Cluster analysis of repeated ordinal data: A model approach based on finite mixtures

Roy Costilla Ivy Liu Richard Arnold

School of Mathematics, Statistics and Operations Research
Victoria University of Wellington

Phd proposal
May 2014

Outline

Aim

To develop probability models based on finite mixture models for cluster analysis of ordinal data that arises repeated measures settings

Why? Traditionally, cluster methods for ordinal data

- Assume continuous data/cardinality
- Do not often use statistical inference to compare models
- Allocate subjects to clusters deterministically.

Introduction

Models for Ordinal Data

- Cumulative Logit Models (Multivariate GLM)
- Proportional Odds Model (POM) by McCullagh (1980)
- For a ordinal response Y with q ordered categories and a set of predictors x

$$\text{Logit}[P(Y \leq k|x)] = \mu_k - \beta' x \qquad k = 1, \ldots, q-1$$

Where $\mu_1 < \mu_2 \cdots < \mu_{q-1}$.

Introduction

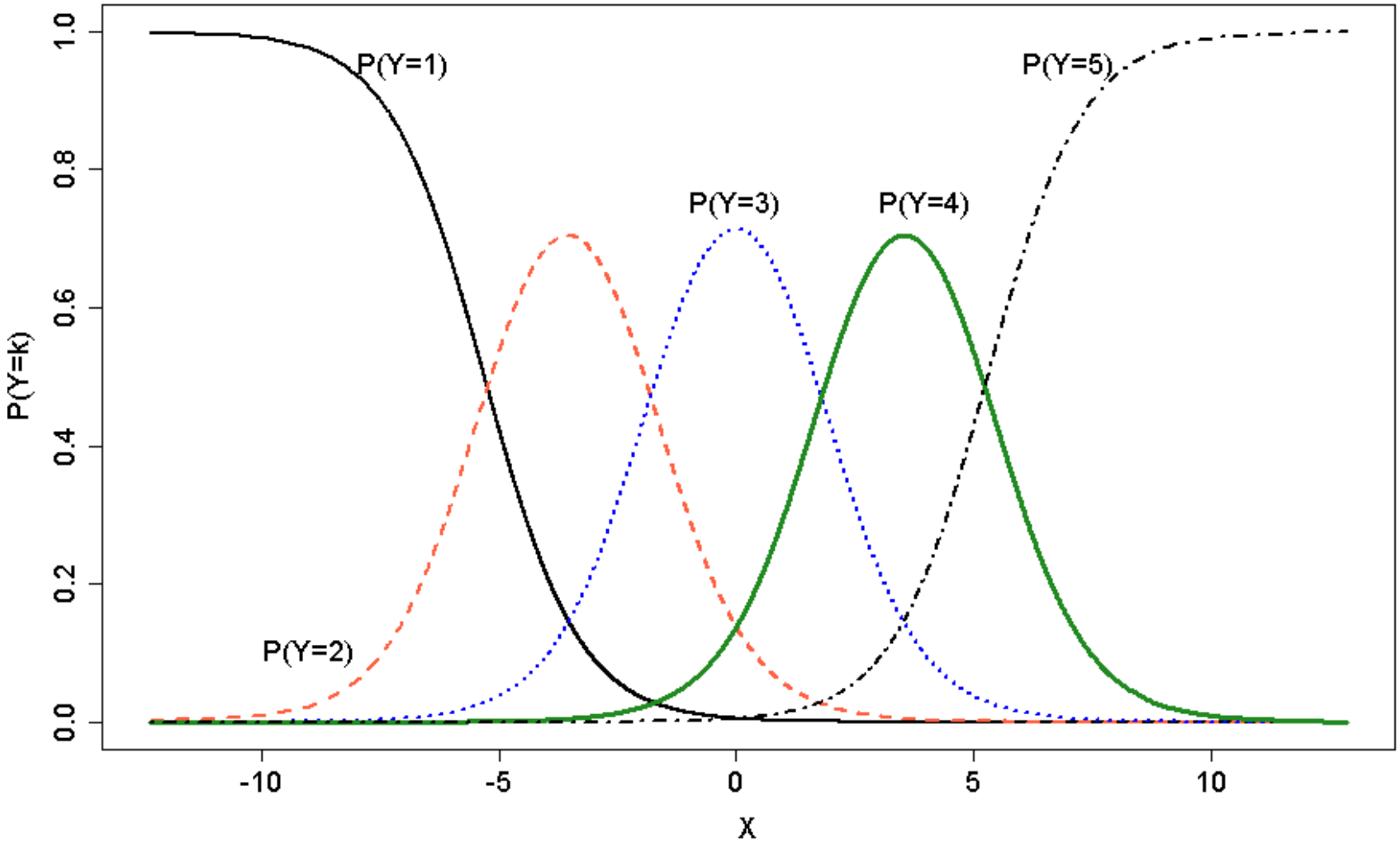


Figure : Individual category probabilities for the POM with five response

Introduction

Models for Ordinal Data

- Trend Odds Model (TOM) by Capuano and Dawson (2012)
- Setting an arbitrary scalar t_k that varies by ordinal outcome (k)

$$\text{Logit}[P(Y \leq k|x)] = \mu_k - (\beta + \gamma t_k)'x \qquad t_k \leq t_{k+1}; k = 1, \ldots, q-1$$

Where

- $\mu_k - \mu_{k-1} \geq \gamma(t_k - t_{k-1})x, \forall x$
- non-decreasing cumulative probabilities constraint

