Model-based clustering of multinomial count data

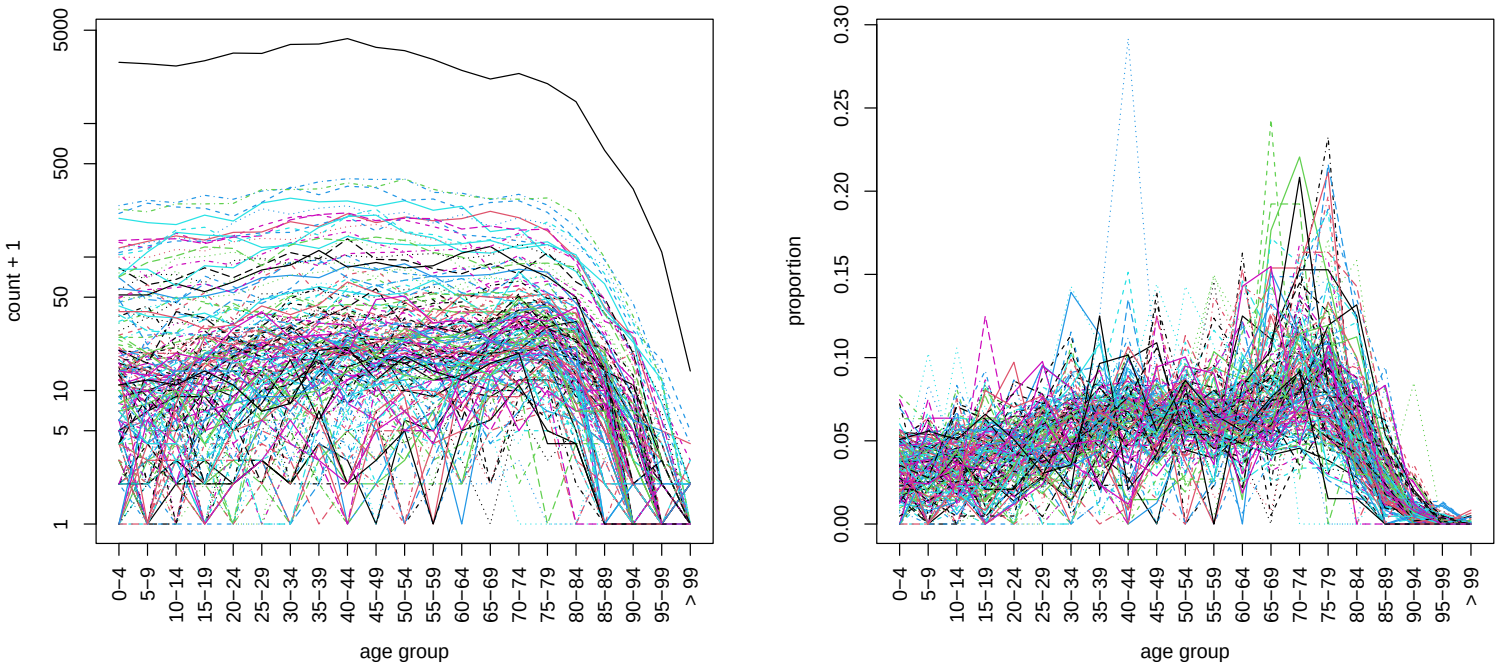
Panagiotis Papastamoulis

Department of Statistics, AUEB

Research seminar, Department of Economics
University of Crete
November 3, 2022

Overview

Motivation: dataset 1



(a) (b)

Figure: Age profiles for $n = 187$ settlements in the Phthiotis regional unit according to the 2011 census of Eurostat. (a): Population counts (increased by 1) displayed in log-scale in the y axis and (b): relative frequency of population counts.

Motivation: dataset 2

- Engagement metrics of Facebook pages for Thai fashion and cosmetics retail sellers

