A Mixture of Experts Latent Position Cluster Model for Social Network Data

Isobel Claire Gormley and Thomas Brendan Murphy

University College Dublin
{claire.gormley, brendan.murphy}@ucd.ie

Abstract

Social network data represent the interactions between a group of social actors. Interactions between colleagues and friendship networks are typical examples of such data.

The latent space model for social network data locates each actor in a network in a latent (social) space and models the probability of an interaction between two actors as a function of their locations. The latent position cluster model extends the latent space model to deal with network data in which clusters of actors exist — actor locations are drawn from a finite mixture model, each component of which represents a cluster of actors.

A mixture of experts model builds on the structure of a mixture model by taking account of both observations and associated covariates when modeling a heterogeneous population. Herein, a mixture of experts extension of the latent position cluster model is developed. The mixture of experts framework allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations.

Estimates of the model parameters are derived in a Bayesian framework using a Markov Chain Monte Carlo algorithm. The algorithm is generally computationally expensive — surrogate proposal distributions which shadow the target distributions are derived, reducing the computational burden.

The methodology is demonstrated through an illustrative example detailing relations between a group of lawyers in the USA.

Key words: Clustering; covariates; latent space; mixture models; mixture of experts models; social network data; surrogate proposal distributions.

1. Introduction

Social network data has recently attracted much attention from a broad spectrum of research communities including sociology, statistics, mathematics, physics and computer science. Social network data record the interactions and relationships between a group of social entities or actors. For example, a social network data set may detail the friendship links among a group of colleagues, or it may detail the level of international trade between countries. Network data may be binary, indicating the presence/absence of a link between two actors, or it may be non-binary indicating the level of interaction between two actors. The aim of social network analysis (SNA) is to explore the structure within the network, to aid understanding of underlying phenomena and the relations that may or may not exist within the network.

Recent advances in social network analysis focus on the idea of using a latent social space in which actors are located to model the interactions between actors. Hoff et al. (2002) develop the idea of a latent social space and define the probability of a link between two actors as a function of their separation in the latent space. Moreover, conditional on the latent locations of two actors, the presence or absence of a link between them is considered to be independent of all other links in the network.

The latent position cluster model (Handcock et al., 2007) further develops the idea of a latent social space by extending the model to accommodate clusters of actors in the latent space. Under the latent position cluster model the latent location of each actor is assumed to be drawn from a finite mixture model, each component of which represents a cluster of actors. The latent position cluster model offers a more flexible version of the latent space model for modeling heterogeneous social networks. Section 2.1 provides an introduction to the latent position cluster model.

Both the latent space model and the latent position cluster model provide a framework in which covariates or attributes of the actors may be explicitly included in the model — in these models the probability of a link between two actors may be modeled as a function of both their separation in the latent space and of their relative covariates; this modeling extension facilitates homophily by attributes. However, the covariates may contribute more to the structure of the network model than solely through the link probabilities. For example, perhaps the covariates influence both the cluster membership of an actor and their link probabilities.

Herein, the latent position cluster model is extended to provide a model

with the flexibility to allow covariates to contribute to the network structure in a number different ways. The inclusion of covariates is achieved using a mixture of experts (Jacobs et al., 1991) modeling framework. In its original form, the mixture of experts model is a mixture of generalized linear models where additionally the probability of belonging to a cluster is modeled as a logistic function of covariates. Under the latent position cluster model the latent location of an actor is assumed to be drawn from a finite mixture of multivariate normal distributions. The mixture of experts latent position cluster model extends the latent position cluster model such that the mixing proportions are modeled as functions of the actor covariates. Thus, the covariates of an actor may influence the probability of a link with another actor, but additionally covariates may influence the cluster membership of the actor; this network modeling idea was originally proposed in Gormley and Murphy (2007a).

The mixture of experts latent position cluster 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; covariates may also influence the link probabilities directly. An outline of the background to the mixture of experts modeling framework is given in Section 2.2. The mixture of experts latent position cluster model is developed in Section 2.3.

Estimation of the mixture of experts latent position cluster model is achieved within a Bayesian framework using a Markov Chain Monte Carlo algorithm. Both the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings algorithm (Metropolis et al., 1953; Chib and Greenberg, 1995) are employed to obtain parameter estimates. Typically, within the context of network models, such methods are computationally expensive and require lengthy run times to achieve sufficient mixing of the chain. An inherent problem is the selection of a suitable proposal distribution when employing the Metropolis-Hastings algorithm. Additionally, tuning the parameters of the selected proposal distribution in order to achieve sufficient mixing can be a difficult and time consuming process, leading to inefficient model fitting. Often the target distribution from which we wish to sample has a complex form and standard proposal distributions perform poorly, necessitating long run times to achieve sufficient mixing. The inclusion of covariates in the latent position cluster model makes such methods even more computationally expensive. Here, multivariate Taylor expansions of terms within the target distributions are employed to derive suitable proposal distributions which are

good surrogates for the target distributions. Details of the derivation of such surrogate proposal distributions are given in Section 3.2.

Modeling issues such as identifiability, label switching and model selection are dealt with in Section 4.

An illustrative example of the methodology is detailed in Section 5 — the mixture of experts latent position cluster model is fitted to two social network data sets detailing friendship and co-worker relations between a set of 71 attorneys in a northeastern USA corporate law firm. A number of covariates are available for each attorney including seniority, age, office location and law school. The mixture of experts latent position cluster model is fitted over a range of dimensions of the latent social space and over a range of numbers of clusters. Influential covariates are selected using a variable selection procedure. Overall model performance is guided by the AICM (the Akaike Information Criterion – Monte (Carlo)) (Raftery et al., 2007). A number of covariates are deemed to be influential, offering further insight to the structure of the network than was available under the default latent position cluster model.

Section 6 comments on the mixture of experts latent position cluster model in general, and suggests directions for future research.

2. The mixture of experts latent position cluster model

Social network data take the form of a set of relations $\{y_{i,j}\}$ from a group of $i, j = 1, \ldots, n$ actors, represented by an $n \times n$ sociomatrix $\mathbf{Y}$. The relation $y_{i,j}$ between actor i and actor j may be a binary relation, indicating the presence or absence of a link between the two actors. In such a case, the matrix $\mathbf{Y}$ can be thought of as a graph in which nodes are actors and edges indicate a link between two nodes. The mixture of experts latent position cluster model is developed within the context of binary valued relations, but the methodology is easily extended to other forms of relation (such as count data). Covariate data $\underline{w}_i^T = (w_{i1}, \ldots, w_{ip})$ associated with actor i may also be available, where p denotes the number of observed covariates.

We focus on the recent advances in social network analysis based on the idea of a latent social space (Hoff et al., 2002) in which actors each have a location. The probability of a link or relation between two actors is modeled as a function of their separation in the latent space. The idea of a latent social space provides the underlying framework for the latent position cluster model Handcock et al. (2007).

2.1. The latent position cluster model

For each actor, $i = 1, 2, \ldots, n$, the latent position cluster model associates a location $\underline{z}_i^T = (z_{i1}, \ldots, z_{id})$ in a d-dimensional latent social space. The model assumes that the probability of a link between two actors is independent of all other links in the network, conditional on the latent locations of the actors. Hence, it follows that the likelihood function of the data is

$$\mathbf{P}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}) = \prod_{i=1}^{n} \prod_{j \neq i} \mathbf{P}(y_{i,j}|\underline{z}_i, \underline{z}_j, \underline{x}_{i,j}, \underline{\beta}) \quad (1)$$

where $\mathbf{Z}$ is the $n \times d$ matrix of latent locations, $\mathbf{X}$ is the matrix of dyadic specific covariates associated with the n actors and $\underline{\beta}$ is a vector of parameters. In this specification $\underline{x}_{i,j}^T = (x_{i,j,1}, \ldots, x_{i,j,p})$ denotes a p vector of dyadic specific covariates where typically $x_{i,j,k} = d(w_{ik}, w_{jk})$, where $d(w_{ik}, w_{jk})$ is a measure of the difference or similarity in the value of the kth covariate for actors i and j, for $k = 1, \ldots, p$. The probability of a link between actors i and j is in turn modeled using a logistic regression model where

$$\log \left\{ \frac{\mathbf{P}(y_{i,j} = 1)}{\mathbf{P}(y_{i,j} = 0)} \right\} = \beta_0 + \beta_1 x_{i,j,1} + \cdots + \beta_p x_{i,j,p} - \parallel \underline{z}_i - \underline{z}_j \parallel . \quad (2)$$

The logistic regression model models the probability of a link between two actors as a function of both their dyadic specific covariates, and the Euclidean distance between them in the latent social space. The logistic regression model is appropriate for modeling binary valued relations; for alternative edge types the logistic regression term can be changed to an appropriate alternative generalized linear model (McCullagh and Nelder, 1983). To identify the model, as specified in Krivitsky and Handcock (2008), the regression coefficient associated with the Euclidean distance is constrained to be -1 while the latent locations remain unconstrained.

To complete the specification of the latent position cluster model, it is assumed that the latent locations $\underline{z}_i$ are drawn from a finite mixture model. This mixture modeling framework accounts for clustering of actor locations in the latent space. Specifically, the latent locations are assumed to be drawn from a mixture of G multivariate normal distributions where G denotes the unknown number of clusters of actors:

$$\underline{z}_i \sim \sum_{g=1}^{G} \lambda_g \mathrm{MVN}(\underline{\mu}_g, \sigma_g^2 \mathbf{I}).$$

The probability of belonging to cluster g, or the gth mixing proportion, is denoted λ_g ($\sum_g \lambda_g = 1$). Each component of the mixture model, or cluster, is characterized by a multivariate normal distribution with a d dimensional mean vector $\underline{\mu}_g$ and a diagonal covariance matrix $\sigma_g^2\mathbf{I}$ where $\mathbf{I}$ is the $d \times d$ identity matrix. Modeling the latent locations in this way relates to model-based clustering of observed variables (Banfield and Raftery, 1993; Fraley and Raftery, 2002).