Introduction

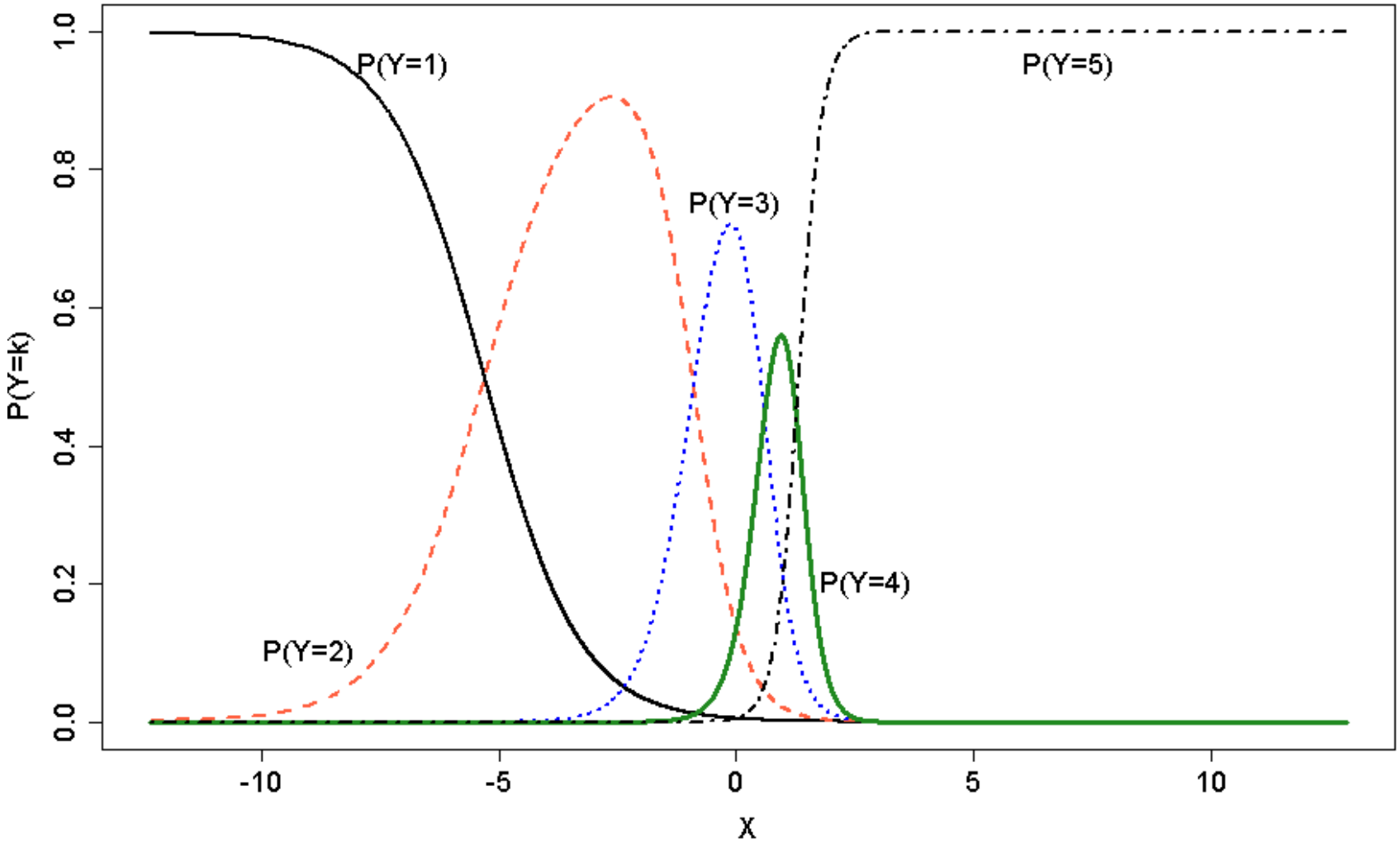


Figure : Individual category probabilities for the TOM with five response

Methods

Model Setup

- Extension of Pledger and Arnold (2013) to ordinal data
- Data **Y** is as n x p matrix where each cell y_{ij} is equal to any of the q ordinal categories
  $i = 1, ..., n$ subjects
  $j = 1, ..., p$ occasions
  $k = 1, ..., q$ ordinal categories

- **Row-clustering**
  - Rows i come from a finite mixture with R components
  - R and row-cluster proportions π_r are unknown.
  - $R < n$ and $\sum_{r=1}^{R} \pi_r = 1$

Methods: TOM row-clustering

- Let $P(y_{ij} = k|i \in r) = \theta_{rjk}$. TOM with row-clustering has the form

$$\text{Logit}[P(y_{ij} \leq k|i \in r)] = \mu_k - \alpha_r - \gamma_r t_k \tag{1}$$

or equivalently

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r - \gamma_r t_k)}{1 + \exp(\mu_k - \alpha_r - \gamma_r t_k)} - \frac{\exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})}{1 + \exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})} \tag{2}$$

Where

- $\alpha_1 = \gamma_1 = 0$
- $\mu_k - \mu_{k-1} \geq \gamma_r \qquad \forall r, k$ when setting $t_k = k - 1$
- non-decreasing cumulative probabilities constraint

Methods: TOM row-clustering

Likelihood

Assuming independence over the rows and, conditional on the rows, independence over the columns, the likelihood is

$$L(\phi, \pi|\mathbf{Y}) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{3}$$

where

- ϕ set of model parameters (μ, α, γ)
- $I(y_{ij} = k)$ indicator function equal to 1 if the condition $y_{ij} = k$ is satisfied and 0 otherwise.
- $L(\phi, \pi|\mathbf{Y})$ aka incomplete data likelihood (unknown cluster memberships, π_r).
- number of model parameters $v = (q - 1) + 3(R - 1)$

Methods: TOM row-clustering

Estimation

ϕ and π estimated using the Expected-Maximisation algorithm (EM)

E step Estimate unknown cluster memberships

$$\hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{ajk}^{I(y_{ij}=k)}}$$

M step Numerically maximise complete data log-likelihood

$$\ell_c(\phi, \pi|\mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \sum_{k=1}^{q} \hat{z}_{ir} I(y_{ij} = k) \log(\theta_{rjk}) + \sum_{i=1}^{n} \sum_{r=1}^{R} \hat{z}_{ir} \log(\pi_r)$$

Methods: TOM row-clustering

Model Selection for finite mixture models

- Comparing models with different number of clusters violates regularity conditions
  1. estimates under the null hypothesis are on boundary of the parameter space
  2. null hypothesis corresponds to a non-identifiable subset of the parameter space

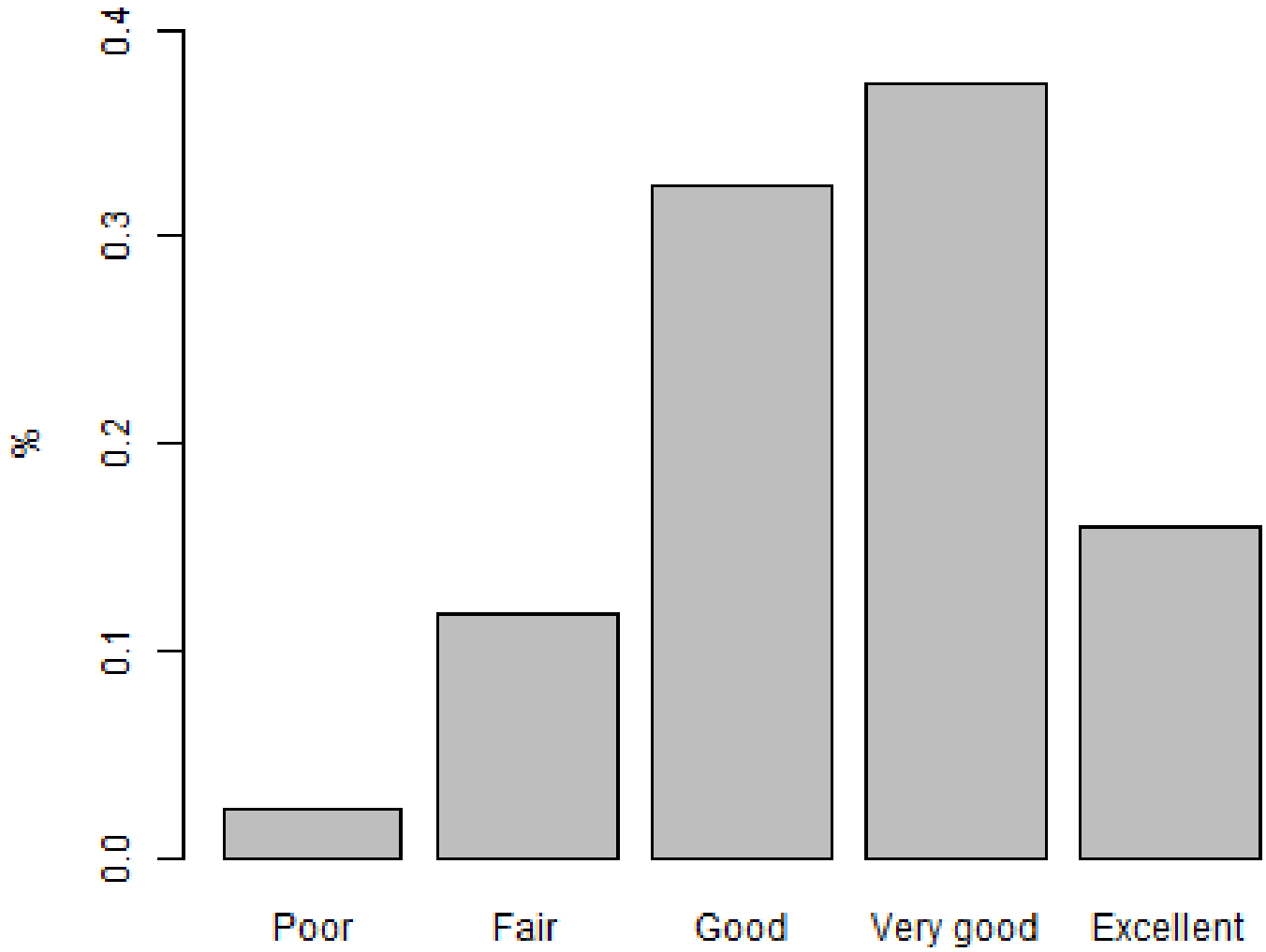
- Simulation studies shown AIC and BIC overestimate number of mixture components
- For this application, we used the ICL-BIC (Biernacki et al 2000)
  - classification-based information criteria
  - also takes into account the degree of separation of the mixture components (entropy).