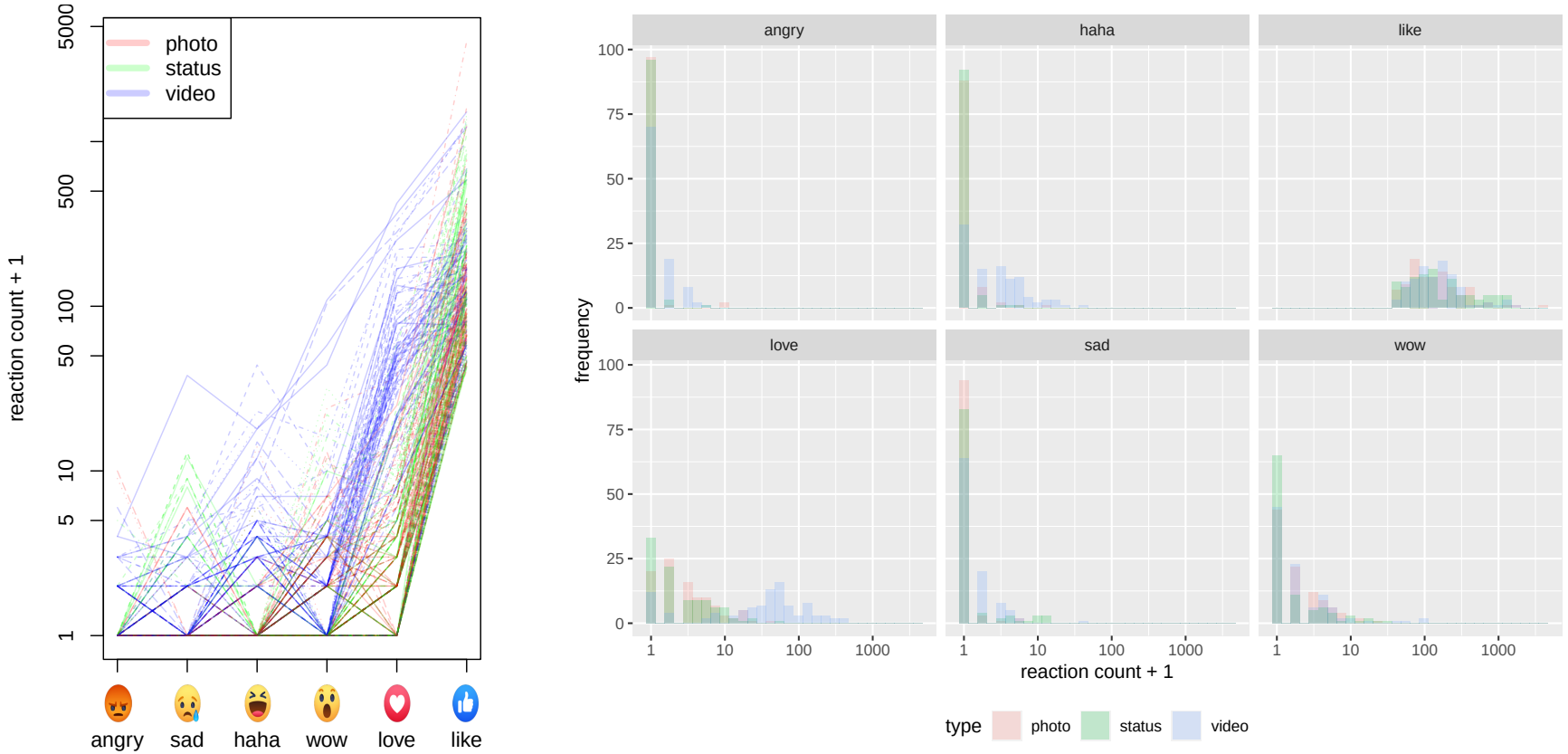


Figure: Reaction counts for 300 posts of the Facebook Live Sellers Dataset. A different colour displays the type of each post (100 video, 100 photos and 100 statuses). Note that the y axis of the left graph and x axis of the right graph is displayed in log-scale after increasing each observed count by one.

Notation

- Multinomial experiment with $J + 1$ possible outcomes
- Category probabilities

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J; \theta_{J+1})$$

  with $0 \leqslant \theta_j \leqslant 1$ and $\sum_{j=1}^{J+1} \theta_j = 1$
- $S \in \mathbb{Z}_+$ number of independent replicates of the experiment
- Y_j: number of occurences of category j, $j = 1, \ldots, j$
- Multinomial distribution

$$\boldsymbol{Y} = (Y_1, \ldots, Y_J; Y_{J+1})^\top \sim \mathcal{M}_{J+1}(S, \boldsymbol{\theta})$$

Heterogeneity

- Assume that there are K (unobserved) heterogeneous populations
- Each cluster is characterized by its own multinomial proportions

$$\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kJ}; \theta_{k;J+1})$$

for $k = 1, \ldots, K$

- (Unknown) cluster weights

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$$

where $0 \leqslant \pi_k \leqslant 1$ and $\sum_{k=1}^{K} \pi_k = 1$

Finite mixture model

- Latent multinomial random variable

$$\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\top \sim \mathcal{M}_K(1, \boldsymbol{\pi})$$

Finite mixture model

- Latent multinomial random variable

$$\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\top \sim \mathcal{M}_K(1, \boldsymbol{\pi})$$

- Conditional distribution

$$\boldsymbol{Y} | Z_k = 1 \sim \mathcal{M}_{J+1}(S, \boldsymbol{\theta}_k)$$

Finite mixture model

- Latent multinomial random variable

$$\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\top \sim \mathcal{M}_K(1, \boldsymbol{\pi})$$

- Conditional distribution

$$\boldsymbol{Y} | Z_k = 1 \sim \mathcal{M}_{J+1}(\mathrm{S}, \boldsymbol{\theta}_k)$$

- Marginal distribution

$$\boldsymbol{Y} \sim \sum_{k=1}^{K} \pi_k f(\boldsymbol{y}|\boldsymbol{\theta}_k).$$

where $f(\cdot|\boldsymbol{\theta}_k)$ denotes the probability mass function of $\mathcal{M}_J(\mathrm{S}, \boldsymbol{\theta})$

$$f(\boldsymbol{y}|\boldsymbol{\theta}_k) = \frac{\mathrm{S}!}{\prod_{j=1}^{J+1} y_j!} \prod_{j=1}^{J+1} y_j^{\theta_{kj}} \mathrm{I}_{\mathcal{Y}_{\mathrm{S},J}}(\boldsymbol{y}),$$

Covariates

Given a vector of P covariates $\boldsymbol{x} = (x_1, \ldots, x_P)$ and assuming that category $J + 1$ is the baseline, we express the log-odds as

$$\text{logit}\theta_j = \log \frac{\theta_j}{\theta_{J+1}} = \boldsymbol{\beta}_j^\top \boldsymbol{x}, \tag{1}$$

for $j = 1, \ldots, J$. The vector $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jP})^\top \in \mathbb{R}^P$ contains the regression coefficients for category j. Thus

$$\theta_j = \frac{\exp\{\boldsymbol{\beta}_j^\top \boldsymbol{x}\}}{1 + \sum_{\ell \leqslant J} \exp\{\boldsymbol{\beta}_\ell^\top \boldsymbol{x}\}}, \tag{2}$$

for $j = 1, \ldots, J$.

