Clustering ranked preference data using sociodemographic covariates

Isobel Claire Gormley & Thomas Brendan Murphy

School of Mathematical Sciences, University College Dublin.

Abstract

Ranked preference data arise when a set of judges rank, in order of their preference, a set of objects. Such data arise in preferential voting systems and market research surveys. Covariate data associated with the judges are also often recorded. Such covariate data should be used in conjunction with preference data when drawing inferences about judges.

To cluster a population of judges, the population is modeled as a collection of homogeneous groups. The Plackett-Luce model for ranked data is employed to model a judge's ranked preferences within a group. A mixture of Plackett-Luce models is employed to model the population of judges, where each component in the mixture represents a group of judges.

Mixture of experts models provide a framework in which covariates are included in mixture models. Covariates are included through the mixing proportions and the component density parameters. A mixture of experts model for ranked preference data is developed by combining a mixture of experts model and a mixture of Plackett-Luce models. Particular attention is given to the manner in which covariates enter the model. The mixing proportions and group specific parameters are potentially dependent on covariates. Model selection procedures are employed to choose optimal models.

Model parameters are estimated via the 'EMM algorithm', a hybrid of the Expectation-Maximization and the Minorization-Maximization algorithms. Examples are provided through a menu survey and through Irish election data. Results indicate mixture modeling using covariates is insightful when examining a population of judges who express preferences.

Keywords: ranked preference data; clustering; covariates; Plackett-Luce model.

1 Introduction

Ranked preference data arise when judges rank some or all of a set of objects in order of their preference. Such data arise in a wide range of contexts including in preferential

voting systems (eg. Irish elections that use a preferential voting system (Coakley and Gallagher, 2004)), in market research surveys (eg. food preference surveys (Ralston et al., 2002)) and in university application procedures (eg. in Ireland, students rank up to ten degree courses in their college application (Gormley and Murphy, 2006)).

Modeling preference data in an appropriate manner is imperative when examining the behavior of the set of judges who gave rise to the data. Additionally, it is often the case that covariates associated with each judge are recorded when a survey of their preferences is taken. Such covariate data can be used in conjunction with preference data to provide a deeper understanding of the preferences and/or structure of the population of judges under investigation. Models for preference data including those that incorporate covariates are discussed in Section 2.1.

Clustering methods are used when it is believed that a heterogeneous population of judges consists of a collection of homogeneous subpopulations and these subpopulations are unknown and need to be characterized. Clustering methods tend to be either algorithmic (eg. hierarchical or k-means clustering) or based on statistical models (eg. Fraley and Raftery, 2002). In the model-based approach to clustering, the population is modeled as a finite collection of homogeneous groups that are modeled individually using appropriate statistical models; that is, a finite mixture model is used to model the data. In this work a model-based approach is taken, where the Plackett-Luce (or exploded logit) model for rank data is employed within each group to model the way in which group members rank preferences. Thus, a mixture of Plackett-Luce models is employed as an appropriate statistical model for the population of judges, where each component in the mixture represents a group of judges with a specific parameterization of the Plackett-Luce model. A more detailed outline of the mixture of Plackett-Luce models is given in Section 2.2.

Mixture of experts models (Jacobs et al., 1991) provide a framework in which covariates are included in mixture models. In these models, covariates are included through the mixing proportions and through the parameters of component densities using generalized linear model theory. In Section 2.3, a mixture of experts model for ranked preference data is developed by combining a mixture of experts model and a mixture of Plackett-Luce models. Particular attention is given to the manner in which covariates enter the model.

The model parameters are estimated via the 'EMM' algorithm, a hybrid of the Expectation-Maximization (EM) and the Minorization-Maximization (MM) algorithms. Model selection procedures are employed to select both the manner in which covariates enter the model and to select the optimal number of groups within the population. This approach provides a framework where the manner in which covariates influence a clustering is selected in a unified, statistically principled manner. Details of model fitting and selection are given in Section 3.

In this paper, two applications are used to illustrate the proposed methodology for clustering ranked preference data in the presence of covariates: a marketing data set collected through the 1996 Menu Census Survey conducted by the Market Research Corporation of America and an Irish election data set where voters rank electoral candidates in order of their preference. In both examples, data have been collected from a heterogeneous set of judges who have expressed their preferences. Interest lies in establishing

the existence of homogeneous subgroups of judges in the population who have similar preferences. Covariates associated with the judges are also available and there is interest in establishing if the covariates provide information about the clustering. The data sets are described in detail in Section 4.

The results of applying the mixture of experts model for ranked preference data to the illustrative examples are given in Section 5. The results indicate that mixture modeling using covariates can be insightful when examining a population of judges who express preferences.

Section 6 concludes with a discussion of the proposed model and the results of its application in the illustrative examples.

2 A Mixture of Experts Model for Ranked Preference Data

The mixture of experts model (Jacobs et al., 1991) accommodates clustering and covariate modeling in a single modeling framework. In this section, we develop a mixture of experts model for ranked preference data, so that covariates can be used in conjunction with ranked preference data for clustering purposes.

2.1 Models for Ranked Preference Data

Many models have been proposed for ranked preference data. Examples of models include distance-based models (Critchlow, 1985; Mallows, 1957) where the probability of a ranking decreases as the distance from a central ranking increases, order statistic (random utility) models (Chapman and Staelin, 1982; Luce and Suppes, 1963; McFadden, 1974; Thurstone, 1927) where the ranking reflects the ordering of latent scores given to each object and multistage ranking models (Benter, 1994; Fligner and Verducci, 1988; Plackett, 1975) where the ranking is modeled as a sequential process of selecting the next most preferred object. Detailed reviews of models for ranking data are given by Critchlow et al. (1991), Fligner and Verducci (1993) and Marden (1995) and references therein.

In this paper, the Plackett-Luce model (or exploded logit model) (Chapman and Staelin, 1982; Plackett, 1975) for ranked preference data is used to model data within a homogeneous set of judges. Suppose that data are collected from M judges who list their preference ordering for a set of N objects. Let $c(i,j)$ denote the object ranked in j^{th} position by judge i. Then $\underline{x}_i = (c(i,1), c(i,2), \ldots, c(i,n_i))$ is an ordered list of the objects as recorded in the ranked preference of judge i, where n_i is the number of preferences expressed by this judge. The Plackett-Luce model with support parameter $\underline{p} = (p_1, p_2, \ldots, p_N)$ is of the form

$$\mathbf{P}(\underline{x}_i|\underline{p}) = \frac{p_{c(i,1)}}{\sum_{s=1}^{N} p_{c(i,s)}} \cdot \frac{p_{c(i,2)}}{\sum_{s=2}^{N} p_{c(i,s)}} \cdots \frac{p_{c(i,n_i)}}{\sum_{s=n_i}^{N} p_{c(i,s)}} = \prod_{t=1}^{n_i} \frac{p_{c(i,t)}}{\sum_{s=t}^{N} p_{c(i,s)}}, \tag{1}$$

where $c(i, n_i+1), \ldots, c(i, N)$ is any permutation of objects not listed in the judge's ranked preference; the choice of this ordering does not influence the probability. In order to make the parameter $\underline{p}$ identifiable, it is usual to restrict $\sum_{j=1}^{N} p_j = 1$; under this restriction, the support parameter p_j can be interpreted as the probability of selecting object j in first place, out of the currently available choice set. Under the Plackett-Luce model the ranking of objects by a judge is modeled as a set of independent choices by the judge, conditional on the cardinality of the choice set being reduced by one after each choice is made.

The Plackett-Luce model can be interpreted as either an order statistic (random utility) model or a multistage model (cf. McFadden, 1974). In particular, let the utility that voter i selects candidate j be $U_{ij} = -\log p_j + \epsilon_{ij}$, where the ϵ_{ij} are independent identically distributed according to an extreme value distribution. Then, the $\mathbf{P}(\underline{x}_i|\underline{p}) = \mathbf{P}(U_{ic(i,1)} > U_{ic(i,2)} > \cdots > U_{ic(i,n_i)})$ and can be written in the same form as (1) (eg. Train, 2003, Section 7.3.1).

The Plackett-Luce model can accommodate covariates using a multinomial logit structure, as proposed in Chapman and Staelin (1982) and Train (2003, Chapter 7). Let $\underline{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iL})$ be the covariates for observation i. The support parameters are modeled as a logistic function of the covariates,

$$\log\left(\frac{p_j(\underline{w}_i)}{p_1(\underline{w}_i)}\right) = \gamma_{j0} + \gamma_{j1}w_{i1} + \gamma_{j2}w_{i2} + \cdots + \gamma_{jL}w_{iL} = \underline{\gamma}_j^T \underline{w}_i.$$

Object 1 is the baseline category, with $\underline{\gamma}_1 = (0, \ldots, 0)$, in order to assure identifiability of the model parameters. This leads to the exploded logit model of the form,

$$\begin{aligned}\mathbf{P}(\underline{x}_i|\underline{p}(\underline{w}_i)) = \mathbf{P}(\underline{x}_i|\boldsymbol{\gamma}, \underline{w}_i) &= \frac{\exp(\underline{\gamma}_{c(i,1)}^T \underline{w}_i)}{\sum_{s=1}^{N} \exp(\underline{\gamma}_{c(i,s)}^T \underline{w}_i)} \cdot \frac{\exp(\underline{\gamma}_{(i,2)}^T \underline{w}_i)}{\sum_{s=2}^{N} \exp(\underline{\gamma}_{c(i,s)}^T \underline{w}_i)} \cdots \frac{\exp(\underline{\gamma}_{c(i,n_i)}^T \underline{w}_i)}{\sum_{s=n_i}^{N} \exp(\underline{\gamma}_{c(i,s)}^T \underline{w}_i)} \\ &= \prod_{t=1}^{n_i} \frac{\exp(\underline{\gamma}_{c(i,t)}^T \underline{w}_i)}{\sum_{s=t}^{N} \exp(\underline{\gamma}_{c(i,s)}^T \underline{w}_i)}. \end{aligned} \tag{2}$$

In this case, the model corresponds to a random utility model with $U_{ij} = \underline{\gamma}_j^T \underline{w}_i + \epsilon_{ij}$ where the ϵ_{ij} are iid according to an extreme value distribution.

