Title	Mixtures of Experts Models
Authors(s)	Gormley, Isobel Claire; Frühwirth-Schnatter, Sylvia
Publication date	2019-01-07
Publication information	Fruhwirth-Schnatter, S. Celeux, G., Robert, C.P. (eds.). Handbook of Mixture Analysis
Publisher	CRC Press
Link to online version	https://www.crcpress.com/Handbook-of-Mixture-Analysis/Fruhwirth-Schnatter-Celeux-Robert/p/boo
Item record/more information	http://hdl.handle.net/10197/10285
Publisher's statement	This is an Accepted Manuscript of a book chapter published by CRC Press in Fruhwirth-Schnatter, S. Celeux, G., Robert, C.P. (eds.). Handbook of Mixture Analysis on 07 January 2019, available online: http://www.crcpress.com/9781498763813
Publisher's version (DOI)	10.1201/9780429055911

Mixtures of Experts Models*

Isobel Claire Gormley† and Sylvia Frühwirth-Schnatter ‡

Abstract

Mixtures of experts models provide a framework in which covariates may be included in mixture models. This is achieved by modelling the parameters of the mixture model as functions of the concomitant covariates. Given their mixture model foundation, mixtures of experts models possess a diverse range of analytic uses, from clustering observations to capturing parameter heterogeneity in cross-sectional data. This chapter focuses on delineating the mixture of experts modelling framework and demonstrates the utility and flexibility of mixtures of experts models as an analytic tool.

1 Introduction

The terminology *mixtures of experts models* encapsulates a broad class of mixture models in which the model parameters are modelled as functions of concomitant covariates. While the response variable y is modelled via a mixture model, model parameters are modelled as functions of other, related, covariates x from the context under study.

The mixture of experts nomenclature (ME) has its origins in the machine-learning literature (Jacobs *et al.* , 1991), but mixtures of experts models appear in many different guises, including switching regression models (Quandt, 1972), concomitant variable latent-class models (Dayton & Macready, 1988), latent class regression models (DeSarbo & Cron, 1988), and mixed models (Wang *et al.* , 1996). Li *et al.* (2011) discuss finite smooth mixtures, a special case of ME modelling. McLachlan & Peel (2000) and Frühwirth-Schnatter (2006) provide background to a range of mixtures of experts models; Masoudnia & Ebrahimpour (2014) survey the ME literature from a machine learning perspective.

The mixture of experts framework facilitates flexible modelling, allowing a wide range of application. ME models for rank data (Gormley & Murphy, 2008b, 2010a), ME models for network data (Gormley & Murphy, 2010b), for time series data (Waterhouse *et al.* , 1996; Huerta *et al.* , 2003; Frühwirth-Schnatter *et al.* , 2012), for non-normal data (Villani *et al.* , 2009; Chamroukhi, 2015) and for longitudinal data (Tang & Qu, 2015), among others, have been developed. Peng *et al.* (1996) employed a hierarchical mixture of experts model in a speech recognition context. The general ME framework has also been incorporated in the mixed membership model setting, giving rise to a mixed membership of experts model (White & Murphy, 2016), and into the infinite mixture model setting (Rasmussen &

*A chapter prepared for the forthcoming *Handbook of Mixture Analysis*

†School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin, Ireland. claire.gormley@ucd.ie

‡Institute for Statistics and Mathematics, Vienna University of Economics and Business, Austria. sfruehwi@wu.ac.at

Ghahramani, 2002). Cluster weighted models (Ingrassia *et al.* , 2015; Subedi *et al.* , 2013; Gershenfeld, 1997) are also closely related to ME models.

This chapter introduces the generic mixture of experts framework, in Section 2, and describes approaches to inference for ME models in Section 3. A broad range of illustrative data analyses are given in Section 4, and an overview of existing softwares which fit ME models is provided. Section 5 discusses identifiability issues for mixtures of experts models. The chapter concludes with some discussion of the benefits and issues of the ME framework, and of some areas ripe for future development.

2 The Mixture of Experts Framework

Any mixture model which incorporates covariates or concomitant variables falls within the mixture of experts framework.

2.1 A mixture of experts model

Let $y_1, \ldots, y_n$ be an independent and identically distributed sample of outcome variables from a population modelled by a G component finite mixture model. Depending on the application context, the outcome variable can be univariate or multivariate, discrete or continuous, or of a more general structure such as time series or network data. Each component g (for $g = 1, \ldots, G$) is modelled by the probability density function $f_g(\cdot|\theta_g)$ with parameters denoted by θ_g, and has weight η_g where $\sum_{g=1}^{G} \eta_g = 1$. Observation y_i $(i = 1, \ldots, n)$ has q associated covariates, which are denoted x_i. The ME model extends the standard finite mixture model introduced in Chapter 1 of this volume by allowing model parameters to be functions of the concomitant variables x_i:

$$p(y_i|x_i) = \sum_{g=1}^{G} \eta_g(x_i) f_g(y_i|\theta_g(x_i)). \quad (1)$$

ME models can be considered as a member of the class of conditional mixture models (Bishop, 2006); for a given set of covariates x_i, the distribution of y_i is a finite mixture model. Jacobs *et al.* (1991) consider the component densities $f_g(y_i|\theta_g(x_i))$ as the *experts*, which model different parts of the input space, and the component weights $\eta_g(x_i)$ as the *gating networks*, hence the *mixture of experts* terminology.

The models for $\eta_g(x_i)$ and for $\theta_g(x_i)$ in (1) vary and are typically application specific. For example, Jacobs *et al.* (1991) model the component weights using a multinomial logit (MNL) regression model, and the component densities using generalized linear models. Young & Hunter (2010) provide further flexibility by allowing the mixing proportions to be modelled nonparametrically, as a function of the covariates.

2.2 An illustration

A simple simulated data set is employed here to introduce the mixture of experts framework. Figure 1 shows $n = 200$ two-dimensional continuously valued observations, $y_1, \ldots, y_n$, simulated from an ME model with $G = 2$ components. A single $(q = 1)$ categorical covariate x_i is associated with each observation representing, for example, gender where level 0 denotes female. Interest lies in clustering the observations and exploring any relations between the resulting clusters and the associated covariate.

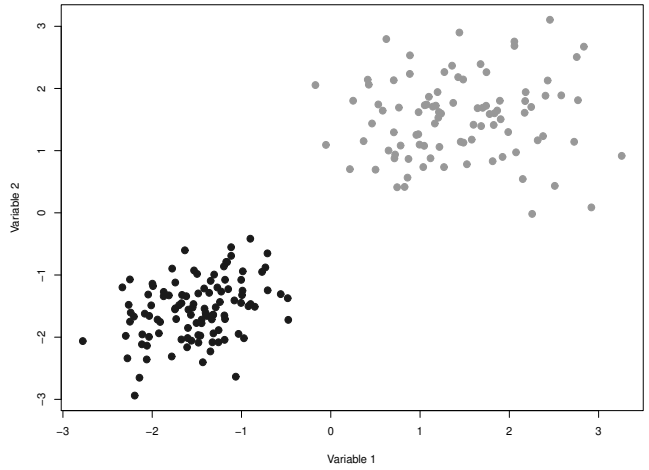


Figure 1: A two-dimensional simulated data set, from a $G = 2$ ME model. Black observations belong to cluster 1, and grey to cluster 2, based on the MAP clustering from fitting a $G = 2$ mixture of bivariate Gaussian distributions.

Table 1: Cross tabulation of MAP cluster memberships and the gender covariate for the simulated data of Figure 1

	Female	Male
Cluster 1	75	17
Cluster 2	23	85

It is common that a clustering method is implemented on the outcome variables of interest, $y_1, \ldots, y_n$, without reference to the covariate information. Once a clustering has been produced, the user typically probes the clusters to investigate their structure. Interpretations of the clusters are produced with reference to values of the model parameters within each cluster and with reference to the covariates that were not used in the construction of the clusters. Therefore, a natural approach to modelling the data in Figure 1 is to cluster them by fitting a two component mixture of bivariate Gaussian distributions to $y_1, \ldots, y_n$. The *maximum a posteriori* (MAP) cluster membership of each observation resulting from fitting such a model is also illustrated in Figure 1.

A cross tabulation of the MAP cluster memberships and the gender covariate is given in Table 1. It is clear that females have a strong presence in cluster 1, and males in cluster 2. However, the mixture of Gaussians model fitted does not incorporate or quantify this relationship or its associated uncertainty. It is in such a setting that an ME model is useful.

The model from which the data in Figure 1 are simulated is an ME model where $f_g(y_i|\theta_g) = \phi(y_i|\mu_g, \Sigma_g)$ is the density of a bivariate Normal distribution and in which the component weights arise from a multinomial logit model with G categories with gender as covariate x_i, i.e.

$$\log\left[\frac{\eta_g(x_i)}{\eta_1(x_i)}\right] = \gamma_{g0} + \gamma_{g1}x_i, \tag{2}$$

where cluster 1 is the baseline cluster with $\gamma_1 = (\gamma_{10}, \gamma_{11})^\top = (0, 0)^\top$, and $g = 2, \ldots, G$. In our example, where $G = 2$, model (2) reduces to a binary logit model. The parameter

γ_{g1} (and its associated uncertainty) quantifies the relationship between the gender covariate and membership of cluster g, with $\gamma_{g1} = 0$ corresponding to independence between cluster membership and the gender covariate. Note that such a model easily extends to $q > 1$ covariates $x_i = (x_{i1}, \ldots, x_{iq})$ with associated parameter $\gamma_g = (\gamma_{g0}, \ldots, \gamma_{gq})^\top$ for cluster g.

Fitting such an ME model to the simulated data results in a MAP clustering unchanged from that reported in Table 1 and gives the maximum likelihood estimate $\hat{\gamma}_{21} = 2.79$, with standard error 0.36. (Details of the maximum likelihood estimation process and standard error derivation follow in Section 3.1.) Thus, the odds of a male belonging to cluster 2 are $\exp(2.79) \approx 16$ times greater than the odds of a female belonging to cluster 2. Thus the ME model has clustering capabilities and provides insight into the type of observation which characterises each cluster.

2.3 The suite of ME models

The ME model outlined in Section 2.2 involves modelling the component weights as a function of covariates. This is one model type (termed a *simple mixture of experts model*) from the ME framework. Figure 2 shows a graphical model representation of the suite of four models in the ME framework, based on a latent variable representation of the mixture model (1), involving the latent cluster membership of each observation, denoted z_i, where $z_i = g$ if observation y_i belongs to cluster g. The indicator variable z_i therefore has a multinomial distribution with a single trial and probabilities equal to $\eta_g(x_i)$ for $g = 1, \ldots, G$ and the latent variable representation reads:

$$y_i|x_i, z_i = g \sim f_g(y_i|\theta_g(x_i)), \qquad \mathrm{P}(z_i = g|x_i) = \eta_g(x_i). \tag{3}$$

This suite of models ranges from a standard mixture of experts regression model (in which all model parameters are functions of covariates) to the special cases where some of the model parameters do not depend on covariates. The four models in the ME framework have the following interpretations, see also Figure 2:

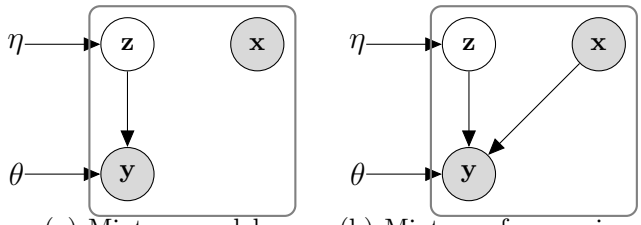
(a) Mixture models, where the outcome variable distribution depends on the latent cluster membership, denoted z. The model is independent of the covariates x; i.e. $p(y_i, z_i|x_i) = f_{z_i}(y_i|\theta_{z_i})\eta_{z_i}$.

(b) Mixtures of regression models, where the outcome variable distribution depends on both the covariates x and the latent cluster membership variable z; the distribution of the latent variable is independent of the covariates; i.e. $p(y_i, z_i|x_i) = f_{z_i}(y_i|\theta_{z_i}(x_i))\eta_{z_i}$.

(c) Simple mixtures of experts models, where the outcome variable distribution depends on the latent cluster membership variable z and the distribution of the latent variable z depends on the covariates x; i.e. $p(y_i, z_i|x_i) = f_{z_i}(y_i|\theta_{z_i})\eta_{z_i}(x_i)$.

(d) Standard mixtures of experts regression models, where the outcome variable distribution depends on both the covariates x and on the latent cluster membership variable z. Additionally the distribution of the latent variable z depends on the covariates x; i.e. $p(y_i, z_i|x_i) = f_{z_i}(y_i|\theta_{z_i}(x_i))\eta_{z_i}(x_i)$.

The manner in which the different models within the ME framework depend on the covariates is typically application specific. The component weights are usually modelled using a MNL model, but this need not be the case; Geweke & Keane (2007) employ a model similar to an ME model, where the component weights have a multinomial probit structure. The form



(a) Mixture model

(b) Mixture of regressions model

(c) Simple mixture of experts model

(d) Standard mixture of experts regression model

Figure 2: The graphical model representation of mixtures of experts models. The differences between the four special cases are due to the presence or absence of edges between the covariates x and the latent variable z and response variable y. For model (a) $p(y, z|x) = p(y|z)p(z)$, for model (b) $p(y, z|x) = p(y|x, z)p(z)$, for model (c) $p(y, z|x) = p(y|z)p(z|x)$, whereas for model (d) $p(y, z|x) = p(y|x, z)p(z|x)$.

of the distribution $f_g(y_i|\theta_g(x_i))$ depends on the type of outcome data under study. The applications of the ME framework outlined in Section 4 include cases where the outcome data range from a categorical time series, to rank data, to network data.

