Title	Mixture of experts modelling with social science applications
Authors(s)	Gormley, Isobel Claire; Murphy, Thomas Brendan
Publication date	2011-04
Publication information	Mengersen, K., Robert, C, Titterington, M. (eds.). Mixture : estimation and applications
Publisher	Wiley
Item record/more information	http://hdl.handle.net/10197/2831
Publisher's statement	This is the author's version of Chapter 9 published in "Mixture: Estimation and Applications" (2011), edited by Christian Robert, Kerrie Mengersen, Mike Titterington.

1

Mixture of Experts Modeling with Social Science Applications

Isobel Claire Gormley & Thomas Brendan Murphy

University College Dublin

1.1 Introduction

Clustering methods are used to group observations into homogeneous subgroups. Clustering methods are usually either algorithmically based (e.g. k-means or hierarchical clustering) (see Hartigan 1975) or based on statistical models (e.g. Fraley and Raftery 2002; McLachlan and Basford 1988).

Clustering methods have been widely used in the social sciences. Examples of clustering applications include market research (Punj and Stewart 1983), archaeology (Hall 2004), education (Aitkin et al. 1981; Gormley and Murphy 2006) and sociology (Lee et al. 2005). In Section 1.2, we outline two applications of clustering in the social sciences: studying voting blocs in elections (Section 1.2.1) and exploring organizational structure in a corporation (Section 1.2.2).

In any cluster analysis application, it is common that clustering is implemented on outcome variables of interest without reference to concomitant covariate information on the objects being clustered. Once a clustering of objects has been produced, the user must probe the clusters to investigate their structure. Interpretations of the clusters can be produced with reference to values of the outcome variables within each cluster and/or with reference to the concomitant covariate information that wasn't used in the construction of the clusters.

The use of a model-based approach to clustering allows for any uncertainty to be accounted for in a probabilistic framework. Mixture models are the basis of many model-based clustering methods. In Section 1.3, we briefly describe the use of mixture models for clustering. The mixture of experts model (Jacobs et al. 1991) provides a framework for extending the mixture model to allow the model parameters to depend on concomitant covariate information; these models are reviewed in Section 1.4.

Examples of mixture of experts models and their application are motivated in Section 1.2 and implemented for the study of voting blocs in Section 1.5 and for studying organizational structure in Section 1.6.

We conclude, in Section 1.7, by discussing mixture of experts models and their interpretation in statistical applications.

1.2 Motivating Examples

1.2.1 Voting Blocs

In any election, members of the electorate exhibit different voting behaviors by choosing to vote for different candidates. Differences in voting behavior may be due to allegiance to a political party or faction, choosing familiar candidates, choosing geographically local candidates or one of many other reasons. Such different voting behaviors lead to a collection of votes from a heterogeneous population.

The discovery and characterization of voting blocs (i.e. groups of voters with similar preferences) is of considerable interest. For example, Tam (1995) studies Asian voting behavior within the American political arena via a multinomial logistic regression model and concludes that Asians should not be treated as a monolithic group. Holloway (1995) examines the differences between voting blocs when analyzing United Nations roll call data using a multidimensional scaling technique. Stern (1993), Murphy and Martin (2003) and Busse et al. (2007) use mixtures of distance-based models to characterize voting blocs in the American Psychological Association presidential election of 1980. Gormley and Murphy (2008a) use a mixture of Plackett-Luce (Plackett 1975) and Benter (Benter 1994) models to characterize voting blocs in the electorate for Irish governmental and Irish presidential elections. Spirling and Quinn (2010) use a Dirichlet process mixture model to study voting blocs in the U.K. House of Commons.

Many of the above studies investigate the existence of voting blocs by clustering voting data and then subsequently investigating the resulting clusters by examining the cluster parameters and by exploring concomitant voter covariates (when available) for members of each cluster. Such explorations can assist in determining what factors influence voting bloc membership and as a result voting behavior.

A more principled approach to investigating which factors influence voting behavior is to incorporate voter covariates in the modeling process which is used to construct the clusters. The mixture of experts framework provides a modeling structure to do this. In Section 1.5, we outline the use of mixture of experts models to characterize intended voting behavior in the 1997 Irish presidential election.

1.2.2 Social and Organizational Structure

The study of social mechanisms which underlie cooperation among peers within an organization is an important area of study within sociology and within organizations in general (Lazega 2001). Social network analysis (e.g. Wasserman and Faust 1994) is a highly active research area which provides an approach to examining structures within an organization or a network of ‘actors’. The study of such networks has recently attracted

attention from a broad spectrum of research communities including sociology, statistics, mathematics, physics and computer science.

Specifically, social network data record the interactions (relationships) between a group of actors or social entities. For example, a social network data set may detail the friendship links among a group of colleagues or it may detail the level of international trade between countries. Network data may be binary, indicating the presence/absence of a link between two actors, or it may be non-binary indicating the level of interaction between two actors. The aim of social network analysis is to explore the structure within the network, to aid understanding of underlying phenomena and the relations that may or may not exist within the network.

Many statistical approaches to modeling the interactions between actors in a network are available (Goldenberg et al. 2009; Kolaczyk 2009; Snijders et al. 2006; Wasserman and Faust 1994); many recent modeling advances tend to employ the idea of locating actors in a latent social space. In particular, Hoff et al. (2002) develop the idea of a latent social space and define the probability of a link between two actors as a function of their separation in the latent social space; this idea has been developed in various directions in Handcock et al. (2007), Krivitsky and Handcock (2008) and Krivitsky et al. (2009) in order to accommodate clusters (or communities) of highly connected actors in the network and other network effects. Airoldi et al. (2008) develop an alternative latent variable model for social network data where a soft clustering of network actors is achieved; this has been further extended by Xing et al. (2010) to model dynamic networks. More recently, Mariadassou et al. (2010) and Latouche et al. (2010) developed novel latent variable models for finding clusters of actors (or nodes) in network data.

In many social network modeling applications, concomitant covariate information on each actor is not used in the clustering of actors in the network. The clusters discovered in the network are explored and explained by examining the actor attributes that were not used in the clustering process. We endorse the use of mixture of experts models to provide a principled framework for clustering actors in a social network when concomitant covariate information is available for the actors.

In Section 1.6, a mixture of experts model for social network data is employed to explore the organizational structure within a northeastern USA corporate law firm.