Mixture of multinomial logit models

Extending the previous model to the case of K latent groups, Equation (1) becomes

$$\text{logit}\theta_{kj} = \boldsymbol{\beta}_{kj}^{\top}\boldsymbol{x}, \tag{3}$$

for category $j = 1, \ldots, J$ and group-specific parameters $\boldsymbol{\beta}_{kj} = (\beta_{kj1}, \ldots, \beta_{kjP})^{\top}$, $k = 1, \ldots, K$. In analogy to (2), define

$$\theta_{kj} = \begin{cases} \dfrac{\exp\{\boldsymbol{\beta}_{kj}^{\top}\boldsymbol{x}\}}{1 + \sum_{\ell \leqslant J} \exp\{\boldsymbol{\beta}_{k\ell}^{\top}\boldsymbol{x}\}}, & j \leqslant J \\ \dfrac{1}{1 + \sum_{\ell \leqslant J} \exp\{\boldsymbol{\beta}_{k\ell}^{\top}\boldsymbol{x}\}}, & j = J + 1 \end{cases} \tag{4}$$

for $k = 1, \ldots, K$.

Data

We observe n pairs $(\boldsymbol{y}_i, \boldsymbol{x}_i)$; $i = 1, \ldots, n$, where the joint-probability mass function of $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\top$ given $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is written as

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{x}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{x}_i) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{S_i!}{\prod_{j=1}^{J+1} y_{ij}!} \prod_{j=1}^{J+1} y_{ij}^{g_{ikj}} \mathrm{I}_{\mathcal{Y}_{S_i,J}}(\boldsymbol{y}_i). \end{aligned} \tag{5}$$

where

$$g_{ikj} = \begin{cases} \dfrac{\exp\{\boldsymbol{\beta}_{kj}^\top \boldsymbol{x}_i\}}{1 + \sum_{\ell \leqslant J} \exp\{\boldsymbol{\beta}_{k\ell}^\top \boldsymbol{x}_i\}}, & j \leqslant J \\ \dfrac{1}{1 + \sum_{\ell \leqslant J} \exp\{\boldsymbol{\beta}_{k\ell}^\top \boldsymbol{x}_i\}}, & j = J+1 \end{cases} \tag{6}$$

for $i = 1, \ldots, n$; $k = 1, \ldots, K$.

Complete log-likelihood

- $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$: latent allocation vector for observation $i = 1, \ldots, n$
- Complete log-likelihood

$$\log f(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log \pi_k + \log c_i + \sum_{j=1}^{J+1} y_{ij} \log g_{ikj} \right\}, \quad (7)$$

where $c_i = s_i! / \prod_{j=1}^{J+1} y_{ij}!$.

Expected complete log-likelihood

- Posterior membership probabilities

$$w_{ik} = \mathrm{P}(Z_{ik} = 1 | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\beta}) = \frac{\pi_k f(\boldsymbol{y}_i | \boldsymbol{g}_{ik})}{\sum_{\ell=1}^{K} \pi_\ell f(\boldsymbol{y}_i | \boldsymbol{g}_{i\ell})}, \quad i = 1, \ldots, n; k = 1, \ldots, K$$

- Expected complete log-likelihood

$$\begin{aligned} Q(\boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{\pi}^{(t)}, \boldsymbol{\beta}^{(t)}) &:= \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\pi}^{(t)},\boldsymbol{\beta}^{(t)}} \left\{ \log f(\boldsymbol{y} | \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{Z}) \right\} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \left\{ \log \pi_k + \log c_i + \sum_{j=1}^{J+1} y_{ij} \log g_{ikj} \right\} \end{aligned} \tag{8}$$

M–step

- In the maximization step (M-step), (8) is maximized with respect to the parameters $\boldsymbol{\pi}, \boldsymbol{\beta}$, that is,

$$\left(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}\right) = \underset{\boldsymbol{\pi},\boldsymbol{\beta}}{\operatorname{argmax}} Q(\boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{\pi}^{(t)}, \boldsymbol{\beta}^{(t)})$$

- The maximization of the expected complete log-likelihood with respect to the mixing proportions leads to

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}, \quad k = 1, \ldots, K.$$

- The maximization with respect to $\boldsymbol{\beta}$ is analytically tractable only when $p = 1$ (that is, no covariates are present)

Partial derivatives

- First order

$$\frac{\partial Q}{\partial \beta_{kjp}} = \sum_{i=1}^{n} w_{ik} \left\{ y_{ij} - S_i g_{ikj} \right\} x_{ip}, \tag{9}$$

- Second order

$$\frac{\partial^2 Q}{\partial \beta_{kjp} \partial \beta_{k'j'p'}} = -\delta_{kk'} \sum_{i=1}^{n} S_i w_{ik} x_{ip} x_{ip'} g_{ikj} \left(\delta_{jj'} - g_{ikj'} \right),$$

where δ_{ij} denotes the Kronecker delta, for $k, k' = 1, \ldots, K$, $j, j' = 1, \ldots, J$ and $p, p' = 1, \ldots, P$.