2.2 Mixture Models

The mixture model (also known as a latent class model (LCM)) assumes that a population can be modeled as a finite collection of subpopulations, where each subpopulation can be characterized by a suitable probability density. Mixture models have been used in a wide range of applications where data are collected from heterogeneous sources. Latent Class Analysis (LCA) (Lazarsfeld and Henry, 1968) uses a mixture model to investigate the dependence between categorical variables, thus providing a discrete version of factor analysis for categorical data. General reviews of mixture modeling are given by Titterington et al. (1985), McLachlan and Basford (1988) and McLachlan and Peel (2000). In

addition, Fraley and Raftery (2002) provide a review of the use of mixture models for clustering.

Suppose that a population consists of K subpopulations and that the probability of belonging to subpopulation k is π_k. The probability density for observation $\underline{x}_i$ from subpopulation k is $f(\underline{x}_i|\theta_k)$, where θ_k are the parameters of the model for subpopulation k. Then the model for an observation of unknown subpopulation is of the form

$$\mathbf{P}(\underline{x}_i) = \sum_{k=1}^{K} \pi_k f(\underline{x}_i|\theta_k), \tag{3}$$

that is a K component mixture model. The π_k values are known as mixing proportions, $f(\underline{x}_i|\theta_k)$ are called component densities and θ_k are the parameters of the component densities.

Mixture models have been successfully applied to the analysis of ranked preference data in Stern (1993), Vigneau et al. (1999), Murphy and Martin (2003), Gormley and Murphy (2006, 2008a), Busse et al. (2007) and Meilă and Bao (2008) amongst others. The mixture of Plackett-Luce models,

$$\mathbf{P}(\underline{x}_i) = \sum_{k=1}^{K} \pi_k f(\underline{x}_i|\underline{p}_k) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}, \tag{4}$$

where $\underline{p}_k$ is the parameter of the Plackett-Luce model that characterizes subpopulation k, is applied to ranked preference data by Gormley and Murphy (2006) who analyze Irish college applications. Additionally, Gormley and Murphy (2008a) analyze Irish election data using both a mixture of Plackett-Luce models and a mixture of Benter models (Benter, 1994). In this article, the mixture of Plackett-Luce models is extended to facilitate the inclusion of covariates. The mixture of Plackett-Luce models has some connections with Latent Class Analysis, in that a mixture model is being used to model the data, but the motivation is quite different. In this work, the mixture modeling framework is being used to find clusters in the data, whereas in Latent Class Analysis the emphasis is to study dependence between variables. However, both approaches use a discrete latent variable to study structure in the population.

2.3 Mixture of Experts Models

Jacobs et al. (1991) introduce the mixture of experts (MoE) model as an extension of the standard mixture model to include covariates. Covariates are incorporated in the mixture model through the use of generalized linear models (GLMs) (Dobson, 2002; McCullagh and Nelder, 1983; Nelder and Wedderburn, 1972). GLMs are used to model both the relationship between the outcome variable and covariates and the relationship between the mixing proportions and covariates. Hence, the general MoE model is of the form

$$\mathbf{P}(\underline{x}_i|\underline{w}_i) = \sum_{k=1}^{K} \pi_k(\underline{w}_i) f(\underline{x}_i|\theta_k(\underline{w}_i)) \tag{5}$$

where the relationship between the mixing proportions and covariates, for example, is modeled as a multinomial logistic regression model of the form

$$\log\left(\frac{\pi_k(\underline{w}_i)}{\pi_1(\underline{w}_i)}\right) = \beta_{k0} + \beta_{k1}w_{i1} + \cdots + \beta_{kL}w_{iL} = \underline{\beta}_k^T\underline{w}_i. \tag{6}$$

The mixture of experts model originates from the machine learning literature and the terminology used for this model is different from mixture modeling. The mixing proportions $\pi_k(\underline{w}_i)$ are called *gating networks* and the component densities $f(\underline{x}_i|\theta_k(\underline{w}_i))$ are called *expert networks*.

The MoE model (5) generalizes the mixture model (3) by allowing both the mixing proportions and the component densities to be functions of the covariates. In this paper four possible models are proposed by either allowing or not allowing terms in the model to depend on the covariates. Specifically, the general MoE model in (5) models both the mixing proportions and the component density parameters as functions of covariates. The mixture model is a special case of the general MoE model in which neither the mixing proportions nor the component density parameters are influenced by covariates. The expert network MoE model allows the component density parameters to depend on the covariates, but not the mixing proportions while in the gating network MoE model covariates influence the mixing proportions but not the component density parameters.

In the context of modeling ranked data, the mixture model has been proposed previously by Gormley and Murphy (2006, 2008a) and the gating network MoE model was proposed by Gormley and Murphy (2008b). In other contexts, Hurn et al. (2003) used the expert network MoE model in the special case where the model reduces to a mixture of regression models. Thompson et al. (1998) used the general MoE model to evaluate diagnostic criteria for diabetes. A unified framework in which the optimal model is chosen using model selection techniques has not been employed in any of these applications.

In the MoE model for ranked preference data, for example, the general MoE model allows both the component densities (or expert networks) and the mixing proportions (or gating networks) depend on the covariates. Explicitly,

$$\begin{aligned}\mathbf{P}(\underline{x}_i|\underline{w}_i) &= \sum_{k=1}^{K}\pi_k(\underline{w}_i)f(\underline{x}_i|\theta_k(\underline{w}_i)) \\ &= \sum_{k=1}^{K}\frac{\exp(\underline{\beta}_k^T\underline{w}_i)}{\sum_{r=1}^{K}\exp(\underline{\beta}_r^T\underline{w}_i)}\left\{\prod_{t=1}^{n_i}\frac{\exp(\underline{\gamma}_{kc(i,t)}^T\underline{w}_i)}{\sum_{s=t}^{N}\exp(\underline{\gamma}_{kc(i,s)}^T\underline{w}_i)}\right\},\end{aligned} \tag{7}$$

which arises from substituting equations (2) and (6) into (5). The mixture model, the gating network MoE model and the expert network MoE model are special cases of (7) where the mixing proportions and/or the component densities are treated as constant with respect to the covariates.

This paper provides an unifying framework for mixture of experts modeling for ranked preference data by including all four models and allowing the most appropriate model to be selected using model selection criteria. This framework addresses the question of how and where covariates can be used in the clustering of ranked preference data.

3 Model Fitting and Selection

3.1 Model Fitting

The mixture of experts model for ranked preference data can be fitted in a maximum likelihood framework using an Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997). The methods for fitting the model closely follow the methods outlined in Gormley and Murphy (2006, 2008a,b); model fitting details for the more general model are outlined in this paper.

Let $\boldsymbol{\gamma} = (\underline{\gamma}_1, \underline{\gamma}_2, \ldots, \underline{\gamma}_N)$ and $\boldsymbol{\beta} = (\underline{\beta}_1, \underline{\beta}_2, \ldots, \underline{\beta}_K)$ be the unknown parameters in the general MoE model for ranked preference data (7). The likelihood function is of the form,

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{M} \left[\sum_{k=1}^{K} \frac{\exp(\underline{\beta}_k^T \underline{w}_i)}{\sum_{r=1}^{K} \exp(\underline{\beta}_r^T \underline{w}_i)} \left\{ \prod_{t=1}^{n_i} \frac{\exp(\underline{\gamma}_{kc(i,t)}^T \underline{w}_i)}{\sum_{s=t}^{N} \exp(\underline{\gamma}_{kc(i,s)}^T \underline{w}_i)} \right\} \right]. \tag{8}$$

This likelihood function (8) is not easy to maximize directly, due to the conditional mixture form of the likelihood. As a result, an EM algorithm is used for model fitting.

