Clustering South African Households Based on their Asset Status Using Latent Variable Models.

Damien McParland
University College Dublin

Isobel Claire Gormley
University College Dublin

Samuel J. Clark
University of Washington
University of the Witwatersrand
INDEPTH Network

Tyler H. McCormick
University of Washington

Chodziwadziwa Whiteson Kabudula
University of the Witwatersrand

Mark A. Collinson
University of the Witwatersrand
INDEPTH Network

Working Paper no. 121
Center for Statistics and the Social Sciences
University of Washington

October 10, 2012

Abstract

The Agincourt Health and Demographic Surveillance System has since 2001 conducted a biannual household asset survey in order to quantify household socio-economic status (SES) in a rural population living in northeast South Africa. The data contain binary, ordinal and nominal items. We aim to describe the SES landscape in the study population by clustering the households into homogeneous groups based on their asset status.

A model-based approach to clustering, based on latent variable models, is proposed. In the case of modeling binary or ordinal items, item response models are employed. For nominal survey items, a factor analysis model, similar in nature to a multinomial probit model, is used. Both model types have an underlying latent variable structure – this similarity is exploited and the models are combined to produce a hybrid model capable of handling mixed data types. Further, a mixture of the hybrid models is considered to provide clustering capabilities within the context of mixed binary, ordinal and nominal response data. The proposed model is termed the mixture for factor analyzers for mixed data (MFA-MD).

The MFA-MD model is applied to the SES data to cluster households into homogeneous groups. The model is estimated within the Bayesian paradigm, using a Markov chain Monte Carlo algorithm. Intuitive groupings result providing insight to the different socio-economic strata within the Agincourt region.

KEY WORDS: clustering; mixed data; item response theory; Metropolis-within-Gibbs.

1 Introduction

Categorical data arise in many different fields and are typically collected as responses to a number of questions or survey items. These responses could be either binary, ordinal or nominal in nature. A model is presented here which facilitates clustering of observations in the context of such mixed categorical data. Latent variable modeling ideas are used as the observed response is viewed as a categorical manifestation of a latent continuous variable(s).

The proposed clustering method is motivated and illustrated by a study of household socio-economic status conducted by the Agincourt health and demographic surveillance system (HDSS) in rural northeast South Africa [30]. Asset-based wealth indices are a common way of quantifying wealth in populations for which alternative methods are not feasible [48]. Households in the study area have been surveyed biannually since 2001 to elicit an accounting of assets similar to that used by the Demographic and Health Surveys [44] to construct a wealth index. We use the most recent measurement for each household. The resulting dataset contains binary, ordinal and nominal items. Examples of typical survey items include whether or not a household owns a television, or the type of toilet facilities used by a household. Based on the asset status data, we are interested in exploring any clustering structure within the set of Agincourt households, and in uncovering whether underlying clusters relate to different socio-economic strata. Hence a clustering method which can appropriately model the mixed categorical response data is required.

Several models for clustering mixed data are detailed in the literature. Early work on modeling such data employed the location model [26, 31, 49], in which the joint distribution of mixed data is decomposed as the product of the marginal distribution of the categorical variables and the conditional distribution of the continuous variables, given the categorical variables. More recently, [27] re-examined these location models in the presence of missing data. Latent factor models in particular have generated recent interest for modeling mixed data; [41] uses such models in a political science context. [11, 12] and [38] provide an early view of clustering mixed data, including the use of latent variable models. More recently [5], [4] and [21] propose clustering models for mixed data based on a latent variable. The clustering model proposed here is also based on latent variable models, but differs from existing models in that it proposes a unifying latent variable framework by elegantly combining ideas from item response theory and from factor analysis models for nominal data.

Item response modeling is an established method for analyzing binary or ordinal response data. First introduced by [46], item response theory has its roots in educational testing. Many authors have contributed to the expansion of this theory since then including [32, 43], [33] and [47]. Extensions include the graded response model [45] and the partial credit model [34]. Bayesian approaches to fitting such models are detailed in [29] and [15]. Item response models assume that each observed ordinal response is a manifestation of a latent continuous variable. The observed response will be level k, say, if the latent continuous variable lies within a specific interval. Further, typical item response models assume that the latent continuous variable is a function of both a respondent specific latent trait variable and item specific parameters.

Modeling nominal response data is typically more complex than modeling binary or ordinal data as the set of possible responses is unordered. A popular model for nominal choice data is the multinomial probit (MNP) model [18]. Bayesian approaches to fitting the MNP

model have been proposed by [2, 35, 39] and [8]. The model has also been extended to include multivariate nominal responses by [51]. Typically, the MNP model models nominal response data as a manifestation of an underlying multidimensional continuous latent variable, which depends on a respondent's covariate information and some item specific parameters. Here a factor analysis model for nominal data, similar in nature to the MNP model, is proposed where the observed nominal response is a manifestation of the multidimensional latent variable which is itself modeled as a function of both a respondent's latent trait variable and some item specific parameters.

The similarities between item response models and the factor analytic version of the MNP model suggest a hybrid model would be advantageous. Both models have a latent variable structure underlying the observed data, which exhibits dependence on item specific parameters. Further, the latent variable in both models has an underlying factor analytic structure, through the dependency on the latent trait. This similarity is exploited and the models are combined to produce a hybrid model capable of modeling mixed categorical data types. This hybrid model can be thought of as a factor analysis model for mixed data.

As stated, the motivating application involves the need to cluster observations based on mixed categorical data. A model-based approach to clustering is proposed, in that a mixture modeling framework provides the clustering machinery. Specifically, a mixture of the factor analytic models for mixed data is considered to provide clustering capabilities within the context of mixed binary, ordinal and nominal response data. The resulting model is termed the mixture of factor analyzers for mixed data (MFA-MD). Here, a Bayesian framework is employed for inference.

The paper proceeds as follows. Background information about the Agincourt region of South Africa as well as the socio-economic status (SES) survey and resulting data set are introduced in Section 2. Item response models, a model for nominal response data and the amalgamation and extension of these models to a MFA-MD model are considered in Section 3. Section 4 is concerned with Bayesian model estimation and inference. The results from fitting the model to the SES data set are presented in Section 5. Finally, discussion of the model and future research areas takes place in Section 6.

2 The Agincourt HDSS SES Data Set

The Agincourt HDSS [30] continuously monitors the population of 21 villages located in the Bushbuckridge subdistrict of Mpumalanga Province in northeast South Africa. This is a rural population living in what was during Apartheid a black 'homeland'. The Agincourt HDSS was established in the early 1990s with the purpose of guiding the reorganization of South Africa's health system. Since then the goals of the HDSS have evolved and now it contributes to evaluation of national policy at population, household and individual levels. For more information on the AHDSS and how the data were collected see `www.agincourt.co.za`.

The HDSS covers and area of 420km^2 consisting of 21 villages with a total population of approximately $82,000$ people. The infrastructure in the area is mixed. The roads in and surrounding the study area are in the process of rapidly being upgraded from dirt to tar. The cost of electricity is prohibitively high for many households though it is available in all villages. A dam has been constructed nearby but to date there is no piped water to dwellings

and sanitation is rudimentary. The soil in the area is generally suited to game farming and there is virtually no commercial farming activity. Most households contain wage earners who purchase maize and other foods which they then supplement with home-grown crops and collected wild foodstuffs.

The dataset analyzed here describes assets of households in the Agincourt study area. The data consist of the responses of $N = 17,617$ households to each of $J = 28$ categorical survey items. There are 22 binary items, 3 ordinal items and 3 nominal items. The binary items are asset ownership indicators for the most part. These items record whether or not a household owns a particular asset (e.g. whether or not they own a working car). An example of an ordinal item is the type of toilet the household uses. This follows an ordinal scale from no toilet at all to a modern flush toilet. Finally, the power used for cooking is an example of a nominal item. The household may use electricity, bottled gas or wood, among others. This is an unordered set since a particular response may not be indicative of socio-economic status. A full list of survey items and possible responses is given in Appendix A.

Previous analyses of data from this study include a principal component analysis-based approach by [13] in which the relationship between educational enrollment and wealth is investigated. Further, [9] construct an asset index for each household which is then used to divide households into clusters based on the quintiles of this index.

3 A Mixture of Factor Analysers Model for Mixed Data

A mixture of factor analyzers model for mixed data (MFA-MD) is proposed to facilitate clustering of the Agincourt households. Each component of the MFA-MD model is a hybrid of an item response model and a factor analytic model for nominal data. In this section item response models for ordinal data and a latent variable model for nominal data are introduced, before they are combined and extended to the MFA-MD model.

3.1 Item Response Models for Ordinal Data

Suppose item j (for $j = 1, \ldots, J$) is ordinal and the set of possible responses is denoted $\{1, 2, \ldots, K_j\}$ where K_j denotes the number of response levels to item j. Item response models assume that, for respondent i, a latent Gaussian variable z_{ij} corresponds to each categorical response y_{ij}. Typically, a Gaussian link function is assumed, though other link functions, such as the logit, are detailed in the item response theory literature [15, 33].