- Gradient vector

$$\nabla Q(\boldsymbol{\beta}) := \left(\sum_{i=1}^{n} w_{i1} \left\{ \boldsymbol{y}_i - S_i \boldsymbol{g}_{i1} \right\} \otimes \boldsymbol{x}_i, \ldots, \sum_{i=1}^{n} w_{iK} \left\{ \boldsymbol{y}_i - S_i \boldsymbol{g}_{iK} \right\} \otimes \boldsymbol{x}_i, \right)^{\top}, \tag{10}$$

- Hessian matrix

$$H(\boldsymbol{\beta}) = \begin{pmatrix} H_1(\boldsymbol{\beta}_1) & 0 & \ldots & 0 \\ 0 & H_2(\boldsymbol{\beta}_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & H_K(\boldsymbol{\beta}_K) \end{pmatrix}$$

is a block diagonal matrix consisting of K blocks H_k, where each one of them being a $JP \times JP$-dimensional matrix, with

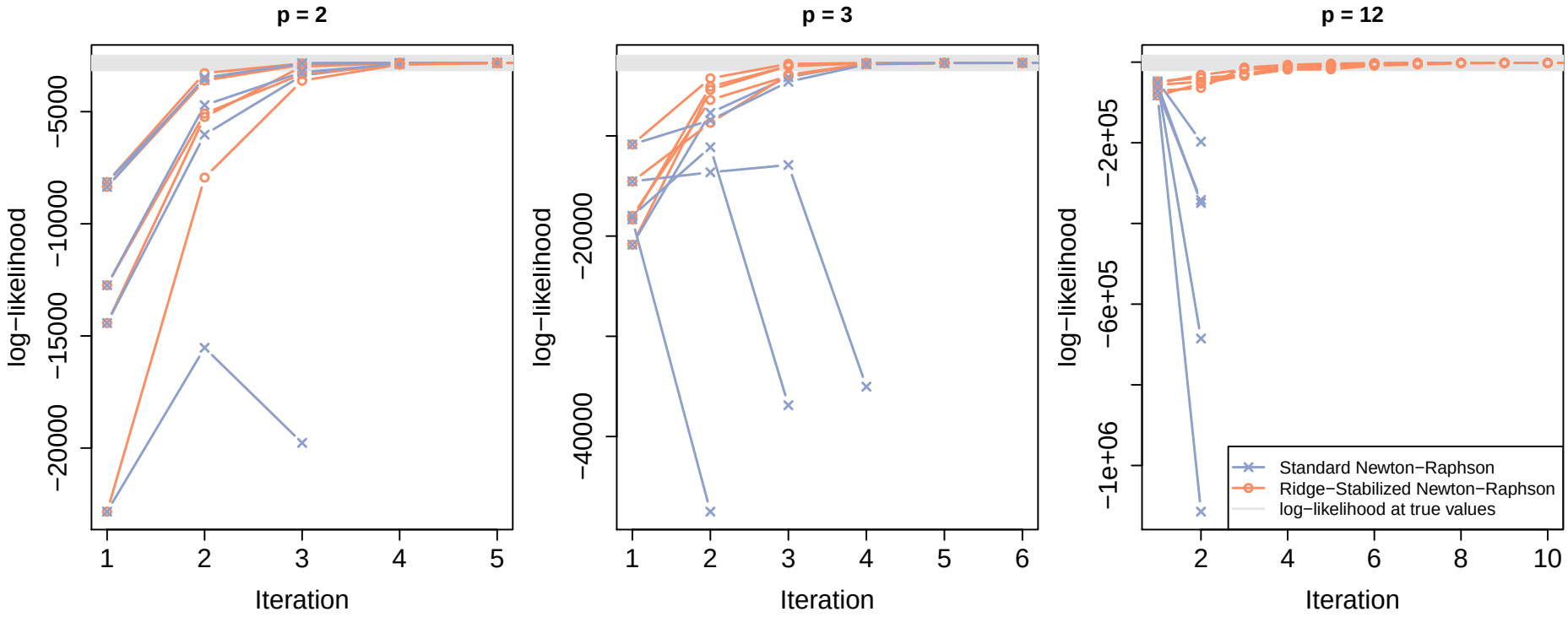
$$H_k = \left\{ \frac{\partial^2 Q}{\partial \beta_{kjp} \partial \beta_{kj'p'}} \right\}_{j=1,\ldots,J;p=1,\ldots,P}.$$

Newton-Raphson update

- Typical Newton-Raphson update at the $m+1$-th iteration

$$\boldsymbol{\beta}^{(t,m+1)} = \boldsymbol{\beta}^{(t,m)} - H^{-1}(\boldsymbol{\beta}^{(t,m)})\nabla Q(\boldsymbol{\beta}^{(t,m)}), \quad m = 1, 2, \ldots. \tag{11}$$

- In case that a second-order Taylor expansion is a good approximation of the underlying function around a maximum, the Newton-Raphson method will converge rapidly (Crockett, Chernoff, et al., 1955).
- Not quite the case here
  - The step of the basic update in Equation (11) will be too large
  - Or the Hessian will be negative definite, in which case the quadratic approximation has no validity