The EM algorithm is used to provide maximum likelihood parameter estimates when some of the data under study is (treated as) missing. In this case a latent indicator variable $\underline{z}_i = (z_{i1}, \ldots, z_{iK})$ is imputed which records the unknown group membership of observation i, where $z_{ik} = 1$ if observation i comes from group k and $z_{ik} = 0$ otherwise. The complete data likelihood (i.e. the likelihood of both the missing and observed data) is then of the form,

$$L_C(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{M} \prod_{k=1}^{K} \left[\frac{\exp(\underline{\beta}_k^T \underline{w}_i)}{\sum_{r=1}^{K} \exp(\underline{\beta}_r^T \underline{w}_i)} \left\{ \prod_{t=1}^{n_i} \frac{\exp(\underline{\gamma}_{kc(i,t)}^T \underline{w}_i)}{\sum_{s=t}^{N} \exp(\underline{\gamma}_{kc(i,s)}^T \underline{w}_i)} \right\} \right]^{z_{ik}}, \tag{9}$$

giving the complete data log-likelihood

$$\begin{aligned} l_C(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \log L_C(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \log \left[\frac{\exp(\underline{\beta}_k^T \underline{w}_i)}{\sum_{r=1}^{K} \exp(\underline{\beta}_r^T \underline{w}_i)} \left\{ \prod_{t=1}^{n_i} \frac{\exp(\underline{\gamma}_{kc(i,t)}^T \underline{w}_i)}{\sum_{s=t}^{N} \exp(\underline{\gamma}_{kc(i,s)}^T \underline{w}_i)} \right\} \right] \\ &= \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \log \left\{ \frac{\exp(\underline{\beta}_k^T \underline{w}_i)}{\sum_{r=1}^{K} \exp(\underline{\beta}_r^T \underline{w}_i)} \right\} \\ &+ \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \sum_{t=1}^{n_i} \log \left\{ \frac{\exp(\underline{\gamma}_{kc(i,t)}^T \underline{w}_i)}{\sum_{s=t}^{N} \exp(\underline{\gamma}_{kc(i,s)}^T \underline{w}_i)} \right\}. \end{aligned} \tag{10}$$

The EM algorithm is an iterative algorithm in which each iteration consists of two steps — an expectation step and a maximization step. At the expectation or E step the expected value of the complete data log-likelihood is estimated; in this case the E step reduces to estimating the expected value of the missing data $\mathbf{z}$. At the maximization or M step the expected complete data log-likelihood is then maximized with respect to the model parameters, producing on convergence (at least local) maximum likelihood parameter estimates.

At the M step of the EM algorithm, maximization of an equation of the form (10) can be achieved by noticing that the first term is of the same form as a multinomial logistic regression likelihood and the second term is of the same form as the likelihood for fitting an exploded logit mixture model. The two terms involve independent parameters, so they can be maximized independently at the M step. A Minorization-Maximization (MM) algorithm (Hunter, 2004; Hunter and Lange, 2004; Lange et al., 2000) is used here as a more efficient way to implement the M step of the EM algorithm. A Minorization-Maximization algorithm proceeds by iteratively maximizing a minorizing surrogate function which approximates the original objective function to be maximized. Full details are provided in Appendix C for the expert network MoE model. Extra details for the implementation of the M step for the other models are contained in Gormley and Murphy (2008a,b).

Approximate standard errors for the model parameters are computed from the empirical information matrix as outlined in McLachlan and Krishnan (1997) and McLachlan and Peel (2000), after the EM algorithm has converged.

Given the definition of the latent variables $\underline{z}_i$ the structure of the different forms of the MoE model for rank data can be clarified, as illustrated using a graphical model in Figure 1.

3.2 Model Selection

In the unifying framework developed here, all of the MoE models for ranked preference data are fitted over a range of values for K. The Bayesian Information Criterion (BIC) (Kass and Raftery, 1995; Schwartz, 1978) is used for model comparison; this criterion is a penalized likelihood criterion which rewards model fit but penalizes unparsimonious models. The BIC value is defined to be

$$-2 \times (\text{maximized log-likelihood}) + \log(M)(\text{number of parameters}),$$

where the first term measures model fit and the second term penalizes for complexity. Small BIC values indicate an optimal model. The use of BIC for model selection in mixture models is supported by theoretical results concerning consistency (Keribin, 2000; Leroux, 1992) and by practical performance (eg. Fraley and Raftery, 2002; Gormley and Murphy, 2008a; McNicholas and Murphy, 2008; Murphy and Martin, 2003). There are a number of other model selection methods available including the Akaike Information Criterion (AIC) (Akaike, 1973), Integrated Completed Likelihood (ICL) (Biernacki et al., 2000) and cross-validated likelihood (Smyth, 2000). Yang and Yang (2007) discuss the use of BIC and other information criteria in the separation of latent classes and conclude that care is advised when separating a large number of latent classes when sample size is small. Additionally Yang and Yang (2007) comment that the inclusion of informative covariates improves the performance of information criteria when separating latent classes. In the applications examined here, we found that BIC gave good clustering results that closely correspond to the findings in Gormley and Murphy (2008a).

The space of potential MoE models for ranked preference data is very large, once variable selection for the covariates entering the mixing proportions and mixture com-

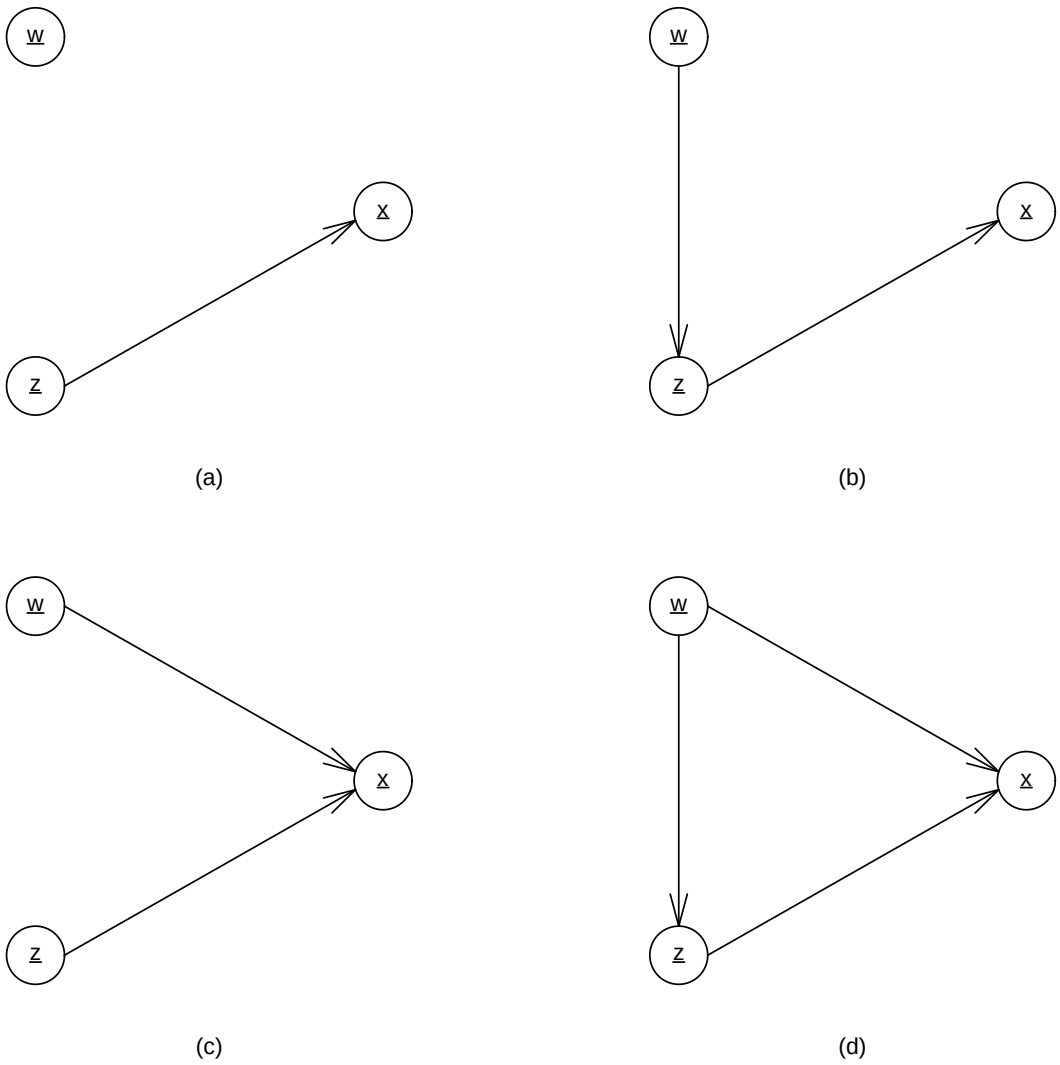


Figure 1: Graphical model representation of the four mixture of experts models: (a) in the mixture model, the ranking distribution depends on the latent variable $\underline{z}$ and the model is independent of the covariates $\underline{w}$; (b) in the gating network MoE model, the ranking distribution depends on the latent variable $\underline{z}$ and the distribution of the latent variable depends on $\underline{w}$; (c) in the expert network MoE model, the ranking distribution depends on the covariates $\underline{w}$ and the latent variable $\underline{z}$; the distribution of the latent variable is independent of the covariates; (d) in the general MoE model, the ranking distribution depends on the covariates $\underline{w}$ and the latent variable $\underline{z}$ and the distribution of the latent variable also depends on the covariates.

ponents is considered. Here only models where covariates enter all mixture components or all mixing proportions are considered in order to restrict the size of the model search space. In fact, even for this reduced model space, there are $K \times 2^L \times 2^L$ possible models

to consider.

A forwards covariate selection procedure was used to find the optimal model within each type of MoE model. Initially all possible models incorporating a single covariate were fitted. The covariate in the optimal model, as determined by BIC, is then retained. The remaining covariates are then added in turn to the optimal model selected at the first stage, and the best model from this set is selected using BIC. This process continues until all covariates are included. All models can then be compared via the BIC.