For each ordinal item j there exists a vector of threshold parameters $\underset{\sim}{\gamma}_j = \left(\gamma_{j,0}, \gamma_{j,1}, \ldots, \gamma_{j,K_j}\right)$. The elements of this vector are constrained such that

$$-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \ldots \leq \gamma_{j,K_j} = \infty.$$

For identifiability reasons [2, 41] typically $\gamma_{j,1} = 0$. The observed ordinal response, y_{ij}, for respondent i is a manifestation of the latent variable z_{ij} i.e.

$$\text{if } \gamma_{j,k-1} \leq z_{ij} \leq \gamma_{j,k} \quad \text{then} \quad y_{ij} = k.$$

That is, if the underlying latent continuous variable lies within a interval bounded by the threshold parameters $\gamma_{j,k-1}$ and $\gamma_{j,k}$, then the observed ordinal response is level k.

In a standard item response model, a factor analytic model is then used to model the underlying latent variable z_{ij}. It is assumed that the mean of the conditional distribution of z_{ij} depends on a q dimensional, respondent specific, latent variable $\underline{\theta}_i$ and on some item specific parameters. The latent variable $\underline{\theta}_i$ is sometimes referred to as the latent trait or a respondent's ability parameter in item response theory. Specifically, the underlying latent variable z_{ij} for respondent i and item j is assumed to be distributed as

$$z_{ij}|\underline{\theta}_i \sim N(\mu_j + \underline{\lambda}_j^T \underline{\theta}_i, 1).$$

The parameters $\underline{\lambda}_j$ and μ_j are usually termed the item discrimination parameters and the negative item difficulty parameter respectively. Typically, the variance of z_{ij} is fixed at 1[2].

Under this model, the conditional probability that a response takes a certain ordinal value can then be expressed as the difference between two standard Gaussian cumulative distribution functions:

$$P(y_{ij} = k|\underline{\lambda}_j, \mu_j, \underline{\gamma}_j, \underline{\theta}_i) = \Phi[\gamma_{j,k} - (\mu_j + \underline{\lambda}_j^T \underline{\theta}_i)] - \Phi[\gamma_{j,k-1} - (\mu_j + \underline{\lambda}_j^T \underline{\theta}_i)].$$

Since a binary item can be viewed as an ordinal item with two levels (0 and 1, say) the item response model can also be used to model binary response data. The threshold parameter for a binary item j is $\underline{\gamma}_j = (-\infty, 0, \infty)$ and hence for a binary item

$$P(y_{ij} = 1|\underline{\lambda}_j, \mu_j, \underline{\gamma}_j, \underline{\theta}_i) = \Phi\left(\mu_j + \underline{\lambda}_j^T \underline{\theta}_i\right).$$

3.2 A Factor Analytic Model for Nominal Data

Modeling nominal response data is typically more complicated than modeling ordinal data since the set of possible responses is no longer ordered. The set of nominal responses for item j is denoted $\{1, 2, \ldots, K_j\}$ such that 1 corresponds to the first response choice while K_j corresponds to the last response choice, but where no inherent ordering among the choices is assumed.

As detailed in Section 3.1, the item response model for ordinal data posits a one dimensional latent variable for each observed ordinal response. In the factor analytic model for nominal data proposed here, a $K_j - 1$ dimensional latent variable is required for each observed nominal response. That is, the latent variable for observation i corresponding to nominal item j is denoted

$$\underline{z}_{ij} = (z_{ij}^1, \ldots, z_{ij}^{K_j - 1}).$$

The observed nominal response is then assumed to be a manifestation of the values of the elements of $\underline{z}_{ij}$ relative to each other and to a cut-off point, assumed to be 0. That is,

$$y_{ij} = \begin{cases} 1 & \text{if } \max_k\{z_{ij}^k\} < 0; \\ \\ k & \text{if } z_{ij}^{k-1} = \max_k\{z_{ij}^k\} \text{ and } z_{ij}^{k-1} > 0 \quad \text{for } k = 2, \ldots, K_j. \end{cases}$$

Similar to the item response model, the underlying latent vector $\underline{z}_{ij}$ is modeled via a factor analytic model. The mean of the conditional distribution of $\underline{z}_{ij}$ depends on a respondent specific, q-dimensional, latent trait variable, $\underline{\theta}_i$, and item specific parameters i.e.

$$\underline{z}_{ij}|\underline{\theta}_i \sim \text{MVN}_{K_j-1}(\underline{\mu}_j + \Lambda_j\underline{\theta}_i, \mathbf{I})$$

where $\mathbf{I}$ denotes the identity matrix. The loadings matrix Λ_j is a $(K_j - 1) \times q$ matrix, analogous to the item discrimination parameter in the item response model of Section 3.1; likewise, the mean vector $\underline{\mu}_j$ is analogous to the item difficulty parameter in the item response model.

3.3 A Factor Analysis Model for Mixed Data

It is clear that the item response model for ordinal data (Section 3.1) and the factor analytic model for nominal data (Section 3.2) are similar in structure. Both model the observed data as a manifestation of an underlying latent variable, which is itself modeled using a factor analytic structure. This similarity is exploited to obtain a hybrid factor analysis model for mixed binary, ordinal and nominal data.

Suppose Y, an $N \times J$ matrix of mixed data, denotes the data from N respondents to J survey items. Let O denote the number of binary items plus the number of ordinal items, leaving $J-O$ nominal items. Without loss of generality, suppose that the binary and ordinal items are in the first O columns of Y while the nominal items are in the remaining columns.

The binary and ordinal items are modeled using an item response model and the nominal items using the factor analytic model for nominal data. Therefore, for each respondent i there are O latent continuous variables corresponding to the ordinal items and $J-O$ latent continuous vectors corresponding to the nominal items. The latent variables and latent vectors for respondent i are collected together in a single D dimensional vector $\underline{z}_i$ where $D = O + \sum_{j=O+1}^{J}(K_j - 1)$. That is, underlying respondent i's set of J binary, ordinal and nominal responses lies the latent vector

$$\underline{z}_i = (z_{i1}, \ldots, z_{iO}, \ldots, z_{iJ}^1, \ldots, z_{iJ}^{K_J-1}).$$

This latent vector is then modeled using a factor analytic structure:

$$\underline{z}_i|\underline{\theta}_i \quad \sim \quad \text{MVN}_D(\underline{\mu} + \Lambda\underline{\theta}_i, \mathbf{I}). \tag{1}$$

The $D \times q$ dimensional matrix Λ is termed the loadings matrix and $\underline{\mu}$ is the D dimensional mean vector. Combining the item response and factor analytic models in this way facilitates the modeling of binary, ordinal and nominal response data in an elegant and unifying latent variable framework.

The model in (1) provides a parsimonious factor analysis model for the high-dimensional latent vector $\underline{z}_i$ which underlies the observed mixed data. As in any model which relies on a factor analytic structure, the loadings matrix details the relationship between the low dimensional latent trait $\underline{\theta}_i$ and the high-dimensional latent vector $\underline{z}_i$. Marginally the latent vector is distributed as

$$\underline{z}_i \sim \text{MVN}_D(\underline{\mu}, \Lambda\Lambda^T + \mathbf{I})$$

resulting in a parsimonious covariance structure for $\underline{z}_i$.

3.4 A Mixture of Factor Analyzers Model for Mixed Data

To facilitate clustering when the observed data are mixed categorical variables, a mixture modeling framework can be imposed on the hybrid model defined in Section 3.3. The resulting model is termed the mixture of factor analyzers model for mixed data. In the MFA-MD model, the clustering is deemed to occur at the underlying latent variable level. That is, under the MFA-MD model the distribution of the latent data $\underline{z}_i$ is modeled as a mixture of G Gaussian densities

$$f(\underline{z}_i) = \sum_{g=1}^{G} \pi_g \text{MVN}_D\left(\underline{\mu}_g,\ \Lambda_g\Lambda_g^T + \mathbf{I}_D\right). \tag{2}$$

The probability of belonging to cluster g is denoted by π_g where $\sum_{g=1}^{G}\pi_g = 1$ and $\pi_g > 0\ \forall\ g$. The mean and loading parameters $\underline{\mu}_g$ and Λ_g are specific to cluster g.

As is standard in a model-based approach to clustering [7, 16], a latent indicator variable, $\underline{\ell}_i = (\ell_{i1}, \ldots, \ell_{iG})$ is introduced for each respondent i. This binary vector indicates the cluster to which individual i belongs i.e. $l_{ig} = 1$ if i belongs to cluster g; all other entries in the vector are 0. Under the model in (2), the augmented likelihood function for the N respondents is then given by

$$\begin{aligned}\mathcal{L}(\underline{\pi}, \tilde{\Lambda}, \Gamma, Z, \Theta, L|Y) = \prod_{i=1}^{N}\prod_{g=1}^{G}\Bigg\{\pi_g\Bigg[\prod_{j=1}^{O}\prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T\tilde{\underline{\theta}}_i, 1)^{\mathbb{I}\{\gamma_{j,k-1}<z_{ij}<\gamma_{j,k}\}}\Bigg] \\ \times\Bigg[\prod_{j=O+1}^{J}\prod_{k=2}^{K_j}\prod_{s=1}^{3} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T}\tilde{\underline{\theta}}_i, 1)^{\mathbb{I}(\text{case } s)}\Bigg]\Bigg\}^{\ell_{ig}}\end{aligned} \tag{3}$$

where $\tilde{\underline{\theta}}_i = (1, \theta_{i1}, \ldots, \theta_{iq})^T$ and $\tilde{\Lambda}_g$ is the matrix resulting from the combination of $\underline{\mu}_g$ and Λ_g so that the first column of $\tilde{\Lambda}_g$ is $\underline{\mu}_g$. Thus the d^{th} row of $\tilde{\Lambda}_g$ is $\tilde{\underline{\lambda}}_{gd} = (\mu_{gd}, \lambda_{gd1}, \ldots, \lambda_{gdq})$.