A Bayesian framework can be employed for parameter estimation. As is standard in Bayesian estimation of mixture models (Diebolt and Robert, 1994; Hurn et al., 2003) the problem is greatly simplified by augmenting the observed data with an indicator variable K_i for each actor i where $K_i = g$ if actor i belongs to cluster g. The indicator variable K_i therefore has a multinomial distribution with a single trial and probabilities equal to λ_g for $g = 1, \ldots, G$.

Parameter estimation could also be achieved via maximum likelihood, but in practice this is challenging. Handcock et al. (2007) derive a two-stage maximum likelihood approach to parameter estimation which first computes the maximum likelihood estimates of the standard latent space model of Hoff et al. (2002), and then computes the maximum likelihood estimates for the mixture model applied to the resulting estimated latent locations. While this approach is fast and algebraically simpler than the Bayesian approach taken herein, it does not take advantage of the clustering information when estimating the latent locations and was observed in practice by Handcock et al. (2007) to be out-performed by the Bayesian approach.

In the original specification of the latent position cluster model, actor covariates only play a part in determining the probability of a link between two actors. It is intuitive to also allow covariates to influence other parts of the model — it is conceivable that the covariates of an actor influence which cluster they belong to and, conditional on their cluster membership, they then have a propensity for forming relations. To facilitate such modeling flexibility, a mixture of experts modeling framework is employed.

2.2. Mixture of experts models

Mixture of experts models originally appear in the computer science literature (Jacobs et al., 1991) as a mixture of generalized linear models, where the mixing proportions are also modeled using generalized linear model theory. The components of the mixture model are termed 'expert networks',

with the mixing proportions known as 'gating networks' leading to the name mixture of experts models.

In the mixture of experts model, the mixing proportions are viewed as success probabilities from a multinomial logistic regression where the probability of belonging to each of $G-1$ clusters compared to a baseline cluster is a function of the covariates of an observation. In this framework each observation has a cluster membership probability, which is a direct parameter of the model. Observation i's mixing proportions $\underline{\lambda}_i^T = (\lambda_1(\underline{w}_i), \ldots, \lambda_G(\underline{w}_i))$ are modeled as a logistic function of their p covariates $\underline{w}_i$ i.e. for $g = 2, \ldots, G$

$$\log\left\{\frac{\lambda_g(\underline{w}_i)}{\lambda_1(\underline{w}_i)}\right\} = \tau_{g0} + \tau_{g1}w_{i1} + \cdots + \tau_{gp}w_{ip} \quad (3)$$

where cluster 1 is treated as a baseline cluster, and τ_{g0} is an intercept term.

In its original form the mixture of experts model also used a generalized linear model as the model within each mixture component. By changing the form of the model in the components to suit the problem at hand a flexible suite of models is available. In a univariate setting a simple linear regression model could be used; employing a logistic regression model offers a classification model. Mixture of experts models have therefore been applied and developed in a wide range of areas ranging from speech recognition problems (Peng et al., 1996), to modeling rank data (Gormley and Murphy, 2008, 2009a). Jordan and Jacobs (1994) estimate the mixture of experts model within a classical framework via the EM algorithm; Peng et al. (1996) employ a Bayesian approach. Additionally Jordan and Jacobs (1994) develop a hierarchical mixture of experts model which is a mixture of experts model which has more than one layer of mixing (ie. more than one layer of gating networks).

2.3. The mixture of experts latent position cluster model

The latent position cluster model can be extended within a mixture of experts framework. Under the latent position cluster model, the latent locations $\underline{z}_i$ $(i = 1, \ldots, n)$ are assumed to be drawn from a finite mixture of multivariate normal distributions with mixing proportions $\underline{\lambda}^T = (\lambda_1, \ldots, \lambda_G)$. In the mixture of experts latent position cluster model framework, the mixing proportions are actor specific and can be modeled as a logistic function of their covariates, that is,

$$\underline{z}_i \sim \sum_{g=1}^{G} \lambda_g(\underline{w}_i)\text{MVN}(\underline{\mu}_g, \sigma_g^2\mathbf{I}) \quad (4)$$

where $\sum_g \lambda_g(\underline{w}_i) = 1$ and by (3)

$$\mathbf{P}\{K_i = g|\underline{w}_i\} = \lambda_g(\underline{w}_i) = \frac{\exp(\tau_{g0} + \tau_{g1}w_{i1} + \cdots + \tau_{gp}w_{ip})}{\sum_{g'=1}^{G}\exp(\tau_{g'0} + \tau_{g'1}w_{i1} + \cdots + \tau_{g'p}w_{ip})} \quad (5)$$

where $\underline{\tau}_1^T = (0, 0, \ldots, 0)$.

Thus the mixture of experts latent position cluster model is the latent position cluster model of Handcock et al. (2007) where the mixing proportions are modeled as a function of the actor covariates. Figure 1 provides an illustration of the dependencies within the original latent position cluster model and the mixture of experts latent position cluster model using graphical models.

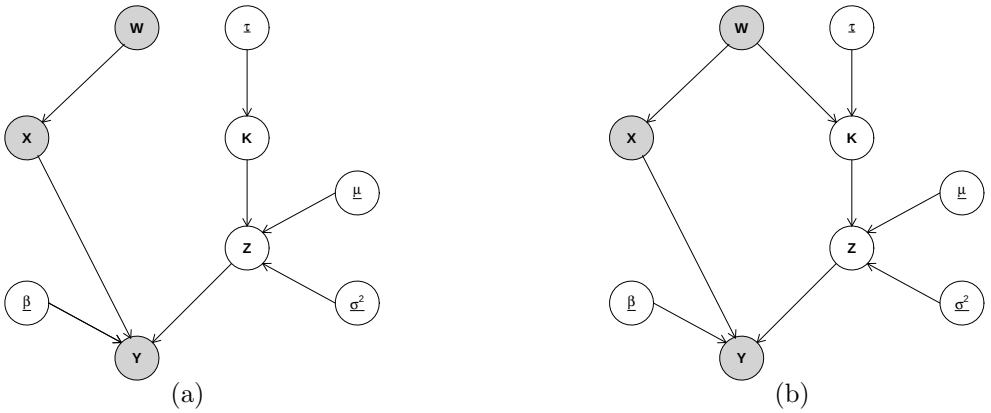


Figure 1: Graphical models illustrating the dependencies (a) within the latent position cluster model and (b) within the mixture of experts latent position cluster model.

The actor covariates can influence the cluster membership model (5), the link probability model (2), both models or neither model. This yields a suite of four possible network models within the mixture of experts framework. The graphical model representations for these four options can be developed

by including or deleting the edges from the **W** node to the **X** node (and from the **X** node to the **Y** node) and from the **W** node to the **K** node in Figure 1(b).

3. Model fitting

A Bayesian approach using Markov Chain Monte Carlo is utilized for fitting the mixture of experts latent position cluster model. Interest lies in inferring the latent locations $\underline{z}_i$ of each actor i, the link probability regression parameters $\underline{\beta}$, the (mixing proportion) cluster membership regression parameters $\underline{\tau}_g$, the cluster mean $\underline{\mu}_g$ and variance σ_g^2 of each cluster g and the actor cluster memberships $\underline{K} = (K_1, \ldots, K_n)$.

3.1. Model specification

Prior distributions on the model parameters are required and are specified as follows:

$$\underline{\beta} \sim \text{MVN}(\underline{\xi},\ \Psi) \qquad \underline{\tau}_g \sim \text{MVN}(\underline{\gamma},\ \Phi) \text{ i.i.d } g = 2, \ldots, G.$$

$$\underline{\mu}_g \sim \text{MVN}(\underline{0},\ \Omega) \text{ i.i.d } g = 1, \ldots, G. \quad \sigma_g^2 \sim \sigma_0^2\, \text{Inv}\chi^2_\alpha \quad \text{i.i.d } g = 1, \ldots, G.$$

where $\underline{\xi}, \Psi, \Omega, \sigma_0^2, \alpha, \underline{\gamma}$ and Φ are hyperparameters which need to be specified by the practitioner. Krivitsky and Handcock (2008) employ a hierarchical Bayesian model for the latent position cluster model and specify hyperpriors on the hyperparameters. In this work, as in Handcock et al. (2007), the hyperparameters are specified to be $\underline{\xi} = \underline{\gamma} = \underline{0}$ and $\Psi = \Phi = \Omega = 2\mathbf{I}$ which postulates a wide range of values for both the link probability regression parameters $\underline{\beta}$ and the mixing proportion regression parameters $\underline{\tau}_g$ and provides a relatively flat prior for the cluster means. For the cluster covariance parameter σ_g^2, the same prior as that in Handcock et al. (2007) is used with $\alpha = 2$ and $\sigma_0^2 = 0.103$.