- Typical ($K = 1$) multinomial logit model
- $D = 6$ categories and p covariates (including constant term)
- 5 random starting values

Ridge-stabilized Newton-Raphson (Goldfeld, Quandt, and Trotter, 1966)

$$\boldsymbol{\beta}^{(t,m+1)} = \boldsymbol{\beta}^{(t,m)} - H_{\alpha}^{-1}(\boldsymbol{\beta}^{(t,m)})\nabla Q(\boldsymbol{\beta}^{(t,m)}), \quad m = 1, 2, \ldots, \tag{12}$$

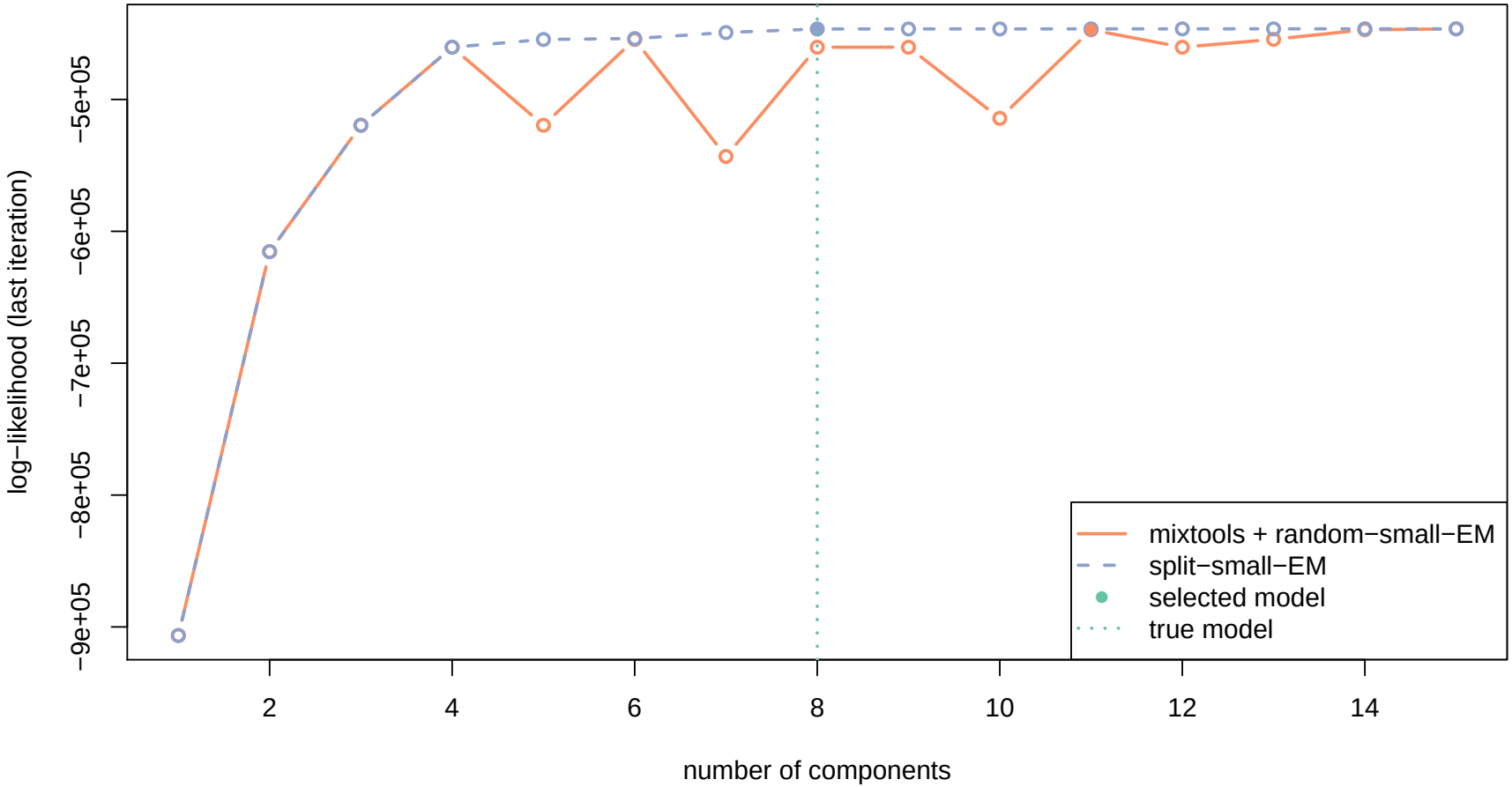
where,

$$\alpha = \lambda_1 + R\|\nabla Q(\boldsymbol{\beta}^{t-1})\| \tag{13}$$

$$H_{\alpha}(\boldsymbol{\beta}) = \begin{cases} H(\boldsymbol{\beta}) - \alpha I, & \text{if } \alpha > 0 \\ H(\boldsymbol{\beta}), & \text{if } \alpha \leqslant 0 \end{cases} \tag{14}$$

- λ_1: the largest eigenvalue of H
- The parameter R controls the step size of the update: smaller values result to larger step sizes
- It is adjusted on the run: the step size increases when the quadratic approximation appears to be satisfactory.