The 'cases' referred to in the nominal part of the model defined in (3) are as defined as follows. For $k = 2, \ldots, K_j$, the latent variable z_{ij}^{k-1} must satisfy one of the following three cases:

- case 1: if $z_{ij}^{k-1} = \max_k\{z_{ij}^k\} < 0$ then $y_{ij} = 1$.
- case 2: if $z_{ij}^{k-1} = \max_k\{z_{ij}^k\}$ and $z_{ij}^{k-1} > 0$ then $y_{ij} = k$.
- case 3: if $z_{ij}^{k-1} < \max_k\{z_{ij}^k\}$ and $z_{ij}^{k-1} > 0$ then $y_{ij} \neq 1 \wedge y_{ij} \neq k$.

The MFA-MD model proposed here is related to the mixture of factor analyzers model [20] which is appropriate when the observed data are continuous in nature. [14] detail a Bayesian treatment of such a model; [36] detail a suite of parsimonious mixture of factor analyzer models within a maximum likelihood framework.

4 Bayesian Model Estimation

A Bayesian approach using Markov chain Monte Carlo is utilized for fitting the MFA-MD model. Interest lies in inferring the underlying latent variables Z, the latent traits Θ, the cluster membership vectors L and the model parameters i.e. the mixing proportions $\underline{\pi}$, the item parameters $\tilde{\Lambda}_g (\forall g = 1, \ldots, G)$ and the threshold parameters Γ.

4.1 Prior and Posterior Distributions

To fit the MFA-MD model in a Bayesian framework prior distributions are required for all unknown parameters. As in [2], a uniform prior is specified for the threshold parameters. Conjugate prior distributions are specified for the other model parameters:

$$p(\underline{\tilde{\lambda}}_{gd}) = \text{MVN}_{(q+1)}(\underline{\mu}_\lambda, \Sigma_\lambda) \qquad p(\underline{\pi}) = \text{Dirichlet}(\underline{\alpha})$$

In terms of latent variables, it is assumed the latent traits $\underline{\theta}_i$ follow a standard multivariate Gaussian distribution while the latent indicator variables, $\underline{\ell}_i$, follow a Multinomial$(1, \underline{\pi})$ distribution. Further, conditional on membership of cluster g, the latent variable $\underline{z}_i | l_{ig} = 1 \sim \text{MVN}_D(\underline{\mu}_g, \Lambda_g \Lambda_g^T + \mathbf{I})$. Combining these latent variable distributions and prior distributions with the likelihood function specified in (3) results in the joint posterior distribution, from which samples of the model parameters and latent variables are drawn using a Markov chain Monte Carlo sampling scheme.

4.2 Estimation via a Markov Chain Monte Carlo Sampling Scheme

As the marginal distributions of the model parameters cannot be obtained analytically a Markov chain Monte Carlo (MCMC) sampling scheme is employed. All parameters and latent variables are sampled using Gibbs sampling, with the exception of the threshold parameters Γ, which are sampled using a Metropolis-Hastings step.

The full conditional distributions for the latent variables and model parameters are detailed in what follows. Full derivations of the full conditional posterior distributions are given in Appendix B.

- Allocation vectors:
  $\underline{\ell}_i | \ldots \sim \text{Multinomial}(\underline{p})$ for $i = 1, \ldots, N$, where $\underline{p} = (p_1, \ldots, p_G)$ and p_g is defined in (4) in Appendix B.

- Latent traits:
  $\underline{\theta}_i | \ldots \sim \text{MVN}_q \left\{ \left[\Lambda_g^T \Lambda_g + \mathbf{I} \right]^{-1} \left[\Lambda_g^T \left(\underline{z}_i - \underline{\mu}_g \right) \right], \left[\Lambda_g^T \Lambda_g + \mathbf{I} \right]^{-1} \right\}$ for $i = 1, \ldots, N$.

- Mixing proportions:
  $\underline{\pi} | \ldots \sim \text{Dirichlet}(n_1 + \alpha_1, \ldots, n_g + \alpha_G)$ where $n_g = \sum_{i=1}^N \ell_{ig}$.

- Item parameters:
  $\underline{\tilde{\lambda}}_{gd} | \ldots \sim \text{MVN}_{(q+1)} \left\{ \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T \tilde{\Theta}_g \right]^{-1} \left[\tilde{\Theta}_g^T \underline{z}_{gd} + \Sigma_\lambda^{-1} \underline{\mu}_\lambda \right], \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T \tilde{\Theta}_g \right]^{-1} \right\}$ for $g =$

$1, \ldots, G$ and $d = 1, \ldots, D$, where $\underline{z}_{gd} = \{z_{id}\}$ for all respondents i in cluster g and $\tilde{\Theta}_g$ is a matrix, the rows of which are $\underline{\tilde{\theta}}_i$ for members of cluster g.

The full conditional distribution for the underlying latent data Z follows a truncated Gaussian distribution. The point(s) of truncation depends on the nature of the corresponding item and the observed response.

- If item j is ordinal and $y_{ij} = k$ then,

$$z_{ij}|\ldots \sim N^T\left(\underline{\tilde{\lambda}}_{gj}^T\underline{\tilde{\theta}}_i, 1\right)$$

  where the distribution is truncated on the interval $(\gamma_{j,k-1}, \gamma_{j,k})$.

- If item j is nominal then:

$$z_{ij}^k|\ldots \sim N^T\left(\underline{\tilde{\lambda}}_{gj}^{kT}\underline{\tilde{\theta}}_i, 1\right)$$

  where $\underline{\tilde{\lambda}}_{gj}^k$ is the row of $\tilde{\Lambda}_g$ corresponding to z_{ij}^k. The distribution is truncated as follows:
  - if $y_{ij} = 1$ then $z_{ij}^k \in (-\infty, 0)\ \forall\ k$.
  - if $y_{ij} = k > 1$ then:
    - $z_{ij}^{k-1} \in (\tau, \infty)$ where $\tau = \max\left(0, \max\limits_{l \neq k-1}\{z_{ij}^l\}\right)$.
    - for $l \neq k-1$ then $z_{ij}^l \in \left(-\infty, z_{ij}^{k-1}\right)$.

As a uniform prior is specified for the threshold parameters, the posterior full conditional distribution of $\underline{\gamma}_j$ is also uniform, facilitating the use of a Gibbs sampler. However, if there are large numbers of observations in adjacent response categories very slow mixing may be observed. If the threshold parameters are slow to converge this may result in slow convergence of other parameters also. Thus, as in [10, 15, 29] a Metropolis-Hastings step is used to sample the threshold parameters; the overall sampling scheme employed is therefore a Metropolis-within-Gibbs sampler.

Briefly, the Metropolis-Hastings step involves proposing candidate values $v_{j,k}$ (for $k = 2, \ldots, K_j - 1$) for $\gamma_{j,k}$ from the Gaussian distribution $\mathrm{N}^T(\gamma_{j,k}^{(t-1)}, \sigma_{MH}^2)$ truncated to the interval $(v_{j,k-1}, \gamma_{j,k+1}^{(t-1)})$ where $\gamma_{j,k+1}^{(t-1)}$ is the value of $\gamma_{j,k+1}$ sampled at iteration $(t-1)$. The threshold vector $\underline{\gamma}_j$ is set equal to the proposed vector, $\underline{v}_j$, with probability $\beta = \min(1, R)$ where R is defined by (5) in Appendix B. The tuning parameter σ_{MH}^2 is selected to achieve appropriate acceptance rates.

This Metropolis-within-Gibbs sampling scheme is iterated until convergence, after which the samples drawn are from the joint posterior distribution of all the model parameters and latent variables of the MFA-MD model.

4.3 Model Identifiability

The MFA-MD model as described is not identifiable. One identifiability aspect of the model concerns the threshold parameters. If a constant is added to the threshold parameters for an ordinal item j and the same constant is added to the corresponding mean parameter(s), the likelihood remains unchanged. Therefore, as outlined in Section 3.1, the second element γ_{j1} of the vector of threshold parameters, $\underline{\gamma}_j$, is fixed at 0 for all ordinal items j.

Another identifiability aspect of the model involves its factor analytic structure. Since a factor analytic structure is imposed on the underlying latent data, the model is also rotationally invariant. Many approaches to this identifiability issue have been proposed in the literature. A popular solution is that proposed by [19] where the loadings matrix is constrained such that the first q rows have a lower triangular form and the diagonal elements are positive. This approach is adopted by [41] and [14] among others. However, this approach enforces an ordering on the variables [1] which is not suitable under the MFA-MD model.