1.3 Mixture Models

Let $y_1, y_2, \ldots, y_N$ be an iid sample of outcome variables from a population that is modeled by a probability density $p(\cdot)$. The mixture model assumes that the population consists of G components or sub-populations. The probability of component g occurring is τ_g and each component is modeled using a probability density $p(y_i|\theta_g)$, for $g = 1, 2, \ldots, G$. Hence, the overall model for a member of the population is of the form

$$p(y_i) = \sum_{g=1}^{G} \tau_g p(y_i|\theta_g).$$

In many mixture modeling contexts, an augmented form of the mixture model which includes the unknown sub-population membership vectors l_i for $i = 1, \ldots, N$, greatly assists

computations. The augmented model is

$$p(y_i, l_i) = \prod_{g=1}^{G} [\tau_g p(y_i|\theta_g)]^{l_{ig}}$$

where $l_{ig} = 1$ if observation i comes from sub-population g and $l_{ig} = 0$ otherwise.

Inference for the mixture model is usually implemented by maximum likelihood using the EM algorithm (Dempster et al. 1977) or in a Bayesian framework using Markov Chain Monte Carlo (Diebolt and Robert 1994). The clustering of observations is based on the posterior probability of component membership for each observation,

$$\mathbf{P}(l_{ig} = 1|y_i) = \mathbf{E}(l_{ig}|y_i) = \frac{\tau_g p(y_i|\theta_g)}{\sum_{g'=1}^{G} \tau_{g'} p(y_i|\theta_{g'})}.$$

The maximum *a posteriori* estimate of cluster membership assigns each observation to its most probable group, thus achieving a clustering of the observations.

Amongst the most commonly studied and applied mixture models are the Gaussian mixture model (e.g. Fraley and Raftery 2002) and the Latent Class Analysis model, which is a mixture of products of independent Bernoulli models (Lazarsfeld and Henry 1968). Extensive reviews of mixture models and their application are given in Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Fraley and Raftery (1998, 2002) and Melnykov and Maitra (2010).

Software for fitting mixture models in R (R Development Core Team 2009) include mclust (Fraley and Raftery 2006), mixtools (Benaglia et al. 2009) and flexmix (Leisch 2004) amongst others. Other software for mixture modeling includes MIXMOD (Biernacki et al. 2006) and EMMIX (McLachlan et al. 1999).

1.4 Mixture of Experts Models

The mixture of experts model (Jacobs et al. 1991) extends the mixture model by allowing the parameters of the model to be functions of an observation's concomitant covariates w_i:

$$p(y_i|w_i) = \sum_{g=1}^{G} \tau_g(w_i) p(y_i|\theta_g(w_i)). \tag{1.1}$$

Bishop (2006, Chapter 14.5) refers to the mixture of experts model as a conditional mixture model, as for a given set of concomitant covariates w_i the distribution of y_i is a mixture model.

The terminology used in the mixture of experts model literature calls the $p(y_i|\theta_g(w_i))$ densities 'experts' and the $\tau_g(w_i)$ probabilities 'gating networks'. In its original formulation in Jacobs et al. (1991), the model for $\tau_1(w_i), \tau_2(w_i), \ldots, \tau_G(w_i)$ is a multinomial logistic regression model and $p(y_i|\theta_g(w_i))$ is a general linear model.

Figure 1.1 illustrates a graphical model representation of the mixture of experts model. This representation aids the interpretation of the full mixture of experts model (in which all model parameters are functions of covariates (Figure1.1(d))) and the special cases where some of the model parameters do not depend on the covariates (Figures1.1(a)-1.1(c)). The four models detailed in Figure 1.1 have the following interpretations:

(a) in the mixture model, the outcome variable distribution depends on the latent cluster membership variable l_i and the model is independent of the covariates w_i.

(b) in the expert network mixture of experts model, the outcome variable distribution depends on both the covariates w_i and the latent cluster membership variable l_i; the distribution of the latent variable is independent of the covariates.

(c) in the gating network mixture of experts model, the outcome variable distribution depends on the latent cluster membership variable l_i and the distribution of the latent variable depends on w_i.

(d) in the full mixture of experts model, the outcome variable distribution depends on both the covariates w_i and on the latent cluster membership variable l_i. Additionally the distribution of the latent variable l_i depends on the covariates w_i.

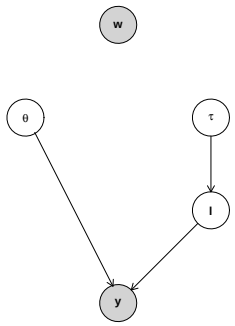
Mixture of experts models have been employed in a wide range of modeling settings — Peng et al. (1996) use a mixture of experts model and a hierarchical mixture of experts model in speech recognition applications. Thompson et al. (1998) use a mixture of experts model for studying the diagnosis of diabetic patients. Rosen and Tanner (1999) develop a mixture of experts proportional hazards model and analyze a multiple myeloma data set. Hurn et al. (2003) use MCMC to fit a mixture of regressions model which is a special case of the mixture of experts model, but where the mixing proportions don't depend on the covariates w_i. Carvalho and Tanner (2007) use a mixture of experts model for non-linear time-series modeling. Geweke and Keane (2007) use a model similar to the mixture of experts model, but where the gating network has a probit structure, in a number of econometric applications.

Further details on mixture of experts models are given in McLachlan and Peel (2000, Chapter 5.13), Tanner and Jacobs (2001) and Bishop (2006, Chapter 14.5), where extensions including the hierarchical mixture of experts model (Jordan and Jacobs 1994) are discussed. Software for fitting mixture of experts models in the R programming environment (R Development Core Team 2009) include hme (Evers 2007), mixtools (Benaglia et al. 2009) and integrativeME (Cao 2010).

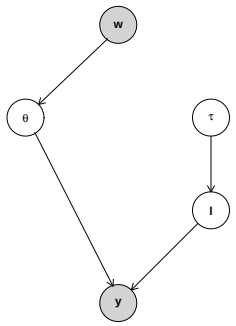
1.5 A Mixture of Experts Model for Ranked Preference Data

The current President of Ireland, Mary McAleese, was first elected in 1997 under the Single Transferable Vote electoral system. Under this system voters rank, in order of their preference, some or all of the electoral candidates. The vote counting system which results in the elimination of candidates and the subsequent election of the President is an intricate process involving the transfer of votes between candidates as specified by the voters' ballots. Details of the electoral system, the counting process and the 1997 Irish presidential election are given in Coakley and Gallagher (2004), Sinnott (1995), Sinnott (1999) and Marsh (1999).