The posterior distribution for the model parameters given the observed data is formed by combining the likelihood in equations (1) – (5) with the specified prior distributions. The full conditional posterior distributions of the model parameters are specified as follows:

$$
\begin{aligned}
\underline{z}_i|K_i=g,\cdots &\propto \mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\}\ \text{MVN}(\underline{z}_i\,|\,\underline{\mu}_g,\ \sigma_g^2\mathbf{I}) && \text{for } i=1,\ldots,n \\
& && \text{and } g=1,\ldots,G \\
\underline{\beta}|\cdots &\propto \mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\}\ \text{MVN}(\underline{\beta}\,|\,\underline{\xi},\ \Psi) && \\
\underline{\tau}_g|\cdots &\propto \mathbf{P}\{\underline{K}=(g,\ldots,g)|\underline{\tau}_g,\mathbf{W}\}\ \text{MVN}(\underline{\tau}_g\,|\,\underline{\gamma},\ \Phi) && \text{for } g=2,\ldots,G \\
\underline{\mu}_g|\cdots &\propto \text{MVN}\left(\tfrac{m_g\bar{z}_g}{m_g+\sigma_g^2/\omega^2},\tfrac{\sigma_g^2}{m_g+\sigma_g^2/\omega^2}\mathbf{I}\right) && \text{for } g=1,\ldots,G \\
\sigma_g^2|\cdots &\propto (\sigma_0^2+s_g^2)\ \text{Inv}\ \chi^2_{\alpha+m_gd} && \text{for } g=1,\ldots,G \\
K_i|\cdots &\propto \text{Multinomial}\Big[1,\Big\{\tfrac{1}{C}\lambda_1(\underline{w}_i)\text{MVN}(\underline{z}_i|\underline{\mu}_1,\sigma_1^2\mathbf{I}), \\
& \qquad \ldots,\tfrac{1}{C}\lambda_G(\underline{w}_i)\text{MVN}(\underline{z}_i|\underline{\mu}_G,\sigma_G^2\mathbf{I})\Big\}\Big] && \text{for } i=1,\ldots,n
\end{aligned}
$$

where $\cdots$ denotes all other parameters, latent locations and cluster memberships in the model and $\mathbf{W}$ is the $n\times p$ matrix of actor specific covariates. Also

$$
\begin{aligned}
m_g &= \sum_{i=1}^{n}\mathbf{1}\{K_i=g\} & s_g^2 &= \sum_{i=1}^{n}(\underline{z}_i-\underline{\mu}_g)^T(\underline{z}_i-\underline{\mu}_g)\mathbf{1}\{K_i=g\} \\
\bar{z}_g &= \tfrac{1}{m_g}\sum_{i=1}^{n}\underline{z}_i\mathbf{1}\{K_i=g\} & C &= \sum_{g'=1}^{G}\lambda_{g'}(\underline{w}_i)\text{MVN}(\underline{z}_i|\underline{\mu}_{g'},\sigma_{g'}^2\mathbf{I}).
\end{aligned}
$$

The full conditional posterior distributions of some model parameters are readily available and sampling parameter values in these cases is straightforward using the Gibbs sampler. However the full conditional posterior distributions of some model parameters are not in a recognizable form and require a Metropolis-Hastings sampling procedure. A Metropolis-within-Gibbs sampler (eg. O'Hagan and Forster, 2004) is therefore required — this algorithm embeds a Metropolis step (or steps) within a Gibbs algorithm. Carlin and Louis (2000) discuss convergence issues associated with such an algorithm.

As with any Metropolis-Hastings algorithm a suitable proposal distribution must be specified from which to draw proposed parameter values. Ideally, such proposal distributions shadow the target distribution from which sampling is required — the rate at which the algorithm converges to sampling

from the target distribution depends crucially on the relationship between the proposal distribution and the target distribution (Gilks et al., 1996). Determining a 'good' proposal distribution often relies on craftsmanship and initial experimentation to determine both the form of a good proposal distribution and its associated parameters. In high-dimensional models, such as the mixture of experts latent position cluster model, such an experimental approach involving choosing a suitable distribution and tuning parameters in the proposal is unpractical and time consuming. Herein, multivariate Taylor expansions of terms within the target distributions are employed to aid the construction of *surrogate proposal distributions*. A surrogate proposal distribution is an approximation of a target distribution, derived by approximating terms within the target distribution via multivariate Taylor expansions. Similar ideas are employed in Gormley and Murphy (2009b).

The idea of surrogate proposal distributions is loosely related to the concepts behind MM algorithms which facilitate optimization of problematic functions by iteratively optimizing a surrogate function (Lange et al., 2000; Hunter and Lange, 2004; Hunter, 2004). Additionally, surrogate proposal distributions and variational methods have a similar ethos in that they both form an approximation to a problematic posterior distribution, but in the case of variational methods no sampling is required. Variational approaches to fitting graphical models are detailed in Wainwright and Jordan (2008); Salter-Townshend and Murphy (2009) specifically develop a variational approach to inference for the original latent position cluster model of Handcock et al. (2007).

3.2. Surrogate proposal distributions

Surrogate proposal distributions are constructed by approximating a term (or terms) within the target distribution by a quadratic multivariate Taylor expansion. The quadratic multivariate Taylor expansion of a function $g(\underline{\theta})$ with parameter vector $\underline{\theta}$ about $\bar{\underline{\theta}}$,

$$g(\underline{\theta}) \approx g(\bar{\underline{\theta}}) + (\underline{\theta} - \bar{\underline{\theta}})^T \{\nabla g(\bar{\underline{\theta}})\} + 0.5(\underline{\theta} - \bar{\underline{\theta}})^T \{\nabla^2 g(\bar{\underline{\theta}})\}(\underline{\theta} - \bar{\underline{\theta}}), \tag{6}$$

provides a quadratic approximation or *surrogate function* for $g(\underline{\theta})$.

The idea of approximating a term within a function to produce a surrogate function is employed to construct surrogate proposal distributions which approximate problematic full conditional distributions from which sampling is required. When sampling a parameter $\underline{\theta}$ at each iteration of the MCMC

algorithm, a surrogate function is formed about a parameter value $\underline{\bar{\theta}}$ which in practice is the previously sampled value of the parameter. Thus the surrogate proposal distribution is updated each time a new parameter value is sampled, maintaining the improved approximation throughout the run time of the algorithm.

Since the choice of both the form and the parameters of the surrogate proposal distribution is automated, model fitting is efficient as the user need not spend time experimenting with different proposal distributions and tuning the associated parameters until sufficient mixing is achieved. The Metropolis-Hastings algorithm no longer involves user input as it is driven by the surrogate proposal distribution and its parameters, automatically determined by the model itself and the observed data.

Within the mixture of experts latent position cluster model, Metropolis-Hastings steps are required to sample the latent locations $\underline{z}_i$, the link probability regression parameters $\underline{\beta}$ and the cluster membership regression parameters $\underline{\tau}_g$. When sampling the latent locations, a multivariate normal proposal distribution centered on the current estimate and with a fixed diagonal covariance matrix performs sufficiently well; employing such a proposal for the regression parameters does not perform well, and convergence is crucially dependent on both starting values and the choice of the parameters employed within the proposal distributions. Thus surrogate proposal distributions are constructed for the regression parameters only.

The mathematical derivation of the surrogate proposal distributions is given in Appendix A; here it suffices to say that the resulting surrogate proposal distributions are multivariate normal distributions. The parameters of these surrogate proposal distributions depend on the current location of the Markov Chain and are therefore updated at each step of the algorithm. Thus the construction of proposal distributions is efficient, as the choice of a distributional form and associated parameter values is automated, reducing the time required to fit the mixture of experts latent position cluster model.

3.3. The Metropolis-within-Gibbs sampler

The Metropolis-within-Gibbs sampler proceeds as follows:

1. *Update the latent locations* $\underline{z}_i$ *for* $i = 1, \ldots, n$ *using a Metropolis-Hastings step.*

   (a) Propose $\underline{z}_i^* \sim \text{MVN}(\underline{z}_i, \sigma_g^2 \mathbf{I})$.

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

$$u \leq \min\left\{\frac{\mathbf{P}\{\mathbf{Y}|\mathbf{Z}^*,\mathbf{X},\underline{\beta}\}\ \text{MVN}(\underline{z}_i^* \mid \underline{\mu}_{K_i},\ \sigma_{K_i}^2\mathbf{I})}{\mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\}\ \text{MVN}(\underline{z}_i \mid \underline{\mu}_{K_i},\ \sigma_{K_i}^2\mathbf{I})},\ 1\right\}.$$

set latent location $\underline{z}_i$ to be $\underline{z}_i^*$.

2. *Update the link probability regression parameters $\underline{\beta}$ using a Metropolis-Hastings step.*

(a) Propose $\underline{\beta}^* \sim \text{MVN}(\underline{\delta}_{\bar{\underline{\beta}}}, \Delta_{\bar{\underline{\beta}}})$ where $\underline{\delta}_{\bar{\underline{\beta}}}$ and $\Delta_{\bar{\underline{\beta}}}$ are as defined in Appendix A, equations (9) and (8) respectively.

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

$$u \leq \min\left\{\frac{\mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}^*\}\ \text{MVN}(\underline{\beta}^* \mid \underline{\xi},\ \Psi)\ \text{MVN}(\underline{\beta} \mid \underline{\delta}_{\underline{\beta}^*},\ \Delta_{\underline{\beta}^*})}{\mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\}\ \text{MVN}(\underline{\beta} \mid \underline{\xi},\ \Psi)\ \text{MVN}(\underline{\beta}^* \mid \underline{\delta}_{\bar{\underline{\beta}}},\ \Delta_{\bar{\underline{\beta}}})},\ 1\right\}.$$

set the link probability regression parameters $\underline{\beta}$ to be $\underline{\beta}^*$.

3. *Update the mixing proportion regression parameters $\underline{\tau}_g$ for $g = 2, \ldots, G$ using a Metropolis-Hastings step.*

(a) Propose $\underline{\tau}_g^* \sim \text{MVN}(\underline{\delta}_{\bar{\underline{\tau}}_g}, \Delta_{\bar{\underline{\tau}}_g})$ where $\underline{\delta}_{\bar{\underline{\tau}}_g}$ and $\Delta_{\bar{\underline{\tau}}_g}$ are as defined in Appendix A, equations (12) and (11) respectively.

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

$$u \leq \min\left\{\frac{\mathbf{P}\{\underline{K} = (g,\ldots,g)|\underline{\tau}_g^*,\mathbf{W}\}\ \text{MVN}(\underline{\tau}_g^* \mid \underline{\gamma},\ \Phi)\ \text{MVN}(\underline{\tau}_g \mid \underline{\delta}_{\underline{\tau}^*},\ \Delta_{\underline{\tau}^*})}{\mathbf{P}\{\underline{K} = (g,\ldots,g)|\underline{\tau}_g,\mathbf{W}\}\ \text{MVN}(\underline{\tau}_g \mid \underline{\gamma},\ \Phi)\ \text{MVN}(\underline{\tau}_g^* \mid \underline{\delta}_{\bar{\underline{\tau}}},\ \Delta_{\bar{\underline{\tau}}})},\ 1\right\}.$$

set the mixing proportion regression parameters $\underline{\tau}_g$ to be $\underline{\tau}_g^*$.