3 Statistical Inference for Mixtures of Experts Models

Before illustrating the breadth of the ME framework through illustrative applications in Section 4, the issue of inference for ME models is addressed. For any ME model that is underpinned by a finite mixture model, the approaches to inference outlined in Chapter 2 and Chapter 5 in this volume are applicable. Jacobs *et al.* (1991) and Jordan & Jacobs (1994) derive maximum likelihood estimates (MLEs) for ME models via the expectation-maximisation (EM) algorithm; Gormley & Murphy (2008a) employ the closely related expectation-minorisation-maximisation (EMM) algorithm. Estimation of the ME model within the Bayesian framework is detailed, among others, in Peng *et al.* (1996), Frühwirth-Schnatter & Kaufmann (2008), Villani *et al.* (2009), Gormley & Murphy (2010a) and in Frühwirth-Schnatter *et al.* (2012) in which Markov chain Monte Carlo methods (Tanner, 1996) are used; Bishop & Svenskn (2003) use variational methods in the Bayesian paradigm to perform inference for a hierarchical mixture of experts model. Hunter & Young (2012) present an algorithm for parameter estimation in a semiparametric mixtures of regressions model setting.

In this section, a general overview of approaches to inference in the ME framework is provided. Throughout the section, $y = (y_1, \ldots, y_n)$ will denote the collection of outcome variables and $\mathbf{x} = (x_1, \ldots, x_n)$ the associated covariates. The latent cluster membership indicators introduced in (3) are denoted by $\mathbf{z} = (z_1, \ldots, z_n)$, whereas $\theta = \{\theta_1, \ldots, \theta_G\}$ refers to the collection of the G component parameters and $\gamma = \{\gamma_2, \ldots, \gamma_G\}$ to the unknown

parameters in the G component weights.

The exact manner in which an ME model is estimated again depends on the nature of the ME model and the outcome variable. The simple simulated data example of Section 2.2 is used here to delineate approaches to inference; more detailed application specific estimation approaches are outlined in Section 4.

3.1 Maximum likelihood estimation

The EM algorithm (Dempster *et al.* , 1977) provides an efficient approach to deriving MLEs in ME models. The EM algorithm is most commonly known as a technique to produce MLEs in settings where the data under study are incomplete or when optimisation of the likelihood would be simplified if an additional set of variables were known. The iterative EM algorithm consists of an expectation (E) step followed by a maximisation (M) step. Generally, during the E step the conditional expectation of the complete (i.e. observed and unobserved) data log likelihood is computed, given the data and current parameter values. In the M step the expected log likelihood is maximised with respect to the model parameters. The imputation of latent variables often makes maximisation of the expected log likelihood more feasible. The parameter estimates produced in the M step are then used in a new E step and the cycle continues until convergence. The parameter estimates produced on convergence are estimates that achieve a stationary point of the likelihood function of the data, which is at least a local maximum but may be a saddle point.

The component weights of the simple mixture of experts model outlined in Section 2.2 are given by

$$\eta_g(x_i|\gamma) = \exp\left(\tilde{x}_i\gamma_g\right) / \sum_{g'=1}^{G} \exp\left(\tilde{x}_i\gamma_{g'}\right) \tag{4}$$

where $\tilde{x}_i = (1, x_i)$ and $\gamma_g = (\gamma_{g0}, \gamma_{g1})^\top$. Note that this is a special case of the multinomial logit model. For the Normal distribution $\theta_g = \{\mu_g, \Sigma_g\}$ and the likelihood function of the simple mixture of experts model is

$$\mathrm{L}(\gamma, \theta; G) \quad = \quad p(y|\mathbf{x}, \gamma, \theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \eta_g(x_i|\gamma)\phi(y_i|\mu_g, \Sigma_g),$$

where $\phi(y_i|\mu_g, \Sigma_g)$ is the pdf of the d-variate Normal distribution and $d = \dim(y_i)$. It is difficult to directly obtain MLEs from this likelihood. To alleviate this, the data are augmented by imputing for each observation $y_i, i = 1, \ldots, n$, the latent group membership indicator z_i. For the EM algorithm, this latent variable is represented through G binary variables $(z_{i1}, \ldots, z_{iG})$ where $z_{ig} = \mathbb{I}(z_i = g)$ takes the value 1 if observation y_i is a member of component g and the value 0 otherwise. This provides the complete data likelihood

$$\mathrm{L_c}(\gamma, \theta, \mathbf{z}; G) = \; p(y, \mathbf{z}|\mathbf{x}, \gamma, \theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left\{\eta_g(x_i|\gamma)\phi(y_i|\mu_g, \Sigma_g)\right\}^{z_{ig}}, \tag{5}$$

the expectation of (the log of) which is obtained in the E step of the EM algorithm. As the complete data log likelihood is linear in the latent variable, the E step simply consists of replacing for each $i = 1, \ldots, n$ the missing data z_i with their expected values $\hat{z}_i$. In the M step the complete data log likelihood, computed with the estimates $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$, is maximised to provide estimates of the component weight parameters $\hat{\gamma}$ and the component parameters $\hat{\theta}$.

Algorithm 1 EM algorithm for a simple Gaussian mixture of experts model

Let $s = 0$. Choose initial estimates for the component weight parameters $\gamma^{(0)} = (0, \gamma_2^{(0)}, \ldots, \gamma_G^{(0)})$ and for the component parameters $\mu_g^{(0)}$ and $\Sigma_g^{(0)}$ for $g = 1, \ldots, G$.

1 **E step**: for $i = 1, \ldots, n$ and $g = 1, \ldots, G$ compute the estimates:

$$z_{ig}^{(s+1)} = \eta_g^{(s)}(x_i|\gamma^{(s)})\phi(y_i|\mu_g^{(s)}, \Sigma_g^{(s)}) / \sum_{g'=1}^{G} \eta_{g'}^{(s)}(x_i|\gamma^{(s)})\phi(y_i|\mu_{g'}^{(s)}, \Sigma_{g'}^{(s)}).$$

2 **M step:** Substituting the $z_{ig}^{(s+1)}$ values obtained in the E step into the log of the complete data likelihood (5) forms the so called 'Q function'

$$\begin{aligned} Q &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}^{(s+1)} \left[\tilde{x}_i\gamma_g - \log\left\{ \sum_{g'=1}^{G} \exp(\tilde{x}_i\gamma_{g'}) \right\} \right. \\ &\quad \left. -d/2\log(2\pi) - 1/2\log|\Sigma_g| - 1/2(y_i - \mu_g)^\top \Sigma_g^{-1}(y_i - \mu_g) \right] \end{aligned}$$

with $d = \dim(y_i)$, which is maximised with respect to the model parameters.

(a) The updates of the $g = 1, \ldots, G$ component means and covariances are, respectively:

$$\begin{aligned} \mu_g^{(s+1)} &= \sum_{i=1}^{n} z_{ig}^{(s+1)} y_i / \sum_{i=1}^{n} z_{ig}^{(s+1)} \\ \Sigma_g^{(s+1)} &= \sum_{i=1}^{n} z_{ig}^{(s+1)} (y_i - \mu_g^{(s+1)})(y_i - \mu_g^{(s+1)})^\top / \sum_{i=1}^{n} z_{ig}^{(s+1)}. \end{aligned}$$

(b) The update for the component weight parameters is obtained via a numerical optimisation step, such as a Newton-Raphson step, where for $g = 2, \ldots, G$

$$\gamma_g^{(s+1)} = \gamma_g^{(s)} - (H(\gamma_g^{(s)}))^{-1} Q'(\gamma_g^{(s)})$$

and Q' and H denote the first and second derivatives of Q with respect to γ_g respectively. Note that this M step is equivalent to fitting a generalised linear model with weights provided by the E step.

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

The EM algorithm for fitting ME models is straightforward in principle, but the M step is often difficult in practice. This is usually due to a complex component density and/or component weights model, or a large parameter set. A modified version of the EM algorithm, the Expectation and Conditional Maximisation (ECM) algorithm (Meng & Rubin, 1993) is therefore often employed. In the ECM algorithm, the M step consists of a series of conditional maximisation steps. In the context of the simple mixture of experts example considered here, these maximisations are not straightforward with regard to the γ parameters; as in any MNL model, no closed form expression for the parameter MLEs is available. Thus, while the conditional M steps for μ_g and Σ_g $\forall g = 1, \ldots, G$ are available in closed form, the conditional M step for γ requires the use of a numerical optimisation technique, or as in Gormley & Murphy (2008b) the MM algorithm (Hunter & Lange, 2004) in which a

minorising function is iteratively maximised and updated. In summary, to fit the simple mixture of experts example outlined in Section 2.2 the EM algorithm proceeds as described in Algorithm 1. In the simulated data example, $d = 2$.

McLachlan & Peel (2000) outline a number of approaches to assessing convergence in Step 3; typically it is assessed by tracking the change in the log likelihood as the algorithm proceeds. Standard errors of the resulting parameter estimates are not automatically produced by the EM algorithm, but they can be approximately computed after convergence, for example, by computing and inverting the observed information matrix (McLachlan & Peel, 2000). For a detailed discussion of EM algorithms in a mixture context see Chapter 2 and 3 of this volume.

3.2 Bayesian estimation

Estimation of ME models can be achieved within the Bayesian paradigm, either using a Markov chain Monte Carlo (MCMC) algorithm or via a variational approach. The reader is directed to Bishop & Svenskn (2003) for details on the variational approach; this section focuses on inference using MCMC methods. Both the Gibbs sampler (Geman & Geman, 1984) and the Metropolis-Hastings algorithm (Chib & Greenberg, 1995; Metropolis *et al.* , 1953) are typically required. Again, the specific MCMC algorithm, and the form of the prior distributions, depend on the nature of the ME model under study and on the type of the response data. As is standard in Bayesian estimation of mixture models (Diebolt & Robert, 1994; Hurn *et al.* , 2003) fitting ME models is greatly simplified by augmenting the observed data with the latent group indicator variable z_i for each observation y_i.

Performing inference on the illustrative simple mixture of experts model of Section 2.2 is again straightforward in principle, but can be difficult in practice. To begin, priors for the model parameters μ_g, Σ_g for $g = 1, \ldots, G$ and γ_g $(g = 2, \ldots, G)$ require specification. Positing a conditional d-variate normal prior $\mathcal{N}(\mu_0, \Lambda_0)$ on the group means μ_g, and an inverse Wishart prior $\mathcal{IW}(\nu_0, S_0)$ on the group covariances Σ_g provides conjugacy for these parameters (Hoff, 2009). The full conditional distributions for these parameters are therefore available in closed form, and thus Gibbs sampling can be used to draw samples.

A $(q+1)$-variate normal $\mathcal{N}(\mu_\gamma, \Lambda_\gamma)$ is an intuitive prior for the component weight parameters γ_g, but it is non-conjugate. Hence the full conditional distribution is not available in closed form and a Metropolis-Hastings (MH) step can be applied to sample the component weight parameters. One sweep of such a Metropolis-within-Gibbs sampler required to fit the simple mixture of experts model of Section 2.2 in a Bayesian framework is outlined below. Note that the full conditional distribution of the latent indicator variable z_i for $i = 1, \ldots, n$ is also available in closed form and thus a Gibbs step is available, see Algorithm 2.

Sampling the component weight parameters in Step 4 through a MH-algorithm brings issues such as choosing suitable proposal distributions $q(\gamma_g^* \,|\, \gamma_g, \gamma_{-g})$ and tuning parameters, which may make fitting ME models troublesome. Gormley & Murphy (2010b) detail an approach to deriving proposal distributions with attractive properties, within the context of an ME model for network data.

Alternatively, Frühwirth-Schnatter *et al.* (2012) exploit data augmentation of the MNL model (4) based on the differenced random utility model representation in the context of ME models to implement Step 4. As shown by Frühwirth-Schnatter & Frühwirth (2010), for each $g = 1, \ldots, G$ the MNL model has the following representation as a binary logit model

Algorithm 2 MH-within-Gibbs MCMC inference for a simple Gaussian mixture of experts model

Iterate the following steps for $m = 1, \ldots, M$:

1 For $g = 1, \ldots, G$, draw μ_g from the d-variate normal posterior $\mathcal{N}(\mu_{ng}, \Lambda_{ng})$ where $\Lambda_{ng} = (\Lambda_0^{-1} + n_g \Sigma_g^{-1})^{-1}$ and $\mu_{ng} = \Lambda_{ng}(\Lambda_0^{-1}\mu_0 + \Sigma_g^{-1} n_g \bar{y}_g)$ and $n_g = \sum_{i=1}^n \mathbb{I}(z_i = g)$ and $n_g \bar{y}_g = \sum_{i=1}^n y_i \mathbb{I}(z_i = g)$.

2 For $g = 1, \ldots, G$, draw Σ_g from $\mathcal{IW}(\nu_{ng}, S_{ng})$ where $\nu_{ng} = \nu_0 + n_g$ and $S_{ng} = S_0 + \sum_{i=1}^n \mathbb{I}(z_i = g)(y_i - \mu_g)(y_i - \mu_g)^\top$.

3 For $i = 1, \ldots, n$ draw z_i from a multinomial distribution $\mathcal{M}(1, p_{i1}, \ldots, p_{iG})$ with success probabilities $(p_{i1}, \ldots, p_{iG})$ where

$$p_{ig} = \eta_g(x_i|\gamma)\phi(y_i|\mu_g, \Sigma_g) / \sum_{g'=1}^{G} \eta_{g'}(x_i|\gamma)\phi(y_i|\mu_{g'}, \Sigma_{g'}).$$

4 For $g = 2, \ldots, G$, the component weight parameters γ_g are updated via a Metropolis-Hastings step, while holding the remaining component weight parameters γ_{-g} fixed. Typically, a multivariate normal proposal distribution $q(\gamma_g^*|\gamma_g, \gamma_{-g})$ is employed:

   (a) Propose $\gamma_g^* \sim \mathcal{N}(\tilde{\mu}_\gamma, \tilde{\Lambda}_\gamma)$ from a $(q+1)$-variate Normal distribution where $\tilde{\mu}_\gamma$ and $\tilde{\Lambda}_\gamma$ are user specified and might depend on the current value of γ.