EM initialization is crucial



- Multinomial mixture (without covariates) with $K = 8$ components
- The information criterion used to select K is the ICL.
- `mixtools` (Benaglia et al., 2009) package + random small EM versus proposed initialization scheme

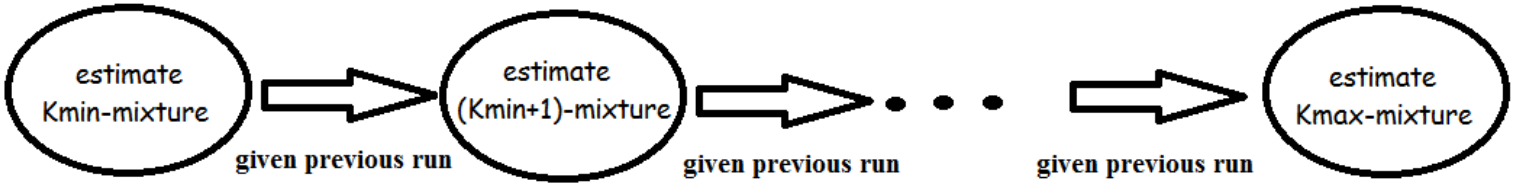
Main Idea

$K \in \{K_{\min}, K_{\min}+1, \ldots, K_{\max}-1, K_{\max}\}$

- Begin the EM algorithm from a model that underestimates the number of clusters and consecutively adding one component (Fraley, Raftery, and Wehrens, 2005)
- In our setup, this procedure begins with estimating the one-component ($K = 1$) mixture model.
- Then, for $g = 2, \ldots, K$, we estimate a g-component mixture by proposing to **randomly split clusters** obtained by the estimated model corresponding to $g - 1$ components (Papastamoulis, Martin-Magniette, and Maugis-Rabusseau, 2016).
- This procedure is embedded within a small EM strategy (Biernacki, Celeux, and Govaert, 2003)

Main Idea

$K \in \{K_{\min}, K_{\min}+1, \ldots, K_{\max}-1, K_{\max}\}$

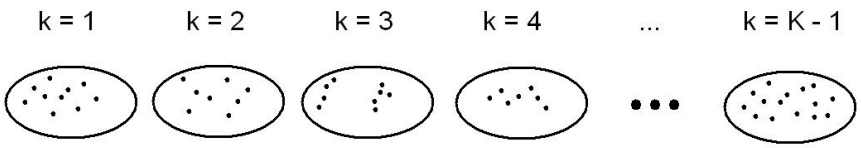


- Begin the EM algorithm from a model that underestimates the number of clusters and consecutively adding one component (Fraley, Raftery, and Wehrens, 2005)
- In our setup, this procedure begins with estimating the one-component ($K = 1$) mixture model.
- Then, for $g = 2, \ldots, K$, we estimate a g-component mixture by proposing to **randomly split clusters** obtained by the estimated model corresponding to $g - 1$ components (Papastamoulis, Martin-Magniette, and Maugis-Rabusseau, 2016).
- This procedure is embedded within a small EM strategy (Biernacki, Celeux, and Govaert, 2003)

Random split initialization

For $K > K_{\min}$

Random splitting of the estimated $K-1$ mixture



1. Simulate $u_i \sim \mathcal{U}(0,1),\ i \in I_k$
2. Run a small EM with starting values of membership probabilities:

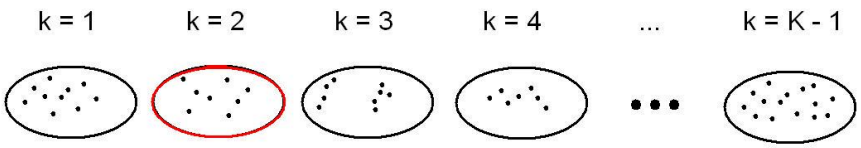
$$\begin{aligned} z_{ik}^{(0)} &= u_i \widehat{z}_{ik;K-1} \\ z_{iK}^{(0)} &= (1-u_i)\widehat{z}_{ik;K-1}, \end{aligned}$$

where $\widehat{z}_{ik;K-1} = \widehat{\mathbb{P}}(z_{ik}=1|\pi,\theta,K-1)$ denotes the estimated membership probability for observation i and $K-1$ components.

Random split initialization

For $K > K_{\min}$

Random splitting of the estimated $K-1$ mixture



1. Simulate $u_i \sim \mathcal{U}(0,1),\ i \in I_k$
2. Run a small EM with starting values of membership probabilities:

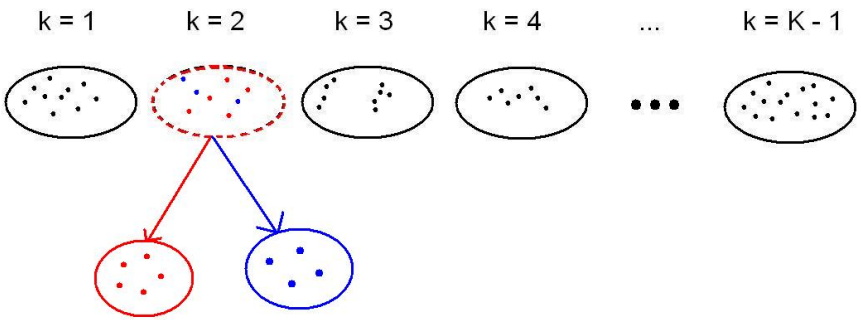
$$\begin{aligned} z_{ik}^{(0)} &= u_i \widehat{z}_{ik;K-1} \\ z_{iK}^{(0)} &= (1-u_i)\widehat{z}_{ik;K-1}, \end{aligned}$$

where $\widehat{z}_{ik;K-1} = \widehat{\mathbb{P}}(z_{ik}=1|\pi,\theta,K-1)$ denotes the estimated membership probability for observation i and $K-1$ components.