4. *Update the remaining parameters using Gibbs sampling.*
Update $\underline{K}$, $\underline{\mu}_g$ and σ_g^2 for $g = 1, \ldots, G$ from the expressions detailed in Section 3.1.

4. Model issues

Prior to examining the approximate posterior distributions output by the Metropolis-within-Gibbs sampler some model issues must be taken into account. Modifications need to be made to the samples drawn prior to summarizing their posterior distributions through the calculation of posterior means or modes. Additionally the issue of model performance must be addressed.

4.1. Model identifiability

The latent locations of actors enter the likelihood function only as a function of their Euclidean distance. The likelihood is therefore invariant to any distance preserving transformation of the configuration of latent locations. If the non-identifiability of the latent locations is not accounted for, posterior estimates of the latent locations will be misleading. Additionally, as in any Bayesian mixture modeling context, the likelihood is invariant to switching the cluster labels. This 'label switching' phenomenon leads to posterior distributions which are multimodal or symmetric and again summarizing the posterior using posterior means is inappropriate. The output from the Markov Chain Monte Carlo must be processed to ensure final estimates of all parameters are valid.

Procrustean methods are used to aid the eradication of likelihood invariance due to transformations of the configuration of latent locations. Procrustean methods (Krzanowski, 1988) match one configuration of points to another as well as possible in a least squares sense. Transformations such as translation, rotation, and dilation are used to create the match. In the current context only translation and orthogonal rotations are applicable due to the definition of the likelihood function — a dilation or oblique rotation of the configuration of latent locations would alter the Euclidean distances between actors therefore changing the likelihood of the model. Orthogonal rotations and translations do not influence the Euclidean distances between actors in the latent space and so are permitted. Thus, a reference configuration of latent locations is selected and each sampled configuration is translated and/or rotated to match it.

In practice, to obtain a reference configuration of locations, the Metropolis-Hastings step to sample latent locations is initially constrained to accept uphill moves only. This will find at least a local optimum on the posterior surface which is employed as a rough approximation of the *maximum a posteriori* (MAP) configuration of latent locations. This approximate MAP configuration of latent locations is then used as a reference to which all future sampled configurations are matched. Posterior estimates of the actors' latent locations derived from the post-uphill samples will therefore be consistently matched to the same approximate MAP configuration.

A local optimum which is far from the global optimum of the posterior surface could unintentionally be employed as the approximate MAP configuration. This is not ideal as a reference configuration, but as the chain converges the sampled configurations will be drawn from the true full conditional

distribution, regardless of the configuration to which they are mapped to aid model identifiability. The distances between actors in the sampled configurations influence the likelihood of the model, and as they are not altered by the transformations, the configuration used as a reference is somewhat irrelevant; the rough MAP approximation is simply used as a reference to identify the model – it does not influence the validity of the sampled configurations.

The issue of label switching in Bayesian mixture models has received much attention in the literature — Richardson and Green (1997) suggest minimizing the label switching problem by imposing artificial identifiability constraints such as ordering the mixing proportions or other model parameters. The selection of the parameters on which to base the ordering and indeed the ordering itself is somewhat ad hoc however. Relabeling strategies (Celeux et al., 2000; Stephens, 2000) based on a decision theoretic approach are an alternative way to deal with label switching. The decision theoretic approach involves defining a loss function $\mathcal{L}(\hat{\theta}|\theta)$ which quantifies the loss incurred by choosing $\hat{\theta}$ when the true parameter value is θ. The aim is therefore to choose $\hat{\theta}$ to minimize the posterior expected loss $\mathbf{E}\{\mathcal{L}(\hat{\theta}|\theta)\}$.

The mixture of experts latent position cluster model has several parameters, any of which could be used as the basis for dealing with the label switching problem. It seems intuitive to base the loss function on a parameter whose full conditional distribution is specifiable, rather than choosing a parameter which requires a Metropolis-Hastings update step. Values sampled within a Gibbs sampler are guaranteed to be drawn from the required target distribution. Values sampled by a Metropolis-Hastings sampler however are only guaranteed to be drawn from the distribution of interest at convergence, which itself is difficult to confirm and also may take some time to achieve. It seems appropriate to base the label switching correction on parameter values sampled from the required target distribution, rather than values which are not guaranteed to be sampled from the target distribution. The full conditional distributions are available for the cluster memberships $\underline{K}$, the cluster means $\underline{\mu}_g$ and the cluster variances σ_g^2. The cluster means and variances appear to be more informative parameters in this context as they are estimated from several observations, whereas each cluster membership is only estimated from a single observation. As the cluster variances, by definition, are constrained to be equal across dimensions, it appears to be desirable to base the loss function on the cluster means $\underline{\mu}_g$.

An approximation to the true value of the cluster means $\boldsymbol{\mu}^R = (\underline{\mu}_1^R, \ldots, \underline{\mu}_G^R)$

is used as a reference. This reference value of the cluster means is chosen to be an approximation of the *maximum a posteriori* (MAP) estimate obtained after the uphill only Metropolis-Hastings steps which were used to determine the reference configuration of latent locations. This approximate MAP value is used as the reference to which each new estimate $\hat{\boldsymbol{\mu}}$ of the means is matched to correct for any label switching which may have occurred during sampling. A sum of squares function is employed as the loss function to be minimized i.e.

$$\mathcal{L}(\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}^R) = \sum_{g=1}^{G}\sum_{j=1}^{d}(\hat{\mu}_{gj} - \mu_{gj}^R)^2.$$

Once the MAP estimate has been obtained and subsequent to a typical burn-in period of the Markov chain, the estimated cluster means $\underline{\hat{\mu}}_g^t$ are permuted until the loss function is minimized. An online algorithm (Stephens, 2000) which calculates valid posterior means then proceeds as follows:

1. Generate all $G!$ permutations ν_l for $l = 1, \ldots, G!$. Set $t = 0$.
2. Subsequent to burn-in of the Markov chain, choose ν_l which minimizes

$$\mathcal{L}(\hat{\boldsymbol{\mu}}_{\nu_l}^t|\boldsymbol{\mu}^R) = \sum_{g=1}^{G}\sum_{j=1}^{d}(\hat{\mu}_{\nu_l(g)j}^t - \mu_{gj}^R)^2.$$

3. Calculate the posterior mean parameter estimates using the computationally efficient online formula:

$$\theta = \frac{t}{t+1}\theta + \frac{1}{t+1}\hat{\theta}_{\nu_l}^t$$

   for $\theta = \underline{\mu}_g, \sigma_g^2, \underline{\tau}_g$ for $g = 1, \ldots, G$. Set $K_i = \nu_l(K_i)$ for each actor.
4. Set $t = t + 1$ and repeat steps 2 and 3 subsequent to each run of the Metropolis-within-Gibbs sampler.

Note that as cluster 1 is used as the baseline cluster the mixing proportion regression parameters $\underline{\tau}_g$ must be re-scaled subsequent to label switching and prior to evaluating the posterior mean. This is achieved by simply subtracting the current $\underline{\tau}_1$ values from $\underline{\tau}_g$ $g = 2, \ldots, G$ and then evaluating the posterior mean.

Employing both the Procrustes and label-switching corrections aims to minimize the influence of identifiability issues on the posterior parameter estimates.

4.2. Model assessment

Model assessment in the Bayesian framework involves evaluating the posterior model probability for all models under consideration and then generally choosing the model with largest posterior probability (Kass and Raftery, 1995). Evaluating the posterior model probability poses difficulties in the context of the mixture of experts latent position cluster model. Handcock et al. (2007) employ a version of the Bayesian information criterion (BIC) (Schwarz, 1978) as an approximation to the posterior model probability, conditional on the latent locations when performing model selection with the latent position cluster model.

Here the AICM (Akaike's information criterion - Monte (Carlo)) (Raftery et al., 2007) is employed as a model assessment tool. The AICM is a simulation based estimate of the AIC (Akaike's information criterion) (Akaike, 1973). The AICM is a penalized version of the posterior mean of the loglikelihoods and is defined to be

$$\text{AICM} = 2(\bar{l} - s_l^2)$$

where $\bar{l}$ and s_l^2 are the mean and variance respectively of the log likelihoods of a posterior sample. Clearly the AICM is easily computed once a posterior sample has been generated. The AICM was successfully employed in Raftery et al. (2007) and in Gormley and Murphy (2007b) to select the dimensionality of latent spaces.

Even though AICM is easily computed, model selection in the mixture of experts latent position cluster model is difficult due to the large model space induced by considering a range of dimensions of the latent space, a range of numbers of clusters, a range of covariates and the range of possible ways in which the covariates may enter the model. Strategic searching of a reduced model space is therefore necessary.

5. Applying the mixture of experts latent position cluster model

Data on the relations between 71 lawyers in a northeastern American law firm were collected in Lazega (2001). A number of network data sets were recorded including a friendship network and a co-workers network. The friendship network was constructed by asking the lawyers to list the names of those lawyers in the firm that they socialize with outside of work. For the co-workers network, lawyers were asked to list all the lawyers in the firm with

whom they had worked during the previous year. Additionally covariates associated with the lawyers were recorded; these are detailed in Table 1.

Table 1: The seven covariates and their respective levels (if categorical variables) recorded on 71 lawyers in an northeastern American law firm. The last category in each categorical covariate is treated as the baseline category in all future analyzes.

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

Interest lies in examining the underlying social structure and processes within the law firm from both the recorded network data sets and the associated covariates. The data has previously been examined in Snijders et al. (2006) and is available as part of the SIENA software (Snijders et al., 2005).

5.1. Exploring the model space

The number of possible mixture of experts latent position cluster models is very large. In addition to choosing the number of clusters and the dimension of the latent space, the covariates can enter the model in several ways, leading to a large model space.