   (b) If $U \sim U[0,1]$ is such that

$$U \leq \min\left\{ \frac{p(\mathbf{z}|\gamma_g^*, \gamma_{-g}, \mathbf{x}) p(\gamma_g^*) q(\gamma_g \mid \gamma_g^*, \gamma_{-g})}{p(\mathbf{z}|\gamma_g, \gamma_{-g}, \mathbf{x}) p(\gamma_g) q(\gamma_g^* \mid \gamma_g, \gamma_{-g})}, \; 1 \right\},$$

   then set $\gamma_g = \gamma_g^*$; otherwise leave γ_g unchanged.

conditional on knowing $\lambda_{hi} = \exp(\tilde{x}_i \gamma_h)$ for all $h \neq g$:

$$\begin{aligned} u_{gi} &= \tilde{x}_i \gamma_g - \log(\sum_{h \neq g} \lambda_{hi}) + \varepsilon_{gi}, \\ D_i^g &= \mathbb{I}(u_{gi} \geq 0) \end{aligned} \tag{6}$$

where u_{gi} is a latent variable, ε_{gi} are i.i.d. errors following a logistic distribution, and $D_i^g = \mathbb{I}(z_i = g)$ is a binary outcome variable indicating whether the group indicator z_i is equal to g. Note that $\gamma_1 = 0$ for the baseline, hence $\lambda_{1i} = 1$. In a data augmented implementation of Step 4, the latent variables $(u_{2i}, \ldots, u_{Gi})$ are introduced for each $i = 1, \ldots, n$ as unknowns. Given $\lambda_{2i}, \ldots, \lambda_{Gi}$ and z_i, $(u_{2i}, \ldots, u_{Gi})$ can be sampled in closed form from exponentially distributed random variables. Following Scott (2011), natural proposal distributions are available to implement an MH-step to sample $\gamma_g|\gamma_{-g}, \mathbf{z}, \mathbf{u}_g$ for all $g = 2, \ldots, G$ conditional on $\mathbf{u}_g = \{u_{g1}, \ldots, u_{gn}\}$ from the linear, non-Gaussian regression model (6).

To avoid any MH-step, Frühwirth-Schnatter *et al.* (2012) apply auxiliary mixture sampling as introduced by Frühwirth-Schnatter & Frühwirth (2010) to (6) and approximate for each ε_{gi} the logistic distribution by a 10-component scale mixture of Normal distributions with zero means and parameters (s_r^2, w_r), $r = 1, \ldots, 10$. In a second step of data augmentation, the component indicator r_{gi} is introduced as yet another latent variable. Conditional on the latent variables $\mathbf{u}_g$ and the indicators $\mathbf{r}_g = \{r_{g1}, \ldots, r_{gn}\}$ the binary logit model (6) reduces to a linear Gaussian regression model. Hence, the posterior $\gamma_g|\gamma_{-g}, \mathbf{z}, \mathbf{u}_g, \mathbf{r}_g$ is Gaus-

sian and a Gibbs step is available to sample γ_g for all $g = 2, \ldots, G$ conditional on $\mathbf{u}_g$ and $\mathbf{r}_g$. Finally, each component indicator r_{gi} is sampled from a discrete distribution conditional on u_{gi} and γ.

Chapter 13 in this volume details Bayesian estimation of informative regime switching models which can be regarded as an extension of ME models to hidden Markov models in time series analysis.

As in any mixture model setting, the so called label switching problem (Stephens, 2000; Frühwirth-Schnatter, 2011a) must be considered when employing such Gibbs based algorithms, see Chapter 5. This identifiability issue, along with others, is discussed in Section 5.

3.3 Model selection

Within the suite of ME models outlined in Section 2.3 the question of which, how and where covariates are used naturally arises. This is a challenging problem as the space of ME models is potentially very large, once variable selection for the covariates entering the component weights and the mixture components is considered. Thus in practice only models where covariates enter all mixture components and/or all component weights as main effects are typically considered in order to restrict the size of the model search space. In fact, even for this reduced model space, there are a maximum of $G \times 2^q \times 2^q$ possible models to consider. In ME models involving generalised linear models of covariates, standard variable selection approaches can be used to find the optimal model. Practical approaches to this issue are detailed in the illustrative applications of Section 4. Note that the manner in which covariates enter the ME model may also be guided by the question of interest in the application under study.

If the number of components G is unknown, the model search space increases again. Approaches such as marginal likelihood evaluation, or information criteria, are useful for choosing the optimal G in ME models; the reader is referred to Chapter 7 in this volume which addresses model selection and selecting the number of components in a mixture model in great detail.

Marginal likelihood computation for mixtures of experts models

As discussed in Chapter 7, Section 7.2.3.2, highly accurate sampling-based approximations to the marginal likelihood are available, if G is not too large. For instance, Frühwirth-Schnatter & Kaufmann (2008) apply bridge sampling (Frühwirth-Schnatter, 2004) to compute marginal likelihoods for a mixture of experts model with a single covariate (that is $q = 1$) with up to four components. Frühwirth-Schnatter (2011b) combines auxiliary mixture sampling (Frühwirth-Schnatter & Wagner, 2008) with importance sampling to compute marginal likelihoods for mixture of experts models. A detailed summary of this approach is provided below.

Permutation sampling is applied to ensure that all equivalent modes of the posterior distribution are visited. Consider a permutation $\sigma \in \mathfrak{S}(G)$, where $\mathfrak{S}(G)$ denotes the set of the $G!$ permutations of $\{1, \ldots, G\}$. To relabel all parameters in a mixture of experts model according to the permutation σ, define $\theta_g^\star = \theta_{\sigma(g)}$ and $\eta_g^\star(\tilde{x}_i) = \eta_{\sigma(g)}(\tilde{x}_i)$ for $g = 1, \ldots, G$. Special attention has to be given to the correct relabelling of the coefficients γ_g in the MNL model when applying the permutation σ. The coefficients $(\gamma_1, \ldots, \gamma_G)$ and $(\gamma_1^\star, \ldots, \gamma_G^\star)$

defining, respectively, the MNL models $\eta_g(\tilde{x}_i)$ and $\eta^\star_g(\tilde{x}_i)$ are related through:

$$\begin{aligned}\tilde{x}_i\gamma^\star_g &= \log\left[\frac{\eta^\star_g(\tilde{x}_i)}{\eta^\star_{g_0}(\tilde{x}_i)}\right] = \log\left[\frac{\eta_{\sigma(g)}(\tilde{x}_i)}{\eta_{\sigma(g_0)}(\tilde{x}_i)}\right] = \log\left[\frac{\eta_{\sigma(g)}(\tilde{x}_i)}{\eta_{g_0}(\tilde{x}_i)}\right] - \log\left[\frac{\eta_{\sigma(g_0)}(\tilde{x}_i)}{\eta_{g_0}(\tilde{x}_i)}\right] \\ &= \tilde{x}_i(\gamma_{\sigma(g)} - \gamma_{\sigma(g_0)}).\end{aligned}$$

To ensure that the baseline g_0 (assumed to be equal to $g_0 = 1$ throughout this chapter) remains the same, despite relabeling, the coefficients are permuted in the following way:

$$\gamma^\star_g = \gamma_{\sigma(g)} - \gamma_{\sigma(g_0)}, \quad g = 1, \ldots, G,$$

which indeed implies that $\gamma^\star_{g_0} = 0$. For $G = 2$, the sign of all coefficients of γ_2 is simply flipped, if $\sigma = (2, 1)$ and remains unchanged, otherwise.

Frühwirth-Schnatter & Wagner (2008) discuss various importance sampling estimators of the marginal likelihood for non-Gaussian models such as logistic models. Using auxiliary mixture sampling, one of their approaches constructs the importance density from the Gaussian full conditional densities appearing in the augmented Gibbs sampler. This approach is easily extended to mixture of experts models. As discussed in Section 3.2, auxiliary mixture sampling yields Gaussian posteriors $p(\gamma_g|\gamma_{-g}, \mathbf{z}, \mathbf{u}_g, \mathbf{r}_g)$ for the MNL coefficients γ_g in a mixture of experts models, conditional on the latent utilities $\mathbf{u}_g$ and the latent indicators $\mathbf{r}_g$. This allows construction of an importance density $q_G(\theta)$ as in Chapter 7, Section 7.2.3.2, however it is essential that $q_G(\theta)$ covers all symmetric modes of the mixture posterior. A successful strategy is to apply random permutation sampling, where each sampling step is concluded by relabelling as described above, using a randomly selected permutation $\sigma \in \mathfrak{S}(G)$. The corresponding importance density reads:

$$q_G(\theta) = \frac{1}{S}\sum_{s=1}^{S}\prod_{g=2}^{G} p(\gamma_g|\gamma^{(s)}_{-g}, \mathbf{u}^{(s)}_g, \mathbf{r}^{(s)}_g, \mathbf{z}^{(s)}) \prod_{g=1}^{G} p(\theta_g|\mathbf{z}^{(s)}, y), \tag{7}$$

where $\{\gamma^{(s)}, \mathbf{u}^{(s)}_2, \ldots, \mathbf{u}^{(s)}_G, \mathbf{r}^{(s)}_2, \ldots, \mathbf{r}^{(s)}_G, \mathbf{z}^{(s)}\}$, $s = 1, \ldots, S$ is a subsequence of posterior draws. Only if S is large compared to $G!$, then all symmetric modes are covered by random permutation sampling, with the number of visits per mode being on average $S/G!$. The construction of this importance density is fully automatic and it is sufficient to store the moments of the various conditional densities (rather than the allocations $\mathbf{z}$ and the latent utilities $\mathbf{u}_g$ and indicators $\mathbf{r}_g$ themselves) during MCMC sampling for later evaluation. This importance density is used to compute importance sampling estimators of the marginal likelihood, see the illustrative application in Section 4.1.

4 Illustrative Applications

The utility of ME models is illustrated in this section through the use of several applications. ME Markov chain models for categorical time series, ME models for ranked preference data, and ME models for network data, all of which are members of the ME model framework, are applied.

4.1 Analysing marijuana use through ME Markov chain models

Lang *et al.* (1999) studied data on the marijuana use of 237 teenagers taken from five annual waves (1976-80) of the National Youth Survey. The respondents were 13 years old in 1976

and reported for five consecutive years their marijuana use in the past year as a categorical variable with the three categories "never", "not more than once a month" and "more than once a month". Hence, for $i = 1, \ldots, 237$, the outcome variable is a categorical time series $y_i = (y_{i0}, y_{i1}, \ldots, y_{i4})$ with three states, labeled 1 for never-user, 2 for light and 3 for heavy users.

To identify groups of teenagers with similar marijuana use behaviour, Frühwirth-Schnatter (2011b) applied a ME approach based on Markov chain models (Frühwirth-Schnatter *et al.* , 2012) and considered each time series y_i as a single entity belonging to one of G underlying classes. Various types of ME Markov chain models were applied to capture dependence in marijuana use over time and to investigate if the gender of the teenagers can be associated with a certain type of marijuana use.

Given the times series nature of the categorical outcome variable y_i, the component density $f_g(\cdot)$ in the mixture of experts model (1) must have an appropriate form and various models are considered. Model $\mathcal{M}_1$ is a standard finite mixture of time-homogeneous Markov chain models of order one (Pamminger & Frühwirth-Schnatter, 2010) where each component-specific density $f_g(\cdot)$ in (1) is characterized by a transition matrix ξ_g with $J = 3$ rows and the weight distribution $\eta_1, \ldots, \eta_G$ is independent of any covariates. Each row $\xi_{g,j\cdot} = (\xi_{g,j1}, \ldots, \xi_{g,j3})$, $j = 1, \ldots, J$, of the matrix ξ_g represents a probability distribution over the three categories of marijuana use with

$$\xi_{g,jk} = \mathrm{P}(y_{it} = k | y_{i,t-1} = j, z_i = g), \qquad k = 1, \ldots, 3.$$

This model is extended in various ways to include covariate information into the transition behaviour. First, an inhomogeneous model (labelled model $\mathcal{M}_2$) is considered, where the transition matrix in each group depends on the gender x_i of the teenager. If all $J = 6$ possible combinations $\mathcal{H}_{it} = (y_{i,t-1}, x_i)$ of the immediate past $y_{i,t-1}$ at time t and the gender x_i are indexed by $j = 1, \ldots, J$, then the component-specific density $f_g(y_i|\xi_g)$ in (1) can be described by a generalized transition matrix ξ_g with six rows, with the jth row $\xi_{g,j\cdot} = (\xi_{g,j1}, \ldots, \xi_{g,j3})$ describing again the conditional distribution of y_{it}, given that the state of the history $\mathcal{H}_{it}$ equals j:

$$\xi_{g,jk} = \mathrm{P}(y_{it} = k | \mathcal{H}_{it} = j, z_i = g), \quad k = 1, \ldots, 3.$$

Evidently, the component specific distribution reads:

$$f_g(y_i|\xi_g) = \prod_{j=1}^{J} \prod_{k=1}^{3} \xi_{g,jk}^{n_{i,jk}} \tag{8}$$

where, for each time series i, $n_{i,jk} = \sum_{t=1}^{4} \mathbb{I}(y_{it} = k, \mathcal{H}_{it} = j)$ is the number of transitions into state k given a history of type j. Note that (8) is formulated conditional on the first observation y_{i0}.

Alternative component-specific distributions can be constructed, by defining the history $\mathcal{H}_{it}$ through different combinations of past values and covariates. Choosing $\mathcal{H}_{it} = (y_{i,t-1}, t)$, for instance, defines a time-inhomogeneous Markov chain model, labeled model $\mathcal{M}_3$, with $J = 12$ different covariate combinations. This model is able to capture the effect that the transition behaviour between the states might change as the teenagers grow older.

The most complex model, labelled model $\mathcal{M}_4$, extends model $\mathcal{M}_3$ by assuming additional dependence on gender, i.e. $\mathcal{H}_{it} = (y_{i,t-1}, t, x_i)$, with $J = 24$ different covariate combinations. Both model $\mathcal{M}_3$ and $\mathcal{M}_4$ are characterised by component-specific generalized transition matrices ξ_g with, respectively, 12 and 24 rows. For each of the models $\mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$, it is

Table 2: Marijuana data; marginal likelihood $\log p(y|\mathcal{M}_k)$ for various finite mixtures of homogeneous ($\mathcal{M}_1$) and inhomogeneous ($\mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$) Markov chain models with an increasing number G of classes (best values for each model in bold font)

			G		
Model	Covariates	J	1	2	3
$\mathcal{M}_1$	-	3	-605.5	**-600.0**	-600.3
$\mathcal{M}_2$	x_i	6	-610.0	**-601.3**	-603.6
$\mathcal{M}_3$	t	12	-613.7	**-596.5**	-599.4
$\mathcal{M}_4$	t, x_i	24	-619.8	-602.7	**-601.1**

assumed that the weight distribution $\eta_1, \ldots, \eta_G$ is independent of any covariate, leading to various finite mixtures of inhomogeneous Markov chain models.