Random split initialization

For $K > K_{\min}$

Random splitting of the estimated $K-1$ mixture



1. Simulate $u_i \sim \mathcal{U}(0,1),\ i \in I_k$
2. Run a small EM with starting values of membership probabilities:

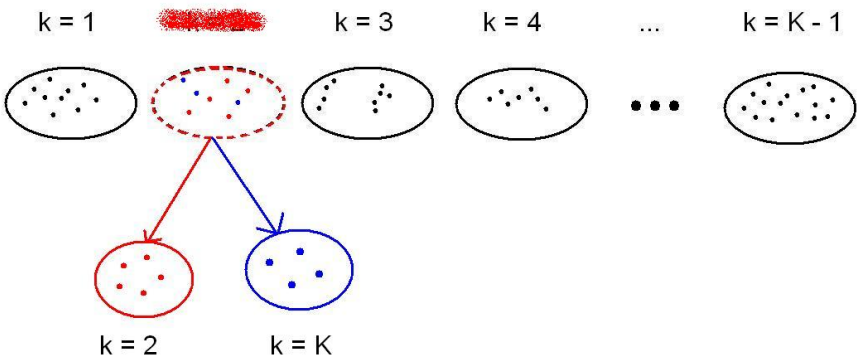
$$\begin{aligned} z_{ik}^{(0)} &= u_i \widehat{z}_{ik;K-1} \\ z_{iK}^{(0)} &= (1-u_i)\widehat{z}_{ik;K-1}, \end{aligned}$$

where $\widehat{z}_{ik;K-1} = \widehat{\mathbb{P}}(z_{ik}=1|\pi,\theta,K-1)$ denotes the estimated membership probability for observation i and $K-1$ components.

Random split initialization

For $K > K_{\min}$

Random splitting of the estimated $K-1$ mixture



1. Simulate $u_i \sim \mathcal{U}(0,1),\ i \in I_k$
2. Run a small EM with starting values of membership probabilities:

$$\begin{aligned} z_{ik}^{(0)} &= u_i \widehat{z}_{ik;K-1} \\ z_{iK}^{(0)} &= (1-u_i)\widehat{z}_{ik;K-1}, \end{aligned}$$

where $\widehat{z}_{ik;K-1} = \widehat{\mathbb{P}}(z_{ik}=1|\pi,\theta,K-1)$ denotes the estimated membership probability for observation i and $K-1$ components.

Bayesian approach

- Prior distribution
  - Mixing proportions

$$\boldsymbol{\pi} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \tag{15}$$

  - Regression coefficients

$$\beta_{kjp} \sim \mathcal{N}(0, \tau^2), \tag{16}$$

Bayesian approach

- Prior distribution
  - Mixing proportions

$$\boldsymbol{\pi} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \tag{15}$$

  - Regression coefficients

$$\beta_{kjp} \sim \mathcal{N}(0, \tau^2), \tag{16}$$

- Joint prior distribution

$$f(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\beta} | K, \boldsymbol{\alpha}, \tau) = f(\boldsymbol{Z} | \boldsymbol{\pi}, K) f(\boldsymbol{\pi} | K, \boldsymbol{\alpha}) f(\boldsymbol{\beta} | K, \tau).$$

Bayesian approach

- Prior distribution
  - Mixing proportions

$$\boldsymbol{\pi} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \tag{15}$$

  - Regression coefficients

$$\beta_{kjp} \sim \mathcal{N}(0, \tau^2), \tag{16}$$

- Joint prior distribution

$$f(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\beta} | K, \boldsymbol{\alpha}, \tau) = f(\boldsymbol{Z} | \boldsymbol{\pi}, K) f(\boldsymbol{\pi} | K, \boldsymbol{\alpha}) f(\boldsymbol{\beta} | K, \tau).$$

- The joint posterior distribution of $\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{x}, K$ is written as

$$f(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{x}, K, \boldsymbol{\alpha}, \tau) \propto f(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{\beta}, K) f(\boldsymbol{Z} | \boldsymbol{\pi}, K) f(\boldsymbol{\pi} | K, \boldsymbol{\alpha}) f(\boldsymbol{\beta} | K, \tau)$$

Hybrid Metropolis-Adjusted-Langevin within Gibbs

- Gibbs update for the allocation variables

$$\boldsymbol{z}_i|\cdots \sim \mathcal{M}(1; w_{i1}, \ldots, w_{iK}), \tag{17}$$

independent for $i = 1, \ldots, n$.

Hybrid Metropolis-Adjusted-Langevin within Gibbs

- Gibbs update for the allocation variables

$$\boldsymbol{z}_i|\cdots \sim \mathcal{M}(1; w_{i1}, \ldots, w_{iK}), \tag{17}$$

  independent for $i = 1, \ldots, n$.

- Gibbs update for mixing proportions

$$\boldsymbol{\pi}|\cdots \sim \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K), \tag{18}$$

  where $n_k = \sum_{i=1}^{n} z_{ik}$.