Initially, the latent position cluster model was fitted for various numbers of clusters and latent space dimensions. Selection of the optimal number of clusters and dimension of the latent space was performed using the methods employed by Handcock et al. (2007) — the posterior model probability of each

of the competing models was considered and that with the largest posterior probability selected. In practice, each model (i.e. each specific combination of G and d) was assigned equal prior probability and conditional (on the estimated latent locations) posterior model probabilities were approximated by the Bayesian Information Criterion (Schwarz, 1978). The model with largest BIC was selected as optimal. Once an optimal choice of model was found within the class of models considered, yielding a value for G and d, the selection of covariates was addressed.

To select a candidate set of covariates for the link probabilities (2), a logistic regression model of the same form was fitted with the $\underline{z}_i$ values fixed to be their posterior mean values from the latent position cluster model analysis. For categorical covariates, a variable indicating that the values of the actor covariates are equal was used as the dyadic specific edge covariate (ie. $x_{i,j,k} = 1$ if $w_{ik} = w_{jk}$). For continuous covariates, the absolute difference and squared differences in covariate values were included as dyadic specific covariates. A backwards stepwise selection procedure was used, with model selection being determined using AIC (Akaike, 1973). This yielded a subset of the candidate variables for further analysis in the mixture of experts latent position cluster model.

In order to find a candidate set of covariates for the cluster membership model (3), a multinomial logistic regression model was fitted with the modal cluster membership values from the latent cluster position model analysis used as the response. A backwards stepwise variable selection procedure was used to select a candidate set of covariates for further analysis in the mixture of experts latent position cluster model.

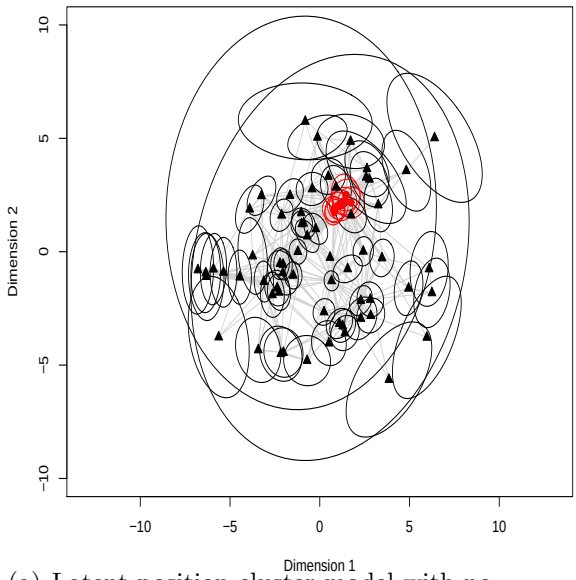
Four versions of the mixture of experts latent position cluster model were fitted for the choices of G, d and covariate selections by considering whether (i) the covariates should enter the link probabilities, (ii) the mixing proportions, (iii) both or (iv) neither. These models were assessed using AICM and the posterior summaries for the model parameters.

5.2. Lawyers friendship network

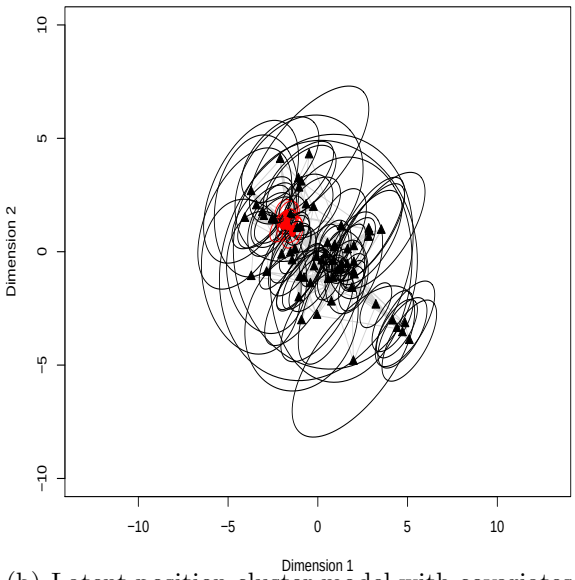
Initially, a latent position cluster model was fitted and a model with $G = 2$ and $d = 2$ was deemed optimal. A number of candidate covariates were then selected using the procedure detailed in Section 5.1. The covariates in the cluster membership model were their office and years working in the firm. The link probability covariates proposed were seniority, gender, office, practice, school and the difference in years worked in the firm. The four cases of

the mixture of experts model were fitted to the data; the estimated latent positions and their uncertainties are shown in Figure 2. The resulting parameter estimates and AICM values are reported in Table 2. The AICM criterion suggests that covariates enter the model through the link probabilities only, however we will explore and compare all of the models.

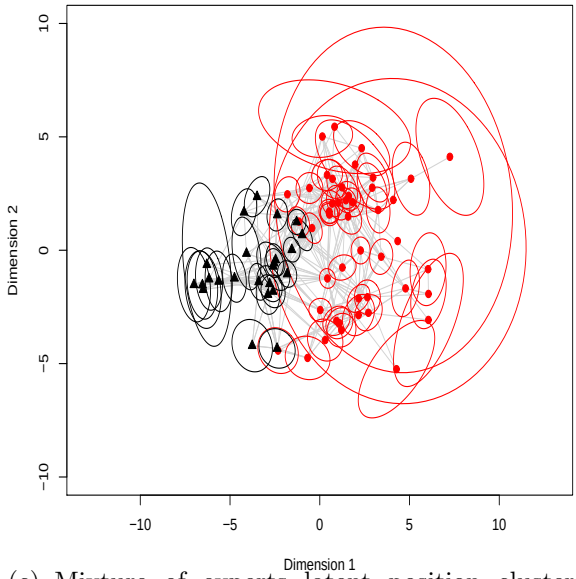
The latent positions of the actors and the cluster memberships vary considerably when comparing the models that do not have covariates that enter the link probabilities to those that do (Figures 2(a) and 2(c) versus Figures 2(b) and 2(d)). This is due to the fact that the latent positions of the actors account for any remaining network structure which has not been explained by the covariates in the link probabilities. So, the latent positions in the models without covariates in the link probabilities are explaining more of the network structure than in the other models. The clustering is quite different in the four models and the clustering structure is clearer in the models where the covariates enter the cluster membership model; this can be explained by the different role that the latent locations play in the two models.



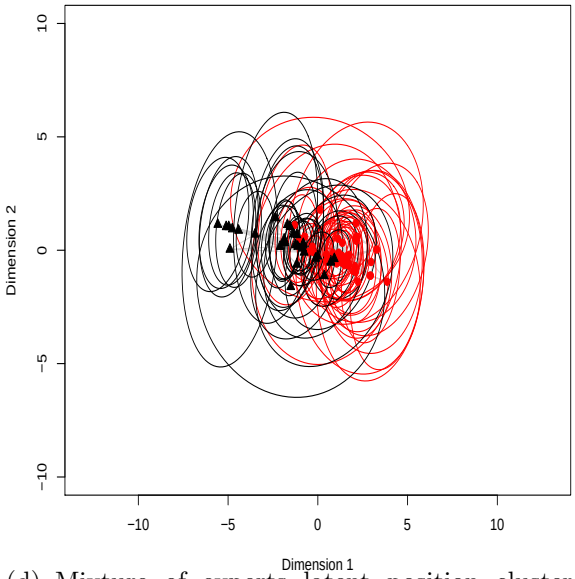
(a) Latent position cluster model with no covariates.



(b) Latent position cluster model with covariates in the link probabilities.



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



(d) Mixture of experts latent position cluster model with covariates in both the cluster memberships and the link probabilities.

Figure 2: Estimates of clusters and latent positions of the lawyers from the friendship network data. The ellipses are 50% posterior sets illustrating the uncertainty in the latent locations. Lawyers who are members of the same cluster are illustrated using the same colour and symbol. Observed links between lawyers are also illustrated.

The coefficients in the link probability function (Table 2) are consistent across model types. The coefficients suggest that lawyers with similar covariates are more likely to have relations, thus demonstrating homophily by attributes in the lawyer friendship data.

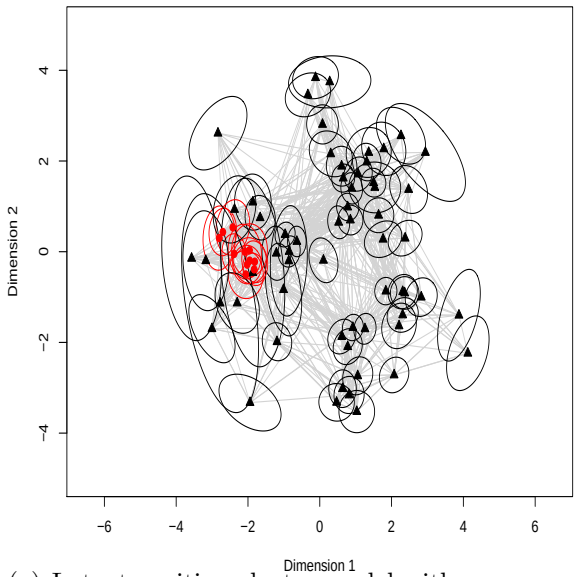
There is a considerable (Boston) office effect in determining the cluster memberships in model (c), but this does not persist in model (d). The number of years that the lawyers have been in the firm is influential in determining cluster membership in model (c) but again this does not persist in (d). The smaller cluster in model (c) appears to be a group of lawyers in the Boston office who have only spent a few years in the firm. Thus the models have given us insight to the characteristics which cause the lawyers to be friends, and have provided an view of the clustering structure within the firm.