Bayesian inference is carried out for all models $\mathcal{M}_1, \ldots, \mathcal{M}_4$ for an increasing number $G = 1, 2, 3$ of classes. MCMC estimation as described in Section 3.2 is easily applied, as the J rows $\xi_{g,j\cdot}$ of ξ_g are conditionally independent under the conditionally conjugate Dirichlet prior $\xi_{g,j\cdot} \sim \mathcal{D}(d_{0,j1}, \ldots, d_{0,j3})$. Given $\mathbf{z}$ and y, the generalized transition matrix ξ_g is sampled row-by-row from a total of JG Dirichlet distributions:

$$\xi_{g,j\cdot}|\mathbf{z}, y \sim \mathcal{D}(d_{0,j1} + n^g_{j1}, \ldots, d_{0,j3} + n^g_{j3}), \quad j = 1, \ldots, J, \; g = 1, \ldots, G, \tag{9}$$

where $n^g_{jk} = \sum_{i:z_i=g} n_{i,jk}$ is the total number of transitions into state k observed in class g given a history of type j.

For model comparison, the marginal likelihood is computed explicitly for $G = 1$, while importance sampling as described in Section 3.3 is applied for $G = 2, 3$, using the importance density:

$$q_G(\theta) = \frac{1}{S} \sum_{s=1}^{S} p(\eta|\mathbf{z}^{(s)}) \prod_{g=1}^{G} \prod_{j=1}^{J} p(\xi_{g,j\cdot}|\mathbf{z}^{(s)}, y),$$

where $p(\xi_{g,j\cdot}|\mathbf{z}, y)$ is equal to the full conditional Dirichlet posterior of $\xi_{g,j\cdot}$ given in (9). Random permutation sampling is applied to ensure that all $G!$ symmetric modes are visited and $S = 10,000$. The marginal likelihoods reported in Table 2 select $G = 2$ for all models except for $\mathcal{M}_4$, where $G = 3$ is selected. Among all models, the marginal likelihood is the highest for model $\mathcal{M}_3$ with $G = 2$ classes.

Hence, a time-inhomogeneous Markov chain model which does not depend on gender best describes the transition behaviour in each class. Table 3 reports the corresponding posterior means $\mathrm{E}(\xi_{g,\cdot}|y)$ and $\mathrm{E}(\eta_g|y)$ for each of the two groups. Label switching was resolved by applying k-means clustering to a vector constructed from all persistence probabilities at all time points. Both groups are roughly of equal size, with the first group being slightly larger. A characteristic difference is evident for the two groups of teenagers. In group 1, never-users have a high probability $\xi_{t,11}$ to remain never-users throughout the whole observation period, whereas this probability is much smaller for the second group right from the beginning and drops to only 45% in the last year.

To investigate if gender is associated with group membership, model $\mathcal{M}_3$ with $G = 2$ classes is combined with the ME model (4), by including gender as subject-specific covariate x_i as in the example in Section 2.2. This model is labelled model $\mathcal{M}_5$. Additionally, a dummy variable D_{i0} is included, indicating if the teenager used marijuana, light or heavy, in the first year. As $G = 2$, the ME model (4) reduces to a binary logit model with regression coefficients $\gamma_2 = (\gamma_{20}, \gamma_{21}, \gamma_{22})$, each assumed to follow a standard normal prior distribution.

Table 3: Marijuana data; finite mixture of time-inhomogeneous Markov chain models (model $\mathcal{M}_3$) with $G = 2$ classes; the estimated posterior mean $\mathrm{E}(\xi_{g,\cdot}|y)$ is arranged for each $t = 1, \ldots, 4$ as a 3×3 matrix; the estimated class sizes $\hat{\eta}_g$ are equal to the posterior mean $\mathrm{E}(\eta_g|y)$

	$t=1$			$t=2$			$t=3$			$t=4$		
Group 1	0.93	0.04	0.03	0.89	0.09	0.02	0.90	0.04	0.06	0.93	0.04	0.03
($\hat{\eta}_1 = 0.56$)	0.50	0.17	0.34	0.10	0.33	0.57	0.17	0.65	0.18	0.20	0.64	0.16
	0.22	0.18	0.60	0.10	0.17	0.73	0.04	0.27	0.69	0.15	0.12	0.74
Group 2	0.76	0.21	0.03	0.70	0.24	0.06	0.75	0.18	0.07	0.45	0.43	0.12
($\hat{\eta}_2 = 0.44$)	0.34	0.15	0.51	0.23	0.39	0.38	0.31	0.41	0.28	0.46	0.43	0.11
	0.18	0.23	0.59	0.10	0.22	0.68	0.13	0.15	0.72	0.05	0.10	0.85

Table 4: Marijuana data; ME model with $\tilde{x}_i = (1, x_i, D_{i0})$ (model $\mathcal{M}_5$), extending model $\mathcal{M}_3$ with $G = 2$ classes. Posterior expectation and 95% HPD region of the component weight parameters γ_{2j} in the ME model (4)

Covariate $\tilde{x}_{ij}$	$\mathrm{E}(\gamma_{2j}\|y)$	95% HPD region of γ_{2j}
constant	-0.69	(-1.75,0.35)
male (baseline: female)	0.28	(-0.71,1.22)
marijuana use in 1976 (baseline: no)	-0.07	(-1.70,1.43)
$\log p(y\|\mathcal{M}_5)$	-598.5	

From posterior inference in Table 4, we find that male teenagers have a slightly higher probability to belong to the second group, because $\mathrm{E}(\gamma_{21}|y) > 0$, however, the coefficient γ_{21} is not significantly different from 0. Similarly, the initial state from which a teenager started in 1976 does not have a significant influence on the probability to belong to the second group. This suggests that the ME time-inhomogeneous Markov chain model actually reduces to a standard mixture of time-inhomogeneous Markov chain models which is confirmed by comparing the log marginal likelihood of both models, being equal to -596.5 for a standard mixture model with $G = 2$ groups, see Table 2, and being equal to -598.5 for an ME model with $G = 2$ groups, see Table 4.

The marginal likelihood estimator for the ME model is based on importance sampling using the importance density (7) derived from auxiliary mixture sampling:

$$q_G(\theta) = \frac{1}{S} \sum_{s=1}^{S} p(\gamma_2|\mathbf{u}_2^{(s)}, \mathbf{r}_2^{(s)}, \mathbf{z}^{(s)}) \prod_{g=1}^{2} \prod_{j=1}^{J} p(\xi_{g,j\cdot}|\mathbf{z}^{(s)}, y),$$

where $p(\gamma_2|\mathbf{u}_2, \mathbf{r}_2, \mathbf{z})$ is conditionally Gaussian and $p(\xi_{g,j\cdot}|\mathbf{z}, y)$ is equal to the full conditional Dirichlet posterior of $\xi_{g,j\cdot}$ given in (9). Again, random permutation sampling is applied to ensure that the two equivalent modes are visited.

To sum up, this investigation shows that teenagers may, indeed, be clustered into two groups with different behaviour with respect to marijuana use, one being a never-user group, while the second group has a much higher risk to become a user. Preference for a standard mixture of Markov chain models over a mixture of experts Markov chain model based on gender shows that the two types of marijuana use cannot be associated with the gender of the teenager. Both male and female teenagers have about the same risk to belong to the second group. Unobserved factors, not the gender, are relevant for membership of a teenager to one group or the other.

Table 5: Covariates recorded for each respondent in the Irish Marketing Surveys poll.

Age	**Area**	**Gender**	**Government satisfaction**	**Marital status**	**Social class**
–	City	Housewife	No opinion	Married	AB
	Rural	Male	Not satisfied	Single	C1
	Town	Non-housewife	Satisfied	Widowed	C2
					DE
					F50+
					F50-

4.2 A mixture of experts model for ranked preference data

Mary McAleese served as the eighth President of Ireland from 1997 to 2011 and was elected 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 & 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 regarded as liberal. Gormley & Murphy (2008a,b, 2010a,b) 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 5.

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 modelling 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 component density $f_g(\cdot)$ in the mixture of experts model (1) must have an appropriate form. The Plackett-Luce model (Plackett, 1975; Gormley & Murphy, 2006) (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

Table 6: The model with smallest BIC within each type of mixture of experts model for ranked preference data applied to the 1997 Irish presidential election data

	BIC	G	Covariates
Simple mixture of experts model	8491	4	η_g: Government satisfaction, Age.
Standard mixture of experts regression model	8512	3	η_g: Government satisfaction, Age. p_g: Age
Mixture model	8513	3	–
Mixture of regressions model	8528	1	p_g: Government satisfaction

voter i's ballot is

$$p(y_i|p_g) = \frac{p_{g,c(i,1)}}{\sum_{s=1}^{M} p_{g,c(i,s)}} \cdot \frac{p_{g,c(i,2)}}{\sum_{s=2}^{M} p_{g,c(i,s)}} \cdots \frac{p_{g,c(i,m_i)}}{\sum_{s=m_i}^{M} p_{g,c(i,s)}},$$

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 standard mixture of experts regression model, the parameters of the component densities are modelled as a function of covariates. Here the support parameters are modelled as a logistic function of the covariates

$$\log\left[\frac{p_{gj}(x_i)}{p_{g1}(x_i)}\right] = \beta_{gj0} + \beta_{gj1}x_{i1} + \cdots + \beta_{gjq}x_{iq}$$

where $x_i = (x_{i1}, \ldots, x_{iq})$ is the set of q covariates associated with voter i and $\beta_{gj} = (\beta_{gj0}, \ldots, \beta_{gjq})^\top$ are unknown parameters for $j = 2, \ldots, M$. Note that for identifiability reasons candidate 1 is used as the baseline choice and $\beta_{g1} = (0, \ldots, 0)$ for all $g = 1, \ldots, G$.

In the standard mixture of experts regression model, the component weights are also modelled as a function of covariates, in a similar vein to the example used in Section 2.2, i.e.

$$\log\left[\frac{\eta_g(x_i)}{\eta_1(x_i)}\right] = \gamma_{g0} + \gamma_{g1}x_{i1} + \cdots + \gamma_{gq}x_{iq},$$

where voting bloc 1 is used as the baseline voting bloc.

The suite of four ME models in the ME framework (Figure 2) arise from modelling the component parameters and/or the component weights as functions of covariates, or as constant with respect to covariates. In this application, each model is fitted in a maximum likelihood framework using the EM algorithm; approximate standard errors for the model parameters are derived from the empirical information matrix (McLachlan & Peel, 2000) after the EM algorithm has converged. Model fitting details for each model are outlined in Gormley & Murphy (2008a,b, 2010a,b).

Each of the four ME models for rank preference data were fitted to the data from the electorate in the Irish presidential election poll. A range of models with $G = 1, \ldots, 5$ was considered and a forward step-wise selection method was employed to choose influential covariates. The Bayesian Information Criterion (BIC) (Kass & Raftery, 1995; Schwarz,

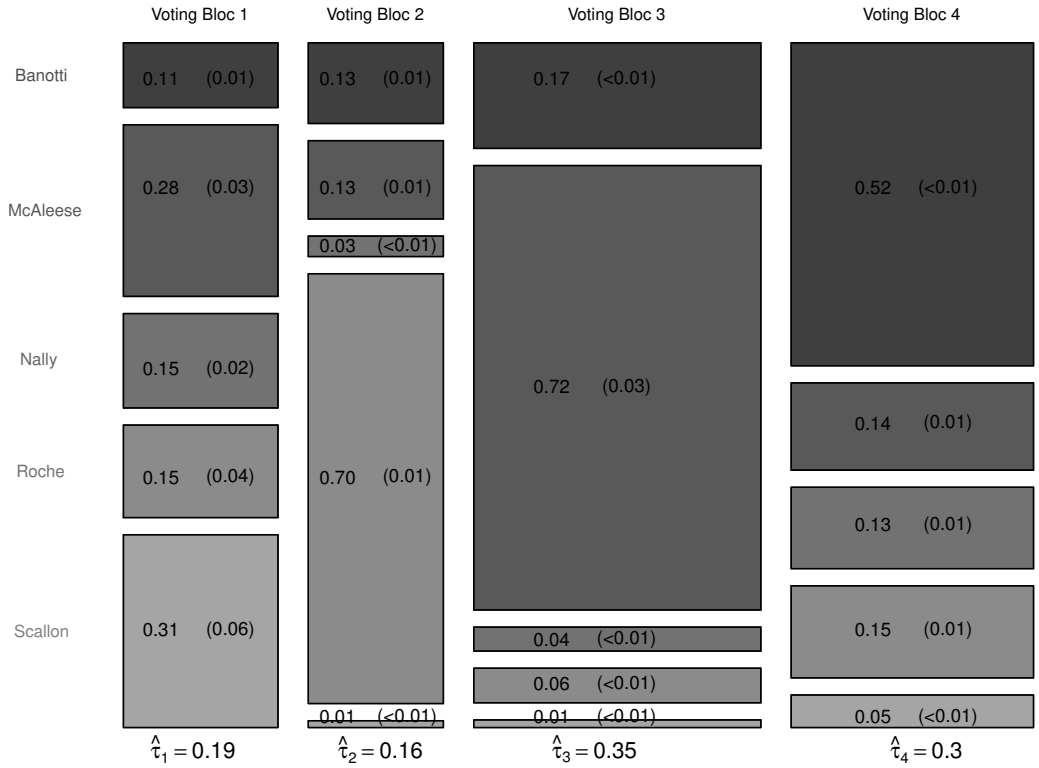


Figure 3: A mosaic plot representation of the parameters of the component densities of the simple mixture of experts model for rank preference data. The width of each block is proportional to the marginal probability of component membership ($\hat{\eta}_g = \sum_{i=1}^{n} \eta_g(x_i|\hat{\gamma})/n$). The blocks are divided in proportion to the Plackett-Luce support parameters which are detailed therein. Standard errors are provided in parentheses.

1978) was used to select the optimal model; this criterion is a penalized likelihood criterion which rewards model fit while penalizing non-parsimonious models, see also Chapter 7, Section 7.2.2 of this volume. Small BIC values indicate a preferable model. Table 6 details the optimal models for each type of ME model fitted.