Here, the approach to identifying the MFA-MD model is based on that suggested by [25] and [23] in relation to latent space models for network data. Instead of imposing a particular form on the loadings matrices the MCMC samples are post processed using Procrustean methods. Each sampled Λ_g is rotated and/or reflected to match as closely as possible to a reference loadings matrix. The latent traits, $\underline{\theta}_i$, are then subjected to the same transformation. The sample mean of these transformed values is then used to estimate the mean of the posterior distribution.

Conditional on the cluster memberships on convergence of the MCMC chain, a factor analysis model is fitted to the underlying latent data within each cluster. The estimated loadings matrix obtained is used as the reference matrix for each cluster. Only the saved MCMC samples need to be subjected to this transformation which is done post hoc and is computationally cheap.

5 Results: fitting the MFA-MD model to the SES data

A number of models with G, the unknown number of clusters, ranging from 1 to 6 and q the dimension of the latent trait ranging from 1 to 2, were fitted to the SES data. The 3 cluster solution with 1 dimensional latent trait which is discussed here was chosen for reasons addressed in Section 5.3.

5.1 Results: Three Component MFA-MD model

The clustering resulting from fitting a 3 component MFA-MD model divides the households into 3 distinct homogeneous subpopulations, with intuitive socio-economic characteristics. The modal responses to items for which the modal response differed across groups are presented in Table 1. These statistics only tell part of the story however, and the distribution of responses will be analyzed later.

It can be seen from Table 1 that cluster 1 is a modern/wealthy group of households. The modal responses indicate that households in this cluster are most likely to possess modern conveniences such as a stove, a fridge and also some luxury items such as a television.

In contrast, cluster 3 is a less wealthy group. Households in this group are likely to have poor sanitary facilities – the modal response to the location of toilet facilities and the type of toilet are "bush" and "none" respectively. Households in cluster 3 are also less likely to own modern conveniences such as a fridge or television.

The socio-economic status of cluster 2 is somewhere between that of the other two groups, but closer to cluster 1 than 3. Households in cluster 2 are likely to have better sewage facilities and larger dwellings than those in cluster 3 but lack some luxury assets such as a video player. They are also likely to keep poultry and cook with wood rather than electricity which suggests this group may be less modern than cluster 1 to some degree.

It is interesting to note that the largest group is the wealthy/modern cluster 1 while the smallest group is cluster 3 who have the lowest living standards.

Table 1: The cardinality of each group and the modal response to items on which the modal response differs across groups.

G	1	2	3
#	7864	6543	3210
# Bedrooms	2	2	$\leq$1
Separate Living Area	Yes	Yes	No
Toilet Facilities	Yard	Yard	Bush
Toilet Type	Pit	Pit	None
Power for Cooking	Electric	Wood	Wood
Stove	Yes	No	No
Fridge	Yes	Yes	No
Television	Yes	Yes	No
Video	Yes	No	No
Poultry	No	Yes	No

A more detailed picture of how the groups differ from each other is presented in Figures 1 and 2. Box plots of the MCMC samples of the cluster specific mean parameter $\underline{\mu}_g$ are shown in these figures. The box plots for the binary/ordinal items (Figure 1) have a different interpretation to those for the nominal items (Figure 2). The binary and ordinal responses have been coded with the convention that larger responses correspond to greater wealth. Thus a higher mean value for the latent data corresponding to these items is indicative of greater wealth. To interpret the box plots for the nominal items all latent dimensions for a particular item must be considered. If the mean of one dimension (k, say) is greater than the means of the others for a particular cluster, then the response corresponding to dimension k is the most likely response within that cluster. If the means for all dimensions for a particular item are less than 0 then the most likely response by households in that cluster is the first choice.

The box plots corresponding to the binary and ordinal items are shown in Figure 1. The elements of the mean of cluster 1 (the wealthy/modern cluster) $\underline{\mu}_1$ can be seen to be greater than those for the other clusters in general; this reflects the greater wealth observed in cluster 1 compared to the other groups. Similarly the elements of $\underline{\mu}_3$ (the least wealthy group) are lower than those for the other groups reflecting the lower socio economic status of households

in cluster 3. The difference between cluster 3 and clusters 1 and 2 is particularly stark on the location of toilet facilities (*ToiletFac*) and the type of toilet facilities (*ToiletType*) items. The means for clusters 1 and 2 are notably higher than the mean for cluster 3 since the responses for groups 1 and 2 are typically a number of categories higher on these items.

Figure 2 shows box plots of the MCMC samples of the dimensions of the cluster mean parameters, $\underline{\mu}_g$, corresponding to the nominal items. Focusing on the latent dimensions corresponding to the *PowerCook* item, say, it can be seen that the highest mean for cluster 1 is on the 'electricity' dimension followed closely by the 'wood' dimension, and that these means are greater than 0. This implies that the most likely response to the *PowerCook* item for cluster 1 is electricity but that a significant proportion of the households in this group cook with wood. The highest means for clusters 2 and 3 are on the 'wood' latent dimension. Thus most of the households in these clusters cook with wood in contrast to the wealthy/modern cluster 1. This difference is indicative of a socio-economic divide. In a similar way, the mean parameters for the *PowerLight* item suggest that electricity is the most likely source of power for lighting for households in all clusters; the parameter estimates associated with the *Roof* item suggest corrugated iron roofs are the predominant roofing type on dwellings in the Agincourt region.

To further investigate the difference between the 3 clusters the response probabilities to individual survey items within a cluster are examined. For example Table 2 shows the probability of observing each possible response to the *Stove* item, conditional on the members of each cluster.

Table 2: Cluster specific response probabilities to the survey item *Stove*.

G	No	Yes
1	0.005	0.995
2	0.509	0.491
3	0.626	0.374

The distances between the cluster specific item response probability vectors can be used to make pairwise comparisons of groups. The distance measure used here is Hellinger distance [3, 6, 42]. Suppose there exist two vectors of probabilities, $\underline{a} = (a_1, \ldots, a_m)$ and $\underline{b} = (b_1, \ldots, b_m)$, where $\sum_{i=1}^{m} a_i = 1$ and $\sum_{i=1}^{m} b_i = 1$, then the Hellinger distance between $\underline{a}$ and $\underline{b}$ is:

$$H(\underline{a}, \underline{b}) = \frac{1}{\sqrt{2}} \left[\sum_{i=1}^{m} \left(\sqrt{a_i} - \sqrt{b_i} \right)^2 \right]^{\frac{1}{2}}$$

The Hellinger distance that may be achieved between 2 vectors of probabilities lies in $[0, 1]$. A distance of 1 is achieved if all the households in one cluster respond in one way while all the households in the other cluster respond in another.

It should be noted however that there is more than one way in which a Hellinger distance of 1 may be achieved if $K_j > 2$. This only causes a problem for ordinal items. For example, in the case of an ordinal item with $K_j = 4$, a distance of 1 will be found if all the households in one group respond with a 1 while all the households in the other group respond with a 4. However, a distance of 1 will also be achieved if all households in one group respond