4 Illustrative Applications

Clustering preference data in the presence of covariates is illustrated through the use of two applications — the first involves clustering a set of respondents from a food preference survey, while the second involves clustering members of the Irish electorate.

4.1 The Hamburger Preparation Quiz

In 1996 the Market Research Corporation of America carried out an extensive national mail survey called the Menu Census Survey. The aim of the survey was to conduct an in-depth study of consumer food safety behavior. As a supplement to the Menu Census Survey respondents were required to complete a 'Hamburger Preparation Quiz'. In this supplement respondents detailed their preferences for hamburgers. Typical survey items involved respondents stating their preferred hamburger order in a restaurant and their taste preferences for hamburger styles. Demographic information such as the age, the population size of their residential area, ethnicity and the gender of each respondent was also recorded, as was the type of diet (if any) that the respondent was currently following. The Hamburger Preparation Quiz (HPQ) was completed by 1133 individuals, of which 594 provided complete responses to the demographic questions. The adult with the most recent birthday in each household completed the HPQ. Appendix A provides full details of the source of the HPQ data.

Interest lies in determining if groups (or *clusters*) of people with similar preferences for hamburgers exist within the population. If such groups do exist, interest lies in determining the types of preferences within each cluster. Moreover, given that demographic information is available, examining the influence demographic factors may have on the clustering structure of the population and/or on the preferences within clusters is of interest.

Question twelve from the HPQ asks respondents to rank hamburger patties in terms of their taste. Specifically, respondents were asked to rank rare (R), medium-rare (MR), medium (M), medium-well (MW) and well-done (WD) patties in order of their preference. A 'don't know/not sure' option was also offered to respondents but no respondents availed of this choice. In this application the response to question twelve is treated as the rank response observation from each respondent. These data were previously analyzed in Ralston et al. (2002) and Bao and Meilă (2008). Details of the demographic information

or covariates recorded are provided in Table 1.

Table 1: Demographic information recorded in the Hamburger Preparation Quiz. The levels of each demographic covariate are detailed as appropriate.

Age	**Residential area population size**	**Diet**	**Ethnicity**	**Gender**
—	Farm	Not on diet	Black	Female
	< 2500	Diet for medical reasons	White	Male
	2,500–9,999	Diet to gain weight	Other	
	10,000–49,999	Diet to maintain weight		
	50,000–99,999	Diet to reduce weight		
	100,000-249,000	Diet for other reasons		
	250,000-499,999			
	500,000-999,999			
	1 million – 2 million			
	> 2 million			

4.2 Irish Election Data

Both governmental and presidential elections in Ireland employ a preferential voting electoral system known as 'proportional representation by means of a single transferable vote' (PR-STV). Under this system voters rank, in order of their preference, some or all of the electoral candidates. The counting process which results in the election or elimination of candidates is an intricate procedure which involves the transfer of votes between candidates as specified by the rank ballots of the voters. Full details of the mechanics of the PR-STV electoral system are given in Sinnott (2004). Further details on the Irish political system in general are given in Coakley and Gallagher (2004) and Sinnott (1995).

In this article the electorate from the 1997 Irish presidential election is analyzed. In 1997 five candidates ran for the office of President of Ireland. Mary Banotti was endorsed by the political party Fine Gael who were the main government opposition party in 1997. She was deemed to be a liberal candidate and was popular throughout the electoral campaign. Mary McAleese was backed by the current government party Fine Fáil and was known as a conservative candidate. McAleese was widely believed to be the favorite for the presidency throughout the campaign and she was subsequently elected as President of Ireland on October 30th 1997. Derek Nally was a late contender for the post, only being nominated as a potential candidate one month prior to election day. He ran on an independent ticket and received the least number of first preference votes on polling day. Adi Roche's involvement in the Irish presidential campaign in 1997 was the most unstable. Roche was backed by another government opposition party, the Labour party, and began the campaign as joint favorite for the presidency along with McAleese.

Her liberal campaign contrasted with McAleese's conservative campaign, but Roche's popularity began to dramatically decline after negative publicity regarding her work affairs emerged in the media. As the campaign developed her support ratings dropped. The fifth candidate Rosemary Scallon was an independent, conservative, candidate. Scallon's support pattern during the campaign was the reverse of Roche's. Scallon began the campaign with an extremely small support base, but as the campaign wore on she emerged as a capable candidate and finished in a respectable third place, behind the favorites McAleese and Banotti. Table 2 summarizes the candidates' details. A full treatment of the vote counting process in the 1997 Irish presidential election, and further details, can be found in Gormley and Murphy (2008b). Additionally, a detailed account of the 1997 Irish presidential election campaign is provided by Marsh (1999).

Table 2: Information about the five candidates who ran for the office of President of Ireland in 1997.

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

Irish Marketing Surveys (IMS) completed a survey one month prior to the 1997 presidential election. At this time, Roche was still a major contender in the election and Nally had only started his campaign a few days earlier. In the IMS poll, a sample of 1100 potential voters were asked to list the candidates in order of preference (as if they were voting on that day); seventeen people who were sampled said that they did not intend to vote, so they were excluded from this analysis. In addition to the voting preferences, a number of socioeconomic variables were recorded for each person sampled in the poll; these are listed in Table 3. Further details on this poll and the covariates are provided in Appendix B and in Gormley and Murphy (2008b).

Table 3: Covariates recorded for each voter sampled in the IMS poll

Age	**Area**	**Gender**	**Government satisfaction**	**Marital status**	**Social class**
—	City	Housewife	Satisfied	Married	AB
	Town	Non-housewife	Dissatisfied	Single	C1
	Rural	Male	No opinion	Widowed	C2
					DE
					F50+
					F50-

In this application interest lies in determining if groups (or 'voting blocs') of voters with similar preferences for the electoral candidates exist within the electorate. If the electorate is heterogeneous, interest lies in determining the preferences for the candidates within each voting bloc. Examining the influence the recorded socioeconomic variables may have on the clustering structure and/or on the preferences within voting blocs is also of interest.

5 Application results

The proposed MoE model for ranked preference data was applied to the illustrative examples in order to determine if and how the covariates enter the model and how their inclusion affects the clustering results.

5.1 The Hamburger Preparation Quiz

All of the MoE models for ranked preference data were fitted to the Hamburger Preparation data (Section 4.1) with $K = 1, 2, \ldots, 10$. The forwards selection strategy for selecting covariates outlined in Section 3.2 was utilized and the model with the highest Bayesian Information Criterion (BIC) was found within each type of MoE model. The results of this analysis are shown in Table 4.

Table 4: The model with the smallest BIC within each type of the mixture of experts model for ranked preference data applied to the Hamburger Preparation Quiz data.

	BIC	K	**Covariates**
The mixture model	3677	6	—
The gating network MoE model	3722	6	π_k: Ethnicity
The expert network MoE model	3922	5	$\underline{p}_k$: Gender
The general MoE model	4682	2	π_k: Ethnicity $\underline{p}_k$: Area size

Based on the BIC, the optimal model in this case (the mixture model) suggests that there are six groups of judges in the population and that none of the recorded covariates are informative in the modeling. The difference in BIC values indicates very strong support (Kass and Raftery, 1995, Section 3.2) for the fact that the covariates are noninformative. The mixing proportions and support parameters of the optimal model are detailed in Table 5 and shown using a mosaic plot (Emerson, 1998; Hartigan and Kleiner, 1981) in Figure 2. In the mosaic plot, the width of the bar shows the mixing proportion for each

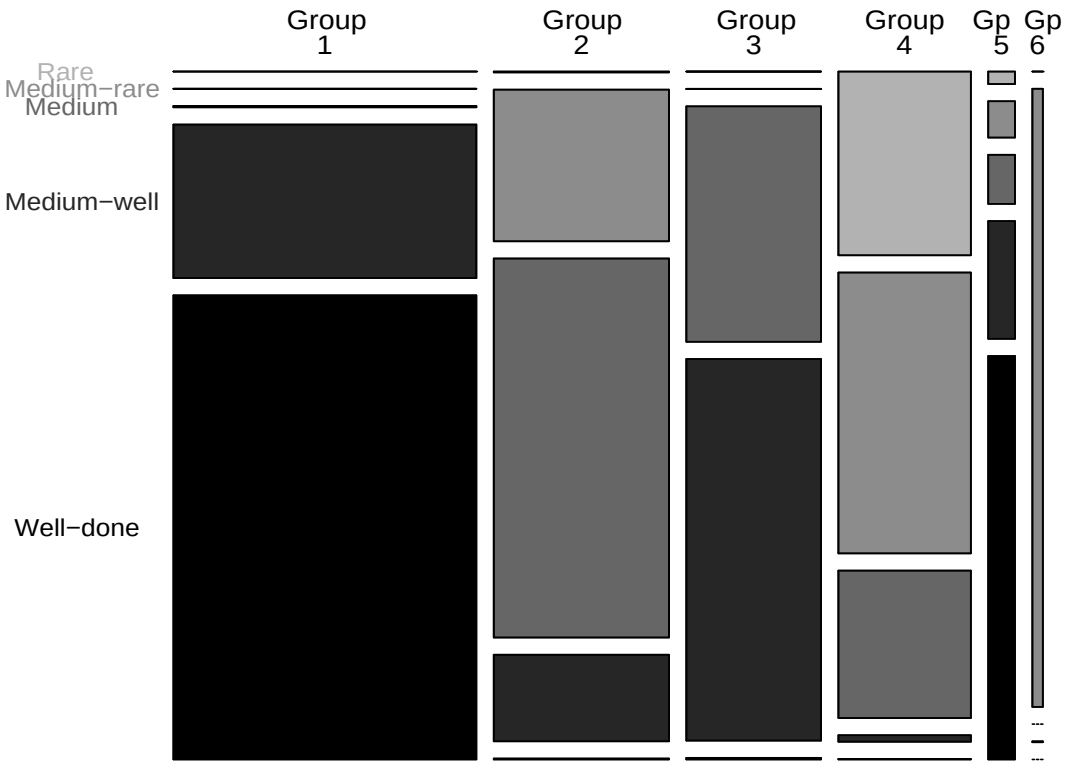


Figure 2: For the Hamburger Preparation Quiz data, a mixture model with 6 groups was deemed the optimal model according to the BIC (see Table 4). A mosaic plot representation of the mixture model parameter estimates is given — the width of a bar illustrates the mixing proportion for each group and the division of a bar shows the support parameter values within each group. Parameter estimates and standard errors are detailed in Table 5.