Applications: Case study

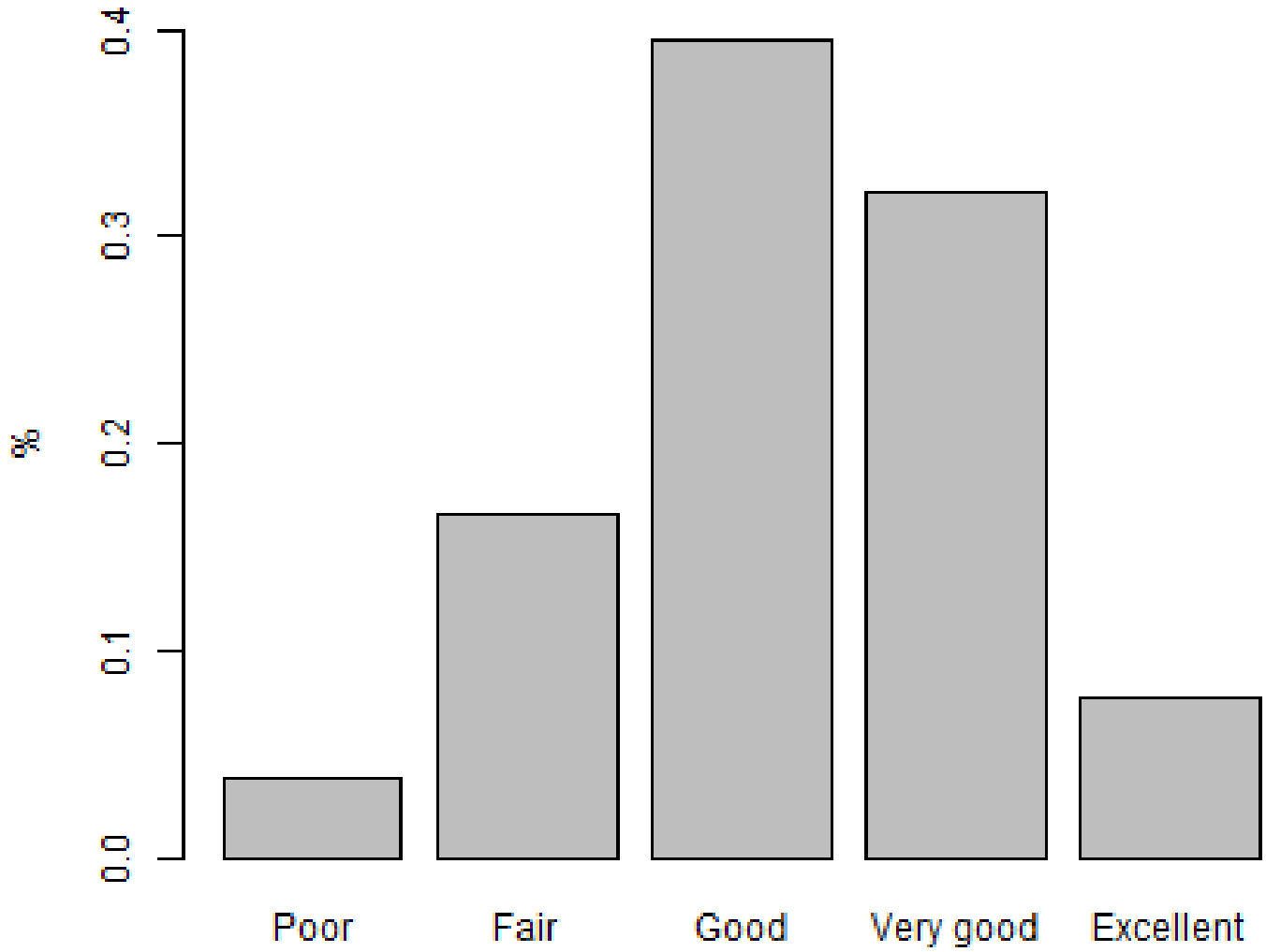
Clustering health status in Australia over 2001-2011

- Household, Income and Labour Dynamics in Australia survey-HILDA
- Self-reported health status (poor, fair, good, very good and excellent)
- 4660 individuals with complete observations over 2001-2011
- POM and TOM results for a subsample of 136

Applications: Self-Reported Health Status 2001



Applications: Self-Reported Health Status 2011



Applications: Case study

Table : Model comparison in the case study: SRHS from HILDA

Model	Linear Predictor	Rows	Cols	Param	AIC	BIC	ICL-BIC
Null	μ_k	1	1	4	4089	4139	
Row fixed effects	$\mu_k - \alpha_i$	136	1	139	2446	4200	
Col fixed effects	$\mu_k - \beta_j$	1	11	14	4094	4271	
Row and Col fixed effects	$\mu_k - \alpha_i - \beta_j$	136	11	149	**2424**	4305	
POM row clustering	$\mu_k - \alpha_r$	2	1	6	3313	3345	3355
		3	1	8	3029	3071	3095
		4	1	10	2920	2973	2999
		5	1	12	2829	2893	2927
		6	1	14	2809	2883	2947
POM col clustering	$\mu_k - \beta_c$	1	2	6	4092	4124	4134
POM biclustering	$\mu_k - \alpha_r - \beta_c$	2	2	8	3313	3355	3376
		3	2	10	3024	3077	3111
		4	2	12	2914	2978	3015
		5	2	14	2827	2901	2951
		6	2	16	2803	2888	2967
TOM row clustering	$\mu_k - \alpha_r - \gamma_r t_k$	2	1	7	3313	3351	3360
		3	1	10	3025	3078	3098
		4	1	13	2912	2981	3006
		5	1	16	2808	2893	**2922**
		6	1	19	2774	**2875**	2925
TOM col clustering	$\mu_k - \beta_c - \delta_c t_k$	1	2	7	4095	4132	4132
TOM biclustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k$	2	2	10	4033	4087	4078
		3	2	13	3882	3951	3938
		4	2	16	3003	3088	3066
		5	2	19	5541	5642	5634
		6	2	22	5776	5893	5885