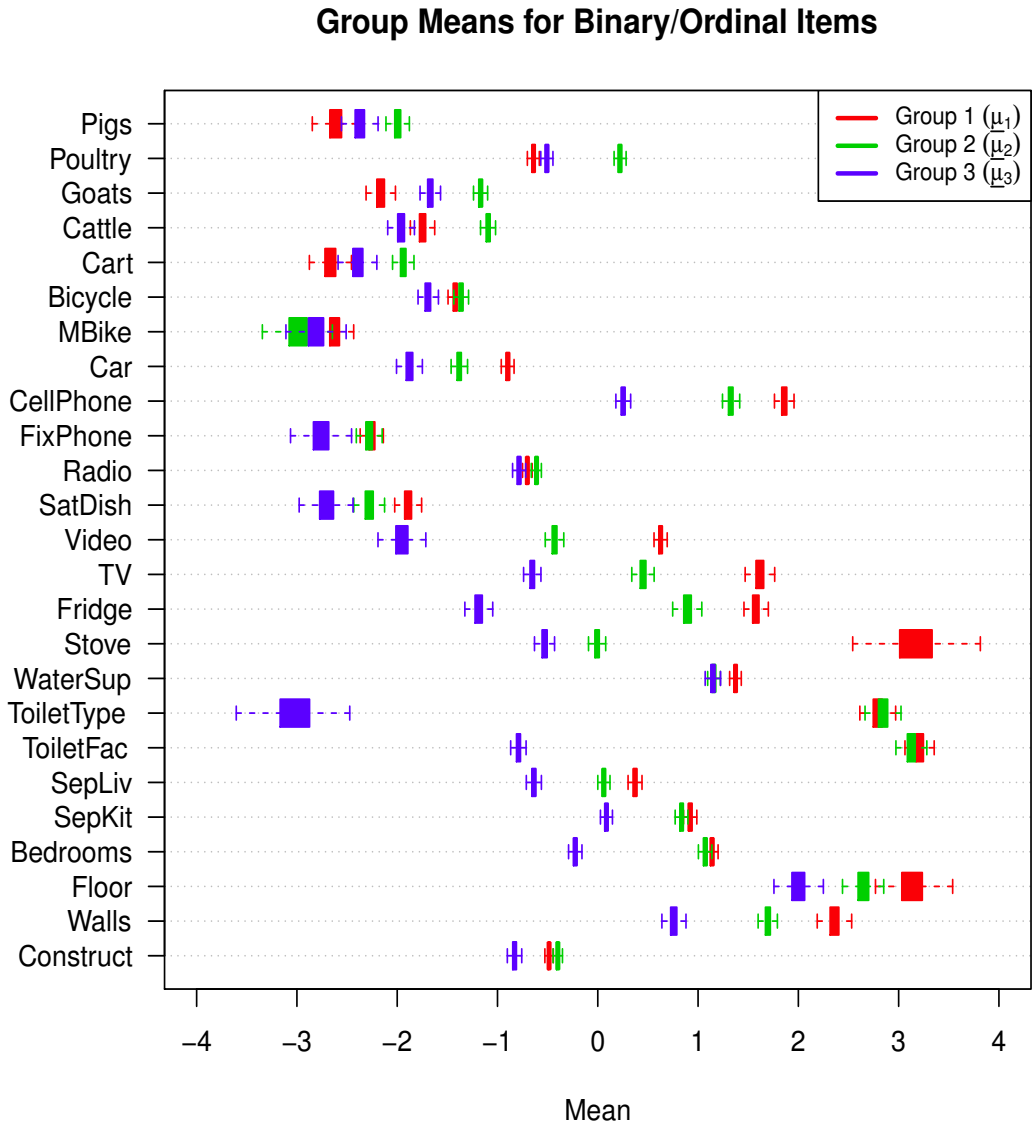


Figure 1: Box plots of MCMC samples of the dimensions of the cluster means, $\underline{\mu}_g$, corresponding to binary and ordinal items.

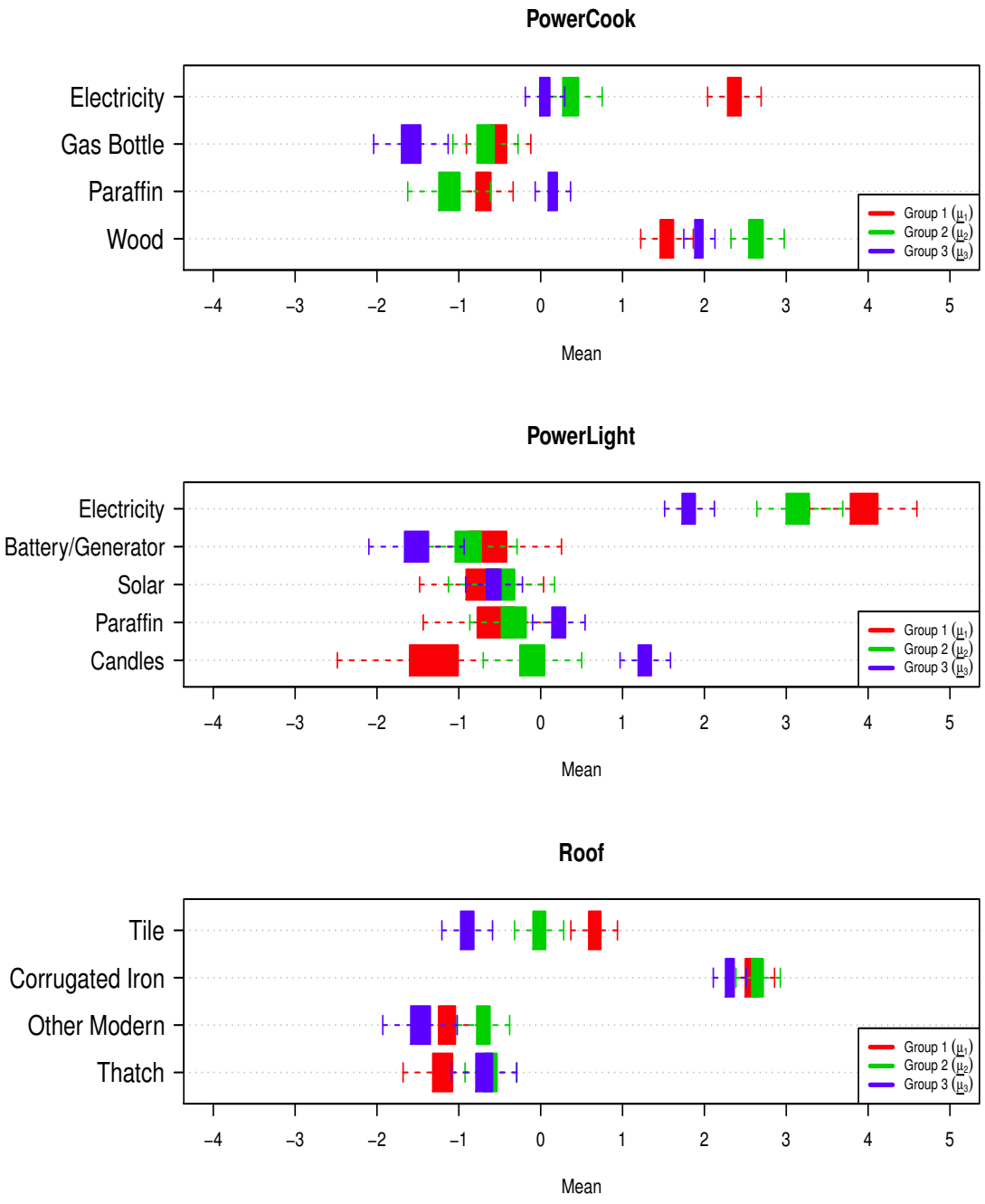
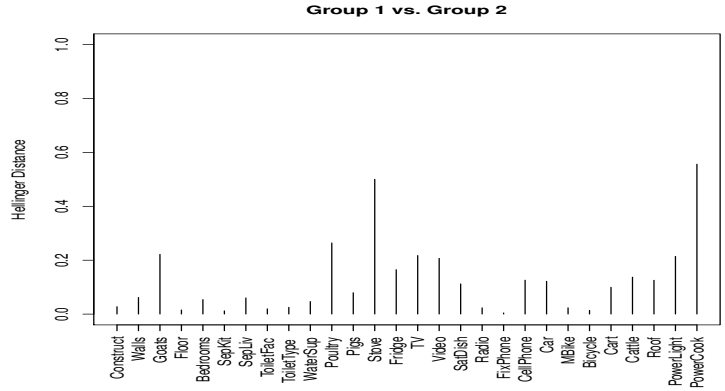
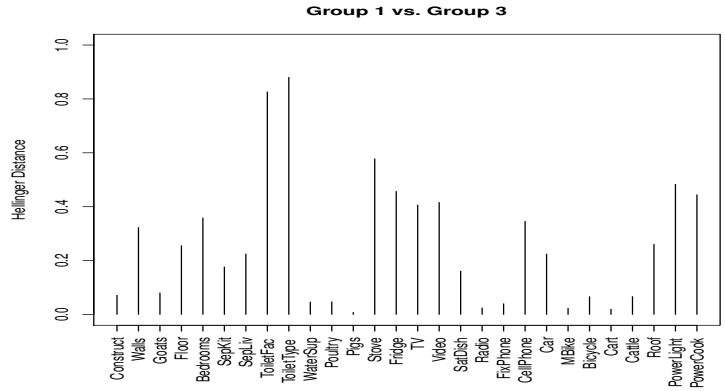


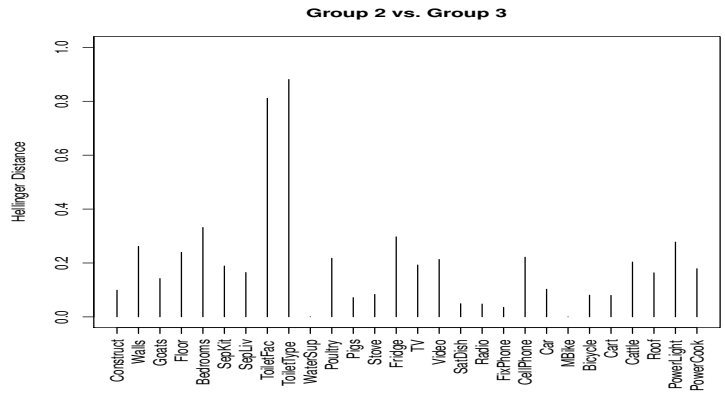
Figure 2: Box plots of MCMC samples of the dimensions of the cluster means, $\underline{\mu}_g$, corresponding to nominal items. The first plot shows box plots of the means of the latent dimensions relating to the *PowerCook* item, the second shows the means of the dimensions representing the *PowerLight* item and the third shows the means of the dimensions corresponding to the *Roof* item.



(a) Hellinger distances between groups 1 and 2. Total distance is 3.544



(b) Hellinger distances between groups 1 and 3. Total distance is 7.316



(c) Hellinger distances between groups 2 and 3. Total distance is 5.643

Figure 3: Pairwise comparisons of groups using Hellinger distance.

with a 2 and all households in the other group respond with a 3. Since the question is ordinal the distance in the former case should be larger than that in the latter case. This is a shortcoming of the metric and will result in the underestimation of distances between groups on ordinal items.