5.3. Lawyers co-workers network

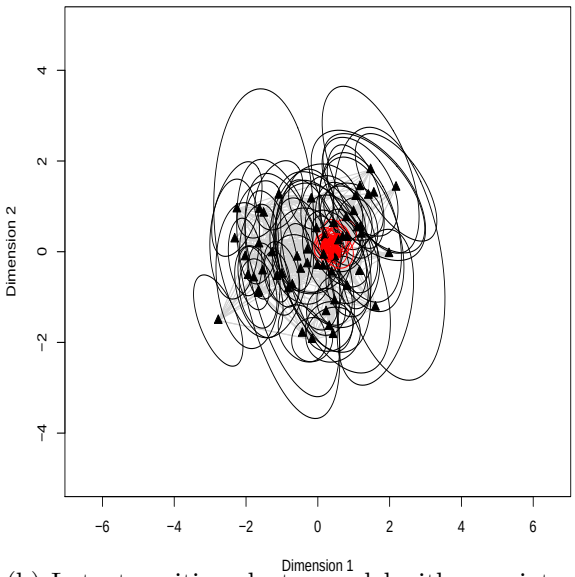
For the co-workers network the initial selection of covariates detailed in Section 5.1 suggested that a lawyer's seniority, their office, their years with the firm and their type of practice were important covariates in the cluster membership model. For the dyadic covariates in the link probabilities seniority, gender, office, (absolute difference in) age and their practice types were deemed influential. Hence, four mixture of experts latent position cluster models with $G = 2$, $d = 2$ (as selected by fitting a latent position cluster model and choosing the optimal model) and these selected covariates were fitted to the co-workers network data set. The posterior mean estimates of the latent locations and the clusters of lawyers are illustrated in Figure 3, for each of the four mixture of experts latent position cluster models. The resulting parameter estimates and AICM values for each model are detailed in Table 3.

Table 2: Posterior mean parameter estimates for the four mixture of experts models fitted to the lawyers friendship data as detailed in Figure 2. Standard deviations are given in parentheses. Note that cluster 1 was used as the baseline cluster in the case of the cluster membership parameters. Baseline categories for the covariates are detailed in Table 1.

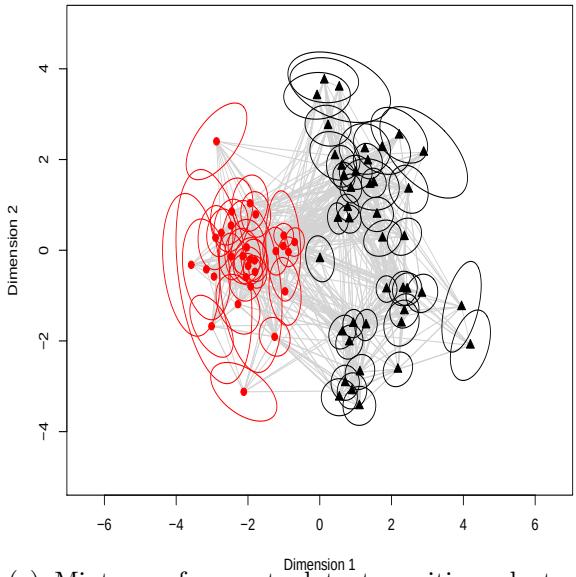
	(a)	(b)	(c)	(d)
Link Probabilities				
Intercept	1.71 (0.14)	-2.20 (0.23)	1.73 (0.13)	-2.18 (0.26)
Seniority		1.15 (0.19)		1.07 (0.20)
Gender		0.66 (0.12)		0.65 (0.12)
Office		2.38 (0.14)		2.41 (0.13)
Practice		0.78 (0.10)		0.78 (0.11)
School		0.37 (0.10)		0.35 (0.11)
Years		-0.01 (0.00)		-0.01 (0.00)
Cluster Memberships				
Intercept	1.63 (1.40)	1.43 (1.67)	1.13 (1.02)	2.11 (1.59)
Office (=1)			2.93 (1.06)	1.00 (1.21)
Office (=2)			-1.18 (1.17)	0.70 (1.48)
Years			-0.77 (0.27)	-0.57 (0.37)
Latent Space Model				
Group 1 Mean	-0.24 (0.60)	0.33 (0.64)	-3.40 (0.53)	-1.71 (0.89)
	-0.21 (0.65)	-0.27 (0.54)	-0.85 (0.66)	0.09 (1.34)
Group 1 Variance	10.59 (3.74)	6.66 (2.49)	3.95 (1.17)	5.03 (1.35)
Group 2 Mean	1.06 (0.53)	-1.35 (0.76)	1.72 (0.47)	1.36 (0.66)
	1.86 (0.79)	0.98 (0.63)	0.44 (0.50)	-0.03 (1.20)
Group 2 Variance	1.40 (3.89)	1.44 (2.53)	8.69 (1.66)	4.17 (1.06)
AICM	-2359	-2325	-2341	-2390



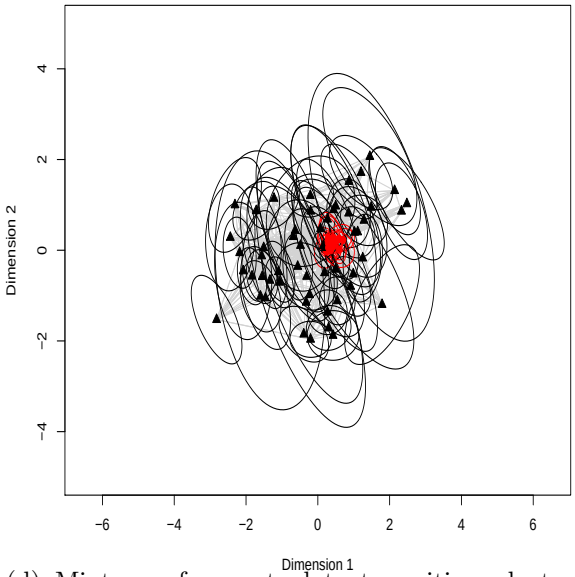
(a) Latent position cluster model with no covariates.



(b) Latent position cluster model with covariates in the link probabilities.



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



(d) Mixture of experts latent position cluster model with covariates in both the cluster memberships and the link probabilities.

Figure 3: Estimates of clusters and latent positions of the lawyers from the co-workers network data. The ellipses are 50% posterior sets illustrating the uncertainty in the latent locations. Lawyers who are members of the same cluster are illustrated using the same colour and symbol. Observed links between lawyers are also illustrated.

In the case of the co-workers network data, AICM suggests the mixture of experts latent position cluster model is optimal, where covariates enter the cluster membership probabilities. All models will be discussed however since, as indicated by Figure 3, all provide interesting insights to both the network of lawyers, and to the mixture of experts latent position cluster model itself.

As in the case of the friendship network, the posterior estimates of the latent locations are relatively consistent within models with covariates in the link probabilities (Figures 3(b) and 3(d)) and within models without covariates in the link probabilities (Figures 3(a) and 3(c)). Again it is clear that the cluster memberships do differ between the models. The two clusters are particularly distinct in the latent space (see Figure 3(c)) where the covariates enter the cluster memberships; this appears to be intuitive in that the covariates of a lawyer influence their cluster membership, and then given their cluster membership, the lawyer forms relations.

The posterior means detailed in Table 3 also provide insight to the network structure and to the mixture of experts latent position cluster model. Of note are the consistent link probability regression parameter estimates. The link probability regression parameter estimates also have intuitive interpretations; for example, the negative seniority coefficient suggests lawyers in the firm appear less likely to work together if they have the same level of seniority. This is conceivable as it is likely that a partner and an associate would work together on a case, rather than two partners working together on a case.

Also of note is that the practice covariate appears to be influential in the co-workers network cluster membership coefficients, but office does not appear to be, as it did in the case of the friendship network. As expected the practice effect has a positive influence on link probability, but it also has a strong influence on cluster membership in model (c) where covariates enter the cluster memberships. The smaller cluster in model (c) is a cluster of litigation lawyers. Thus the models have given us insight to the characteristics which cause the lawyers to be be co-workers, and have provided an view of the clustering structure within the firm.

6. Discussion

The mixture of experts latent position cluster model offers an important extension to the latent position cluster model of Handcock et al. (2007).

Table 3: Posterior mean parameter estimates for the four mixture of experts models fitted to the lawyers co-worker data as detailed in Figure 3. Standard deviations are given in parentheses. Note that cluster 1 was used as the baseline cluster in the case of the cluster membership parameters. Baseline categories for the covariates are detailed in Table 1.

	(a)	(b)	(c)	(d)
Link Probabilities				
Intercept	1.6 (0.1)	-1.46 (0.16)	1.56 (0.09)	-1.42 (0.16)
Seniority		-0.45 (0.09)		-0.49 (0.09)
Gender		0.40 (0.10)		0.41 (0.10)
Office		1.91 (0.10)		1.91 (0.10)
Practice		1.64 (0.10)		1.64 (0.09)
Age		-0.01 (0.00)		-0.00 (0.00)
Cluster Memberships				
Intercept	-1.2 (1.7)	1.41 (0.66)	-2.07 (1.17)	1.29 (0.88)
Seniority			-0.47 (0.97)	-3.20 (1.33)
Office (=1)			0.52 (0.94)	1.49 (0.87)
Office (=2)			-2.34 (1.50)	0.17 (0.84)
Practice			3.30 (0.78)	0.72 (0.59)
Years			-0.04 (0.08)	0.06 (0.05)
Latent Space Model				
Cluster 1 mean	0.47 (0.48)	-0.09 (0.19)	1.19 (0.31)	-0.11 (0.20)
	0.21 (0.69)	-0.03 (0.21)	0.26 (0.52)	-0.03 (0.20)
Cluster 1 variance	3.14 (1.45)	1.98 (0.45)	2.93 (0.76)	2.15 (0.46)
Cluster 2 mean	-1.65 (0.92)	0.40 (0.21)	-1.88 (0.48)	0.40 (0.13)
	-0.12 (0.66)	0.12 (0.21)	-0.25 (0.28)	0.13 (0.18)
Cluster 2 variance	0.92 (1.40)	0.09 (0.32)	1.29 (0.73)	0.07 (0.28)
AICM	-4074	-4095	-4066	-4117

The model has the flexibility to allow actor covariates to influence network structure in a number of ways, yielding a rich family of network models.

A major issue with this model is searching the potential models, so that an optimal model can be found. An efficient procedure is proposed in this work, but it does not guarantee finding an optimal model. In addition to this, the choice of model needs to be guided using a criterion like AICM but also the interpretation of the model outputs. The manner in which covariates enter the mixture of experts model has a major impact on the interpretation of the latent positions of the actors. If the aim of the analysis is to capture much of the network structure using the latent positions, then the covariates should enter the cluster membership model. However, if the latent positions are being used to find network structure beyond that which can be explained by the covariates, then it is important that the covariates enter the link probabilities.