Based on the BIC values, the optimal model is a simple mixture of experts model with four groups where "age" and "government satisfaction" are important covariates for determining group or "voting bloc" membership. Under this simple mixture of experts 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 3 and in Table 7.

The support parameter estimates illustrated in Figure 3 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 bloc 2 has large support for the liberal candidate Adi Roche. Voting bloc 3 is the

Table 7: Odds ratios ($\exp(\gamma_g)/\exp(\gamma_1)$) for the component weight parameters in the simple ME model for rank preference data (95% confidence intervals are given in parentheses). The covariates 'age' and 'government satisfaction level' were selected as influential

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

largest voting bloc in terms of marginal component weights 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 & Murphy (2008b).

Table 7 details the odds ratios computed from the component weight parameters $\gamma = \{\gamma_2, \gamma_3, \gamma_4\}$. In the model, voting bloc 1 (the conservative voting bloc) is the baseline voting bloc and $\gamma_1 = (0, \ldots, 0)^\top$. 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 characterises 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 ($\exp(\gamma_{21}) = 0.01$). Also, voters with some interest in government are more likely to belong to voting bloc 3 ($\exp(\gamma_{32}) = 3.81$ and $\exp(\gamma_{33}) = 3.12$), the bloc favouring 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 component weight 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 ($\exp(\gamma_{42}) = 3.50$). This is again intuitive as voting bloc 4 favours Mary Banotti who was backed by the main government opposition party, while voting bloc 1 favours the government backed Mary McAleese. Further interpretation of the component weight parameters are given in Gormley & Murphy (2008b).

4.3 A mixture of experts latent position cluster model

The latent position cluster model (Handcock *et al.* , 2007) develops the idea of the latent social space (Hoff *et al.* , 2002) by extending it 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) assumes that the latent positions were normally distributed. Thus, the latent position cluster model offers a more flexible version of the latent space model for modelling 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 modelled 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 modelled as a function of their covariates lies within

Table 8: 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 analyses.

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

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 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 $x_i = (x_{i1}, \ldots, x_{iq})$ associated with actor i are assumed to be available, where q denotes the number of observed covariates.

Each actor i is assumed to have a location $w_i = (w_{i1}, \ldots, w_{iD})$ 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 $x_{i,j} = (x_{ij1}, \ldots, x_{ijq})$ denote an q vector of dyadic specific covariates where $x_{ijk} = d(x_{ik}, x_{jk})$ 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

$$p(y|\mathbf{w}, \mathbf{x}, \beta) = \prod_{i=1}^{n} \prod_{j \neq i} p(y_{i,j}|w_i, w_j, x_{i,j}, \beta)$$

where $\mathbf{w}$ 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 modelled using a logistic regression model where both dyadic specific covariates and Euclidean distance in the latent space are covariates:

$$\log\left[\frac{\mathrm{P}(y_{i,j}=1)}{\mathrm{P}(y_{i,j}=0)}\right] = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_q x_{ijq} - ||w_i - w_j||.$$

To account for clustering of actor locations in the latent space, it is assumed that the latent locations w_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:

$$w_i \sim \sum_{g=1}^{G} \eta_g(x_i|\gamma)\phi(w_i|\mu_g, \sigma_g^2 I)$$

where

$$\eta_g(x_i|\gamma) = \frac{\exp(\gamma_{g0} + \gamma_{g1}x_{i1} + \cdots + \gamma_{gq}x_{iq})}{\sum_{g'=1}^{G} \exp(\gamma_{g'0} + \gamma_{g'1}x_{i1} + \cdots + \gamma_{g'q}x_{iq})}$$

and $\gamma_1 = (0, \ldots, 0)^\top$. 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.

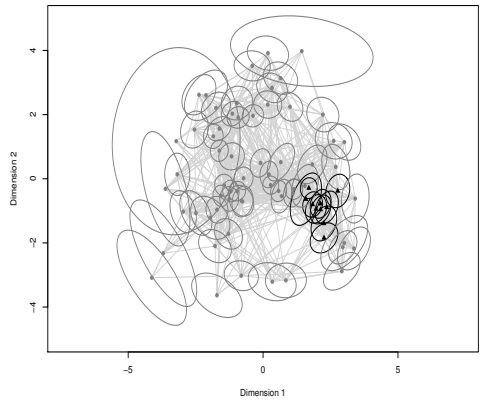
The mixture of experts latent position cluster model can be fitted within the Bayesian paradigm; as outlined in Section 3.2 a Metropolis-within-Gibbs sampler can be employed to draw samples from the posterior distribution of interest. Model issues such as likelihood invariance to distance preserving transformations of the latent space and label switching must be considered during the model fitting process – an approach to dealing with such model identifiability and full model fitting details are available in Gormley & Murphy (2010b). In this application, model choice concerns not only the number G of clusters, but also the dimension D of the latent space.

An example of the mixture of experts latent position cluster model methodology is provided here 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 in the firm are also included and are detailed in Table 8. Interest lies in identifying social processes within the firm such as knowledge sharing and organisational structures, and examining the potential influence of covariates on such processes.

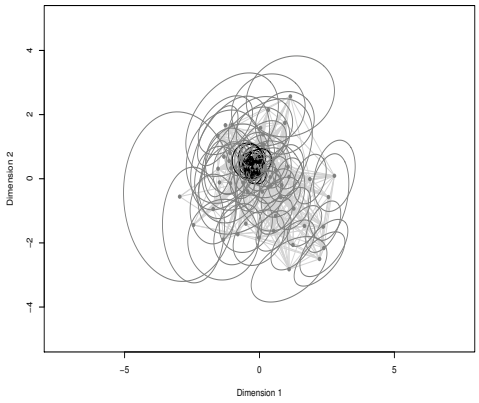
Under the ME model framework outlined in Section 2.3, a suite of four mixtures of experts latent position cluster models is available. This suite of models was fitted to the advice network; data in this network detail links between lawyers who sought basic professional advice from each other over the previous twelve months. Gormley & Murphy (2010b) explore the coworkers network data set and the friendship network data set using similar methodology. Figure 4 illustrates the resulting latent space locations of the lawyers under each fitted model with $(G, D) = (2, 2)$. These 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 9 details the resulting regression parameter estimates and their associated uncertainty for the four fitted models.

The models are compared through the AICM, the posterior simulation-based analogue of Akaike's Information Criterion (AIC) (Akaike, 1973; Raftery *et al.* , 2007). In this implementation the optimal model is that with the highest AICM and is the model with covariates in the link probabilities and in the component weights. The results of the analysis show some interesting patterns. The coefficients of the covariates in the link probabilities are very similar in the models (b) and (d) in Table 9. 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 are correlated with seniority and with each other, so their marginal effects are more difficult to interpret.

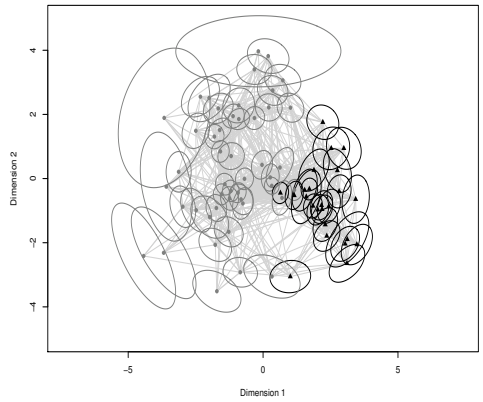
Importantly, the latent positions are very similar in models (a) and (c) which do not have covariates in the link probabilities and models (b) and (d) which do have covariates



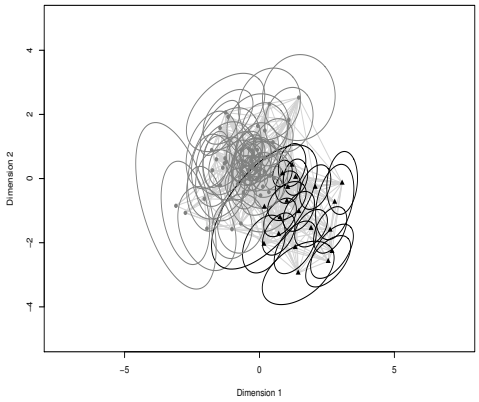
(a) Latent position cluster model.



(b) Mixture of regressions latent position cluster model.



(c) Simple latent position cluster model.



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

Figure 4: 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 shade and symbol. Observed links between lawyers are also illustrated.

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 those that do not have such covariates. When the covariates are in the link probabilities, the latent space is modelling the network structure that could not be explained by the link covariates, whereas in the other case the latent space is modelling much 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 modelling what could not 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 in this application the signs of the coefficients are identical because the cluster memberships shown for these models in Figure 4(c) and Figure 4(d) are similar; this phenomenon

Table 9: Posterior mean parameter estimates for the four mixtures of experts models fitted to the lawyers advice data as detailed in Figure 4. 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 8

	Model (a)	Model (b)	Model (c)	Model (d)
Link Probabilities				
Intercept	1.26 (0.10)	-2.87 (0.17)	1.23 (0.10)	-2.65 (0.17)
Age		-0.02 (0.004)		-0.02 (0.004)
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)
Seniority		0.89 (0.11)		0.81 (0.11)
Years		-0.04 (0.005)		-0.04 (0.005)
Cluster Memberships				
Intercept	-1.05 (1.75)	0.94 (0.79)	-0.62 (1.23)	1.27 (1.29)
Age			-0.09 (0.04)	-0.14 (0.06)
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)
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

does not hold generally (see Gormley & 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 the lawyer advice network. 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 latent space is being used to find interesting or anomalous structure in the network that cannot be explained by the covariates, then one should consider allowing the covariates enter the cluster membership probabilities.

4.4 Software

As demonstrated in this section, the approach to fitting an ME model depends on the application setting and on the form of the ME model itself. Therefore, a single software capable of fitting any ME model is not currently available.

In R (R Core Team, 2018), the `MEclustnet` package (Gormley & Murphy, 2018) fits the mixture of experts latent position cluster model detailed in Section 4.3. The `flexmix` package (Grün & Leisch, 2008b) has model fitting capabilities for a range of mixture of regression models, which include covariates (or concomitant variables), as does the `mixreg` package (Turner, 2014). Additionally, `mixtools` (Benaglia *et al.* , 2009) facilitates fitting of a $G = 2$ mixture of regressions model in which the component weights are modelled as an inverse logit function of the covariates. The cluster weighted models which are closely related to ME models can be fitted using the `flexCWM` package (Mazza *et al.* , 2017). All packages are freely available through the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org`.

In MATLAB, the `bayesf` package (Frühwirth-Schnatter, 2018) allows to estimate a broad range of mixture models using either finite mixtures, mixtures of experts or Markov switching models as a model for the hidden group indicators $\mathbf{z}$.

In terms of other softwares, the `FMM` procedure in SAS also facilitates ME model fitting, and stand alone softwares such as `Latent GOLD` (Vermunt & Magidson, 2005) and `Mplus` (Muthén & Muthén, 2011) fit closely related latent class models.

5 Identifiability of Mixtures of Experts Models

For a finite mixture distribution one has to distinguish three types of non-identifiability (Frühwirth-Schnatter, 2006, Section 1.3): invariance to relabelling the components of the mixture distribution (the so-called label switching problem), non-identifiability due to potential overfitting and generic non-identifiability which occurs only for certain classes of mixture distributions.

Consider a standard mixture distribution with G components with non-zero weights $\eta_1, \ldots, \eta_G$ generated by distinct parameters $\theta_1, \ldots, \theta_G$. Assume that for all possible realisations y from this mixture distribution the identity

$$\sum_{g=1}^{G} \eta_g f_g(y|\theta_g) = \sum_{g=1}^{G^\star} \eta_g^\star f_g(y|\theta_g^\star)$$

holds where the right-hand side is a mixture distribution from the same family with $G^\star$ components with non-zero weights $\eta_1^\star, \ldots, \eta_{G^\star}^\star$ generated by distinct parameters $\theta_1^\star, \ldots, \theta_{G^\star}^\star$. Then generic identifiability implies that $G^\star = G$ and the two mixtures' parameters $\theta = (\eta_1, \ldots, \eta_G, \theta_1, \ldots, \theta_G)$ and $\theta^\star = (\eta_1^\star, \ldots, \eta_G^\star, \theta_1^\star, \ldots, \theta_G^\star)$ are identical up to relabelling the component indices. Common finite mixture distributions such as Gaussian and Poisson mixtures are generically identified, see Teicher (1963), Yakowitz & Spragins (1968), and Chandra (1977) for a detailed discussion.

Discrete mixtures often suffer from generic non-identifiability for certain parameter configurations, well-known examples being mixtures of binomial distributions (see Section 5.1) and mixtures of multinomial distributions (Grün & Leisch, 2008c). Somewhat unexpectedly, mixtures of regression models suffer from generic non-identifiability (Hennig, 2000; Grün & Leisch, 2008a), as will be discussed in more detail in Section 5.2. Little is known about generic identifiability of mixtures of experts models and some results are presented in Section 5.3. However, ensuring generic identifiability for general ME models remains a challenging issue.

Identifiability problems for mixture with nonparametric components are discussed in Chapter 14 of this volume.

5.1 Identifiability for mixtures of binomials

For binomial mixtures the component densities arise from $\mathcal{B}(N, \pi)$-distributions, where N is commonly assumed to be known, whereas π is heterogeneous across the components:

$$Y \sim \eta_1 \mathcal{B}(N, \pi_1) + \cdots + \eta_G \mathcal{B}(N, \pi_G). \tag{10}$$

The probability mass function (pmf) of this mixture takes on $N+1$ different support points:

$$p(y|\theta) = \mathrm{P}(Y = y|\theta) = \sum_{g=1}^{G} \eta_g \begin{pmatrix} N \\ y \end{pmatrix} \pi_g^y (1-\pi_g)^{N-y}, \quad y = 0, 1, \ldots, N, \tag{11}$$

with $2G-1$ independent parameters $\theta = (\pi_1, \ldots, \pi_G, \eta_1, \ldots, \eta_G)$, with $\eta_G = 1 - \sum_{g=1}^{G-1} \eta_g$.