Pairwise comparisons between clusters are illustrated in Figure 3. The Hellinger distance between response probability vectors for each item are plotted. The groups that are most different are clusters 1 (the wealthy/modern cluster) and 3 (the least wealthy cluster). The sum of the Hellinger distances between these groups across all items is 7.316. The items for which the Hellinger distance between the response probability vectors is largest are *ToiletType*, *ToiletFac*, *Stove* and *PowerLight*, highlighting the areas in which households in these clusters differ most. There are noteworthy Hellinger distances for many other items also. The difference in response patterns for these items is also evident in the box plots in Figures 1 and 2.

The sum of the Hellinger distances between clusters 1 and 2 (the wealthier two clusters) across all items is 3.544 making these two groups the most similar. There are some notable differences however; the Hellinger distance between the groups on the items *Stove* and *PowerCook* are 0.501 and 0.556 respectively which accounts for almost 30% of the total distance.

Clusters 2 and 3 are quite different and the sum of the Hellinger distances between these groups is 5.643. As was the case for clusters 1 and 3 the items *ToiletType* and *ToiletFac* provide the largest Hellinger distances between groups 2 and 3. In contrast however there are much smaller differences for the items *Stove* and *PowerCook*. Again these results highlight the specific areas in which the socio-economic status of households within each cluster differ. A similar pattern was observed in Table 1 and Figures 1 and 2.

5.2 Model Fit

In order to assess model fit we used a cross validation approach based on ideas originally introduced by [50] and examined more recently by [40]. In the case of the SES data this involves removing a portion of the data points (not households) at random across the data matrix, then running the model and imputing these missing values based on the parameter values at each iteration. The categorical values are imputed by simulating values for the latent variable Z then using the decision rules explained in Section 3 to determine the categorical value Y. This procedure is repeated until all data points have been removed once. The percentage of correct imputations for each data point is used to assess how well the model fits the data.

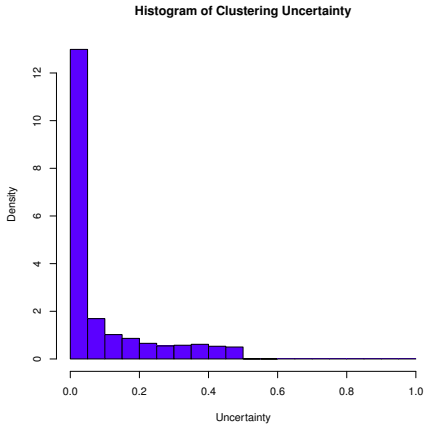
The three component, one-dimensional latent trait model performs very well with respect to this criterion. The SES data was divided into 10 groups with each data point assigned to one of the groups with equal probability. The model then ran 10 times, leaving out one of the groups of data points each time. The mean percentage of correct imputations across all data points was 0.78 and the median was 0.89 indicating that the three component, one dimensional latent trait model is a good fit for the SES data.

5.3 Model Selection

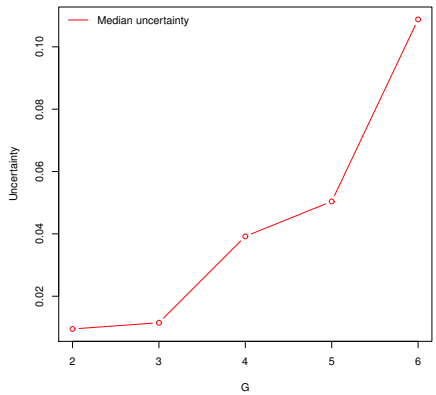
There are two separate aspects to model selection for the MFA-MD model since both the dimension of the latent trait, q, and the number of mixture components, G, must be chosen. Traditional likelihood based tools for model selection are not available here since the likelihood is intractable.

The choice of q is addressed first. Models with the number of mixture components ranging from 1 to 6 were fitted for both q equal to 1 and q equal to 2. Having fitted these models it was found that both the clustering and the interpretation of the clusters were broadly similar whether q was 1 or 2. The Rand index comparing clusterings with the same number of components was typically about 0.75. However, adding an extra dimension to the latent trait adds over $17,500$ extra parameters to the model since when q is 2 one extra dimension of $\underline{\theta}_i$ must be estimated for each household in the data set. The more parsimonious model with a one dimensional latent trait was chosen as there was little to be gained from adding an extra layer of complexity.

Having chosen the dimension of the latent trait it remains to choose the most appropriate number of mixture components. Clustering uncertainty is one model attribute which helps choose G. A histogram of the clustering uncertainty of each household under the 3 component model is shown in Figure 4. It is clear to see that these values are very low in general indicating that households are assigned to groups with a high degree of confidence. Histograms for other numbers of components were quite similar. Although clustering uncertainty was low across all models the lowest median values were observed for the 2 and 3 component models. The median clustering uncertainty for these models were 0.009 and 0.011 respectively. The median value for the 4 component model was 0.039 and this value increased as more components were added, this can be seen in Figure 4. The clustering uncertainty for the 2 and 3 component models are virtually identical but the 3 component model was favored since it provided a much more intuitive interpretation of the clusters.



(a) Histogram of clustering uncertainties when a 3 component, 1 dimensional latent trait model is fitted to the SES data.



(b) Median clustering uncertainty as the number of components, G, varies from 2 to 6.

Figure 4: Examining clustering uncertainty.

6 Discussion

The MFA-MD model described in this paper successfully clusters households into groups of differing socio-economic status. Which households are in each group and what differentiates these clusters from each other can be examined in the model output. This information is potentially of great benefit to various authorities in the Agincourt region. The interpretation of these groups could aid decision making with regard to infrastructural development and other social policy.

There are many other potential uses for the MFA-MD model. The cluster memberships could be used for targeted sampling of a particular cluster with new questions taking into account what is known about that group already. The clustering output from this model could also be used as covariate input to other models such as mortality models. There may be important differences in mortality rates in different socio-economic strata within the region. New health policies may need to take these differences into account.

The MFA-MD model is also a novel model-based clustering approach to clustering mixed categorical data. The SES data used here is a mix of binary, ordinal and nominal data. The MFA-MD model provides clustering capabilities in the context of such mixed data without mistreating any one data type. A factor analytic model is fitted to each group individually which may be interpreted in the usual manner.

Future research directions are plentiful and varied. The lack of a formal model selection criterion for the MFA-MD model is the most pressing, and challenging. Currently the model relies on expert knowledge and some informative heuristics to choose the most appropriate number of components. A formal criterion which selects the most appropriate number of components and also the dimension of the latent trait would be very beneficial. Model selection tools based on the marginal likelihood [17] are a natural approach to model selection within the Bayesian paradigm; such approaches are potentially useful within the context of the MFA-MD model, but the intractable likelihood poses difficulties. One potential future avenue would be to approximate such model selection tools using the underlying latent data, but this also brings difficulties.

Additionally, there are several ways in which MFA-MD model itself could be extended. The data set analyzed here is survey data from a single time point. There have been several waves of this particular survey – extending the MFA-MD model to appropriately model longitudinal data would be beneficial. In this way households could be tracked across time as they may or may not move between socio-economic strata. As with most clustering models, the variables included in the model are potentially influential. The addition of a variable selection method within the context of the MFA-MD model could significantly improve clustering performance. A reduction in the number of variables would also decrease the computational time required to fit such models. In a similar vein, the Metropolis-Hastings step required to sample the threshold parameters in the current model fitting approach could potentially be removed by using a rank likelihood approach – [24] details such an approach in relation to ordinal regression. This could also offer an improvement in computational time.

Other areas of ongoing and future work include the inclusion of modeling continuous data by the MFA-MD model. This would facilitate the clustering of mixed data consisting of both continuous and categorical data [37], and requires little extension to the MFA-MD model proposed here. Finally, covariate information could naturally be incorporated in the

MFA-MD model in the framework of a mixture of experts model [22, 28].

A Survey Items

Table 3 displays a list of all survey items and the possible responses. The final three items in the table are regarded as nominal, all other items are binary or ordinal.

Table 3: A list of all survey items and the possible responses. The final three items (*Roof*, *PowerLight* and *PowerCook*) in the table are regarded as nominal, all other items are binary or ordinal.