Applications: Case study

Table : Model comparison in the case study: SRHS from HILDA

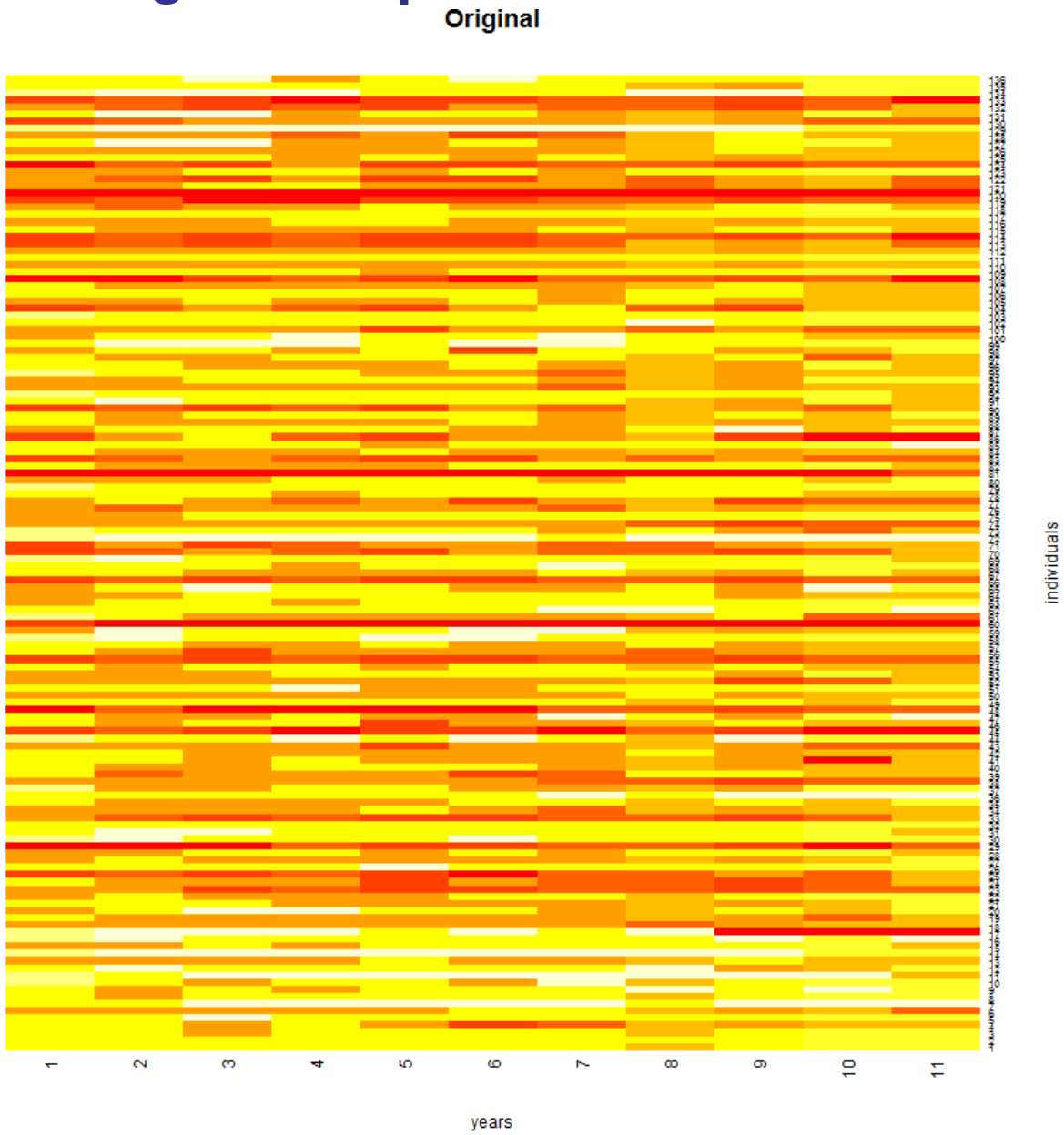
Model	Linear Predictor	Rows	Cols	Param	AIC	BIC	ICL-BIC
Null	μ_k	1	1	4	4089	4139	
Row fixed effects	$\mu_k - \alpha_i$	136	1	139	2446	4200	
Col fixed effects	$\mu_k - \beta_j$	1	11	14	4094	4271	
Row and Col fixed effects	$\mu_k - \alpha_i - \beta_j$	136	11	149	**2424**	4305	
POM row clustering	$\mu_k - \alpha_r$	2	1	6	3313	3345	3355
		3	1	8	3029	3071	3095
		4	1	10	2920	2973	2999
		5	1	12	2829	2893	2927
		6	1	14	2809	2883	2947
POM col clustering	$\mu_k - \beta_c$	1	2	6	4092	4124	4134
POM biclustering	$\mu_k - \alpha_r - \beta_c$	2	2	8	3313	3355	3376
		3	2	10	3024	3077	3111
		4	2	12	2914	2978	3015
		5	2	14	2827	2901	2951
		6	2	16	2803	2888	2967
TOM row clustering	$\mu_k - \alpha_r - \gamma_r t_k$	2	1	7	3313	3351	3360
		3	1	10	3025	3078	3098
		4	1	13	2912	2981	3006
		5	1	16	2808	2893	**2922**
		6	1	19	**2774**	**2875**	2925
TOM col clustering	$\mu_k - \beta_c - \delta_c t_k$	1	2	7	4095	4132	4132
TOM biclustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k$	2	2	10	4033	4087	4078
		3	2	13	3882	3951	3938
		4	2	16	3003	3088	3066
		5	2	19	5541	5642	5634
		6	2	22	5776	5893	5885

Applications: Case study

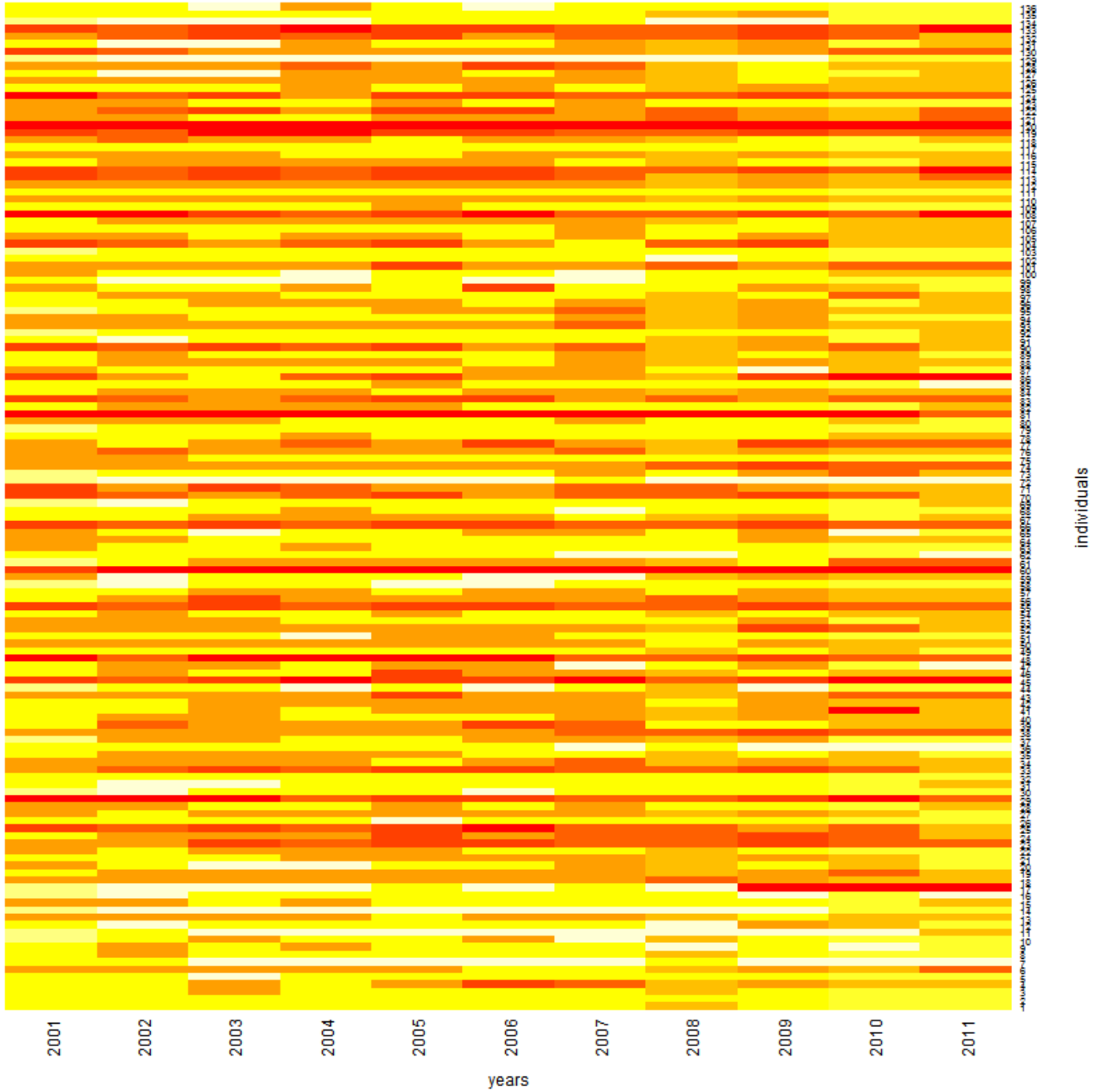
Table : Model comparison in the case study: SRHS from HILDA