Given data $y = (y_1, \ldots, y_n)$ from mixture (10), the only information available to estimate θ are N (among the $N+1$ observed) relative frequencies $h_n(Y = y)$ $(y = 0, 1, \ldots, N)$. As $n \to \infty$ (while N is fixed), $h_n(Y = y)$ converges to $\mathrm{P}(Y = y|\theta)$ by the law of large numbers, but the number of support points remains fixed. Hence, the data provide only N statistics, given by the relative frequencies, to estimate $2G-1$ parameters. Simple counting yields the following necessary condition for identifiability for a binomial mixture, which has been shown by Teicher (1961) to be also sufficient:

$$2G - 1 \leq N \qquad \Leftrightarrow \qquad G \leq (N+1)/2. \tag{12}$$

Consider, for illustration, a mixture of two binomial distributions,

$$Y \sim \eta \times \mathcal{B}(N, \pi_1) + (1-\eta) \times \mathcal{B}(N, \pi_2), \tag{13}$$

with three unknown parameters $\theta = (\eta, \pi_1, \pi_2)$ and assume that the population indeed contains two different groups, i.e. $\pi_1 \neq \pi_2$ and $\eta > 0$. Assuming $N = 2$ obviously violates condition (12). Lack of identification can be verified directly from the pmf which is different from zero only for the three outcomes $y \in \{0, 1, 2\}$:

$$\begin{aligned} \mathrm{P}(Y = 0|\theta) &= \eta(1-\pi_1)^2 + (1-\eta)(1-\pi_2)^2, \\ \mathrm{P}(Y = 1|\theta) &= 2\eta\pi_1(1-\pi_1) + 2(1-\eta)\pi_2(1-\pi_2), \\ \mathrm{P}(Y = 2|\theta) &= \eta\pi_1^2 + (1-\eta)\pi_2^2. \end{aligned} \tag{14}$$

Since $\sum_y \mathrm{P}(Y = y|\theta) = 1$, only two linearly independent equations remain to identify the three parameters (η, π_1, π_2). Hence parameters $\theta = (\pi_1, \pi_2, \eta) \neq \theta^\star = (\pi_1^\star, \pi_2^\star, \eta^\star)$ fulfilling equations (14) exist which imply the same distribution for Y, i.e.: $\mathrm{P}(Y = y|\theta) = \mathrm{P}(Y = y|\theta^\star)$, $\forall y = 0, 1, 2$, but are not related to each other by simple relabelling of the component indices.

Such generic non-identifiability severely impacts statistical estimation of the mixture parameters θ from observations $y = (y_1, \ldots, y_n)$, even if G is known, and goes far beyond label switching. Assume, for illustration, that y is the realisation of a random sample $(Y_1, \ldots, Y_n)$ from the two-component binomial mixture (13) with $N = 2$ and true parameter $\theta^{\text{true}} = (\pi_1^{\text{true}}, \pi_2^{\text{true}}, \eta^{\text{true}})$ and consider the corresponding observed-data likelihood $p(y|\theta) = \prod_{i=1}^n \mathrm{P}(Y_i = y_i|\theta)$. Generic non-identifiability of the underlying mixture distribution implies that the observed-data likelihood is the same for any pair $\theta \neq \theta^\star$ of distinct parameters satisfying (14), for any possible sample y in the sampling space $\mathcal{Y} = \{0, 1, 2\}^n$, i.e.: $p(y|\theta) = p(y|\theta^\star), \forall y \in \mathcal{Y}$. Since this holds for arbitrary sample size $n = 1, 2, \ldots$, the true parameter θ^{true} cannot be recovered, even if $n \to \infty$, and both maximum likelihood estimation as well Bayesian inference suffer from non-identifiability problems for such a mixture.

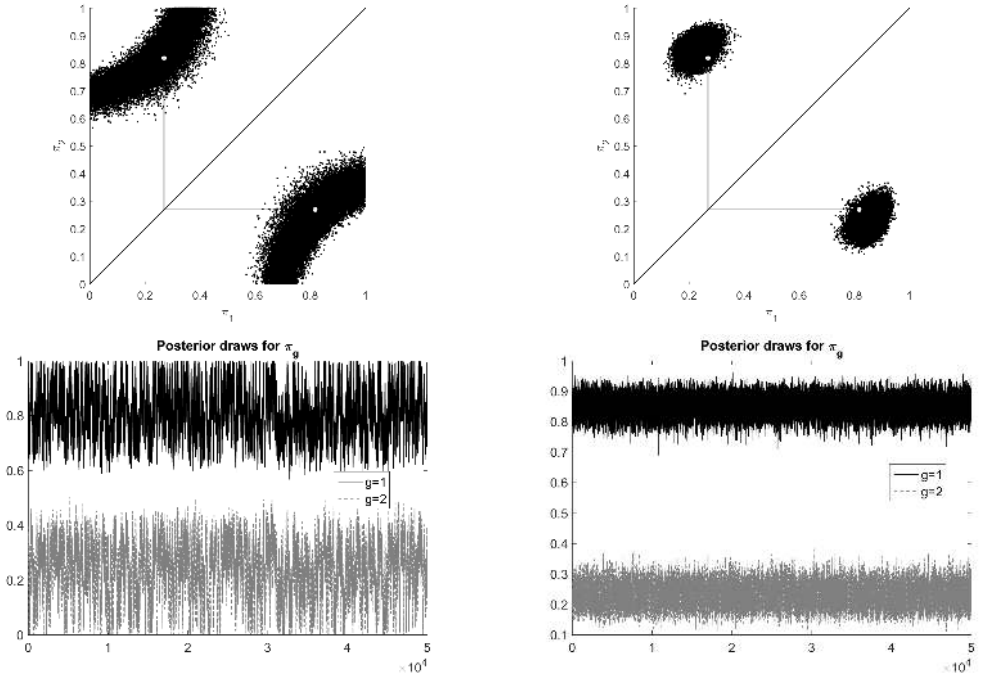


Figure 5: MCMC inference for data simulated from a mixture of two binomial distributions with $N = 2$ (left-hand side) and $N = 5$ (right-hand side). Top: scatter plot of π_1 versus π_2 (true values indicated by a circle). Bottom: posterior draws of the group-specific probabilities π_1 and π_2 after resolving label switching in the scatter plot of π_1 versus π_2 through k-means clustering.

This example motivates the following more formal definition of generic non-identifiability. For a given θ, any subset $U(\theta)$ of the parameter space Θ of a mixture model, defined as $U(\theta) = \{\theta^\star \in \Theta : p(y|\theta^\star) = p(y|\theta), \forall y \in \mathcal{Y}\}$, is called a non-identifiability set, if it contains at least one point $\theta^\star$ which is not related to θ by simple relabelling of the component indices. Let θ^{true} be the true parameter value of a mixture model with G distinct parameters (i.e. $\theta_g \neq \theta_{g'}$, for $g \neq g'$). If $U(\theta^{\text{true}})$ is a non-identifiability set in the sense defined above, then θ^{true} cannot be recovered from data, even as n goes to infinity.

Such generic non-identifiability has important implications for practical mixture analysis. For finite n, the observed-data likelihood function $p(y|\theta)$ has a ridge close to $U(\theta^{\text{true}})$ instead of $G!$ isolated modes and no unique maximum, leading to inconsistent estimates of θ^{true}. In a Bayesian framework, this leads to a posterior distribution that does not concentrate around $G!$ isolated, equivalent modes as n increases, as for identifiable models (see Chapter 4, Section 4.3). Rather, the posterior concentrates over the entire non-identifiability set $U(\theta^{\text{true}})$ which has a complex geometry and can be represented as the union of $G!$ symmetric subspaces, see e.g. Figure 5 for a binomial mixture with $G = 2$ and $N = 2$. The prior $p(\theta)$ provides information beyond the data and might influence how the posterior concentrates on each of these $G!$ subspaces $U(\theta^{\text{true}})$, in particular, if the prior $p(\theta)$ is not constant over $U(\theta)$.

While generic non-identifiability has important practical implications for mixture analysis, it is rarely as easily diagnosed as for mixtures of binomial distributions and can easily go unnoticed for more complex mixture models, in particular for maximum likelihood estimation, whereas MCMC based Bayesian inference often provides indications of potential identifiability problems, as the following example demonstrates.

MCMC inference for an example: a mixture of binomial distributions

For further illustration, we perform MCMC inference (based on 10,000 draws after a burn-in of 5,000 iterations) for two data sets simulated from a mixture of two binomial distributions with logit $\pi_1 = -1$ and logit $\pi_2 = 1.5$ using random permutation sampling as explained in Chapter 5, Section 5.2. We assume that $G = 2$ and N is known, whereas all other parameters in mixture (13) are unknown. Bayesian inference is based on the following priors: $\pi_g \sim \mathcal{U}(0,1)$, and $(\eta_1, \eta_2) \sim \mathcal{D}(1,1)$. The two data sets were generated with, respectively, $N = 2, n = 250$ and $N = 5, n = 100$, implying the same total number $n \times N = 500$ of experiments.

MCMC inference is summarised in Figure 5, showing scatter plot of π_1 versus π_2 for both values of N. For $N = 5$, the mixture is generically identified and the posterior draws concentrate around two symmetric modes, centered at the true values $(0.269, 0.818)$ and $(0.818, 0.269)$. Non-identifiability due to label switching is resolved by applying k-means clustering to the posterior draws, see the lower part of Figure 5 showing identified posterior draws of the group-specific success probabilities π_1 and π_2.

For $N = 2$, a similar scatter plot of π_1 versus π_2 clearly indicates severe identifiability issues, showing that the posterior draws arise from two symmetric unidentifiability sets, rather than concentrating around two symmetric modes centered at the true values. When we apply k-means clustering to resolve label switching, we obtain the posterior draws of the success probabilities π_1 and π_2 shown in the lower part of Figure 5, also indicating problems with identifying π_1 and π_2 from the data for $N = 2$.

5.2 Identifiability for mixtures of regression models

Consider a mixture of G regression models for $i = 1, \ldots, n$ outcomes y_i, arising from G different groups,

$$y_i|\tilde{x}_i \sim \sum_{g=1}^{G} \eta_g \phi(y|\mu_{i,g}(\tilde{x}_i), \sigma_g^2) \tag{15}$$

where for each $g = 1, \ldots, G$, the group-specific mean $\mu_{i,g}(\tilde{x}_i) = \tilde{x}_i \beta_g$ depends on a group-specific regression parameter β_g and on the $(1 \times (q+1))$-dimensional row vector $\tilde{x}_i$ containing the q covariates x_i and a constant. For a fixed design point $x = \tilde{x}_i$, (15) is a standard finite Gaussian mixture distribution and as such generically identified. Hence, if the identity

$$\sum_{g=1}^{G} \eta_g \phi(y|\mu_{i,g}(x), \sigma_g^2) = \sum_{g=1}^{G} \eta_g^\star \phi(y|\mu_{i,g}^\star(x), \sigma_g^{2,\star}), \tag{16}$$

holds, then the two mixtures are related to each other by relabelling, i.e. $\mu_{i,g}^\star(x) = \mu_{i,\sigma_x(g)}(x) = x\beta_{\sigma_x(g)}$, $\sigma_g^{2,\star} = \sigma_{\sigma_x(g)}^2$, and $\eta_g^\star = \eta_{\sigma_x(g)}$ for $g = 1, \ldots, G$, for some permutation $\sigma_x \in \mathfrak{S}(G)$, where $\mathfrak{S}(G)$ denotes the set of the $G!$ permutations of $\{1, \ldots, G\}$. Note that σ_x depends on the covariate x and that there is no guarantee that σ_x is identical across different values of x which can cause *intra-component label switching*. One such example is displayed on the left-hand side of Figure 6.

Nevertheless, assume for the moment that $\sigma_x \equiv \sigma_\star$ is the same for all possible covariates x. Then (16) implies $\tilde{x}_i \beta_g^\star = \tilde{x}_i \beta_{\sigma_\star(g)}$ for all $i = 1, \ldots, n$ and $X\beta_g^\star = X\beta_{\sigma_\star(g)}$ where the rows of the matrix X are equal to $\tilde{x}_1, \ldots, \tilde{x}_n$. If the usual condition in regression modelling is satisfied that $X^\top X$ has full rank, then it follows immediately that the regression coefficients are determined up to relabelling: $\beta_g^\star = \beta_{\sigma_\star(g)}$.

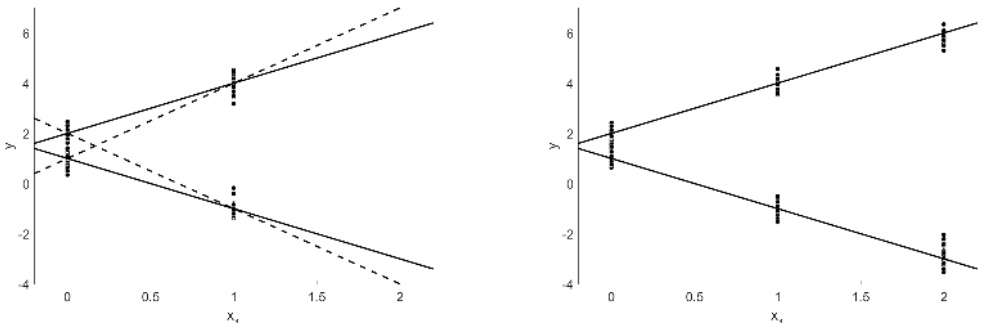


Figure 6: Data simulated from a mixture of two regression lines under *Design 1* (left-hand side) and *Design 2* (right-hand side). The full lines indicate the true underlying model used to generate 100 data points (black dots). For the unidentified *Design 1*, a second solution exists which is indicated by the dashed lines.

Hence, generic identifiability for a mixture of regressions model can be verified through sufficient conditions guaranteeing that σ_x is indeed identical across all values of x. Mathematically, one such condition is the assumption that either the error variances $\sigma_1^2, \ldots, \sigma_G^2$ or the weights $\eta_1, \ldots, \eta_G$ satisfy a strict order constraint. However, in practice such constraints are rarely fulfilled and forcing an order constraint on one coefficient does not necessarily prevent label switching for the other coefficients in Bayesian posterior sampling, see e.g. Frühwirth-Schnatter (2006, Section 2.4).

