journal of the American Statistical Association*, 98: 282–291. 268
- Haberman, S. J. (1995). “Book review of ‘Statistical Applications Using Fuzzy Sets’, by Manton, K. G., Woodbury, M. A. and Corder, L.S.” *Journal of the American Statistical Association*, 90: 1131–1133. 271
- Hunter, D. R. and Lange, K. (2004). “A tutorial on MM algorithms.” *The American Statistician*, 58(1): 30–37. 275, 285
- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). “Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling.” *Statistical Science*, 20: 50–67. 276, 277
- Joutard, C., Arioldi, E., Fienberg, S., and Love, T. (2008). “Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice.” In Poncet, P., Masseglia, F., and Teisseire, M. (eds.), *Data Mining Patterns: New Methods and Applications*. Pennsylvania: IGI Global. 270, 277, 287

- Lange, K., Hunter, D. R., and Yang, I. (2000). “Optimization transfer using surrogate objective functions.” *Journal of Computational and Graphical Statistics*, 9(1): 1–59. With discussion, and a rejoinder by Hunter and Lange. 266, 275
- Lazarsfeld, P. and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin. 271
- Mallows, C. L. (1957). “Non-null ranking models. I.” *Biometrika*, 44: 114–130. 268
- Marden, J. I. (1995). *Analyzing and modeling rank data*. London: Chapman & Hall. 268
- Marsh, M. (1999). “The Making of the Eighth President.” In Marsh, M. and Mitchell, P. (eds.), *How Ireland Voted 1997*, 215–242. Boulder, CO: Westview and PSAI Press. 267
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture models*. New York: John Wiley & Sons. 265
- O’Hagan, A. and Forster, J. (2004). *Kendall’s Advanced Theory of Statistics: Volume 2B Bayesian Inference*. London, UK: Arnold, second edition. 273
- Plackett, R. L. (1975). “The analysis of permutations.” *Applied Statistics*, 24(2): 193–202. 268
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics*, 155: 945–959. 273, 287
- Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). “Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity (with Discussion).” In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 8*, 1–45. Oxford: Oxford University Press. 287
- Regenwetter, M., Grofman, B., Marley, A. A. J., and Tsetlin, I. M. (2006). *Behavioral Social Choice. Probabilistic Models, Statistical Inference and Applications.*. New York: Cambridge University Press. 287
- Richardson, S. and Green, P. J. (1997). “On Bayesian Analysis of Mixtures With An Unknown Number of Components.” *Journal of the Royal Statistical Society, Series B*, 59: 731–758. 276
- Rosén, B. (1972). “Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I.” *The Annals of Mathematical Statistics*, 43(2): 373–397. 274
- Sinnott, R. (1995). *Irish voters decide: Voting behaviour in elections and referendums since 1918*. Manchester: Manchester University Press. 267
- (1999). “The Electoral System.” In Coakley, J. and Gallagher, M. (eds.), *Politics in the Republic of Ireland*, 99–126. London: Routledge & PSAI Press, 3rd edition. 267

- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). “Bayesian measures of model complexity and fit.” *Journal of the Royal Statistical Society, Series B*, 64(4): 583–639. 277
- Stephens, M. (2000). “Dealing with label-switching in mixture models.” *Journal of the Royal Statistical Society, Series B*, 62(4): 795–810. 276, 277
- Thurstone, L. L. (1927). “A law of comparative judgement.” *Psychological Review*, 34: 273–286. 268
- Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press. 268, 269
- van der Brug, W., van der Eijk, C., and Marsh, M. (2000). “Exploring Uncharted Territory: The Irish Presidential Election 1997.” *British Journal of Political Science*, 30(4): 631–650. 267
- Woodbury, M., Clive, J., and Garson, A. (1978). “Mathematical typology: A grade of membership technique for obtaining disease definition.” *Computers and Biomedical Research*, 11: 277–298. 270

#### Acknowledgments

Isobel Claire Gormley was supported by a Government of Ireland Research Scholarship in Science, Engineering and Technology provided by the Irish Research Council for Science, Engineering and Technology, funded by the National Development Plan. Thomas Brendan Murphy was supported by a Science Foundation of Ireland Research Frontiers Grant (06/RFP/M040). A substantial part of this work was completed when the authors were in the Department of Statistics, Trinity College Dublin. We thank Professor Michael Marsh for supplying the data set used in this study. We thank Professors Stephen Fienberg and Adrian Raftery for useful comments on this work.