Model	Linear Predictor	Rows	Cols	Param	AIC	BIC	ICL-BIC
Null	μ_k	1	1	4	4089	4139	
Row fixed effects	$\mu_k - \alpha_i$	136	1	139	2446	4200	
Col fixed effects	$\mu_k - \beta_j$	1	11	14	4094	4271	
Row and Col fixed effects	$\mu_k - \alpha_i - \beta_j$	136	11	149	**2424**	4305	
POM row clustering	$\mu_k - \alpha_r$	2	1	6	3313	3345	3355
		3	1	8	3029	3071	3095
		4	1	10	2920	2973	2999
		5	1	12	2829	2893	2927
		6	1	14	2809	2883	2947
POM col clustering	$\mu_k - \beta_c$	1	2	6	4092	4124	4134
POM biclustering	$\mu_k - \alpha_r - \beta_c$	2	2	8	3313	3355	3376
		3	2	10	3024	3077	3111
		4	2	12	2914	2978	3015
		5	2	14	2827	2901	2951
		6	2	16	2803	2888	2967
TOM row clustering	$\mu_k - \alpha_r - \gamma_r t_k$	2	1	7	3313	3351	3360
		3	1	10	3025	3078	3098
		4	1	13	2912	2981	3006
		5	1	16	2808	2893	**2922**
		6	1	19	**2774**	**2875**	2925
TOM col clustering	$\mu_k - \beta_c - \delta_c t_k$	1	2	7	4095	4132	4132
TOM biclustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k$	2	2	10	4033	4087	4078
		3	2	13	3882	3951	3938
		4	2	16	3003	3088	3066
		5	2	19	5541	5642	5634
		6	2	22	5776	5893	5885

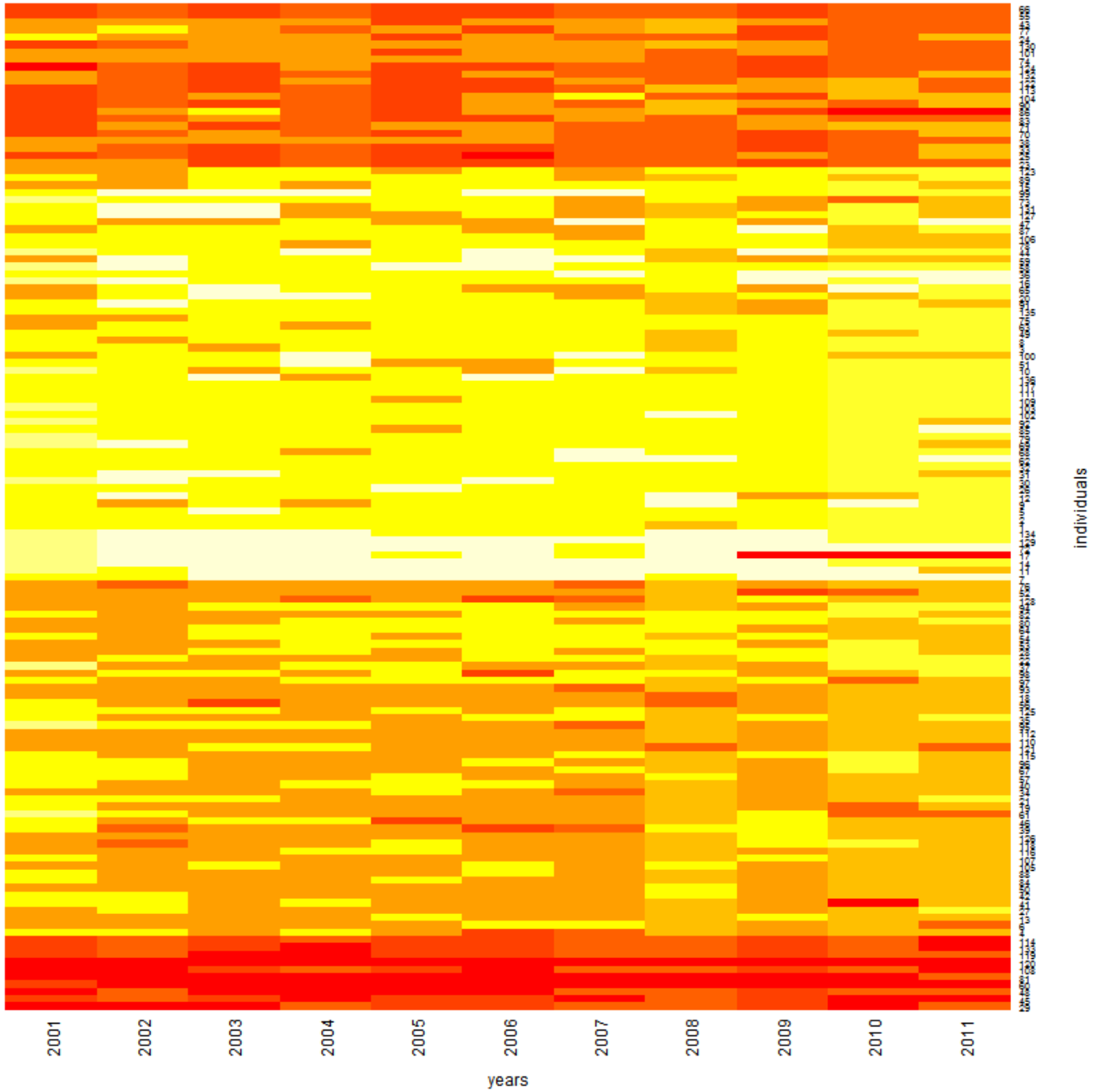
Visualisation using Heatmaps



Raw data


2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
years
individuals

Best model (R=5)