component and the division of the bar shows the support parameter values within each component.

Intuition on the suitability of the modeling techniques employed can be provided through model diagnostics. Here a comparison is made between the expected number of first preferences for each choice category, given the estimated model parameters, and the observed number of first preferences for each choice category. The resulting χ^2 test statistic (detailed in Table 6) demonstrates the suitability of the employed modeling techniques (p-value = 0.56).

Interestingly, the six groups found in this analysis correspond closely to the six groups discovered by Bao and Meilă (2008) using a different modeling framework. It is also notable that within each group the support parameters take large values for a contiguous subset of the available choices. This is intuitive as there is a natural ordering to the tastes being ranked. Also, support only tends to be high for one, two or three tastes within each component suggesting that there are precise preferences within each group.

Table 5: Mixture model support parameter estimates (given as percentages) for the Hamburger Preparation Quiz data. Standard errors are given in parentheses. Figure 2 provides an illustration of the estimates.

	Group 1		**Group 2**		**Group 3**		**Group 4**		**Group 5**		**Group 6**	
Rare	0	(< 1)	0	(< 1)	0	(< 1)	30	(< 1)	2	(< 1)	0	(< 1)
Medium-rare	0	(< 1)	25	(< 1)	0	(< 1)	45	(< 1)	6	(< 1)	100	(< 1)
Medium	0	(< 1)	61	(< 1)	38	(< 1)	24	(< 1)	8	(< 1)	0	(< 1)
Medium-well	25	(< 1)	14	(< 1)	62	(2)	1	(< 1)	19	(< 1)	0	(< 1)
Well-done	75	(< 1)	0	(< 1)	0	(< 1)	0	(< 1)	65	(4)	0	(< 1)
Mixing proportion	39	(3)	22	(2)	17	(2)	17	(2)	3	(< 1)	2	(< 1)

Table 6: Observed and expected number of first preferences for each choice category in the Hamburger Preparation Quiz data. The χ^2 statistic with four degrees of freedom is not significant, suggesting a good model fit.

	Rare	**Medium rare**	**Medium**	**Medium well**	**Well done**	
Observed number of first preferences	36	75	148	149	186	$\chi^2_4 = 2.99$
Expected number of first preferences	31	87	146	144	186	$p = 0.56$

Ralston et al. (2002) examined the Hamburger Preparation Quiz respondents' cooking and food ordering habits and found that the covariate "area size" had a significant effect on ordering, with respondents from large cities having a higher probability of ordering lightly cooked burgers; this analysis did not find this effect in the taste preference data. The analysis in Ralston et al. (2002) uses the taste preference data as a predictor for cooking and food ordering habits rather than as an outcome variable.

5.2 The Irish Presidential Election

All of the MoE models for ranked preference data were fitted for $K = 1, 2, \ldots, 5$ and the forwards covariate selection method was employed when selecting the optimal model using the BIC. The optimal models for each type of MoE model are reported in Table 7.

Based on the BIC values, the optimal model overall is a gating network MoE model with four components where 'age' and 'government satisfaction' are important covariates for determining group or 'voting bloc' membership. Interestingly, the covariates are not

Table 7: The model with smallest BIC within each type of the mixture of experts model for ranked preference data applied to the 1997 Irish Presidential Election data.

	BIC	K	**Covariates**
The gating network MoE model	8491	4	π_k: Age, Government satisfaction
The general MoE model	8512	3	π_k: Age, Government satisfaction $\underline{p}_k$: Age
The mixture model	8513	3	—
The expert network MoE model	8528	1	$\underline{p}_k$: Government satisfaction

informative within voting blocs, but only in determining voting bloc membership. This model corresponds to the model applied to these data in Gormley and Murphy (2008b).

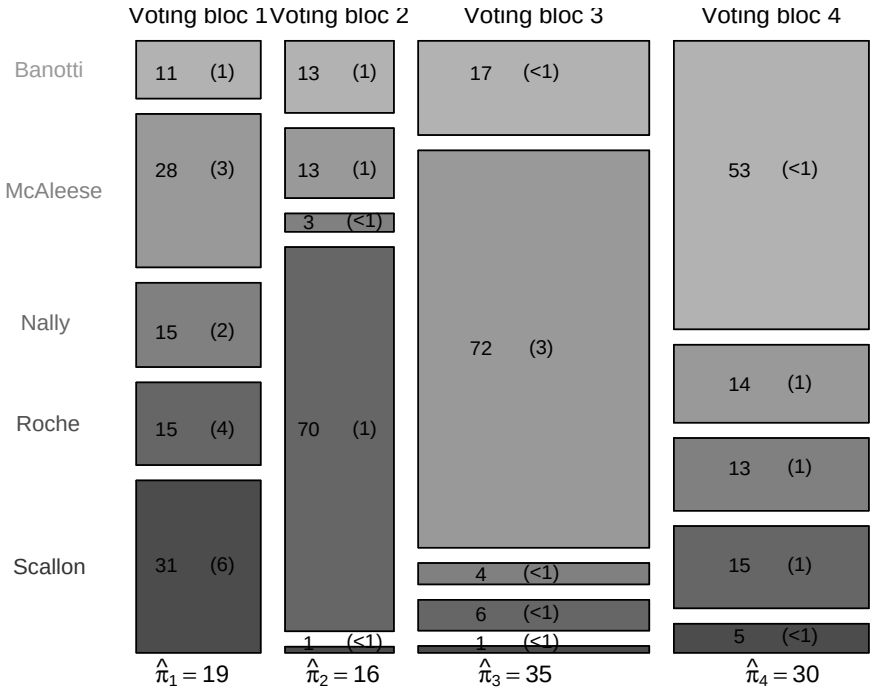
The support parameters for the optimal model are presented in mosaic plot form in Figure 3(a). For comparison purposes, a mosaic plot of the support parameters for the best mixture model are shown in Figure 3(b).

The support parameter estimates have an interpretation in terms of political party competition and in terms of a conservative-liberal competition. Voting bloc 1 is the 'conservative voting bloc' with larger support parameters for McAleese and Scallon. Voting bloc 2 has large support for the liberal candidate Adi Roche. This voting bloc indicates that the model has uncovered some of the observed characteristics of the presidential campaign at the time the poll was taken in that Adi Roche has large support. Voting bloc 3 is the largest voting bloc in terms of marginal mixing proportions and intuitively has larger support parameters for the high profile candidates McAleese and Banotti. Voters belonging to voting bloc 4 favor Banotti and have more uniform levels of support for the other candidates. A detailed discussion of this optimal model is also given in Gormley and Murphy (2008b).

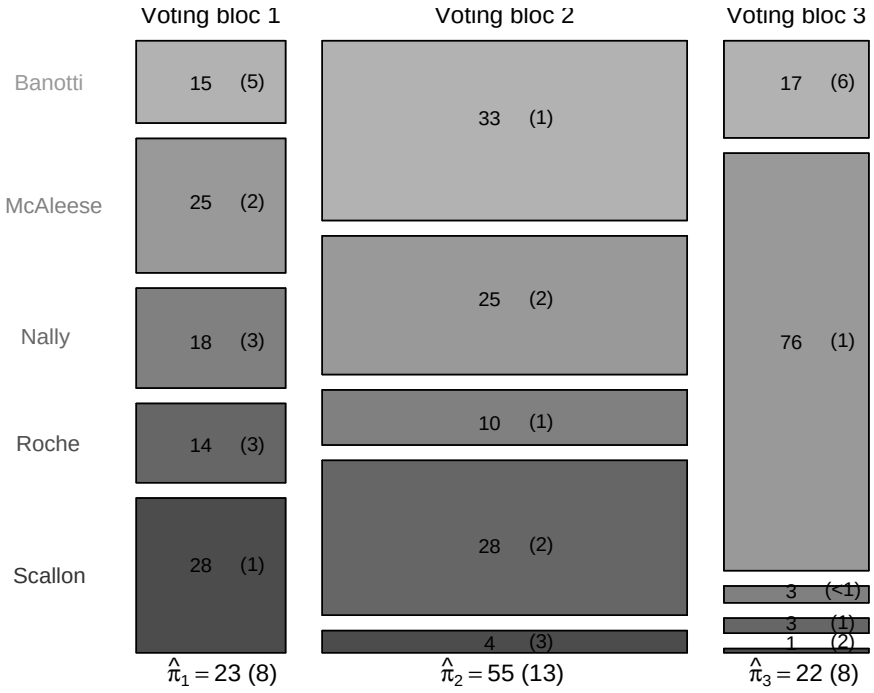
The groups found in the mixture of experts model and the mixture model show some correspondence. Voting blocs 2 and 4 in the mixture of experts model (Figure 3(a)) appear to be a division of voting bloc 3 in the mixture model (Figure 3(b)). This suggests that the mixture of experts model for ranked preference data was able to discover a finer division of the voters into voting blocs than the mixture model.