Item	**Description**	**Response Options**
Construct	Indicates whether main dwelling is still under construction.	(No, Yes)
Walls	Construction materials used for household walls.	(Informal, Modern)
Floor	Construction materials used for household floor.	(Informal, Modern)
Bedrooms	Number of bedrooms in the household.	($\leq$1, 2, 3, 4, 5, $\geq$ 6)
SepKit	Indicates whether kitchen is separate from sleeping area.	(No, Yes)
SepLiv	Indicates whether living room is separate from sleeping area.	(No, Yes)
ToiletFac	Reports the physical location of toilet in the household.	(Bush, Other House, In Yard, In House)
ToiletType	Reports the type of toilet used in the household.	(None, Pit, VIP, Modern)
WaterSup	Reports the water supply source for the household.	(From a tap, Other)
Stove	Reports stove ownership status of the household.	(No, Yes)
Fridge	Reports fridge ownership status of the household.	(No, Yes)
TV	Reports television ownership status of the household.	(No, Yes)
Video	Reports video player ownership status of the household.	(No, Yes)
SatDish	Reports satellite dish ownership status of the household.	(No, Yes)
Radio	Reports radio ownership status of the household.	(No, Yes)
FixPhone	Reports fixed phone ownership status of the household.	(No, Yes)
CellPhone	Reports mobile phone ownership status of the household.	(No, Yes)
Car	Reports car ownership status of the household.	(No, Yes)
MBike	Reports motor bike ownership status of the household.	(No, Yes)
Bicycle	Reports bicycle ownership status of the household.	(No, Yes)
Cart	Reports animal drawn cart ownership status of the household.	(No, Yes)
Cattle	Reports cattle ownership status of the household.	(No, Yes)
Goats	Reports goats ownership status of the household.	(No, Yes)
Poultry	Reports poultry ownership status of the household.	(No, Yes)
Pigs	Reports pig ownership status of the household.	(No, Yes)
Roof	Construction materials used for household roof.	(Other informal, Thatch, Other modern, Corrugated iron, Tile)
PowerLight	Main power supply for lights and appliances.	(Other, Candles, Paraffin, Solar, Battery/Generator, Electricity)
PowerCook	Main power supply for cooking.	(Other, Wood, Paraffin, Gas Bottle, Electricity)

B Full Conditional Posterior Distributions

B.1 Mixing Weights

A Dirichlet prior distribution is used for the mixing weights $\underline{\pi}$. This is a conjugate prior which leads to a Dirichlet full conditional posterior.

$$\begin{aligned} h(\underline{\pi}|\ldots) &\propto \left[\prod_{i=1}^{N}\prod_{g=1}^{G}\pi_g^{\ell_{ig}}\right]\left[\prod_{g=1}^{G}\pi_g^{\alpha_g-1}\right] \\ &= \prod_{g=1}^{G}\pi_g^{(n_g+\alpha_g)-1} \end{aligned}$$

$$\Rightarrow \underline{\pi}|\ldots \sim \text{Dirichlet}(n_1+\alpha_1,\ldots,n_g+\alpha_G)$$

where $n_g = \sum_{i=1}^{N} l_{ig}$.

B.2 Allocation Vectors

A priori the allocation vectors, $\underline{\ell}_i$ are assumed to be Multinomial$(1,\underline{\pi})$ distributed. Thus the full conditional posterior is also multinomial.

$$\begin{aligned} h(\underline{\ell}_i|\ldots) \propto \prod_{g=1}^{G}\Bigg\{\pi_g &\left[\prod_{j=1}^{O}\prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^{T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}\{\gamma_{j,k-1}<z_{ij}<\gamma_{j,k}\}}\right] \\ &\times\left[\prod_{j=O+1}^{J}\prod_{k=2}^{K_j-1}\prod_{s=1}^{3} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}(\text{case } s)}\right]\Bigg\}^{\ell_{ig}} \end{aligned}$$

$$\Rightarrow \underline{\ell_i}|\ldots \sim \text{Multinomial}(\underline{p})$$

where

$$\begin{aligned} p_g = &\left\{\pi_g\left[\prod_{j=1}^{O}\prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^{T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}\{\gamma_{j,k-1}<z_{ij}<\gamma_{j,k}\}}\right]\right. \\ &\left.\times\left[\prod_{j=O+1}^{J}\prod_{k=2}^{K_j-1}\prod_{s=1}^{3} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}(\text{case } s)}\right]\right\} \\ &\times\left\{\sum_{g=1}^{G}\pi_g\left[\prod_{j=1}^{O}\prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^{T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}\{\gamma_{j,k-1}<z_{ij}<\gamma_{j,k}\}}\right]\right. \\ &\left.\times\left[\prod_{j=O+1}^{J}\prod_{k=2}^{K_j-1}\prod_{s=1}^{3} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T}\tilde{\underline{\theta}}_i,1)^{\mathbb{I}(\text{case } s)}\right]\right\}^{-1} \end{aligned} \tag{4}$$

B.3 Threshold Parameters

A uniform prior is specified for the threshold parameters, $\underline{\gamma}_j$, for ordinal items j. This leads to a uniform full conditional posterior distribution.

$$\begin{aligned} h(\gamma_{j,k}|\ldots) \quad \propto \quad & \prod_{i=1}^{N} [\mathbb{I}(y_{ij}=k)\mathbb{I}(\gamma_{j,k-1}<z_{ij}<\gamma_{j,k}) \\ & +\mathbb{I}(y_{ij}=k+1)\mathbb{I}(\gamma_{j,k}<z_{ij}<\gamma_{j,k+1})] \end{aligned}$$

Note that $\underline{\gamma}_j$ is specified so that $\gamma_{j,0}=-\infty$, $\gamma_{j,1}=0$ and $\gamma_{j,K_j}=\infty$ for all ordinal items j. Also, $\underline{\gamma}_j$ is also constrained so that $\gamma_{j,0}<\gamma_{j,1}<\ldots<\gamma_{j,K_j}$.

The uniform full conditional distribution has support on the interval:

$$\left(\max_i\{z_{ij}|y_{ij}=k\}, \min_i\{z_{ij}|y_{ij}=k+1\}\right)$$

Samples from this distribution may be obtained using Gibbs sampling, however, to improve mixing [10] proposes a Metropolis-Hastings approach which is adopted here.

Values $v_{j,k}$ are proposed for $\gamma_{j,k}$ (for $k=2,\ldots,K_j-1$) from $\mathrm{N}^T(\gamma_{j,k}^{(t-1)},\sigma_{MH}^2)$ truncated to the interval $(v_{j,k-1},\gamma_{j,k+1}^{(t-1)})$ where $\gamma_{j,k+1}^{(t-1)}$ is the value for $\gamma_{j,k+1}$ at iteration $(t-1)$. The threshold vector , $\underline{\gamma}_j$, is set equal to the proposed vector, $\underline{v}_j$, with probability $\beta=\min(1,R)$ where:

$$\begin{aligned} R \quad = \quad & \frac{\pi(\underline{v}_j)}{\pi(\underline{\gamma}_j^{(t-1)})}.\frac{q(\underline{\gamma}_j^{(t-1)}|\underline{v}_j)}{q(\underline{v}_j|\underline{\gamma}_j^{(t-1)})} \\ = \quad & \prod_{i=1}^{N}\prod_{g=1}^{G}\left[\frac{\Phi(v_{j,y_{ij}}-\tilde{\underline{\lambda}}_{gj}^T\tilde{\underline{\theta}}_i)-\Phi(v_{j,y_{ij}-1}-\tilde{\underline{\lambda}}_{gj}^T\tilde{\underline{\theta}}_i)}{\Phi(\gamma_{j,y_{ij}}^{(t-1)}-\tilde{\underline{\lambda}}_{gj}^T\tilde{\underline{\theta}}_i)-\Phi(\gamma_{j,y_{ij}-1}^{(t-1)}-\tilde{\underline{\lambda}}_{gj}^T\tilde{\underline{\theta}}_i)}\right]^{\ell_{ig}} \\ & \times\prod_{k=2}^{K_j-1}\frac{\Phi\left[(\gamma_{j,k+1}^{(t-1)}-\gamma_{j,k}^{(t-1)})/\sigma_{MH}\right]-\Phi\left[(v_{j,k-1}-\gamma_{j,k}^{(t-1)})/\sigma_{MH}\right]}{\Phi\left[(v_{j,k+1}-v_{j,k})/\sigma_{MH}\right]-\Phi\left[(\gamma_{j,k-1}^{(t-1)}-v_{j,k})/\sigma_{MH}\right]} \end{aligned} \tag{5}$$

The tuning parameter σ_{MH}^2 is selected to achieve appropriate acceptance rates.

B.4 Latent Traits

A priori the latent traits, $\underline{\theta}_i$, are assumed to be standard multivariate Gaussian distributed. The resulting full conditional posterior is also multivariate Gaussian.