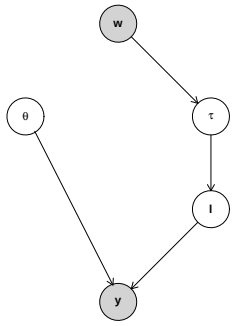
The 1997 presidential election race involved five candidates: Mary Banotti, Mary McAleese, Derek Nally, Adi Roche and Rosemary Scallon. Derek Nally and Rosemary Scallon were independent candidates while Mary Banotti and Adi Roche were endorsed by the then current opposition parties Fine Gael and Labour respectively. Mary McAleese was endorsed by the Fianna Fáil party who were in power at that time. In terms of candidate type, McAleese and Scallon were deemed to be conservative candidates with the other candidates



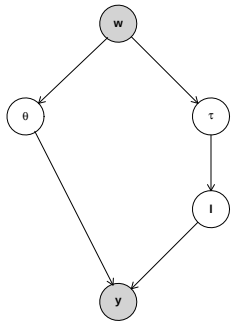
(a) Mixture model



(b) Expert network mixture of experts model



(c) Gating network mixture of experts model



(d) Full mixture of experts model

Figure 1.1 The graphical model representation of the mixture of experts model. The differences between the four special cases are due to the presence or absence of edges between the covariates w and the parameters τ and θ.

Table 1.1 Covariates recorded for each respondent in the Irish Marketing Surveys 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-

regarded as liberal. Gormley and Murphy (2008a,b, 2010a) provide further details on the 1997 presidential election and on the candidates.

One month prior to election day a survey was conducted by Irish Marketing Surveys on 1083 respondents. Respondents were asked to list some or all of the candidates in order of preference, as if they were voting on the day of the poll. In addition, pollsters gathered data on attributes of the respondents as detailed in Table 1.1.

Interest lies in determining if groups of voters with similar preferences (i.e. voting blocs) exist within the electorate. If such voting blocs do exist, the influence the recorded socio-economic variables may have on the clustering structure and/or on the preferences which characterize a voting bloc is also of interest. Jointly modeling the rank preference votes and the covariates through a mixture of experts model for rank preference data when clustering the electorate provides this insight.

Given the rank nature of the outcome variables or votes y_i ($i = 1, \ldots, N = 1083$) the probability density $p(\cdot)$ in the mixture of experts model (1.1) must have an appropriate form. The Plackett-Luce model (Plackett 1975) (or exploded logit model) for rank data provides a suitable model; Benter's model (Benter 1994) provides another alternative. Let $y_i = [c(i,1), \ldots, c(i,m_i)]$ denote the ranked ballot of voter i where $c(i,j)$ denotes the candidate ranked in jth position by voter i and m_i is the number of candidates ranked by voter i. Under the Plackett-Luce model, given that voter i is a member of voting bloc g and given the 'support parameter' $p_g = (p_{g1}, \ldots, p_{gM})$, the probability of voter i's ballot is

$$\mathbb{P}(y_i|p_g) = \frac{p_{gc(i,1)}}{\sum_{s=1}^{M} p_{gc(i,s)}} \cdot \frac{p_{gc(i,2)}}{\sum_{s=2}^{M} p_{gc(i,s)}} \cdots \frac{p_{gc(i,m_i)}}{\sum_{s=m_i}^{M} p_{gc(i,s)}}$$

where $M = 5$ denotes the number of candidates in the electoral race. The support parameter p_{gj} (typically restricted such that $\sum_{j=1}^{M} p_{gj} = 1$) can be interpreted as the probability of ranking candidate j first, out of the currently available choice set. Hence, the Plackett-Luce model models the ranking of candidates by a voter as a set of independent choices by the voter, conditional on the cardinality of the choice set being reduced by one after each choice is made.

In the full mixture of experts model, the parameters of the group densities are modeled as a function of covariates. Here the support parameters are modeled as a logistic function of the covariates:

$$\log\left[\frac{p_{gj}(w_i)}{p_{g1}(w_i)}\right] = \gamma_{gj0} + \gamma_{gj1} w_{i1} + \cdots + \gamma_{gjL} w_{iL}$$

Table 1.2 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. In the 'government satisfaction' variable the 'no opinion' level was used as the baseline category.

	BIC	G	Covariates
The gating network MoE model	8491	4	τ_g: Age, Government satisfaction
The full MoE model	8512	3	τ_g: Age, Government satisfaction p_g: Age
The mixture model	8513	3	–
The expert network MoE model	8528	1	p_g: Government satisfaction

where $w_i = (w_{i1}, \ldots, w_{iL})$ is the set of L covariates associated with voter i. Note that for identifiability reasons candidate 1 is used as the baseline choice and $\gamma_{g1} = (0, \ldots, 0)$ for all $g = 1, \ldots, G$. The intuition behind this version of the model is that a voter's covariates may potentially influence their support for each candidate beyond what explained by their membership of a voting bloc.

In the full mixture of experts model, the gating networks (or mixing proportions) are also modeled as a function of covariates. In a similar vein to the support parameters, the mixing proportions are modeled via a multinomial logistic regression model

$$\log\left[\frac{\tau_g(w_i)}{\tau_1(w_i)}\right] = \beta_{g0} + \beta_{g1}w_{i1} + \cdots + \beta_{gL}w_{iL}$$

where voting bloc 1 is used as the baseline voting bloc. Here, the motivation for this model term is that a voter's covariates may influence their voting bloc membership.

Modeling the group parameters and/or the mixing proportions as functions of covariates, or as constant with respect to covariates, results in the four types of mixture of experts models as illustrated in Figure 1.1. Each model can be fitted in a maximum likelihood framework using an EM algorithm (Dempster et al. 1977). Model fitting details for each model are outlined in Gormley and Murphy (2008a,b, 2010a).

Each of the four mixture of experts models for rank preference data illustrated in Figure 1.1 were fitted to the data from the electorate in the Irish presidential election poll. A range of groups $G = 1, \ldots, 5$ was considered and a forwards selection method was employed to select influential covariates. The Bayesian Information Criterion (BIC) (Kass and Raftery 1995; Schwartz 1978) was used to select the optimal model; this criterion is a penalized likelihood criterion which rewards model fit while penalizing non-parsimonious models. Small BIC values indicate a preferable model. Table 1.2 details the optimal models for each type of mixture of experts model fitted.

Based on the BIC values, the optimal model is a gating network MoE model with four groups where age and government satisfaction are important covariates for determining group or voting bloc membership. Under this gating network MoE model, the covariates are not informative within voting blocs, but only in determining voting bloc membership. The maximum likelihood estimates of the model parameters are reported in Figure 1.2 and in Table 1.3.