Table 8 details the odds ratios computed for the mixing proportion (or gating network) parameters $\boldsymbol{\beta} = (\underline{\beta}_1, \ldots, \underline{\beta}_K)$. In the model, voting bloc 1 (which is the conservative voting bloc) is the baseline voting bloc. In addition, in the 'government satisfaction' covariate, the baseline was chosen to be 'no opinion'.

The odds ratios for the mixing proportions parameters also provide intuitive results within the context of the Irish presidential election. For example, older (and generally more conservative) voters are much less likely to belong to the liberal voting bloc 2 than to the conservative voting bloc 1. Also, voters with some interest in government are more



(a)The gating network MoE model



(b)The mixture model

Figure 3: A mosaic plot representation of the support parameters (given as percentages) for (a) the gating network MoE model for ranked preference data and (b) the mixture model fitted to the Irish Presidential Election data. The width of each column denotes the marginal mixing proportions.

Table 8: Odds ratios for the mixing proportion parameters in the gating network MoE model for ranked preference data.

		Age	**Satisfied**	**Not satisfied**
Voting bloc 2	Odds ratio	0.01	1.14	2.80
	95% CI	[0.00, 0.05]	[0.42, 3.11]	[0.77, 10.15]
Voting bloc 3	Odds ratio	0.95	3.12	3.81
	95% CI	[0.32, 2.81]	[0.94, 10.31]	[0.90, 16.13]
Voting bloc 4	Odds ratio	1.56	0.35	3.50
	95% CI	[0.35, 6.91]	[0.12, 0.98]	[1.07, 11.43]

likely to belong to voting bloc 3, the bloc favoring candidates backed by large government parties, than to belong to the conservative voting bloc 1. Voting bloc 1 had high levels of support for the independent candidate Scallon. The mixing proportions parameter estimates further indicate that voters dissatisfied with the current government are more likely to belong to voting bloc 4 than to voting bloc 1. This is again intuitive as voting bloc 4 favors Mary Banotti who was backed by the main government opposition party, while voting bloc 1 favors the government backed Mary McAleese. Further interpretation of the mixing proportion parameters are given in Gormley and Murphy (2008b).

To diagnose the suitability of the modeling techniques employed, a comparison is made between the expected number of first preferences for each electoral candidate, given the estimated model parameters, and the observed number of first preferences for each candidate. For the optimal (in terms of BIC) gating network MoE model, the resulting χ^2 test statistic (detailed in Table 9 (a)) suggests good model fit; this is not the case for the less preferable mixture model (Table 9 (b)).

6 Discussion

A novel model has been developed to accommodate the use of covariates when clustering ranked preference data. The model developed offers the flexibility to allow covariates influence the clustering by allowing covariate dependence to enter different parts of the model. Efficient model fitting procedures are developed using a hybrid of the EM and MM algorithms. Optimal models are selected in a statistically principled manner via a model selection criterion.

In the application of the model to the Hamburger Preparation Quiz data six clusters were found and the covariates were found to be noninformative in the modeling. However, in the analysis of the Irish election data the covariates were found to be informative and their inclusion provided a deeper picture of the voting bloc structure in the electorate than a standard mixture model does.

In other applications, the covariates may enter the model in different manners and the resulting models have the potential to provide a deeper understanding of the population

Table 9: Table (a) details the observed number of first preferences and the expected number for the gating network MoE model illustrated in Figure 3 (a). Table (b) details the expected values for the mixture model illustrated in Figure 3 (b). For the gating network MoE model the χ^2 statistic suggests a good model fit; this is not the case for the suboptimal (in terms of BIC) mixture model.

(a)

	Banotti	**McAleese**	**Nally**	**Roche**	**Scallon**	
Observed number of first preferences	277	411	89	222	84	$\chi^2_4 = 0.85$
Expected number of first preferences	278	399	94	224	88	$p = 0.93$

(b)

	Banotti	**McAleese**	**Nally**	**Roche**	**Scallon**	
Observed number of first preferences	277	411	89	222	84	$\chi^2_4 = 88.9$
Expected number of first preferences	218	424	129	157	156	$p = 0.00$

than standard clustering methods that do not incorporate covariates. The proposed model formalizes the practice of trying to understand the cluster structure using covariates by including the covariates in the model directly.

The Plackett-Luce model was employed as the rank data model within each homogeneous group in the MoE model for ranked preference data. Alternative rank data models could also be employed in the general MoE model framework developed here. Benter's model for rank data (Benter, 1994) is one obvious alternative — Benter's model is similar to the Plackett-Luce model in that it is parameterized by the same support parameters, but it also has an additional dampening parameter. The dampening parameter models the way in which judges may make choices at different levels within their ranking with differing amounts of certainty. A mixture of Benter models was employed in Gormley and Murphy (2008a) to analyze Irish election data; this model could be extended to include covariates. Many other rank data models are available; see Marden (1995) for further details.

The MoE models for rank data developed here are ideal for specifically modeling stated ranked preference data (eg. from surveys). However, these models are essentially variations on standard discrete choice models, tailored for ranked data. For example, the expert network MoE model for rank data is simply a mixture of standard logit choice models which have been tailored to model rank data (Train, 2003); this can also be thought of a mixed logit model (McFadden and Train, 2000) for rank data where the mixing density is discrete. Hence, the general framework detailed here can be applied

to other forms of choice data by changing the component densities $f(\underline{x}_i|\theta_k(\underline{w}_i))$ to the appropriate form. The parameters of the appropriate density may then be modeled as a function of the covariates.

The generalization of the MoE model for rank data to MoE models for any form of choice data highlights a link with the popular latent class model for choice data. Greene and Hensher (2003) contrast the latent class model with the mixed logit model using an illustrative study in which preferences for road environments are recorded. Latent class models have also been extended to include covariates; for example Dayton and Macready (1988) develop the concomitant-variable latent class model where covariates enter both the mixing proportions and the class specific probabilities using a logistic framework. Reboussin et al. (2008) also incorporate covariates in the latent class model when modeling data from a large scale survey of under-age drinking.

Preference ranking data arises in a wide range of contexts and the proposed model has potential applications in these contexts. For example, marketing surveys such as the Hamburger Preparation Quiz examined in this paper are widespread. The modeling framework developed here can be employed to not only highlight clusters of consumers, but also the covariates which influence, or perhaps significantly do not influence, the clustering structure. The model allows for a detailed analysis of clustering and the effect of covariates on rankings. However a limitation of the proposed MoE model for rank data is its unsuitability for evaluating standard choice modeling outputs such as forecasts or 'willingness to pay' measures (Hensher et al., 2005). Due to the inherent nature of ranked data output measures are difficult to evaluate. Moreover, even diagnosing the suitability of ranked data models poses problems.

The model could be extended to include object covariates as well as covariates for the judges. This would offer an even deeper understanding of the preference ranking procedure. In the Irish election context presented here, for example, including candidate covariates such as their area of residence may provide deeper insight to structure of the electorate and/or to the electorates' preferences.

More advanced methods for selecting the covariates could be considered and there is the possibility of expanding the model space so the covariates only enter some of the mixing proportions or some of the component densities rather than all; this approach was used in a different context by Gustafson and Lefebvre (2008).

Acknowledgments

This research was funded by a Science Foundation Ireland Research Frontiers Programme Grant (06/RFP/M040). The authors would like to thank Professor Adrian Raftery, the members of the Center for Statistics and the Social Sciences and the members of the Working Group on Model-based Clustering at the University of Washington for numerous suggestions that contributed enormously to this work. The authors would also like to thank the anonymous referees for helpful suggestions that have added to the overall quality of this work.

A The Hamburger Preparation Quiz Data Source

The Hamburger Preparation Quiz data set was collected in 1996 by the Market Research Corporation of America through the Menu Census Survey. The Hamburger Preparation Quiz form and the data set are freely available from
http://www.ers.usda.gov/Data/Hamburger/.

B Irish Election Data Source

The 1997 Irish presidential opinion poll data set was collected by Irish Marketing Surveys and is available through the Irish Elections Data Archive
http://www.tcd.ie/Political Science/elections/elections.html
which is maintained by Professor Michael Marsh in the Department of Political Science, Trinity College Dublin, Ireland.