$$
\begin{aligned}
h(\underline{\theta}_i|\ldots) &\propto \prod_{d=1}^{D} \mathrm{N}\left(z_{id}|\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i, 1\right) \mathrm{MVN}_q\left(\underline{0}, \mathbf{I}_q\right) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left(z_{id} - \tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i\right)^2\right\}\exp\left\{-\frac{1}{2}\underline{\theta}_i^T\underline{\theta}_i\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{d=1}^{D}\left(-2z_{id}\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i + \left(\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i\right)^2\right) + \underline{\theta}_i^T\underline{\theta}_i\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{d=1}^{D}\left(-2z_{id}\underline{\lambda}_{gd}^T\underline{\theta}_i + 2\mu_{gd}\underline{\lambda}_{gd}^T\underline{\theta}_i + \underline{\theta}_i^T\underline{\lambda}_{gd}\underline{\lambda}_{gd}^T\underline{\theta}_i\right) + \underline{\theta}_i^T\underline{\theta}_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[-2\sum_d z_{id}\underline{\lambda}_{gd}^T\underline{\theta}_i + 2\sum_d \mu_{gd}\underline{\lambda}_{gd}^T\underline{\theta}_i + \underline{\theta}_i^T\sum_d \underline{\lambda}_{gd}\underline{\lambda}_{gd}^T\underline{\theta}_i + \underline{\theta}_i^T\underline{\theta}_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[-2\underline{z}_i^T\Lambda_g\underline{\theta}_i + 2\underline{\mu}_g^T\Lambda_g\underline{\theta}_i + \underline{\theta}_i^T\Lambda_g^T\Lambda_g\underline{\theta}_i + \underline{\theta}_i^T\underline{\theta}_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\underline{\theta}_i^T\left[\Lambda_g^T\Lambda_g + \mathbf{I}_q\right]\underline{\theta}_i + \left[\left(\underline{z}_i - \underline{\mu}_g\right)^T\Lambda_g\right]\underline{\theta}_i\right\} \\
&\Rightarrow \underline{\theta}_i|\ldots \sim \mathrm{MVN}_q\left\{\left[\Lambda_g^T\Lambda_g + \mathbf{I}_q\right]^{-1}\left[\Lambda_g^T\left(\underline{z}_i - \underline{\mu}_g\right)\right], \left[\Lambda_g^T\Lambda_g + \mathbf{I}_q\right]^{-1}\right\}
\end{aligned}
$$

B.5 Loadings Matrix and Mean

A multivariate Gaussian prior is specified for $\tilde{\underline{\lambda}}_{gd}$. This is a conjugate prior and so leads to a multivariate Gaussian full conditional posterior.

$$
\begin{aligned}
h(\tilde{\underline{\lambda}}_{gd}|\ldots) &\propto \prod_{i=1}^{N}\left[\mathrm{N}\left(z_{id}|\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i, 1\right)\right]^{\ell_{ig}}\left[\mathrm{MVN}_{(q+1)}\left(\tilde{\underline{\lambda}}_{gd}|\underline{\mu}_\lambda, \Sigma_\lambda\right)\right] \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i:\ell_{ig}=1}\left(z_{id} - \tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i\right)^2\right\}\exp\left\{-\frac{1}{2}\left(\tilde{\underline{\lambda}}_{gd} - \underline{\mu}_\lambda\right)^T\Sigma_\lambda^{-1}\left(\tilde{\underline{\lambda}}_{gd} - \underline{\mu}_\lambda\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{i:\ell_{ig}=1}\left(-2z_{id}\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i + \tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i\tilde{\underline{\theta}}_i^T\tilde{\underline{\lambda}}_{gd}\right) + \tilde{\underline{\lambda}}_{gd}^T\Sigma_\lambda^{-1}\tilde{\underline{\lambda}}_{gd} - 2\underline{\mu}_\lambda^T\Sigma_\lambda^{-1}\tilde{\underline{\lambda}}_{gd}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[-2\underline{z}_{gd}^T\tilde{\Theta}_g\tilde{\underline{\lambda}}_{gd} + \tilde{\underline{\lambda}}_{gd}^T\tilde{\Theta}_g^T\tilde{\Theta}_g\tilde{\underline{\lambda}}_{gd} + \tilde{\underline{\lambda}}_{gd}^T\Sigma_\lambda^{-1}\tilde{\underline{\lambda}}_{gd} - 2\underline{\mu}_\lambda^T\Sigma_\lambda^{-1}\tilde{\underline{\lambda}}_{gd}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\tilde{\underline{\lambda}}_{gd}^T\left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T\tilde{\Theta}_g\right]\tilde{\underline{\lambda}}_{gd} + \tilde{\underline{\lambda}}_{gj}^T\left[\tilde{\Theta}_g^T\underline{z}_{gd} + \Sigma_\lambda^{-1}\underline{\mu}_\lambda\right]\right\} \\
\Rightarrow \quad & \tilde{\underline{\lambda}}_{gd}|\ldots \sim \mathrm{MVN}_{(q+1)}\left\{\left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T\tilde{\Theta}_g\right]^{-1}\left[\tilde{\Theta}_g^T\underline{z}_{gd} + \Sigma_\lambda^{-1}\underline{\mu}_\lambda\right], \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T\tilde{\Theta}_g\right]^{-1}\right\}
\end{aligned}
$$

where

$$\underline{z}_{gd} = \{z_{id}\} \text{ for all individuals in group } g.$$

References

[1] O. Aguilar and M. West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

[2] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[3] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] R. P. Browne and P. D. McNicholas. Model-based clustering and classification of data with mixed type. *Journal of Statistical Planning and Inference*, 142:2976–2984, 2012.

[5] J.H. Cai, X.Y. Song, K.H. Lam, and E.H.S. Ip. A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics and Data Analysis*, 55:2889–2907, 2011.

[6] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: some basic concepts*. Springer, 1990.

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

[8] S. Chib, E. Greenberg, and Y. Chen. Mcmc methods for fitting and comparing multinomial response models. Technical report, Washington University, 1998.

[9] M. A. Collinson, S. J. Clark, A. A. M. Gerritsen, P. Byass, K. Kahn, and S. M. Tollmann. The dynamics of poverty and migration in a rural south african community, 2001–2005. Technical report, Center for Statistics and the Social Sciences University of Washington, 2009.

[10] M. K. Cowles. Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2):101–111, 1996.

[11] B. S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters*, 6:305–309, 1988.

[12] B. S. Everitt and C. Merette. The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, 17:283–297, 1988.

[13] D. Filmer and L.H. Pritchett. *Demography*, 38(1):115–132, 2001.

[14] E. Fokoue and D.M. Titterington. Mixtures of factor analysers. bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50(1):73–94, 2003.

[15] J.P. Fox. *Bayesian Item Response Modeling*. Springer, 2010.

[16] C. Fraley and A. E. Raftery. How many clusters? which clustering methods? answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

[17] N. Friel and J. Wyse. Estimating the evidence – a review. *Statistica Neerlandica*, pages no–no, 2011.

[18] J. Geweke, M. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, 76(4):609–632, 1994.

[19] J.F. Geweke and G. Zhou. Measuring the pricing error of arbitrage pricing theory. *Review of Financial Studies*, 9:557–587, 1996.

[20] Z. Ghahramani and G.E. Hinton. The em algorithm for mixtures of factor analyzers. Technical report, University of Toronto, 1997.

[21] I. Gollini and T. B. Murphy. Mixture of latent trait analyzers for model-based clustering of categorical data. Technical report, University College Dublin, 2012.

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

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

[24] P.D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[25] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

[26] L. Hunt and M. Jorgensen. Mixture model clustering using the multimix program. *Australia and New Zealand Journal of Statistics*, 41:153–171, 1999.

[27] L. Hunt and M. Jorgensen. Mixture model clustering for mixed data with missing information. *Computational Statistics and Data Analysis*, 41:429–440, 2003.

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

[29] V.E. Johnson and J.H. Albert. *Ordinal Data Modeling*. Springer, 1999.

[30] K. Kahn, S.M. Tollman, M.A. Collinson, S.J. Clark, R. Twine, B.D. Clark, M. Shabangu, F.X. Gómez-Olivé, O. Mokoena, and M.L. Garenne. Research into health, population and social transitions in rural south africa: Data and methods of the agincourt health and demographic surveillance system1. *Scandinavian Journal of Public Health*, 35(69 suppl):8–20, 2007.

[31] C. J. Lawrence and W. J. Krzanowski. Mixture separation for mixed-mode data'. *Statistics and Computing*, 6:85–92, 1996.

[32] F.M. Lord. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17:181–194, 1952.

[33] F.M. Lord and M.R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

[34] G. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47:149–174, 1982.

[35] R.E. McCulloch and P.E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64:207–240, 1994.

[36] P.D. McNicholas and T.B. Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18:285–296, 2008.

[37] D. McParland, I.C. Gormley, L. Brennan, and H. M. Roche. Clustering mixed continuous and categorical data from the lipgene study: examining the interaction of nutrients and genotype in the metabolic syndrome. Technical report, University College Dublin, 2012.

[38] B. Muthén and K. Shedden. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55:463–469, 1999.

[39] A. Nobile. A hybrid markov chain for the bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8:229–242, 1998.

[40] Patrick O. Perry. *Cross-Validation for Unsupervised Learning*. PhD thesis, Stanford University, September 2009.

[41] K.M. Quinn. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, 12(4):338–353, 2004.

[42] C. R. Rao. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Questiio*, 19:23 – 63, 1995.

[43] G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, The Danish Institute for Educational Research, 1960.

[44] Shea O. Rutstein and Kiersten Johnson. The dhs wealth index. DHS Comparative Reports 6, ORC Macro, Calverton, Maryland, 2004.

[45] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 17, 1969.

[46] L.L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7):433–451, 1925.

[47] J. K. Vermunt. The use restricted latent class models for defining and testing nonparametric and parametric irt models. *Applied Psychological Measurement*, 25:283–294, 2001.

[48] S. Vyas and L. Kumaranayake. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21(6):459–468, 2006.

[49] A. Willse and R. J. Boik. Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9:111–121, 1999.

[50] S. Wold. Cross-validatory estimationof the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978.

[51] X. Zhang, W.J. Boscardin, and T.R. Belin. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational Statistics and Data Analysis*, 52:3697–3708, 2008.