The support parameter estimates illustrated in Figure 1.2 have an interpretation in the context of the 1997 Irish presidential election. Voting bloc 1 could be characterized as the conservative voting bloc due to its large support parameters for McAleese and Scallon. Voting

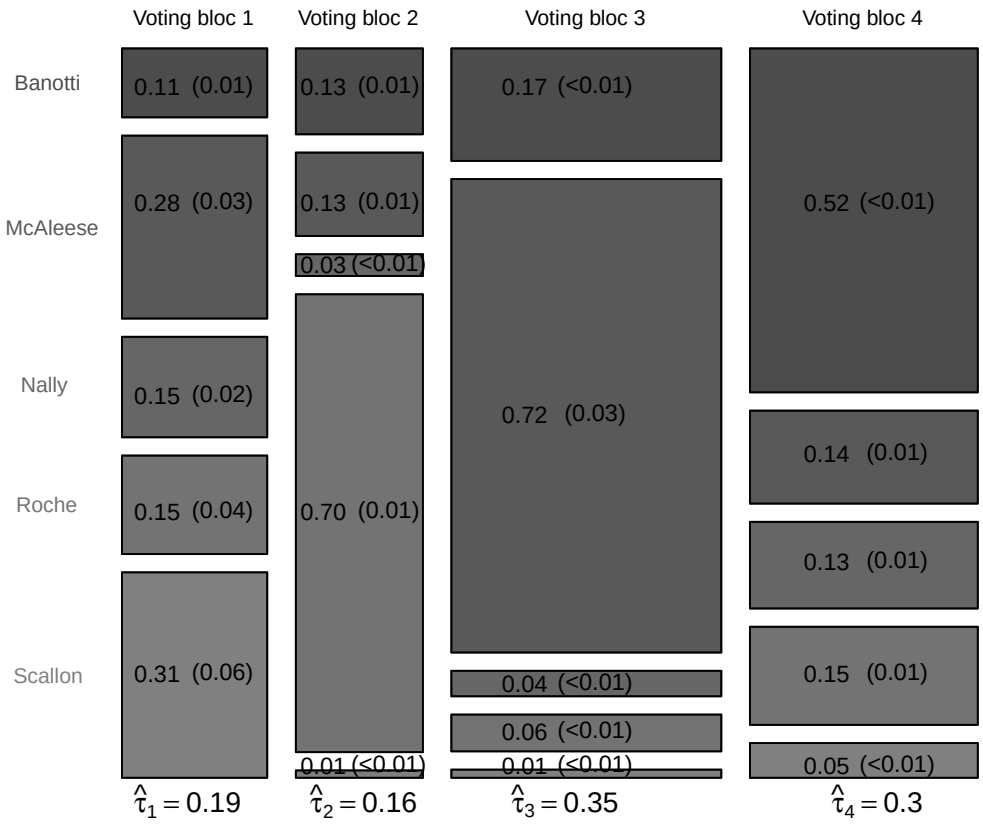


Figure 1.2 A mosaic plot representation of the parameters of the group densities of the gating network mixture of experts model for rank preference data. The width of each block is proportional to the marginal probability of belonging to that group and the blocks are divided in proportion to the Plackett-Luce support parameters.

Table 1.3 Odds ratios for the mixing proportion parameters in the gating network MoE model for rank preference data. The covariates 'age' and 'government satisfaction level' were selected as influential.

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

bloc 2 has large support for the liberal candidate Adi Roche. 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. These candidates were endorsed by the two largest political parties in the country at that time. 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).

Table 1.3 details the odds ratios computed for the mixing proportion (or gating network) parameters $\beta = (\beta_1, \ldots, \beta_G)$. In the model, voting bloc 1 (the conservative voting bloc) is the baseline voting bloc. Two covariates were selected as influential: age and government satisfaction levels. In the government satisfaction covariate, the baseline was chosen to be no opinion.

Interpreting the odds ratios provides insight to the type of voter which characterizes each voting bloc. 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 ($\beta_{21} = 0.01$). Also, voters with some interest in government are more likely to belong to voting bloc 3 ($\beta_{32} = 3.12$ and $\beta_{33} = 3.81$), 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 ($\beta_{43} = 3.50$). 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).

1.5.1 Examining The Clustering Structure

It is important that the clusters found by the mixture of experts model correspond to distinct voting blocs. Baudry et al. (2010) propose a method to check if mixture components are really modeling distinct clusters or whether multiple mixture components are being used to model each cluster because the component density in the mixture model is overly restrictive. Hennig (2010) proposes an alternative approach to this problem specifically for normal mixture models.

The method developed by Baudry et al. (2010) uses the estimated *a posteriori* cluster membership probabilities

$$\hat{l}_{ig} = \frac{\hat{\tau}_g(w_i) f(y_i|\hat{\theta}_g(w_i))}{\sum_{g'=1}^{G} \hat{\tau}_{g'}(w_i) f(y_i|\hat{\theta}_{g'}(w_i))}.$$

In particular, suppose the mixture components $\{1, 2, \ldots, G\}$ are partitioned into sets $\{\rho_1, \rho_2, \ldots, \rho_K\}$ where ρ_k are the mixture components used to model distinct cluster k. Further, let $t_{ik} = \sum_{g \in \rho_k} \hat{l}_{ig}$ be the estimated *a posteriori* probability of membership in cluster k. Then, the entropy of a particular clustering is given as

$$\mathcal{E}_K = -\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik} \log t_{ik}.$$

A greedy algorithm is used to combine mixture components to reduce K from G to $G-1$, from $G-1$ to $G-2$ and so on, until $K=1$. A plot of $\mathcal{E}_K$ versus K gives an indication of the number of clusters in the population, where a large drop in $\mathcal{E}_K$ when K is decreased indicates that multiple components are modeling a cluster in the population. The results of applying the component merging algorithm to the mixture of experts model are shown in Figure 1.3.

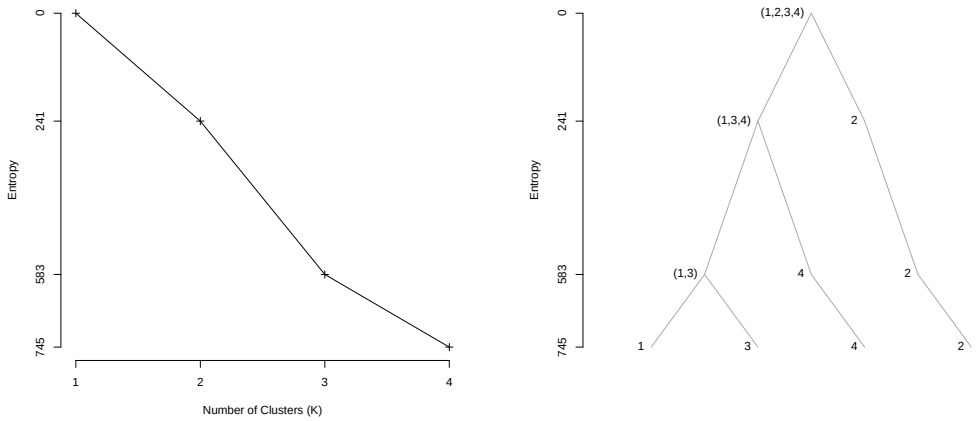


Figure 1.3 (a) The value of $\mathcal{E}_K$ plotted as a function of K. (b) A dendrogram representation of the combination of mixture of experts components when clustered into K clusters. Note that the vertical scale on both plots is inverted.