C Mathematical Details for the EMM algorithm

In this section, the expert network MoE model for ranked preference data will be employed to illustrate parameter estimation via the EMM algorithm. In the expert network MoE model, the support parameters within each group are modeled as a function of the judges' covariates. Specifically, for $j = 1, \ldots, N$, $k = 1, \ldots, K$ and the covariates of judge i, $\underline{w}_i$

$$p_{kj}(\underline{w}_i) = \exp(\underline{\gamma}_{kj}^T \underline{w}_i)$$

where $\underline{\gamma}_{kj} = (\gamma_{kj0}, \ldots, \gamma_{kjL})$ is a vector of parameters. To ensure identifiability in the expert network MoE model $\underline{\gamma}_{k1} = (0, \ldots, 0)$ meaning $p_{k1} = 1$ in all groups. Under this definition $p_{k1} + \cdots + p_{kN} \neq 1$, but the structure of the Plackett-Luce density ensures valid probabilities of the final preference orderings.

The complete data log-likelihood for the expert network MoE model is

$$l_C(\underline{\pi}, \boldsymbol{\gamma}) = \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \underline{\gamma}_{kc(i,t)}^T \underline{w}_i - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} \exp\{\underline{\gamma}_{kc(i,s)}^T \underline{w}_i\} \right\}. \quad (11)$$

Maximizing the complete data log likelihood (11) via the EMM algorithm provides maximum likelihood estimates (MLEs) for $\underline{\pi}$ and $\boldsymbol{\gamma}$.

The EMM algorithm

The EM algorithm (Dempster et al., 1977) is an iterative algorithm consisting of two steps per iteration, an 'E' or expectation step and a 'M' or maximization step. In the EMM algorithm maximization at the M step is achieved by employing ideas from MM algorithms (Hunter and Lange, 2004).

In the E step of the EMM algorithm the expected value of the complete data log-likelihood is calculated; essentially this step involves calculating the expected value of the latent variables, $\mathbf{z}$. The form of the E step is independent of the type of MoE model for ranked preference data; the appendix in Gormley and Murphy (2008a) provides full details of the E step.

In the M step of the EMM algorithm the expected complete data log-likelihood derived in the E step is maximized with respect to the model parameters, $\underline{\pi}$ and $\boldsymbol{\gamma}$. In the expert network MoE model for ranked preference data the mixing proportions $\underline{\pi}$ are treated as independent of the voter covariates. The form of the estimate $\hat{\underline{\pi}}$ derived at the M step is therefore the same as that derived under a mixture of Plackett-Luce models; details can be found in Gormley and Murphy (2008a).

In the expert network MoE model the Plackett-Luce support parameters $\mathbf{p}$ are treated as functions of the voter covariates with parameters $\boldsymbol{\gamma}$. Maximization of the expected value of (11) with respect to $\boldsymbol{\gamma}$ is complex due to the intricate form of the Plackett-Luce density. Ideas from optimization transfer algorithms or 'MM algorithms' are therefore employed to maximize the expected value of (11) with respect to $\boldsymbol{\gamma}$.

Constructing linear surrogate functions

Differentiating the expected value of (11) with respect to γ_{kjl} for $k = 1, \ldots, K$, $j = 2, \ldots, N$ and $l = 0, \ldots, L$ is problematic due to the term $-\log \sum_{s=t}^{N} \exp\{\underline{\gamma}_{kc(i,s)}^T \underline{w}_i\}$. Such maximization issues may be overcome by implementing an optimization transfer algorithm in which optimization is transferred from the problematic objective function to a suitable surrogate function (Lange et al., 2000). Surrogate functions are constructed by exploiting mathematical properties of (part of) the problematic objective function. One approach to constructing linear surrogate functions employs the supporting hyperplane property of a convex function. If $f(\theta)$ is a convex function with differential $f'(\theta)$ and $\bar{\theta}$ denotes a constant value of the generic parameter θ, then

$$f(\theta) \geq f(\bar{\theta}) + f'(\bar{\theta})(\theta - \bar{\theta}).$$

Since $-\log(\cdot)$ is a convex function

$$-\log \sum_{s=t}^{N} \exp\{\underline{\gamma}_{kc(i,s)}^T \underline{w}_i\} \geq -\log \sum_{s=t}^{N} \exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T \underline{w}_i\} + 1 - \frac{\sum_{s=t}^{N} \exp\{\underline{\gamma}_{kc(i,s)}^T \underline{w}_i\}}{\sum_{s=t}^{N} \exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T \underline{w}_i\}}$$

where $\bar{\boldsymbol{\gamma}}$ denotes a constant value of $\boldsymbol{\gamma}$. In practice, this value is the value of the parameter from the previous iteration of the EMM algorithm. Hence the expected complete data

log likelihood becomes, up to a constant,

$$\mathbf{E}\{l_C(\underline{\pi},\boldsymbol{\gamma})\} \geq \sum_{i=1}^{M}\sum_{k=1}^{K}\hat{z}_{ik}\sum_{t=1}^{n_i}\left\{\underline{\gamma}_{kc(i,t)}^T\underline{w}_i - \frac{\sum_{s=t}^{N}\exp\{\underline{\gamma}_{kc(i,s)}^T\underline{w}_i\}}{\sum_{s=t}^{N}\exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T\underline{w}_i\}}\right\}. \quad (12)$$

This surrogate objective function still poses challenges (due to the term $-\exp\{\underline{\gamma}_{kc(i,s)}^T\underline{w}_i\}$) when used to provide estimates of $\boldsymbol{\gamma}$. Optimization transfer algorithms can be implemented again to simplify the maximization.

Constructing quadratic surrogate functions.

The term $-\exp\{\underline{\gamma}_{kc(i,s)}^T\underline{w}_i\}$ is concave and employing a quadratic surrogate function in this case would improve the approximation of the surrogate function to the objective function. A concave function $g(\theta)$ can be bounded around $\bar{\theta}$ using a quadratic function via

$$g(\theta) \geq g(\bar{\theta}) + \{g'(\bar{\theta})\}^T(\theta-\bar{\theta}) + 0.5(\theta-\bar{\theta})^T\mathbf{B}(\theta-\bar{\theta})$$

for negative definite $\mathbf{B}$ where $\mathbf{B} > \frac{d^2g(\bar{\theta})}{d\theta^2}$. Hence

$$-\exp(\underline{\gamma}_{kj}^T\underline{w}_i) \geq -\exp(\bar{\underline{\gamma}}_{kj}^T\underline{w}_i) - \underline{w}_i^T\exp(\bar{\underline{\gamma}}_{kj}^T\underline{w}_i)(\underline{\gamma}_{kj}-\bar{\underline{\gamma}}_{kj}) - 0.5(\underline{\gamma}_{kj}-\bar{\underline{\gamma}}_{kj})^T\mathbf{B}(\underline{\gamma}_{kj}-\bar{\underline{\gamma}}_{kj})$$

where $\mathbf{B} = \underline{w}_i^T\underline{w}_i$. The covariates are constrained such that $0 \leq w_{il} \leq 1$ for computational ease. Hence (12) becomes

$$\mathbf{E}\{l_C(\underline{\pi},\boldsymbol{\gamma})\} \geq \sum_{i=1}^{M}\sum_{k=1}^{K}\hat{z}_{ik}\sum_{t=1}^{n_i}\left[\underline{\gamma}_{kc(i,t)}^T\underline{w}_i - \left[\sum_{s=t}^{N}\exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T\underline{w}_i\}\right]^{-1}\sum_{s=t}^{N}\left\{\underline{w}_i^T\exp(\bar{\underline{\gamma}}_{kc(i,s)}^T\underline{w}_i)\underline{\gamma}_{kc(i,s)} + 0.5\underline{\gamma}_{kc(i,s)}^T(\underline{w}_i^T\underline{w}_i)\underline{\gamma}_{kc(i,s)} - \underline{\gamma}_{kc(i,s)}^T(\underline{w}_i^T\underline{w}_i)\bar{\underline{\gamma}}_{kc(i,s)}\right\}\right].$$

This surrogate to the expected complete data log-likelihood is now simply a quadratic function in γ_{kjl} and maximization is straightforward. Maximizing with respect to γ_{kjl} for $k = 1,\ldots,K$ and $j = 2,\ldots,N$ and $l = 0,\ldots,L$ provides the estimate of the MLE $\hat{\gamma}_{kjl}$

$$\hat{\gamma}_{kjl} = \frac{\sum_{i=1}^{M}\hat{z}_{ik}\sum_{t=1}^{n_i}\left[w_{il}\mathbf{1}\{j=c(i,t)\} - \frac{\sum_{s=t}^{N}\left[w_{il}\exp\{\bar{\underline{\gamma}}_{kj}^T\underline{w}_i\} - (\underline{w}_i^T\underline{w}_i)\bar{\gamma}_{kjl}\right]\mathbf{1}\{j=c(i,s)\}}{\sum_{s=t}^{N}\exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T\underline{w}_i\}}\right]}{\sum_{i=1}^{M}\hat{z}_{ik}\sum_{t=1}^{n_i}\left[\frac{\sum_{s=t}^{N}\{(\underline{w}_i^T\underline{w}_i)\}\mathbf{1}\{j=c(i,s)\}}{\sum_{s=t}^{N}\exp\{\bar{\underline{\gamma}}_{kc(i,s)}^T\underline{w}_i\}}\right]}. \quad (13)$$

The EMM algorithm for the MoE model for ranked preference data

In summary, the steps of the EMM algorithm to estimate the MLEs of the parameters of the expert network MoE model for ranked preference data are:

0. Let $h = 0$ and choose initial parameter estimates $\boldsymbol{\gamma}^{(0)}$ and $\underline{\pi}^{(0)}$.

1. **E-Step:** Compute the quantities $z_{ik}^{(h+1)}$ for $i = 1, \ldots, M$ and $k = 1, \ldots, K$ as detailed in Gormley and Murphy (2008a).