Hence, several papers focused on conditions for generic identifiability through the regression part of the model (Hennig, 2000; Grün & Leisch, 2008c,a). Assume that the covariates $\tilde{x}_i$ take p different values in a design space $\{x_1, \ldots, x_p\}$ for the observed outcome y_i, for $i = 1, \ldots, n$. Identifiability through the regression part requires enough variability in the design space and is guaranteed under so-called *coverage conditions*. These conditions require that the number of clusters G is exceeded by the minimum number of distinct q-dimensional hyperplanes needed to cover the covariates (excluding the constant). For $q = 1$, for instance, the coverage condition is satisfied, if the number of design points p (i.e. the number of distinct values of the univariate covariate) is larger than the number of clusters G. These identifiability conditions go far beyond the usual condition that $X^\top X$ has full rank and are often violated for regression models with too few design points, a common example being regression models with 0/1 dummy variables as covariates which are identifiable for $G = 1$, but not for $G > 1$, as the following examples with $q = 1$ demonstrate.

For illustration, we consider the following special case of the mixture of regressions model (15) investigated in Grün & Leisch (2008a, Section 3.1):

$$y_i \sim 0.5\phi(y|\mu_{i,1}(\tilde{x}_i), 0.1) + 0.5\phi(y|\mu_{i,2}(\tilde{x}_i), 0.1), \tag{17}$$

with covariate vector $\tilde{x}_i = (1 \; d_i)$ and group-specific regression parameters $\beta_1 = (2 \;\; 2)^\top$ and $\beta_2 = (1 \;\; -2)^\top$. We consider two different regression designs, *Design 1* where d_i is a 0/1 dummy variable capturing the effect of gender (with female as baseline) and *Design 2* where d_i captures a time effect over 3 periods (with $t = 0$ serving as baseline):

$$\begin{aligned}
&\textit{Design 1}\text{: } x_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}, \; x_2 = \begin{pmatrix} 1 & 1 \end{pmatrix}, \\
&\textit{Design 2}\text{: } x_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}, \; x_2 = \begin{pmatrix} 1 & 1 \end{pmatrix}, \; x_3 = \begin{pmatrix} 1 & 2 \end{pmatrix}.
\end{aligned}$$

In the following, it is verified that mixture (17) is generically identified under *Design 2* (which contains three design points), but generically unidentified under *Design 1* (which contains only two design points).

We first consider *Design 1*. According to (16), $\mu_{j,1}$ and $\mu_{j,2}$ are identified for $j = 1$ and $j = 2$ up to label switching arising from two permutations σ_1 and σ_2, where we may assume without loss of generality that σ_1 is equal to the identity:

$$\begin{aligned} x_1\beta_1 &= \mu_{1,1}, & x_1\beta_2 &= \mu_{1,2}, \\ x_2\beta_1 &= \mu_{2,\sigma_2(1)}, & x_2\beta_2 &= \mu_{2,\sigma_2(2)}. \end{aligned} \tag{18}$$

If σ_2 is identical to σ_1, then the original values β_1 and β_2 are recovered through:

$$\beta_1 = X_{1,2}^{-1} \begin{pmatrix} \mu_{1,1} \\ \mu_{2,1} \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \beta_2 = X_{1,2}^{-1} \begin{pmatrix} \mu_{1,2} \\ \mu_{2,2} \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \tag{19}$$

since the design matrix

$$X_{1,2} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

is invertible. However, as mentioned above, σ_2 need not be identical to σ_1, in which case $\sigma_2(1) = 2, \sigma_2(2) = 1$, and a second solution emerges:

$$\beta_1^\star = X_{1,2}^{-1} \begin{pmatrix} \mu_{1,1} \\ \mu_{2,2} \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \quad \beta_2^\star = X_{1,2}^{-1} \begin{pmatrix} \mu_{1,2} \\ \mu_{2,1} \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

Evidently, the group-specific slopes of this second solution are different from the original ones and the un-identifiability set $U(\theta^{\text{true}})$ contains two points. The two possible solutions are depicted in the left-hand side of Figure 6 which also shows a balanced sample of $n = 100$ observations simulated from mixture (17) under *Design 1*.

For *Design 2*, the first two design points are as before and a third point is added with $\mu_{3,1}$ and $\mu_{3,2}$ being identified up to label switching according to a permutation σ_3:

$$\begin{aligned} x_3\beta_1 &= \mu_{3,\sigma_3(1)}, & x_3\beta_2 &= \mu_{3,\sigma_3(2)}. \end{aligned} \tag{20}$$

As only two different permutations exist for $G = 2$, at least two of the three permutations σ_1, σ_2, and σ_3 in (18) and (20) have to be identical (assuming again without loss of generality that σ_1 is equal to the identity). Assume, for example, that $\sigma_1 = \sigma_2$. Then the true parameters β_1 and β_2 are recovered from $(\mu_{j,1}, \mu_{j,2}), j = 1, 2$, as in (19) and can be used to uniquely predict $\mu_{3,1} = x_3\beta_1$ and $\mu_{3,2} = x_3\beta_2$ in both groups. Comparing these predictions with (20), it is clear that $\sigma_3(1) = 1$ and $\sigma_3(2) = 2$, hence $\sigma_1 = \sigma_2 = \sigma_3$. A similar proof can be performed for any pair of identical permutations $\sigma_j = \sigma_l, j \neq l$, as long as the matrix $X_{j,l}^\top = (x_j^\top \; x_l^\top)$ is invertible and generic identifiability of *Design 2* follows.

The only possible solution under *Design 2* is depicted in the right-hand side of Figure 6 which also shows a balanced sample of $n = 100$ observations simulated from mixture (17) under this design.

MCMC inference for an example: a mixture of regressions model

For further illustration, we perform MCMC inference (based on 10,000 draws after a burn-in of 5,000 iterations) for both data sets shown in Figure 6 using random permutation sampling as explained in Chapter 5, Section 5.2. We assume that $G = 2$ is known, whereas all other parameters in mixture (15) are unknown. Bayesian inference is based on the priors $(\eta_1, \eta_2) \sim \mathcal{D}(4, 4)$ and $\beta_g \sim \mathcal{N}(0, 100 \times I), \sigma_g^2 \sim \mathcal{IG}(2.5, 1.25 s_y^2)$ for $g = 1, 2$, where

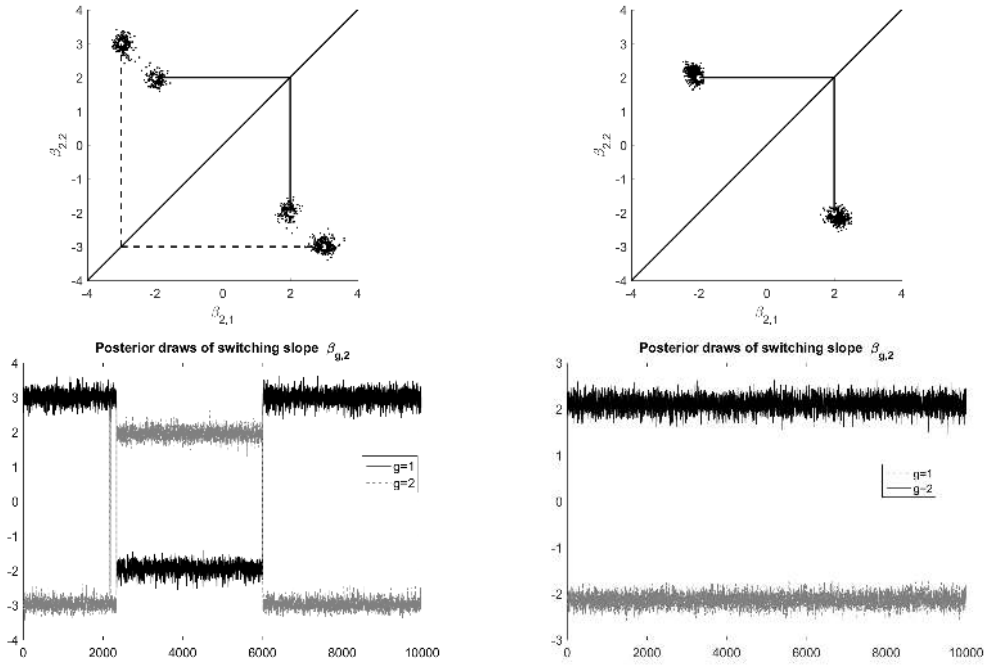


Figure 7: MCMC inference for data simulated from a mixture of two regression models under the generically un-identified *Design 1* (left-hand side) and under the generically identified *Design 2* (right-hand side). Top: scatter plot of the group-specific slopes $\beta_{1,2}$ versus $\beta_{2,2}$. Bottom: posterior draws of the group-specific slopes $\beta_{1,2}$ and $\beta_{2,2}$ after resolving label switching through k-means clustering in the posterior draws.

s_y^2 is the data variance of $(y_1, \ldots, y_n)$. The upper part of Figure 7 shows scatter plots of the group-specific slopes $\beta_{1,2}$ versus $\beta_{2,2}$ for both designs which are symmetric due to label switching.

As expected from Chapter 4, Section 4.3, the posterior draws shown in the upper right-hand part for the generically identified *Design 2* concentrate around two symmetric modes corresponding to the true values $(2, -2)$ and $(-2, 2)$. Label switching is easily resolved by applying k-means clustering to the posterior draws, see the lower right-hand part of Figure 7 showing identified posterior draws of the group-specific slopes $\beta_{1,2}$ and $\beta_{2,2}$ for this design.

For *Design 1*, a similar scatter plot of $\beta_{1,2}$ versus $\beta_{2,2}$ in the upper left-hand part clearly indicates severe identifiability issues. The posterior draws concentrate around four rather than two modes, with two of them being the symmetric modes corresponding to the true values $(2, -2)$ and $(-2, 2)$. The other two symmetric modes correspond to the second solution $(-3, 3)$ and $(3, -3)$, resulting from generic non-identifiability. When we apply k-means clustering to these posterior draws to resolve label switching, we obtain the posterior draws of the group-specific slopes $\beta_{1,2}$ and $\beta_{2,2}$ in the lower left-hand part of Figure 7, showing *intra-component label switching* and indicating identifiability problems for this design. Since $\beta_{g,2}$ switches sign between the two solutions in both groups, it is not possible to recover that "gender" has a strong positive effect on the outcome in one group and a strong negative effect in the other group.

5.3 Identifiability for mixtures of experts models

Hennig (2000) considers mixtures of regression models where the component sizes can arbitrarily depend on covariates and establishes identifiability results in the case that the joint observations of covariates and dependent variable are assumed to be iid and gives sufficient

identifiability conditions for this model. For such a model, the covariates are not assumed fixed or to occur for a fixed design, but random with a specific distribution. As opposed to this, the ME model is defined conditional on the covariates without specific assumptions concerning their distribution. We will discuss identification for this case.

Consider, as a first example, a simple mixture of experts model of G univariate Gaussian distributions for $i = 1, \ldots, n$ outcomes y_i arising from G different groups,

$$y_i|\tilde{x}_i \sim \sum_{g=1}^{G} \eta_g(\tilde{x}_i)\phi(y|\mu_g, \sigma_g^2) \tag{21}$$

where the group weights η_g depend on a covariate $\tilde{x}_i$ with group-specific regression parameters, i.e.:

$$\log\left[\frac{\eta_g(\tilde{x}_i)}{\eta_{g_0}(\tilde{x}_i)}\right] = \tilde{x}_i\gamma_g, \quad g = 1, \ldots, G, \tag{22}$$

with baseline g_0, where $\gamma_{g_0} = 0$. Assume that the component densities differ, i.e. $\theta_g \neq \theta_g'$, for $g \neq g'$, where $\theta_g = (\mu_g, \sigma_g^2)$.

For each fixed design point $x = \tilde{x}_i$, (21) is a standard finite Gaussian mixture and therefore generically identified. Therefore, if the identity

$$\sum_{g=1}^{G} \eta_g(x)\phi(y|\mu_g, \sigma_g^2) = \sum_{g=1}^{G} \eta_g^\star(x)\phi(y|\mu_g^\star, \sigma_g^{2,\star}), \tag{23}$$

holds, then the two mixtures are related to each other by relabelling, i.e. $\mu_g^\star = \mu_{\sigma_x(g)}, \sigma_g^{2,\star} = \sigma_{\sigma_x(g)}^2$, and $\eta_g^\star(x) = \eta_{\sigma_x(g)}(x)$ for $g = 1, \ldots, G$, for some permutation $\sigma_x \in \mathfrak{S}(G)$. As opposed to mixtures of regression models, one can show that $\sigma_x \equiv \sigma_\star$ for all covariate values x.

Assume that $\sigma_{x_i} \neq \sigma_{x_j}$ for two covariates $\tilde{x}_i \neq \tilde{x}_j$ and assume, without loss of generality, that σ_{x_i} is equal to the identity. Consider first the case of $G = 2$. Then (23) implies for $x = \tilde{x}_i$:

$$\theta_1^\star = \begin{pmatrix} \mu_1^\star \\ \sigma_1^{2,\star} \end{pmatrix} = \theta_1, \quad \theta_2^\star = \begin{pmatrix} \mu_2^\star \\ \sigma_2^{2,\star} \end{pmatrix} = \theta_2,$$

whereas for $x = \tilde{x}_j$:

$$\theta_1^\star = \begin{pmatrix} \mu_2^\star \\ \sigma_2^{2,\star} \end{pmatrix} = \theta_2, \quad \theta_2^\star = \begin{pmatrix} \mu_1^\star \\ \sigma_1^{2,\star} \end{pmatrix} = \theta_1,$$

contradicting the assumptions that $\theta_1 \neq \theta_2$. A similar proof is possible for $G > 2$, where the assumption $\sigma_{x_i} \neq \sigma_{x_j}$ (assuming again that σ_1 is equal to the identity) implies for $x = \tilde{x}_i$ that $\theta_g^\star = \theta_g$ for all components $g = 1, \ldots, G$, whereas for $x = \tilde{x}_j$ at least one component g_i exists with $\theta_{g_i}^\star = \theta_{g_j}$, where $g_j = \sigma_{x_j}(g_i) \neq g_i$. Hence, $\theta_{g_i} = \theta_{g_j}$, which contradicts the assumptions that $\theta_{g_i} \neq \theta_{g_j}$. This implies that $\sigma_x \equiv \sigma_\star$ for all covariate values x.