These results suggest that the entropy doesn't decrease substantially when combining the two closest components (ie, 1 and 3) to form a single cluster. So, the two components are distinct from each other. Hence, it appears from this analysis that the four components are modeling distinct clusters in the data. Dean and Nugent (2010) develop a mixture component tree to visualize the relationship between Gaussian mixture components where the similarity of mixture components is measured using the minimum density on a path joining the mixture component means. Following their work, we use the entropy based dendrogram in Figure 1.3(b) to give a visualization of the connections between the four voting blocs found in this analysis.

1.6 A Mixture of Experts Latent Position Cluster Model

The latent position cluster model (LPCM) (Handcock et al. 2007) develops the idea of the latent social space model (Hoff et al. 2002) by extending the model to accommodate clusters of actors in the latent space. Under the latent position cluster model, the latent location of

each actor is assumed to be drawn from a finite normal mixture model, each component of which represents a cluster of actors. In contrast, the model outlined in Hoff et al. (2002) assumed that the latent positions were normally distributed. Thus, the latent position cluster model offers a more flexible version of the latent space model for modeling heterogeneous social networks.

The latent position cluster model provides a framework in which actor covariates may be explicitly included in the model – the probability of a link between two actors may be modeled as a function of both their separation in the latent space and of their relative covariates. However, the covariates may contribute more to the structure of the network than solely through the link probabilities – the covariates may influence both the cluster membership of an actor and their link probabilities. A latent position cluster model in which the cluster membership of an actor is modeled as a function of their covariates lies within the mixture of experts framework.

Specifically, social network data take the form of a set of relations $\{y_{i,j}\}$ between a group of $i, j = 1, \ldots, N$ actors, represented by an $N \times N$ sociomatrix $\mathbf{Y}$. Here it is assumed that the relation $y_{i,j}$ between actors i and j is a binary relation, indicating the presence or absence of a link between the two actors; the mixture of experts latent position cluster model is easily extended to other forms of relation (such as count data). Covariate data $w_i = (w_{i1}, \ldots, w_{iL})$ associated with actor i is assumed to be available, where L denotes the number of observed covariates.

Each actor i is assumed to have a location $z_i = (z_{i1}, \ldots, z_{id})$ in the d dimensional latent social space. The probability of a link between any two actors is assumed to be independent of all other links in the network, given the latent locations of the actors. Let $x_{i,j} = (x_{ij1}, \ldots, x_{ijL})$ denote an L vector of dyadic specific covariates where $x_{ijk} = d(w_{ik}, w_{jk})$ is a measure of the similarity in the value of the kth covariate for actors i and j. Given the link probabilities parameter vector β the likelihood function is then

$$\mathbb{P}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta) = \prod_{i=1}^{N} \prod_{j \neq i} \mathbb{P}(y_{i,j}|z_i, z_j, x_{i,j}, \beta)$$

where $\mathbf{Z}$ is the $N \times d$ matrix of latent locations and $\mathbf{X}$ is the matrix of dyadic specific covariates. The probability of a link between actors i and j is then modeled using a logistic regression model where both dyadic specific covariates and Euclidean distance in the latent space are dependent variables:

$$\log \left\{ \frac{\mathbb{P}(y_{i,j} = 1)}{\mathbb{P}(y_{i,j} = 0)} \right\} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_L x_{ijL} - ||z_i - z_j||.$$

To account for clustering of actor locations in the latent space, it is assumed that the latent locations z_i are drawn from a finite mixture model. Moreover, in the mixture of experts latent position cluster model, the latent locations are assumed drawn from a finite mixture model in which actor covariates may influence the mixing proportions:

$$z_i \sim \sum_{g=1}^{G} \tau_g(w_i) \text{MVN}(\mu_g, \sigma_g^2 \mathbf{I})$$

where

$$\tau_g(w_i) = \frac{\exp(\tau_{g0} + \tau_{g1}w_{i1} + \cdots + \tau_{gL}w_{iL})}{\sum_{g'=1}^{G} \exp(\tau_{g'0} + \tau_{g'1}w_{i1} + \cdots + \tau_{g'L}w_{iL})}$$

and $\tau_1 = (0, \ldots, 0)$. This model has an intuitive motivation: the covariates of an actor may influence their cluster membership, their cluster membership influences their latent location, and in turn their latent location determines their link probabilities.

The mixture of experts latent position cluster model can be fitted within the Bayesian paradigm; a Metropolis-within-Gibbs sampler can be employed to draw samples from the posterior distribution of interest. As is standard in Bayesian estimation of mixture models (Diebolt and Robert 1994; Hurn et al. 2003) the problem is greatly simplified by augmenting the observed data with an indicator variable K_i for each actor i where $K_i = g$ if actor i belongs to cluster g. The indicator variable K_i therefore has a multinomial distribution with a single trial and probabilities equal to $\tau_g(w_i)$ for $g = 1, \ldots, G$. Model issues such as likelihood invariance to distance preserving transformations of the latent space and label switching must be be considered during the model fitting process — an approach to dealing with such model identifiability and full model fitting details are available in Gormley and Murphy (2010b).

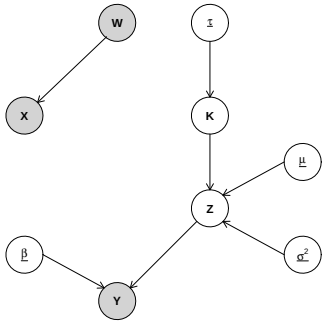
Figure 1.4 illustrates a graphical model representation of the mixture of experts latent position cluster model. Similarly to Figure 1.1, four different models are available by allowing or disallowing covariates to influence the mixing proportions and/or the link probabilities.

An illustrative example of the mixture of experts latent position cluster model methodology is provided through the analysis of a network data set detailing interactions between a set of 71 lawyers in a corporate law firm in the USA (Lazega 2001). The data include measurements of the coworker network, an advice network and a friendship network. Covariates associated with each lawyer on the firm are also included and are detailed in Table 1.4. Interest lies in identifying social processes within the firm such as knowledge sharing and organizational structures.