2. **M-Step:**

   (a) Compute $\pi_k^{(h+1)}$ for $k = 1, \ldots, K$ as detailed in Gormley and Murphy (2008a).

   (b) Compute $\gamma_{kjl}^{(h+1)}$ for $k = 1, \ldots, K$, $j = 2, \ldots, N$ and $l = 0, \ldots, L$ as detailed in (13).

3. If converged, then stop. Otherwise, increment h and return to Step 1.

Convergence is assessed in this case using Aitken's acceleration (Böhning et al., 1994; Lindsay, 1995). The M step changes for the gating network MoE model and for the general MoE model where the mixing proportions $\underline{\pi}$ are treated as functions of the covariates; full details of the calculations required in these M steps are detailed in Gormley and Murphy (2008b). For the mixture model, an EMM algorithm is also required; details are provided in Gormley and Murphy (2006).

References

Akaike, H. (1973), "Information theory and an extension to the maximum likelihood principle," *Second International Symposium on on Information Theory*, 267–281.

Bao, L. and Meilă, M. (2008), "Clustering permutations by Exponential Blurring Mean-Shift algorithm," Tech. Rep. 524, Department of Statistics, University of Washington.

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

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 22, 719–725.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994), "The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family," *Annals of the Institute of Statistical Mathematics*, 46, 373–388.

Busse, L. M., Orbanz, P., and Buhmann, J. M. (2007), "Cluster Analysis of Heterogeneous Rank Data," in *Proceedings of the 24th International Conference on Machine Learning*, ed. Ghahramani, Z., vol. 227 of *ACM International Conference Proceeding Series*, pp. 113–120.

Chapman, R. and Staelin, R. (1982), "Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model," *Journal of Marketing Research*, 19, 288–301.

Coakley, J. and Gallagher, M. (2004), *Politics in the Republic of Ireland*, London: Routledge in association with PSAI Press, 4th ed.

Critchlow, D. E. (1985), *Metric methods for analyzing partially ranked data*, Lecture Notes in Statistics, 34, Berlin: Springer-Verlag.

Critchlow, D. E., Fligner, M. A., and Verducci, J. (1991), "Probability Models on Rankings," *Journal of Mathematical Psychology*, 353, 294–318.

Dayton, C. M. and Macready, G. B. (1988), "Concomitant-Variable Latent-Class Models," *Journal of the American Statistical Association*, 83, 173–178.

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, 1–38, with discussion.

Dobson, A. J. (ed.) (2002), *An introduction to generalized linear models*, London: Chapman and Hall, 2nd ed.

Emerson, J. W. (1998), "Mosaic Displays in S-PLUS: A General Implementation and a Case Study," *Statistical Computing and Statistical Graphics Newsletter*, 9, 17–23.

Fligner, M. A. and Verducci, J. S. (1988), "Multistage Ranking Models," *Journal of the American Statistical Association*, 83, 892–901.

Fligner, M. A. and Verducci, J. S. (eds.) (1993), *Probability models and statistical analyses for ranking data*, New York: Springer-Verlag, papers from the conference held at the University of Massachusetts, Amherst, Massachusetts, June 8–13, 1990.

Fraley, C. and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–612.

Gormley, I. C. and Murphy, T. B. (2006), "Analysis of Irish third-level college applications data," *Journal of the Royal Statistical Society, Series A*, 169, 361—379.

— (2008a), "Exploring Voting Blocs Within the Irish Electorate: A Mixture Modeling Approach," *Journal of the American Statistical Association*, 103, 1014–1027.

— (2008b), "A mixture of experts model for rank data with applications in election studies," *The Annals of Applied Statistics*, 1452–1477.

Greene, W. H. and Hensher, D. A. (2003), "A latent class model for discrete choice analysis: contrasts with mixed logit," *Transportation Research Part B: Methodological*, 37, 681 – 698.

Gustafson, P. and Lefebvre, G. (2008), "Bayesian multinomial regression with class-specific predictor selection," *The Annals of Applied Statistics*, To appear.

Hartigan, J. A. and Kleiner, B. (1981), "Mosaics for Contingency Tables," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer-Verlag, pp. 268–273.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2005), *Applied Choice Analysis: A Primer*, Cambridge University Press.

Hunter, D. R. (2004), "MM algorithms for generalized Bradley-Terry models," *The Annals of Statistics*, 32, 384–406.

Hunter, D. R. and Lange, K. (2004), "A tutorial on MM algorithms," *The American Statistician*, 58, 30–37.

Hurn, M., Justel, A., and Robert, C. P. (2003), "Estimating mixtures of regressions," *Journal of Computational and Graphical Statistics*, 12, 55–79.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive mixture of local experts," *Neural Computation*, 3, 79–87.

Kass, R. E. and Raftery, A. E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā Ser. A*, 62, 49–66.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, 9, 1–59, with discussion, and a rejoinder by Hunter and Lange.

Lazarsfeld, P. F. and Henry, N. W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin.

Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *The Annals of Statistics*, 20, 1350–1360.

Lindsay, B. (1995), *Mixture Models: Theory, Geometry and Applications*, Hayward, CA: Institute of Mathematical Statistics.

Luce, R. D. and Suppes, P. (1963), "Preference, utility and subjective probability," in *Handbook of Mathematical Psychology*, eds. Luce, R., Bush, R., and Galanter, F., New York: Wiley, vol. 3.

Mallows, C. L. (1957), "Non-null ranking models. I," *Biometrika*, 44, 114–130.

Marden, J. I. (1995), *Analyzing and modeling rank data*, London: Chapman & Hall.

Marsh, M. (1999), "The Making of the Eighth President," in *How Ireland Voted 1997*, eds. Marsh, M. and Mitchell, P., Boulder, CO: Westview and PSAI Press, pp. 215–242.

McCullagh, P. and Nelder, J. (1983), *Generalized Linear Models*, London: Chapman and Hall.

McFadden, D. (1974), "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. Zarembka, P., New York: Academic Press, pp. 105–142.

McFadden, D. and Train, K. (2000), "Mixed MNL models of discrete responses," *Journal of Applied Econometrics*, 447–470.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, New York: Marcel Dekker Inc.

McLachlan, G. J. and Krishnan, T. (1997), *The EM algorithm and extensions*, New York: John Wiley & Sons Inc.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

McNicholas, P. D. and Murphy, T. B. (2008), "Parsimonious Gaussian Mixture Models," *Statistics and Computing*, 18, 285–296.

Meilă, M. and Bao, L. (2008), "Estimation and clustering with infinite rankings," in *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, eds. McAllester, D. A. and Myllymäki, P., AUAI Press, pp. 393–402.

Murphy, T. B. and Martin, D. (2003), "Mixtures of Distance-Based Models for Ranking Data," *Computational Statistics and Data Analysis*, 41, 645–655.

Nelder, J. A. and Wedderburn, R. W. (1972), "Generalized linear models," *Journal of the Royal Statistical Society Series A*, 135, 370–384.

Plackett, R. L. (1975), "The analysis of permutations," *Applied Statistics*, 24, 193–202.

Ralston, K., Brent, C. P., Starke, Y., Riggins, T., and Lin, C. J. (2002), *Consumer Food Safety Behavior: A Case Study in Hamburger Cooking and Ordering*, no. AER804 in Agricultural Economic Report, Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture.

Reboussin, B. A., Ip, E. H., and Wolfson, M. (2008), "Locally dependent latent class models with covariates: an application to under-age drinking in the USA," *Journal of the Royal Statistical Society, Series A*, 171, 877–897.

Schwartz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Sinnott, R. (1995), *Irish voters decide: Voting behaviour in elections and referendums since 1918*, Manchester: Manchester University Press.

— (2004), "The Rules of the Electoral Game," in *Politics in the Republic of Ireland*, eds. Coakley, J. and Gallagher, M., London: Routledge & PSAI Press, 4th ed., pp. 105–134.

Smyth, P. (2000), "Model selection for probabilisitc clustering using cross-validated likelihood," *Statistics and Computing*, 9, 63—72.

Stern, H. S. (1993), "Probability Models on Rankings and the Electoral Process," in *Probability Models and Statistical Analyses For Ranking Data*, eds. Fligner, M. A. and Verducci, J. S., New York: Springer-Verlag, pp. 173–195.

Thompson, T. J., Smith, P. J., and Boyle, J. P. (1998), "Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47, 393–404.

Thurstone, L. L. (1927), "A law of comparative judgement," *Psychological Review*, 34, 273–286.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Chichester: Wiley.

Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press.

Vigneau, E., Courcoux, P., and Semenou, M. (1999), "Analysis of ranked preference data using latent class models," *Food Quality and Preference*, 10, 201–207.

Yang, C.-C. and Yang, C.-C. (2007), "Separating latent classes by information criteria," *Journal of Classication*, 24, 183–203.