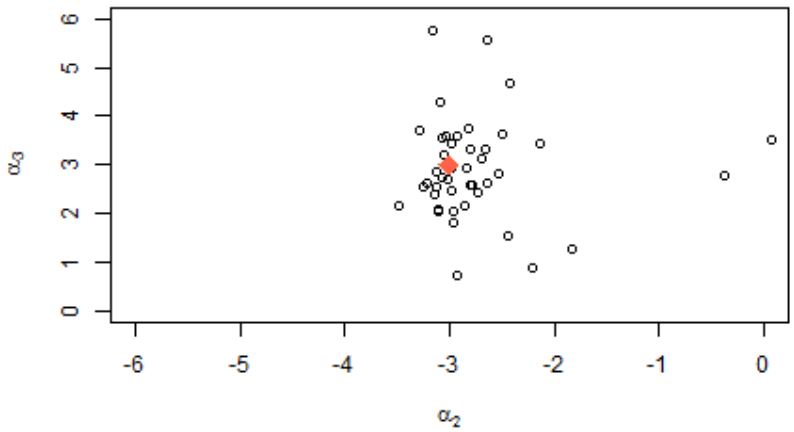


Applications: Simulation

- We simulated data and estimated the model 50 times for each scenario.
- TOM model with three row and two column clusters ($R = 3$ and $C = 2$) and $\alpha = (0, -3, 3)$, $\gamma = (0, -0.2, 0.5)$, $\beta = (0, 2)$, $\delta = (0, 0.5)$
- Same mixture proportions $\pi = (0.3\overline{3}, 0.3\overline{3}, 0.3\overline{3})$ and $\kappa = (0.5, 0.5)$.
- Five ordinal categories ($q = 5$), ten occasions $p = 10$
- Increasing sample size (n)

Applications: Simulation

n=60



n=1200

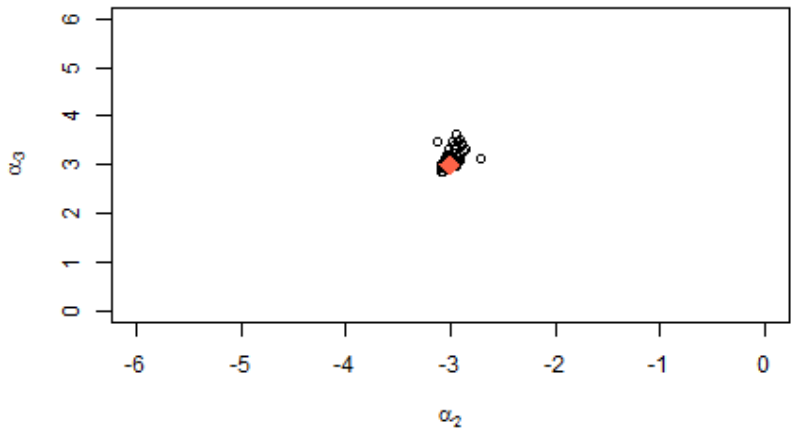
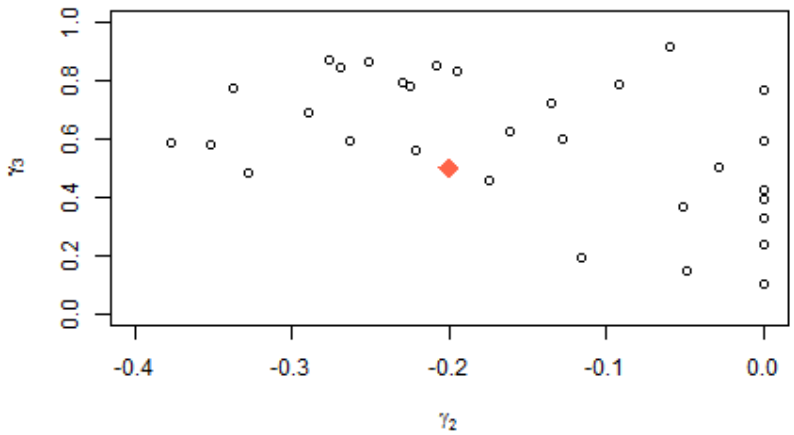


Figure : Simulation results $\alpha_2 = -3$ and $\alpha_3 = 3$

Applications: Simulation

n=60



n=1200

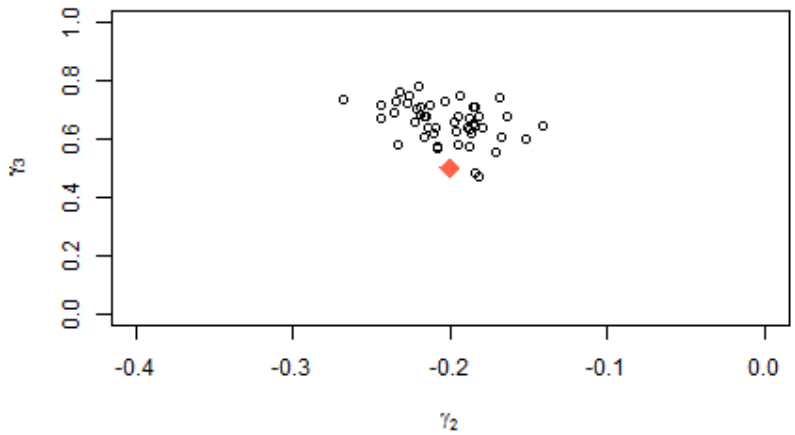
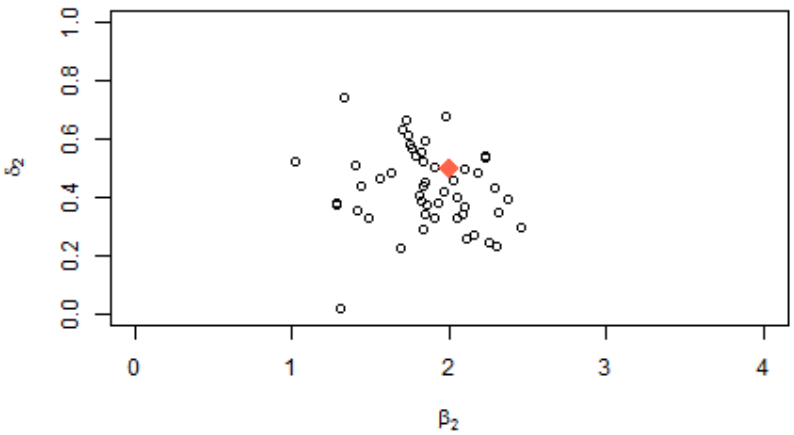


Figure : Simulation results $\gamma_2 = -0.2$ and $\gamma_3 = 0.5$

Applications: Simulation

n=60



n=1200

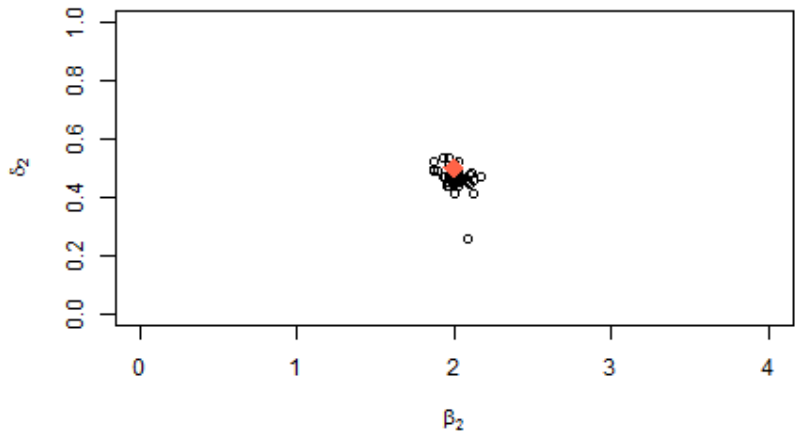


Figure : Simulation results $\beta_2 = 2$ and $\delta_2 = 0.5$

Research Goals (RG)