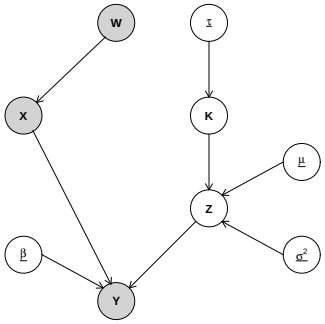
The four mixture of experts latent position cluster models illustrated in Figure 1.4 were fitted to the advice network; data in this network detail links between lawyers who sought basic professional advice from each other over the previous twelve months. Gormley and Murphy (2010b) explore the coworkers network data set and the friendship network data set using similar methodology. Figure 1.5 illustrates the resulting latent space locations of the lawyers under each fitted model with $(G, d) = (2, 2)$. These parameter values were selected using BIC after fitting a range of latent position cluster models (with no covariates) to the network data only (Handcock et al. 2007). Table 1.5 details the resulting regression parameter estimates and their associated uncertainty for the four fitted models.

The results of the analysis show some interesting patterns. The model with the highest AICM (Raftery et al. 2007) value is the model that has covariates in the link probabilities and in the cluster membership probabilities.

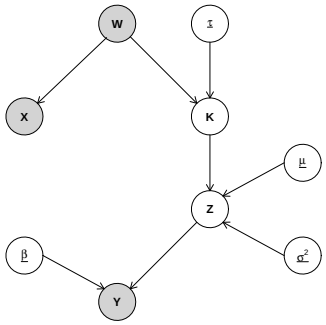
The coefficients of the covariates in the link probabilities are very similar in the models (b) and (d) in Table 1.5. These coefficients indicate that a number of factors have a positive or negative effect on whether a lawyer asks another for advice. In summary, lawyers who are similar in seniority, gender, office location and practice type are more likely to ask each other for advice. The effects of years and age seem to have a negative effect, but these variables



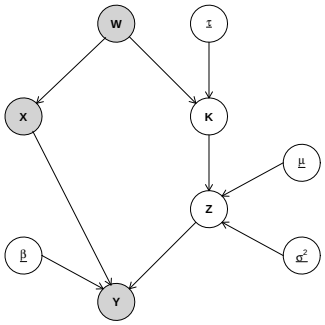
(a) LPCM (No Covariates)



(b) LPCM (Edge Covariates)

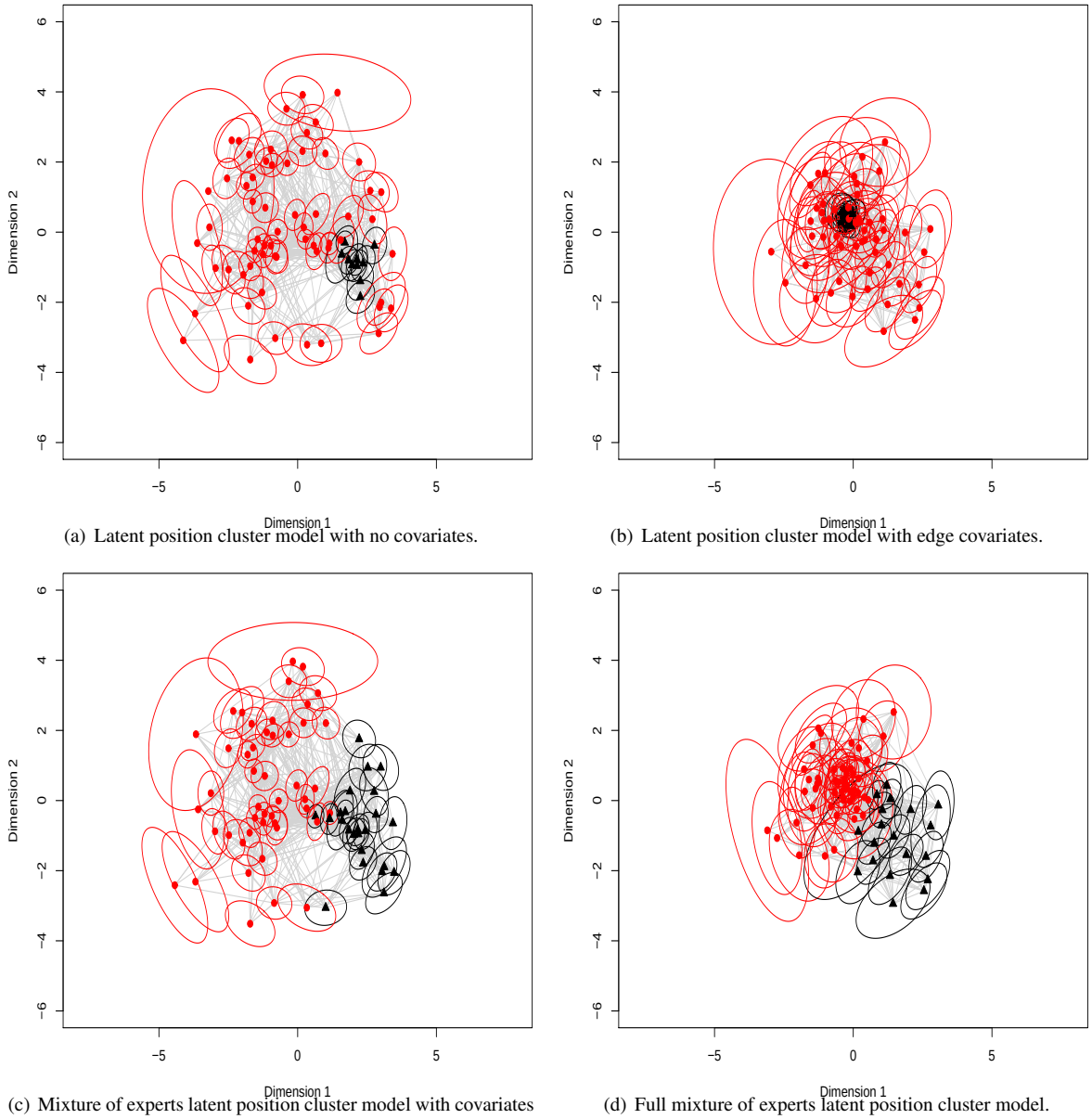


(c) LPCM (Cluster Membership)



(d) LPCM (Full Model)

Figure 1.4 The graphical model representation of the mixture of experts latent position cluster model. The differences between the four special cases are due to the presence or absence of edges between the covariates X and W and the parameters Y and K.



(a) Latent position cluster model with no covariates.

(b) Latent position cluster model with edge covariates.

(c) Mixture of experts latent position cluster model with covariates in the cluster memberships.

(d) Full mixture of experts latent position cluster model.