The latent position cluster model in its original and mixture of experts extended form can accommodate alternative link types. In order to do this, the form of (2) would need to be changed to match the link type. This shows that this modeling framework has potential for applications beyond the analysis of binary link data.

The method of fitting the model using Markov Chain Monte Carlo is computationally expensive, but is much improved by employing surrogate proposal distributions. However, as the number of actors increases computational issues pose problems. There is considerable scope for finding more efficient model fitting procedures for these models and the development of these would greatly extend the scope of the latent position cluster methodology to the analysis of larger networks.

Acknowledgements

We would like to thank the reviewers who suggested many changes that have greatly improved this work. We would like to acknowledge Prof. Adrian Raftery for his discussions on this research. This work is partially supported by Science Foundation Ireland Research Frontiers Programme grant (06/RFP/M040) and Science Foundation Ireland Strategic Research Cluster grant (08/SRC/I1407).

A. Derivation of surrogate proposal distributions

As detailed in Section 3.3, surrogate proposal distributions are employed in the Metropolis-within-Gibbs sampler when sampling the regression parameters $\underline{\beta}$ and $\underline{\tau}_g$ for $g = 2, \ldots, G$. Such proposal distributions are derived by approximating terms within the full conditional distributions. A quadratic approximation to terms within the full conditional distributions are employed.

For clarity in the following derivations, the covariate vector associated with actor i is now defined to be the $p+1$ vector $\underline{w}_i^T = (1, w_{i1}, \ldots, w_{ip})$. Similarly, the dyadic specific covariate vector is defined to be the $p+1$ vector $\underline{x}_{i,j}^T = (1, x_{i,j,1}, \ldots, x_{i,j,p})$.

A.1. Derivation of the surrogate proposal distribution for the link probability regression parameters $\underline{\beta}$.

The full conditional distribution of the link probability regression parameters $\underline{\beta}$ is

$$\underline{\beta}|\ldots \;\propto\; \mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\}\, \mathrm{MVN}(\underline{\beta} \,|\, \underline{\xi},\ \Psi). \tag{7}$$

Let $d_{ij} = \| \underline{z}_i - \underline{z}_j \|$. Then

$$\begin{aligned}
\mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} &= \prod_{i=1}^{n} \prod_{i \neq j} \left\{ \frac{\exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})} \right\}^{y_{i,j}} \left\{ \frac{1}{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})} \right\}^{1 - y_{i,j}} \\
\log \mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} &= \sum_{i=1}^{n} \sum_{j \neq i} y_{i,j} (\underline{\beta}^T \underline{x}_{i,j} - d_{ij}) - \sum_{i=1}^{n} \sum_{j \neq i} \log \left\{ 1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij}) \right\}.
\end{aligned}$$

Examination of the form of the (log) likelihood combined with the form of the multivariate normal prior does not clearly suggest the distributional form of a good proposal distribution, or its associated parameters for $\underline{\beta}$.

Within the log likelihood, we can approximate the term

$$g(\underline{\beta}) = -\sum_{i=1}^{n} \sum_{j \neq i} \log\{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})\}$$

via a quadratic multivariate Taylor expansion (6) in $\underline{\beta}$ about $\bar{\underline{\beta}}$:

$$g(\underline{\beta}) \approx g(\bar{\underline{\beta}}) + (\underline{\beta} - \bar{\underline{\beta}})^T \{\nabla g(\bar{\underline{\beta}})\} + 0.5 (\underline{\beta} - \bar{\underline{\beta}})^T \{\nabla^2 g(\bar{\underline{\beta}})\} (\underline{\beta} - \bar{\underline{\beta}}).$$

671 It follows that

$$\nabla g(\underline{\beta}) \;=\; -\sum_{i=1}^{n}\sum_{j\neq i}\frac{\underline{x}_{i,j}\exp(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})}{1+\exp(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})}$$

672 and

$$\nabla^2 g(\underline{\beta}) \;=\; -\sum_{i=1}^{n}\sum_{j\neq i}\underline{x}_{i,j}\frac{\exp(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})}{[1+\exp(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})]^2}\underline{x}_{i,j}^T$$

673 Substituting the quadratic approximation of $g(\underline{\beta})$ into the log likelihood gives

$$\begin{aligned}\log \mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\} \;\approx\; & \sum_{i=1}^{n}\sum_{j\neq i}y_{i,j}(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})-\sum_{i=1}^{n}\sum_{j\neq i}\log\{1+\exp(\underline{\beta}^T\underline{x}_{i,j}-d_{ij})\}\\ & -(\underline{\beta}-\bar{\underline{\beta}})^T\left[\sum_{i=1}^{n}\sum_{j\neq i}\frac{\underline{x}_{i,j}\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}{1+\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}\right]\\ & -0.5(\underline{\beta}-\bar{\underline{\beta}})^T\left[\sum_{i=1}^{n}\sum_{j\neq i}\underline{x}_{i,j}\frac{\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}{[1+\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})]^2}\underline{x}_{i,j}^T\right](\underline{\beta}-\bar{\underline{\beta}})\end{aligned}$$

674 when in practice $\bar{\underline{\beta}}$ is the previously sampled value of $\underline{\beta}$. Ignoring terms con-
675 stant in $\underline{\beta}$ and introducing the term $-\sum_{i=1}^{n}\sum_{j\neq i}y_{i,j}\bar{\underline{\beta}}^T\underline{x}_{i,j}$ to facilitate algebraic
676 completeness gives

$$\begin{aligned}\log \mathbf{P}\{\mathbf{Y}|\mathbf{Z},\mathbf{X},\underline{\beta}\} \;\approx\; & \sum_{i=1}^{n}\sum_{j\neq i}y_{i,j}(\underline{\beta}-\bar{\underline{\beta}})^T\underline{x}_{i,j}\\ & -(\underline{\beta}-\bar{\underline{\beta}})^T\left[\sum_{i=1}^{n}\sum_{j\neq i}\frac{\underline{x}_{i,j}\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}{1+\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}\right]\\ & -0.5(\underline{\beta}-\bar{\underline{\beta}})^T\left[\sum_{i=1}^{n}\sum_{j\neq i}\underline{x}_{i,j}\frac{\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})}{[1+\exp(\bar{\underline{\beta}}^T\underline{x}_{i,j}-d_{ij})]^2}\underline{x}_{i,j}^T\right](\underline{\beta}-\bar{\underline{\beta}})\end{aligned}$$

which is a quadratic function in $(\underline{\beta}-\bar{\underline{\beta}})$. The quadratic form of the approximated log likelihood suggests a multivariate normal distribution with

covariance matrix

$$\Sigma_L = \left[\sum_{i=1}^{n} \sum_{j \neq i} \underline{x}_{i,j} \frac{\exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})}{[1 + \exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})]^2} \underline{x}_{i,j}^T \right]^{-1}$$

and mean

$$\underline{\mu}_L = \Sigma_L \left[\sum_{i=1}^{n} \sum_{j \neq i} \underline{x}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})} \right\} \right].$$

The prior on $\underline{\beta}$ is MVN($\underline{\xi}, \Psi$), meaning that $(\underline{\beta} - \underline{\bar{\beta}}) \sim$ MVN($\underline{\xi} - \underline{\bar{\beta}}, \Psi$). Hence the sum of the approximated log likelihood and the log of the prior distribution on $(\underline{\beta} - \underline{\bar{\beta}})$ is the sum of (the log of) two multivariate normal distributions, suggesting that the distribution of $(\underline{\beta} - \underline{\bar{\beta}})$ is approximately MVN($\underline{\delta}$, $\Delta_{\underline{\bar{\beta}}}$) where

$$\Delta_{\underline{\bar{\beta}}} = \left[\left\{ \sum_{i=1}^{n} \sum_{j \neq i} \underline{x}_{i,j} \frac{\exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})}{[1 + \exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})]^2} \underline{x}_{i,j}^T \right\} + \Psi^{-1} \right]^{-1} \tag{8}$$

and

$$\underline{\delta} = \Delta_{\underline{\bar{\beta}}} \left[\sum_{i=1}^{n} \sum_{j \neq i} \underline{x}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})} \right\} + \Psi^{-1} (\underline{\xi} - \underline{\bar{\beta}}) \right].$$

Thus the full conditional distribution of $\underline{\beta}$ (7) can be approximated by a multivariate normal distribution with covariance matrix $\Delta_{\underline{\bar{\beta}}}$ and mean

$$\underline{\delta}_{\underline{\beta}} = \underline{\bar{\beta}} + \Delta_{\underline{\bar{\beta}}} \left[\sum_{i=1}^{n} \sum_{j \neq i} \underline{x}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{x}_{i,j} - d_{ij})} \right\} + \Psi^{-1} (\underline{\xi} - \underline{\bar{\beta}}) \right] \tag{9}$$

A multivariate normal distribution with mean $\underline{\delta}_{\underline{\bar{\beta}}}$ and covariance matrix $\Delta_{\underline{\bar{\beta}}}$ is therefore employed as a surrogate proposal distribution in the Metropolis-Hastings step to sample $\underline{\beta}$ of the Metropolis-within-Gibbs sampler.

A.2. Derivation of the surrogate proposal distribution for the cluster membership regression parameters $\underline{\tau}_g$ for $g = 2, \ldots, G$.

The full conditional distribution of the mixing proportions regression parameters $\underline{\tau}_g$ is

$$\underline{\tau}_g | \ldots \quad \propto \quad \mathbf{P}\{\underline{K} = (g, \ldots, g) | \underline{\tau}_g, \mathbf{W}\} \, \mathrm{MVN}(\underline{\tau}_g \,|\, \underline{\gamma}, \, \Phi). \tag{10}$$

Then

$$\log \mathbf{P}\{\underline{K} = (g, \ldots, g) | \underline{\tau}_g, \mathbf{W}\} \quad = \quad \sum_{i=1}^{n} \mathbf{1}\{K_i = g\} (\underline{\tau}_g^T \underline{w}_i) - \sum_{i=1}^{n} \log \left(\sum_{g'=1}^{G} \exp(\underline{\tau}_{g'}^T \underline{w}_i) \right)$$

Examination of the form of the (log) likelihood combined with the form of the multivariate normal prior does not clearly suggest the distributional form of a good proposal distribution, or its associated parameters for $\underline{\tau}_g$.