For the next two years:

- RG1: To determine the impact of different correlation structures on estimated clusters for repeated ordinal data
- RG2: to compare the performance of different model selection criteria
- RG3: To explore new visualisation methods for data, parameters and predictions
- RG4: To jointly model repeated ordinal and continuous data

RG1: To determine the impact of correlation structures on estimated clusters for repeated ordinal data

- Models for two-way data (unit by time period)
  - random effects by cluster and occasion
  - autoregressive order of order one (AR1)
- Models for three-way data (unit by question by time period)
  - bi-clustering in individuals and questions and an AR1 temporal correlation

RG1: Correlation structure example

- Random effects by cluster and occasion
- Repeated measures Proportional Odds Model with clustering

$$\text{Logit}[P(y_{ij} \leq k | i \in r)] = \mu_k - \alpha_r - \beta_{rj} \tag{4}$$

- Where $\beta_{rj} \sim N(0, \sigma_r^2)$

RG2: to compare the performance of different model selection criteria

- Simulation and case studies using models from RG1
- Frequentist: AIC, BIC, and ICL
- Bayesian
  - DIC and Bayes Factors, "fixed" number of mixture components
  - Reversible Jump Markov Chain Montecarlo (RJMCMC), "variable" number of mixture components

RG3: To explore new visualisation methods for data, parameters and predictions

- More heatmaps!
- Transition probability plots (initial and final distributions)
- Others?

RG4: To jointly model repeated ordinal and continuous data

- A common latent variable that could be causing both
- For instance, jointly modelling of household work status and household income overtime
  - household work status: ordinal (none member working, female working, male working, both working)
  - household income (continuous)
- Potential colaboration with SEF-VUW (who will be using simulated methods of moments)

Mosaics by Gaudi, Park Güell, Barcelona

Gracias por su atencion!!

References

- Biernacki, C., Celeux, G., Govaert, G. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". IEEE Transactions on Pattern Analysis and Machine Learning 22, 719-725
- Capuano, A., Dawson, J. (2012). "The trend odds model for ordinal data". Statistics in Medicine 32(13):2250-2261
- Mateochu, E., Liu, I., Farias, M. and Gjelsvik, B. (2013). "Biclustering models for ordinal data". Computational Statistics and Data Analysis. submitted
- McCullach, A. (1980). "Regression Models for Ordinal Data". Statistical Methodology 42, 109-142
- Pledger, S. and Arnold, R. (2013). "Multivariate methods using mixtures:Correspondence analysis, scaling and pattern-detection". Computational Statistics and Data Analysis. In-press

Column-clustering

$$\theta_{ick} = \frac{\exp(\mu_k - \beta_c - \delta_c t_k)}{1 + \exp(\mu_k - \beta_c - \delta_c t_k)} - \frac{\exp(\mu_{k-1} - \beta_c - \delta_c t_{k-1})}{1 + \exp(\mu_{k-1} - \beta_c - \delta_c t_{k-1})} \tag{5}$$

E step:

$$\hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)}}{\sum_{a=1}^{C} \kappa_a \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{iak}^{I(y_{ij}=k)}}$$

M step: Numerically maximise:

$$\ell_c = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{c=1}^{C} \sum_{k=1}^{q} \hat{x}_{jc} I(y_{ij} = k) \log(\theta_{ick}) + \sum_{j=1}^{p} \sum_{c=1}^{C} \hat{x}_{jc} \log(\kappa_c)$$

Bi-clustering

$$\theta_{rck} = \frac{\exp(\mu_k - \alpha_r - \gamma_r t_k - \beta_c - \delta_c t_k)}{1 + \exp \mu_k - \alpha_r - \gamma_r t_k - \beta_c - \delta_c t_k)} - \frac{\exp \mu_k - \alpha_r - \gamma_r t_k - \beta_c - \delta_c t_k)}{1 + \exp(\mu_k - \alpha_r - \gamma_r t_k - \beta_c - \delta_c t_k)}$$

$$\hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \kappa_c \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p} \left\{ \sum_{b=1}^{C} \kappa_b \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}}$$

$$\hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \pi_r \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{b=1}^{C} \kappa_b \prod_{i=1}^{n} \left\{ \sum_{a=1}^{R} \pi_a \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}}$$

Bi-clustering

$$\begin{aligned}\ell_c &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{q}\hat{z}_{ir}\hat{x}_{jc}I(y_{ij}=k)\log(\theta_{rck}) \\ &\quad + \sum_{i=1}^{n}\sum_{r=1}^{R}\hat{z}_{ir}\log(\pi_r) + \sum_{j=1}^{p}\sum_{c=1}^{C}\hat{x}_{jc}\log(\kappa_c).\end{aligned}$$

Repeated measures Proportional Odds Model with row-clustering

Estimation Bayesian hierarchical model with three levels: row groups (latent), individuals and occasions by individual. Given a set of parameters ϕ, an response $\mathbf{Y}$, and prior $\omega(\phi|\mathbf{Y})$, the posterior is

$$\omega(\phi|\mathbf{Y}) = \frac{\ell(\phi|\mathbf{Y})\omega(\phi)}{\int \ell(\phi|\mathbf{Y})\omega(\phi)d\phi}$$

Repeated measures Proportional Odds Model with row-clustering

$$
\begin{aligned}
y_{ij} \mid \mu, \alpha_{r_i}, \beta_{r_i j}, r_i &\sim \text{POM}(\mu, \alpha_{r_i}, \beta_{r_i j}) && i = 1 \ldots n,\ j = 1 \ldots p,\ r = 2 \ldots \\
\mu'_k \mid \sigma^2_\mu &\overset{iid}{\sim} \text{Normal}(0, \sigma^2_\mu) && \mu_k = \mu'_{(k)},\ \text{if } \mu_k > \mu_{k-1},\ k = 1 \ldots \\
\alpha_r \mid \sigma^2_\alpha &\sim \text{Normal}(0, \sigma^2_\alpha) && r = 2 \ldots R,\ \alpha_1 = 0 \\
\beta_{rj} &\sim \text{Normal}(0, \sigma^2_r) && r = 2 \ldots R,\ j = 1 \ldots p,\ \beta_{1j} = 0\ \forall \\
r_i \mid \pi &\sim \text{Categorical}(\pi) && i = 1 \ldots n \\
\sigma^2_\mu &\sim \text{Inverse Gamma}(a_\mu, b_\mu) && \\
\sigma^2_\alpha &\sim \text{Inverse Gamma}(a_\alpha, b_\alpha) && \\
\sigma^2_r &\sim \text{Inverse Gamma}(a_r, b_r) && r = 2 \ldots R \\
\pi &\sim \text{Dirichlet}(\psi) &&
\end{aligned}
$$

With (fixed known) hyperparameters: $\psi, a_\alpha, b_\alpha, a_\mu, b_\mu, a_r$, and b_r.