Figure 1.5 Estimates of clusters and latent positions of the lawyers from the advice network data. The ellipses are 50% posterior sets illustrating the uncertainty in the latent locations. Lawyers who are members of the same cluster are illustrated using the same colour and symbol. Observed links between lawyers are also illustrated.

Table 1.4 Covariates associated with the 71 lawyers in the US corporate law firm. The last category in each categorical covariate is treated as the baseline category in all subsequent analyses.

Covariate	Levels
Seniority	1 = partner
	2 = associate
Gender	1 = male
	2 = female
Office	1= Boston
	2 = Hartford
	3 = Providence
Practice	1 = litigation
	2 = corporate
Law school	1 = Harvard or Yale
	2 = University of Connecticut
	3 = other
Years with the firm	–
Age	–

are correlated with seniority and with each other, so their marginal effects are more difficult to interpret.

Importantly, the latent positions are very similar in models (a) and (c) which don't have covariates in the link probabilties and models (b) and (d) which do have covariates in the link probabilities. This can be explained because of the different role that the latent space plays in the models with covariates in the link probabilities and in those that do not have such covariates. When the covariates are in the link probabilities, the latent space is modeling the part of the network structure that could not be explained by the link covariates, whereas in the other case the latent space is modeling much more of the network structure.

Interestingly, in the model with the highest AICM value, there are covariates in the cluster membership probabilities as well as in the link probabilities. This means that the structure in the latent space, which is modeling what couldn't be explained directly in the link probabilities, has structure that can be further explained using the covariates. The office location, practice and age of the lawyers retain explanatory power in explaining the clustering found in the latent social space.

The difference in the cluster membership coefficients in models (c) and (d) is due to the different interpretation of the latent space in these models. However, it is interesting to note that the signs of the coefficients are identical, this is because the cluster memberships shown for these models Figure 1.5(c) and Figure 1.5(d) are similar; this phenomenon does not hold generally (see Gormley and Murphy 2010b, Section 5.3).

The results of this analysis offer a cautionary message in automatically selecting the type of mixture of experts latent position cluster model for analyzing social network data. The role of the latent space in the model is very different depending on how the covariates enter the model. So, if the latent space is to be interpreted as a social space that explains network structure, then the covariates should not directly enter the link probabilities. However, if the

Table 1.5 Posterior mean parameter estimates for the four mixture of experts models fitted to the lawyers advice data as detailed in Figure 1.5. Standard deviations are given in parentheses. Note that cluster 1 was used as the baseline cluster in the case of the cluster membership parameters. Baseline categories for the covariates are detailed in Table 1.4.

	(a)	(b)	(c)	(d)
Link Probabilities				
Intercept	1.26 (0.10)	-2.87 (0.17)	1.23 (0.10)	-2.65 (0.17)
Seniority		0.89 (0.11)		0.81 (0.11)
Gender		0.60 (0.09)		0.62 (0.09)
Office		2.02 (0.10)		1.97 (0.10)
Practice		1.63 (0.10)		1.57 (0.10)
Years		-0.04 (0.005)		-0.04 (0.005)
Age		-0.02 (0.004)		-0.02 (0.004)
Cluster Memberships				
Intercept	-1.05 (1.75)	0.94 (0.79)	-0.62 (1.23)	1.27 (1.29)
Office (=1)			1.94 (1.02)	2.40 (1.14)
Office (=2)			-2.08 (1.09)	-0.97 (1.19)
Practice			3.18 (0.85)	2.14 (1.08)
Age			-0.09 (0.04)	-0.14 (0.06)
Latent Space Model				
Cluster 1 mean	- 0.50 (0.52)	0.09 (0.19)	-1.09 (0.31)	-0.54 (0.21)
	0.21 (0.58)	-0.09 (0.26)	0.40 (0.28)	0.40 (0.20)
Cluster 1 variance	3.35 (1.29)	2.12 (0.77)	3.19 (0.58)	1.25 (0.34)
Cluster 2 mean	1.66 (0.92)	-0.24 (0.20)	2.10 (0.30)	1.32 (0.51)
	-0.67 (0.58)	0.35 (0.23)	-0.77 (0.30)	-0.98 (0.47)
Cluster 2 variance	1.29 (1.58)	0.27 (0.68)	1.16 (0.40)	1.63 (0.69)
AICM	-3644.24	-3346.87	-3682.71	-3325.95

latent space is being used to find interesting or anomalous structure in the network that can't be explained by the covariates, then one should consider allowing the covariates enter the link probabilities and cluster membership probabilities.

1.7 Discussion

This chapter has illustrated the utility of mixture of experts models in two social science clustering applications where concomitant covariate information is available. The mixture of experts framework provides a systematic method for describing and exploring the clustering found in the outcome variables.

This inclusion of covariates through the mixture of experts model can give different clustering results than when a two stage process of clustering followed by cluster interrogation is taken. This is because both the outcome variables and the concomitant covariates provide information that is relevant in defining the clustering. The result of the use of both sources of information is often a clearer clustering structure.

When using the mixture of experts model, it is important to consider how the covariates enter the model. The interpretation of the latent structure (clustering or other latent variables) in the mixture of experts model depends heavily on how the covariates enter the model. So, this choice needs to be directed by the interpretation of the latent structure in the context of the application.

Acknowledgements

We would like to thank the participants of the ICMS Workshop on Mixture Estimation and Applications for their insightful feedback on this work. This work has been supported by Science Foundation Ireland Research Frontiers grants (06/RFP/M040 and 09/RFP/MTH2367) and Clique, a Science Foundation Ireland Strategic Research Cluster grant (08/SRC/I1407).

References

Airoldi EM, Blei DM, Fienberg SE and Xing E 2008 Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Aitkin M, Anderson D and Hinde J 1981 Statistical Modelling of Data on Teaching Styles (with Discussion). *Journal of the Royal Statistical Society, Series A: General* **144**, 419–461.

Baudry JP, Raftery AE, Celeux G, Lo K and Gottardo RG 2010 Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* **19**(2), 332–353.

Benaglia T, Chauveau D, Hunter DR and Young D 2009 mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32**(6), 1–29.

Benter W 1994 Computer-based horse race handicapping and wagering systems: A report In *Efficiency of Racetrack Betting Markets* (ed. Ziemba WT, Lo VS and Haush DB) Academic Press San Diego and London pp. 183–198.

Biernacki C, Celeux G, Govaert G and Langrognet F 2006 Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis* **51**(2), 587–600.

Bishop CM 2006 *Pattern Recognition and Machine Learning*. Springer, New York.

Busse LM, Orbanz P and Buhmann JM 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.

Cao KAL 2010 *integrativeME: integrative mixture of experts*. R package version 1.2.