Therefore, the weight distribution $\eta_1(x), \ldots, \eta_G(x)$ is identified up to relabelling the components and identification depends on whether γ_g can be recovered from the corresponding MNL model (22) given the design matrix X, constructed row-wise from the covariates $\tilde{x}_i, i = 1, \ldots, n$. Standard conditions for identification in a MNL model apply, e.g. that $(X^\top X)^{-1}$ exists (McCullagh & Nelder, 1999). It is well-known that identification in a logit and more generally in a MNL model fails under complete separation, see e.g. Heinze (2006). Hence, a situation where a mixture of experts model is not generically identified occurs, if

certain clusters do not share covariate values with other clusters, see Example 4.2 in Hennig (2000) for illustration. A rather strong condition ensuring generic identifiability for this type of models is an extended coverage condition (Hennig, 2000) requiring that the number of clusters G is exceeded by the minimum number of distinct q-dimensional hyperplanes needed to cover the covariate values (excluding the constant) for *each* cluster.

Similar arguments as above apply in general for simple mixtures of experts models of G probability distributions,

$$y_i \sim \sum_{g=1}^{G} \eta_g(\tilde{x}_i) p(y_i|\theta_g). \tag{24}$$

Provided that the parameters in the MNL model (22) are identified, it can be shown that a mixture of experts model is generically identified, if the corresponding standard finite mixture distribution is generically identified. In this case, any other mixture representation (24) with parameters $\theta_g^\star$ and $\eta_g^\star(\tilde{x}_i)$ is identified up to (the same) label switching according to a permutation σ for all possible values $\tilde{x}_i$: $\theta_g^\star = \theta_{\sigma(g)}$ and $\eta_g^\star(\tilde{x}_i) = \eta_{\sigma(g)}(\tilde{x}_i)$ for $g = 1, \ldots, G$.

It follows that mixtures of experts of multivariate Gaussian distributions (as considered in Section 2.2) and Poisson distributions, among many others, are generically identified, provided that parameters in the MNL model (22) are identified. Since a standard finite mixture model is that special case of a mixture of experts model where $\tilde{x}_i \equiv 1$ is equal to the intercept, special care must be exercised when the underlying standard finite mixture distribution is generically unidentified, as might be the case when modelling discrete data. It is interesting to note that including $\tilde{x}_i$ into the weight function $\eta_g(\tilde{x}_i)$ in mixtures of experts models is possible for models where including $\tilde{x}_i$ in the component density $p(y_i|\tilde{x}_i, \theta_g)$ yields a generically non-identified model, an example being the regressor $\tilde{x}_i = (1\ d_i)$, where d_i is a 0/1 dummy variable, see Section 5.2.

The situation gets rather complex, when covariates $\tilde{x}_i$ (or subsets of these) are included as regressors both in the outcome distribution $p(y_i|\tilde{x}_i, \theta_g)$ as well as in the weight distribution $\eta_g(\tilde{x}_i)$. The presence of a covariate $\tilde{x}_i$ in $\eta_g(\tilde{x}_i)$ could introduce high discriminative power among the groups and might lead to identification of mixture of regression models which are not identified, if η_g is assumed to be independent of the covariates. To our knowledge, generic identification for general mixtures of experts models has not been studied systematically and would be an interesting venue for future research.

As it is, the only way to investigate, if the chosen mixture model suffers from identifiability problems is to analyze the results obtained from fitting these models to the data carefully. As the examples in Section 5.1 and 5.2 have shown, weird behaviour of the MCMC draws in a Bayesian framework are often a sign of identifiability problems. On the other hand, marginal posterior concentration around pronounced modes, verified for instance through appropriate scatter plots of MCMC draws for the parameters of interest, indicates that identification might not be an issue for that specific application.

6 Concluding Remarks

This chapter has outlined the definition, estimation and application of ME models in a number of settings clearly demonstrating their utility as an analytical tool. Their demonstrated use to cluster observations, and to appropriately capture heterogeneity in cross sectional data, provides only a glimpse of their potential flexibility and utility in a wide range of settings. The ability of ME models to jointly model response and concomitant variables provides deeper and more principled insight into the relations between such data in a mixture model based analysis.

On a cautionary note however, when an ME model is employed as an analytic tool, care must be exercised in how and where covariates enter the ME model framework. The interpretation of the analysis fundamentally depends on which of the suite of ME models is invoked. Further, as outlined herein, the identifiability of an ME model must be carefully considered; establishing identifiability for ME models is an outstanding, challenging problem.

References

Akaike, Hirotogu. 1973. Information theory and an extension of the maximum likelihood principle. *Pages 267–281 of:* Petrov, B. N., & Csáki, F. (eds), *2nd International Symposium Symp. Information Theory*. Budapest: Akadémiai Kiadó.

Benaglia, T., Chauveau, D., Hunter, D.R., & 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. *Pages 183–198 of:* Ziemba, William T., Lo, Victor S., & Haush, Donald B. (eds), *Efficiency of Racetrack Betting Markets*. San Diego and London: Academic Press.

Bishop, C. M., & Svenskn, M. 2003. Bayesian Hierarchical Mixtures of Experts. *Pages 57–64 of: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. UAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Bishop, C.M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Chamroukhi, F. 2015. Non-Normal Mixtures of Experts. *ArXiv preprints* `1506.06707`, June.

Chandra, Satish. 1977. On the Mixtures of Probability Distributions. *Scandinavian Journal of Statistics*, **4**, 105–112.

Chib, S., & Greenberg, E. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**, 327–335.

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

Dayton, C. M., & Macready, G. B. 1988. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, **83**(401), 173–178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1–38.

DeSarbo, W.S., & Cron, W.L. 1988. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 248–282.

Diebolt, J., & Robert, Christian P. 1994. Estimation of Finite Mixture Distributions by Bayesian Sampling. *Journal of the Royal Statistical Society Series B*, **56**, 363–375.

Frühwirth-Schnatter, S. 2004. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, **7**(1), 143–167.

Frühwirth-Schnatter, S. 2011a. Dealing with label switching under model uncertainty. *Chap. 10, pages 213–239 of:* Mengersen, K., Robert, C. P., & Titterington, D. (eds), *Mixture Estimation and Applications*. Chichester: Wiley.

Frühwirth-Schnatter, S. 2011b. Panel Data Analysis - A Survey on Model-Based Clustering of Time Series. *Advances in Data Analysis and Classification*, **5**, 251–280.

Frühwirth-Schnatter, S., & Frühwirth, R. 2010. Data augmentation and MCMC for binary and multinomial logit models. *Pages 111–132 of:* Kneib, Thomas, & Tutz, Gerhard (eds), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*. Heidelberg: Physica-Verlag.

Frühwirth-Schnatter, S., & Kaufmann, Sylvia. 2008. Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, **26**, 78–89.

Frühwirth-Schnatter, S., & Wagner, Helga. 2008. Marginal Likelihoods for Non-Gaussian Models Using Auxiliary Mixture Sampling. *Computational Statistics and Data Analysis*, **52**, 4608–4624.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., & Winter-Ebmer, R. 2012. Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*, **27**(7), 1116–1137.

Frühwirth-Schnatter, Sylvia. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag.

Frühwirth-Schnatter, Sylvia. 2018. *Applied Bayesian Mixture Modelling. Implementations in MATLAB using the package* `bayesf Version 4.0`.

Geman, S., & Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gershenfeld, Neil. 1997. Nonlinear Inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

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

Gormley, I. C., & Murphy, T. B. 2006. Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society Series A*, **169**(2), 361–379.

Gormley, I. C., & Murphy, T. B. 2008a. Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, **103**(483), 1014–1027.

Gormley, I. C., & Murphy, T. B. 2008b. A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, **2**(4), 1452–1477.

Gormley, I. C., & Murphy, T. B. 2010a. Clustering ranked preference data using sociodemographic covariates. *Pages 543–569 of:* Hess, S., & Daly, A. (eds), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. United Kingdom: Emerald.

Gormley, I. C., & Murphy, T. B. 2010b. A Mixture of Experts Latent Position Cluster Model for Social Network Data. *Statistical Methodology*, **7**(3), 385–405.

Gormley, I. C., & Murphy, T. B. 2018. *MEclustnet: fitting the mixture of experts latent position cluster model.* R package version 1.0.

Grün, B., & Leisch, F. 2008a. Finite Mixtures of Generalized Linear Regression Models. *Pages 205–230 of:* Shalabh, & Heumann, Christian (eds), *Recent Advances in Linear Models and Related Areas*. Springer.

Grün, B., & Leisch, F. 2008b. FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**, 1–35.

Grün, B., & Leisch, F. 2008c. Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects. *Journal of Classification*, **25**, 225–247.

Handcock, M., Raftery, A.E., & Tantrum, J. M. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A*, **170**(2), 301 – 354.

Heinze, Georg. 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, **25**, 4216–4226.

Hennig, C. 2000. Identifiability of models for clusterwise linear regression. *Journal of Classification*, **17**, 273–296.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. 2002. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, **97**, 1090–1098.

Hoff, P.D. 2009. *A First Course in Bayesian Statistical Methods.* Springer-Verlag, New York.

Huerta, G., Jiang, W., & Tanner, M. A. 2003. Time series modeling via hierarchical mixtures. *Statistica Sinica*, **13**(4), 1097–1118.

Hunter, D. R., & Lange, K. 2004. A tutorial on MM algorithms. *The American Statistician*, **58**(1), 30–37.

Hunter, David R, & Young, Derek S. 2012. Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, **24**(1), 19–38.

Hurn, M., Justel, A., & Robert, C.P. 2003. Estimating mixtures of regressions. *Journal of Compututional and Graphical Statistics*, **12**, 1–25.

Ingrassia, Salvatore, Punzo, Antonio, Vittadini, Giorgio, & Minotti, Simona C. 2015. The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**(1), 85–113.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., & Hinton, G.E. 1991. Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

Jordan, M.I., & Jacobs, R.A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

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

Lang, J. B., McDonald, J. W., & Smith, P. W. F. 1999. Association-marginal modelling of multivariate categorical responses: A maximim likelihood approach. *Journal of the American Statistical Association*, **94**, 1161–71.

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

Li, F., Villani, M., & Kohn, R. 2011. Modeling conditional densities using finite smooth mixtures. *Chap. 6, pages 123–144 of:* Mengersen, K., Robert, C., & Titterington, M. (eds), *Mixtures: Estimation and Applications.* Wiley.

Marsh, M. 1999. The Making of the Eighth President. *Pages 215–242 of:* Marsh, Michael, & Mitchell, Paul (eds), *How Ireland Voted 1997.* Boulder, CO: Westview and PSAI Press.

Masoudnia, S., & Ebrahimpour, R. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, **42**(2), 275–293.

Mazza, A., Punzo, A., & Ingrassia, S. 2017. *flexCWM: Flexible Cluster-Weighted Modeling.* R package version 1.7.

McCullagh, P., & Nelder, John A. 1999. *Generalized Linear Models.* London: Chapman & Hall.

McLachlan, G, & Peel, D. 2000. *Finite Mixture Models.* New York: John Wiley.

Meng, X.-L., & Rubin, D. B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. 1953. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Muthén, L. K., & Muthén, B. O. 2011. *Mplus User's Guide.* 6 edn. Los Angeles, CA: Muthén and Muthén.

Pamminger, Christoph, & Frühwirth-Schnatter, Sylvia. 2010. Model-based Clustering of Categorical Time Series. *Bayesian Analysis*, **5**, 345–368.

Peng, F., Jacobs, R. A., & Tanner, M. A. 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, R. L. 1975. The analysis of permutations. *Applied Statistics*, **24**(2), 193–202.

Quandt, R.E. 1972. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, **67**, 306–310.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A.E., Newton, M.A., Satagopan, J.M., & Krivitsky, P. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). *Pages 371–416 of:* Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., , & West, M. (eds), *Bayesian Statistics 8.* Oxford University Press.

Rasmussen, C. E., & Ghahramani, Z. 2002. The infinite mixtures of Gaussian process experts. *Pages 554–560 of: Advances in Neural Information Processing Systems*, vol. 12. MIT Press.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Scott, Steven L. 2011. Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, **52**, 87–109.

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

Sinnott, R. 1999. The Electoral System. *Pages 99–126 of:* Coakley, John, & Gallagher, Michael (eds), *Politics in the Republic of Ireland*, 3rd edn. London: Routledge & PSAI Press.

Stephens, M. 2000. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

Subedi, Sanjeena, Punzo, Antonio, Ingrassia, Salvatore, & McNicholas, Paul D. 2013. Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.

Tang, X., & Qu, A. 2015. Mixture Modeling for Longitudinal Data. *Journal of Computational and Graphical Statistics*, **25**, 1117–1137.

Tanner, M. A. 1996. *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. 3 edn. New York: Springer-Verlag.

Teicher, Henry. 1961. Identifiability of mixtures. *The Annals of Mathematical Statistics*, **32**, 244–248.

Teicher, Henry. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **34**, 1265–1269.

Turner, Rolf. 2014. *mixreg: functions to fit mixtures of regressions.* R package version 0.0-5.

Vermunt, Jeroen K, & Magidson, Jay. 2005. *Latent GOLD 4.0 User's Guide*. Statistical Innovations Inc.

Villani, Mattias, Kohn, Robert, & Giordani, Paolo. 2009. Regression density Estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics*, **153**, 155–173.

Wang, P., Puterman, M.L., Cockburn, I., & Le, N. 1996. Mixed Poisson regression models with covariate dependent rates. *Biometrics*, **52**, 381–400.

Waterhouse, S., MacKay, D., & Robinson, T. 1996. Bayesian methods for mixtures of experts. *Pages 351–357 of: Advances in Neural Information Processing Systems*. Morgan Kaufmann Publishers.

White, A., & Murphy, T. B. 2016. Mixed-Membership of Experts Stochastic Blockmodel. *Network Science*, **4**(Apr.), 48–80.

Yakowitz, S. J., & Spragins, J. D. 1968. On the Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **39**, 209–214.

Young, D. S., & Hunter, D. R. 2010. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, **54**(10), 2253–2266.