model's DAG

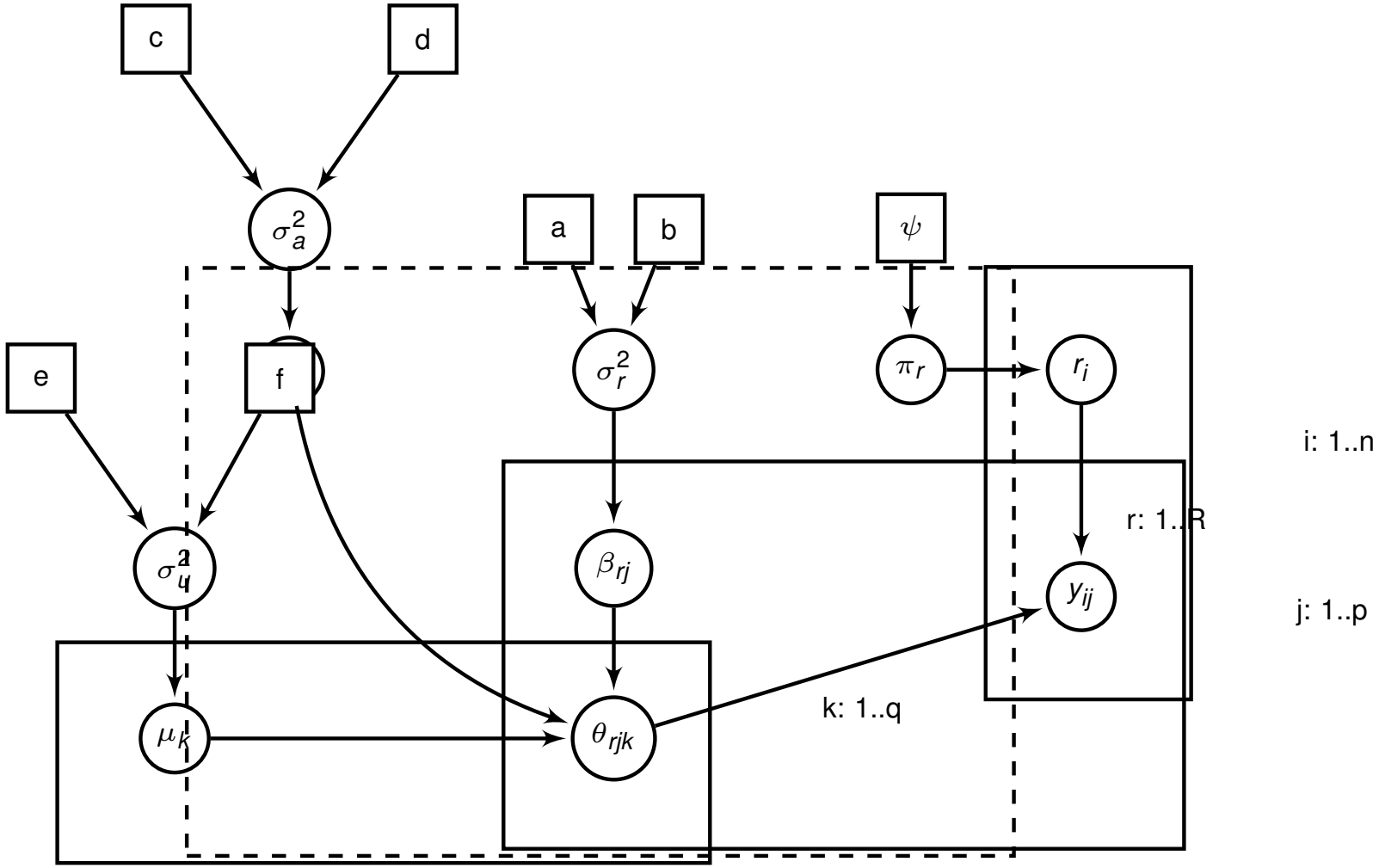
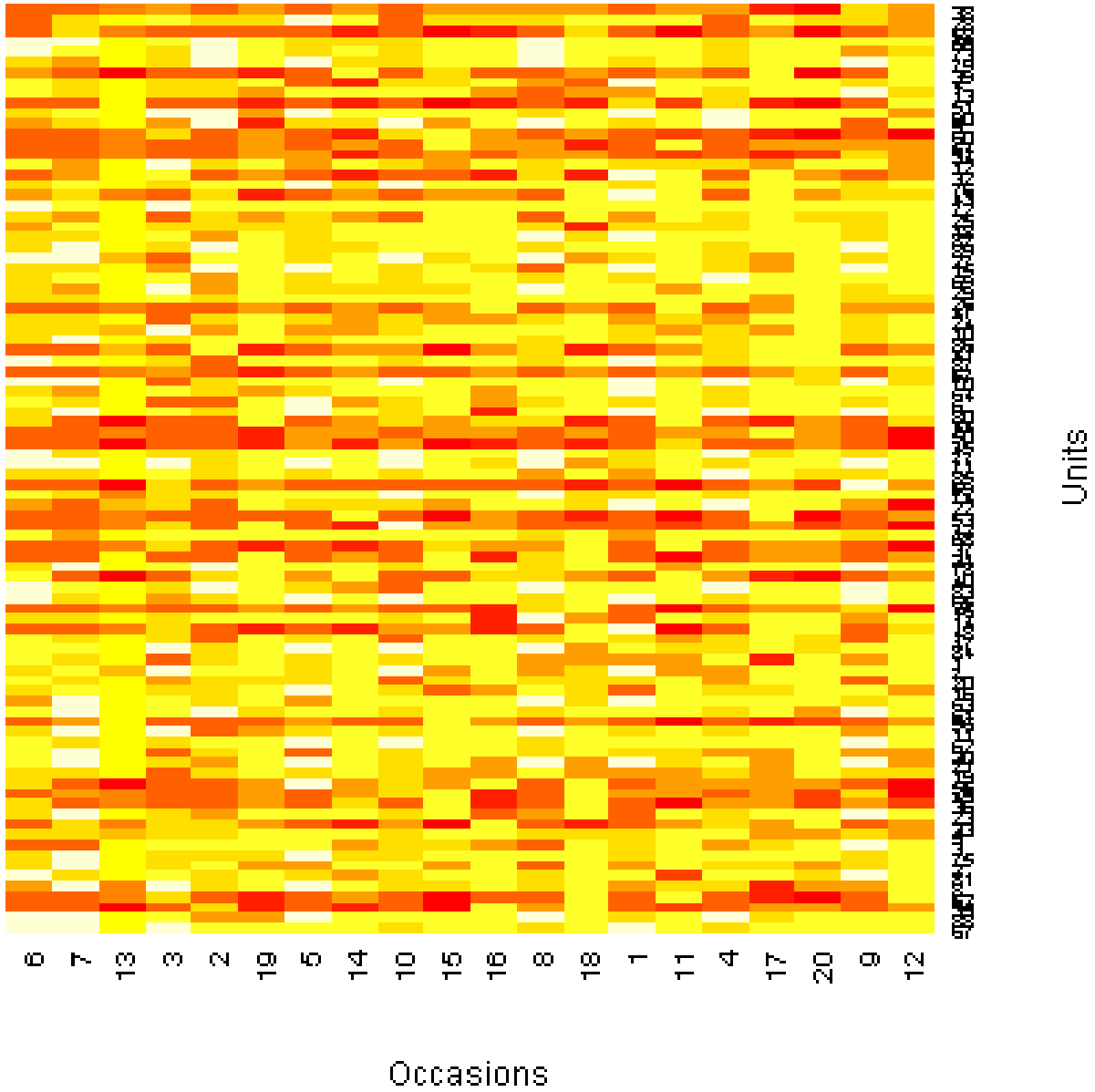


Figure : Graphical representation of REPOMC with row-clustering

Heatmap simulated bi-clustered data (R=3, C=2)



Heatmap simulated bi-clustered data (R=3, C=2)