Carvalho AX and Tanner MA 2007 Modelling nonlinear count time series with local mixtures of Poisson autoregressions. *Computational Statistics and Data Analysis* **51**(11), 5266–5294.

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

Dean N and Nugent R 2010 Mixture Model Component Trees: Visualizing the Hierarchical Structure of Complex Groups. *Technical Report*.

Dempster AP, Laird NM and Rubin DB 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38. With discussion.

Diebolt J and Robert CP 1994 Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B* **56**(2), 363–375.

Everitt BS and Hand DJ 1981 *Finite mixture distributions*. Chapman & Hall, London. Monographs on Applied Probability and Statistics.

Evers L 2007 *hme: Methods for Fitting Hierarchical Mixtures of Experts (HMEs)*. R package version 0.1-0. This packages reuses some GPL'ed code from the package mvtnorm (by Thorsten Hothorn using code from Alan Genz and Frank Bretz).

Fraley C and Raftery AE 1998 How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis. *Computer Journal* **41**, 578–588.

Fraley C and Raftery AE 2002 Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**(458), 611–631.

Fraley C and Raftery AE 2006 MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington.

Geweke J and Keane M 2007 Smoothly mixing regressions. *Journal of Econometrics* **136**(1), 252–290.

Goldenberg A, Zheng AX, Fienberg SE and M. AE 2009 A survey of statistical network models. *Foundations and Trends in Machine Learning* **2**, 129–233.

Gormley IC and Murphy TB 2006 Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society, Series A* **169**(2), 361—379.

Gormley IC and Murphy TB 2008a Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association* **103**(483), 1014–1027.

Gormley IC and Murphy TB 2008b A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics* **2**(4), 1452–1477.

Gormley IC and Murphy TB 2010a Clustering ranked preference data using sociodemographic covariates In *Choice Modelling: The State-of-the-Art and the State-of-Practice* (ed. Hess S and Daly A) Emerald United Kingdom pp. 543–569.

Gormley IC and Murphy TB 2010b A mixture of experts latent position cluster model for social network data. *Statistical Methodology* **7**(3), 385–405.

Hall ME 2004 Pottery production during the late Jomon period: Insights from the chemical analyses of Kasori B pottery. *Journal of Archaeological Science* **31**, 1439–1450.

Handcock M, Raftery A and Tantrum JM 2007 Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(2), 301 – 354.

Hartigan JA 1975 *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.

Hennig C 2010 Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* **4**(1), 3–34.

Hoff PD, Raftery AE and Handcock MS 2002 Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association* **97**, 1090–1098.

Holloway S 1995 Forty years of United Nations General Assembly voting. *Canadian Journal of Political Science* **17**(2), 223–249.

Hurn M, Justel A and Robert CP 2003 Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* **12**(1), 55–79.

Jacobs RA, Jordan MI, Nowlan SJ and Hinton GE 1991 Adaptive mixture of local experts. *Neural Computation* **3**(1), 79–87.

Jordan MI and Jacobs RA 1994 Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181–214.

Kass RE and Raftery AE 1995 Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kolaczyk ED 2009 *Statistical Analysis of Network Data: Methods and Models*. Springer, New York.

Krivitsky PN and Handcock MS 2008 Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software* **24**, 1–23.

Krivitsky PN, Handcock MS, Raftery AE and Hoff PD 2009 Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31**(3), 204–213.

Latouche P, Birmelé E and Ambroise C 2010 Overlapping stochastic block models. *Annals of Applied Statistics* To appear.

Lazarsfeld PF and Henry NW 1968 *Latent Structure Analysis*. Houghton Mifflin, Boston.

Lazega E 2001 *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press.

Lee CK, Lee YK, Bernhard BJ and Yoon YS 2005 Segmenting casino gamblers by motivation: A cluster analysis of Korean gamblers. *Tourism Management* **27**(5), 856–866.
Leisch F 2004 FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **11**(8), 1–18.
Mariadassou M, Robin S and Vacher C 2010 Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics* To appear.
Marsh M 1999 The Making of the Eighth President In *How Ireland Voted 1997* (ed. Marsh M and Mitchell P) Westview and PSAI Press Boulder, CO pp. 215–242.
McLachlan G, Peel D, Basford K and Adams P 1999 The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*.
McLachlan GJ and Basford KE 1988 *Mixture models: Inference and applications to clustering*. Marcel Dekker Inc., New York.
McLachlan GJ and Peel D 2000 *Finite Mixture Models*. John Wiley & Sons, New York.
Melnykov V and Maitra R 2010 Finite mixture models and model-based clustering. *Statistics Surveys* **4**, 80–116.
Murphy TB and Martin D 2003 Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis* **41**(3–4), 645–655.
Peng F, Jacobs RA and Tanner MA 1996 Bayesian inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts models with an application to speech recognition. *Journal of the American Statistical Association* **91**(435), 953—960.
Plackett RL 1975 The analysis of permutations. *Applied Statistics* **24**(2), 193–202.
Punj G and Stewart DW 1983 Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research* **20**, 134–148.
R Development Core Team 2009 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.
Raftery AE, Newton MA, Satagopan JM and Krivitsky PN 2007 Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with Discussion) *Bayesian Statistics, 8* Oxford Univ. Press UK pp. 371–415.
Rosen O and Tanner M 1999 Mixtures of proportional hazards regression models. *Statistics in Medicine* **18**, 1119–1131.
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 University Press, Manchester.
Sinnott R 1999 The electoral system In *Politics in the Republic of Ireland* (ed. Coakley J and Gallagher M) 3rd edn Routledge & PSAI Press London pp. 99–126.
Snijders T, Pattison P, Robins G and Handcock M 2006 New specifications for exponential random graph models. *Sociological Methodology* pp. 99–153.
Spirling A and Quinn K 2010 Identifying intraparty voting blocs in the U.K. House of Commons. *Journal of the American Statistical Association* **105**(490), 447–457.
Stern HS 1993 Probability models on rankings and the electoral process In *Probability Models and Statistical Analyses For Ranking Data* (ed. Fligner MA and Verducci JS) Springer-Verlag New York pp. 173–195.
Tam WK 1995 Asians — A monolithic voting bloc? *Political Behaviour* **17**(2), 223–249.
Tanner MA and Jacobs RA 2001 Neural networks and related statistical latent variable models In *International Encyclopedia of the Social and Behavioral Sciences* (ed. Smelser NJ and Baltes PB) Elsevier pp. 10526–10534.
Thompson TJ, Smith PJ and Boyle JP 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.
Titterington DM, Smith AFM and Makov UE 1985 *Statistical analysis of finite mixture distributions*. Wiley, Chichester.
Wasserman S and Faust K 1994 *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.
Xing E, Fu W and Song L 2010 A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics* To appear.