Within the log likelihood, we can approximate the term

$$g(\underline{\tau}_g) = - \sum_{i=1}^{n} \log \left(\sum_{g'=1}^{G} \exp(\underline{\tau}_{g'}^T \underline{w}_i) \right)$$

via a quadratic multivariate Taylor expansion (6) in $\underline{\tau}_g$ about $\bar{\underline{\tau}}_g$:

$$g(\underline{\tau}_g) \approx g(\bar{\underline{\tau}}_g) + (\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \{\nabla g(\bar{\underline{\tau}}_g)\} + 0.5 (\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \{\nabla^2 g(\bar{\underline{\tau}}_g)\} (\underline{\tau} - \bar{\underline{\tau}}_g).$$

It follows that

$$\nabla g(\underline{\tau}_g) \quad = \quad - \sum_{i=1}^{n} \frac{\underline{w}_i \exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\tau}_{g'}^T \underline{w}_i)}$$

and

$$\nabla^2 g(\underline{\tau}_g) \quad = \quad - \sum_{i=1}^{n} \underline{w}_i \frac{\exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\tau}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\tau}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T$$

Substituting the quadratic approximation of $g(\underline{\tau}_g)$ into the log likelihood gives that $\log \mathbf{P}\{\underline{K} = (g, \ldots, g) | \underline{\tau}_g, \mathbf{W}\}$ is

$$\approx \quad \sum_{i=1}^{n} \mathbf{1}\{K_i = g\} (\underline{\tau}_g^T \underline{w}_i) - \sum_{i=1}^{n} \log \left(\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i) \right)$$

$$-(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \left\{ \sum_{i=1}^{n} \frac{\underline{w}_i \exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right\}$$
$$-0.5(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \left\{ \sum_{i=1}^{n} \underline{w}_i \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right\} (\underline{\tau} - \bar{\underline{\tau}}_g)$$

when in practice $\bar{\underline{\tau}}_g$ is the previously sampled value of $\underline{\tau}_g$. Ignoring terms constant in $\underline{\tau}_g$ and introducing the term $-\sum_{i=1}^{n} \mathbf{1}\{K_i = g\}(\bar{\underline{\tau}}_g^T \underline{w}_i)$ to facilitate algebraic completeness gives that $\log \mathbf{P}\{\underline{K} = (g, \ldots, g) | \underline{\tau}_g, \mathbf{W}\}$ is

$$\begin{aligned} \approx \quad & \sum_{i=1}^{n} \mathbf{1}\{K_i = g\}(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \underline{w}_i \\ & -(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \left\{ \sum_{i=1}^{n} \frac{\underline{w}_i \exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right\} \\ & -0.5(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \left\{ \sum_{i=1}^{n} \underline{w}_i \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right\} (\underline{\tau} - \bar{\underline{\tau}}_g) \end{aligned}$$

which is a quadratic function in $(\underline{\tau}_g - \bar{\underline{\tau}}_g)$. The quadratic form of the approximated log likelihood suggests a multivariate normal distribution with covariance matrix

$$\Sigma_L = \left[\sum_{i=1}^{n} \underline{w}_i \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right]^{-1}$$

and mean

$$\underline{\mu}_L = \Sigma_L \left[\sum_{i=1}^{n} \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\bar{\underline{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i)} \right\} \right].$$

The prior on $\underline{\tau}_g$ is MVN$(\underline{\gamma}, \Phi)$, meaning $(\underline{\tau}_g - \bar{\underline{\tau}}_g) \sim$ MVN$(\underline{\gamma} - \bar{\underline{\tau}}_g, \Phi)$. Hence the sum of the approximated log likelihood and the log of the prior distribution on $(\underline{\tau}_g - \bar{\underline{\tau}}_g)$ is the sum of (the log of) two multivariate normal distributions, suggesting that the distribution of $(\underline{\tau}_g - \bar{\underline{\tau}}_g)$ is also approximately MVN$(\underline{\delta}, \Delta_{\bar{\underline{\tau}}_g})$ where

$$\Delta_{\underline{\bar{\tau}}_g} = \left[\sum_{i=1}^{n} \left\{ \underline{w}_i \frac{\exp(\underline{\bar{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\bar{\tau}}_{g'}^T \underline{w}_i)} \left(1 - \frac{\exp(\underline{\bar{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\bar{\tau}}_{g'}^T \underline{w}_i)} \right) \underline{w}_i^T \right\} + \Phi^{-1} \right]^{-1} \quad (11)$$

and

$$\underline{\delta} \;=\; \Delta_{\underline{\bar{\tau}}_g} \left[\sum_{i=1}^{n} \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\underline{\bar{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\bar{\tau}}_{g'}^T \underline{w}_i)} \right\} + \Phi^{-1}(\underline{\gamma} - \underline{\bar{\tau}}_g) \right].$$

Thus the full conditional distribution of $\underline{\tau}_g$ (10) can be approximated by a multivariate normal distribution with covariance matrix $\Delta_{\underline{\bar{\tau}}_g}$ and mean

$$\underline{\delta}_{\underline{\bar{\tau}}_g} \;=\; \underline{\bar{\tau}}_g + \Delta_{\underline{\bar{\tau}}_g} \left[\sum_{i=1}^{n} \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\underline{\bar{\tau}}_g^T \underline{w}_i)}{\sum_{g'=1}^{G} \exp(\underline{\bar{\tau}}_{g'}^T \underline{w}_i)} \right\} + \Phi^{-1}(\underline{\gamma} - \underline{\bar{\tau}}_g) \right] \quad (12)$$

Thus a multivariate normal distribution with mean $\underline{\delta}_{\underline{\bar{\tau}}_g}$ and covariance matrix $\Delta_{\underline{\bar{\tau}}_g}$ is employed as a surrogate proposal distribution in the Metropolis-Hastings step of the Metropolis-within-Gibbs sampler to sample $\underline{\tau}_g$.

References

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

Banfield, J. D. and Raftery, A. E., Model-based Gaussian and non-Gaussian clustering. Biometrics. 49 (1993) 803–821.

Carlin, B. P. and Louis, T. A., Bayes and empirical bayes methods for data analysis. Chapman & Hall, New York, 2000.

Celeux, G. and Hurn, M. and Robert, C.P., Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association., 95 (2000) 957–970

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

Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Ser. B., 39 (1977) 1–38.

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

Fraley, C. and Raftery, A. E., Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 97 (2002) 611-631

Geman, S. and Geman, D., Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 6 (1984) 721–741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J.,Markov chain Monte Carlo in practice. Chapman & Hall, London, 1996

Gormley, I. C. and Murphy, T. B., Discussion of Handcock et al.'Model-based clustering for social networks'. Journal of the Royal Statistical Society, Series A. 170 (2007) 327–327.

Gormley, I. C. and Murphy, T. B., Discussion of Raftery et al.' Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity'. In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.), Bayesian Statistics 8, Oxford University Press, 2007, pp. 38–40.

Gormley, I. C. and Murphy, T. B., A Mixture of Experts Model for Rank Data with Applications in Election Studies. The Annals of Applied Statistics. 2 (2008) 1452–1477.

Gormley, I. C. and Murphy, T. B., Clustering ranked preference data using sociodemographic covariates. In: *Choice Modelling: The State-of-the-Art and the State-of-Practice*. S. Hess and A. Daly (Eds). Emerald (2009) In Press.

Gormley, I. C. and Murphy, T. B., A grade of membership model for rank data. Bayesian Analysis. 4 (2009b) 265–296.

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

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

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

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

Hurn, M., Justel, A. and Robert, C. P., Estimating Mixtures of Regressions. Journal of Computational and Graphical Statistics. 12 (2003) 55–79.

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

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

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

Krivitsky, P. N. and Handcock, M. S., Fitting Position Latent Cluster Models for Social Networks with latentnet. Journal of Statistical Software. 24 (2008) 1–23.

Krzanowski, W. J., Principles of Multivariate Analysis: A User's Perspective., Clarendon Press, 1988.

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

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

McCullagh, P. and Nelder, J.A., Generalized Linear Models. Chapman and Hall, London, 1983.

Metropolis, N., Rosenbluth A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. Equations of state calculations by fast computing machine. Journal of Chemical Physics. 21 (1953) 1087 – 1091.

O'Hagan, A. and Forster, J. Kendall's Advanced Theory of Statistics: Volume 2B Bayesian Inference. Arnold, London, UK, 2004.

Peng, F. and Jacobs, R. A. and Tanner, M. A., Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With and Application to Speech Recognition. Journal of the American Statistical Association. 91 (1996) 953–960.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N., Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. In: *Bayesian Statistics 8*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.), Oxford University Press (2007), pp. 1–45.

Richardson, S. and Green, P. J., On Bayesian Analysis of Mixtures With An Unknown Number of Components., Journal of the Royal Statistical Society, Series B. 59 (1997) 731–758.

Salter-Townshend, M. and Murphy, T. B., Variational Bayesian inference for the Latent Position Cluster Model. (2009) *Submitted*.

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

Snijders, T. A. B., Christian E. G. S., Schweinberger, M. and Huisman, M., Manual for SIENA version 2.1. Groningen, Netherlands (2005)

Snijders, T.A., Pattison, P.E., Robins, G.L. and Handcock, M.S., New specifications for exponential random graph models. Sociological Methodology. (2006) 99–153

Stephens, M., Dealing with label-switching in mixture models. Journal of the Royal Statistical Society, Series B., 62 (2000) 795–810.

Wainwright, M. J. and Jordan, M. I., Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends® in Machine Learning., 1 (2008) 1–305.