Hybrid Metropolis-Adjusted-Langevin within Gibbs

- Gibbs update for the allocation variables

$$\boldsymbol{z}_i|\cdots \sim \mathcal{M}(1; w_{i1}, \ldots, w_{iK}), \tag{17}$$

  independent for $i = 1, \ldots, n$.
- Gibbs update for mixing proportions

$$\boldsymbol{\pi}|\cdots \sim \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K), \tag{18}$$

  where $n_k = \sum_{i=1}^{n} z_{ik}$.
- Metropolis Adjusted Langevin proposal (Girolami and Calderhead, 2011; Roberts and Rosenthal, 1998; Roberts and Tweedie, 1996)

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)} + \nu \nabla \log f(\boldsymbol{\beta}^{(t)}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{p}) + \sqrt{2\nu}\boldsymbol{\varepsilon}, \tag{19}$$

  - $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  - $\nabla \log f(\boldsymbol{\beta}^{(t)}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{p})$: gradient of the logarithm of the full conditional of $\boldsymbol{\beta}$
  - $\nu > 0$ is adaptively tuned as the MCMC sampler progresses

Overfitting mixture model

$$\text{“True” model:} \quad x \sim \sum_{k=1}^{K^*} w_k^* f(\cdot|\theta_k^*)$$

$$\text{Fitted model:} \quad x \sim \sum_{k=1}^{K} w_k f(\cdot|\theta_k), \quad K > K^*$$

- $\theta_k \in \Theta \subseteq \mathbb{R}^d$, $k = 1, \ldots, K$
- Asymptotic behaviour of the posterior distribution on the weights
  - The posterior behaviour of overfitting mixtures depends on the chosen prior on the weights and on the number of free parameters d:

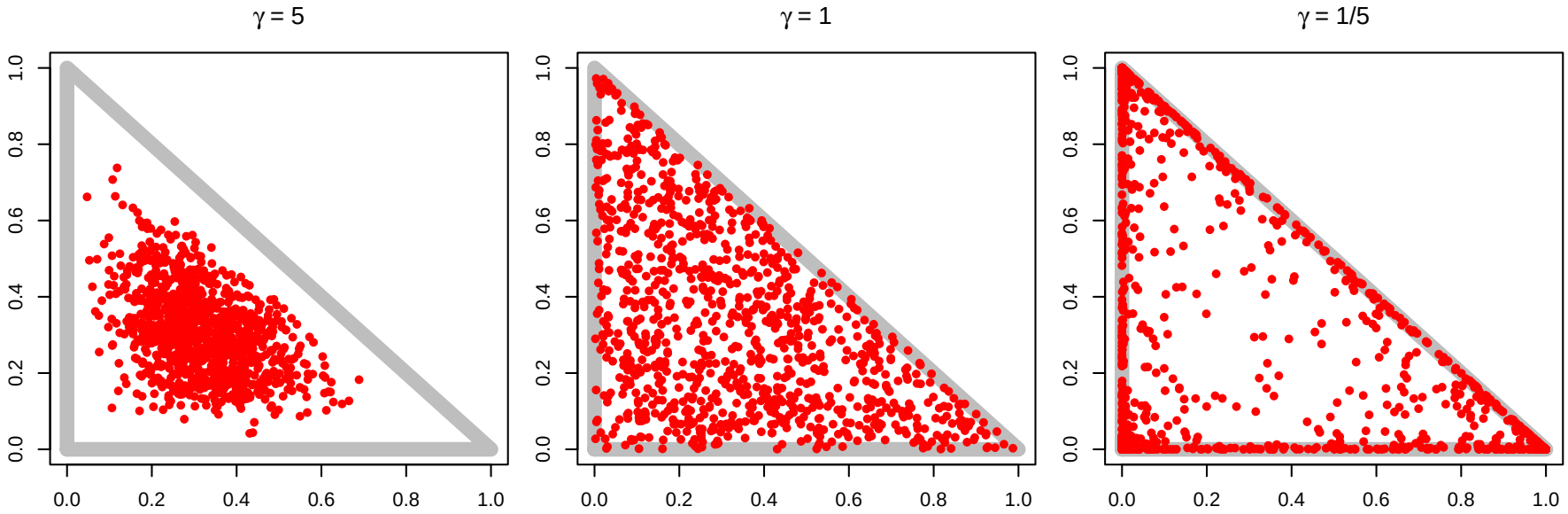
$$\boldsymbol{w} \sim \mathcal{D}(\gamma_1, \ldots, \gamma_K)$$

  - if $\max\{\gamma_k, k \leqslant K\} < d/2$ the extra mixture components are emptied at a rate of $O(n^{-1/2})$
- Inference is based on the “alive” components.

[Rousseau and Mengersen, 2011]

The importance of the Dirichlet prior

- $\boldsymbol{w} = (w_1, \ldots, w_K) \sim \mathcal{D}(\gamma, \ldots, \gamma)$
- Overfitting mixture models should favour sparsity a-priori



Simulations

- Setup

n	K	P	$J+1$
$\{125, 250, 500, 1000\}$	$\{1, \ldots, 8\}$	$\{2, 4, 6\}$	$\{6, 9, 12\}$

Table: Values for sample size (n), number of clusters (K), covariates (P) and number of categories ($J+1$) in the simulation study.

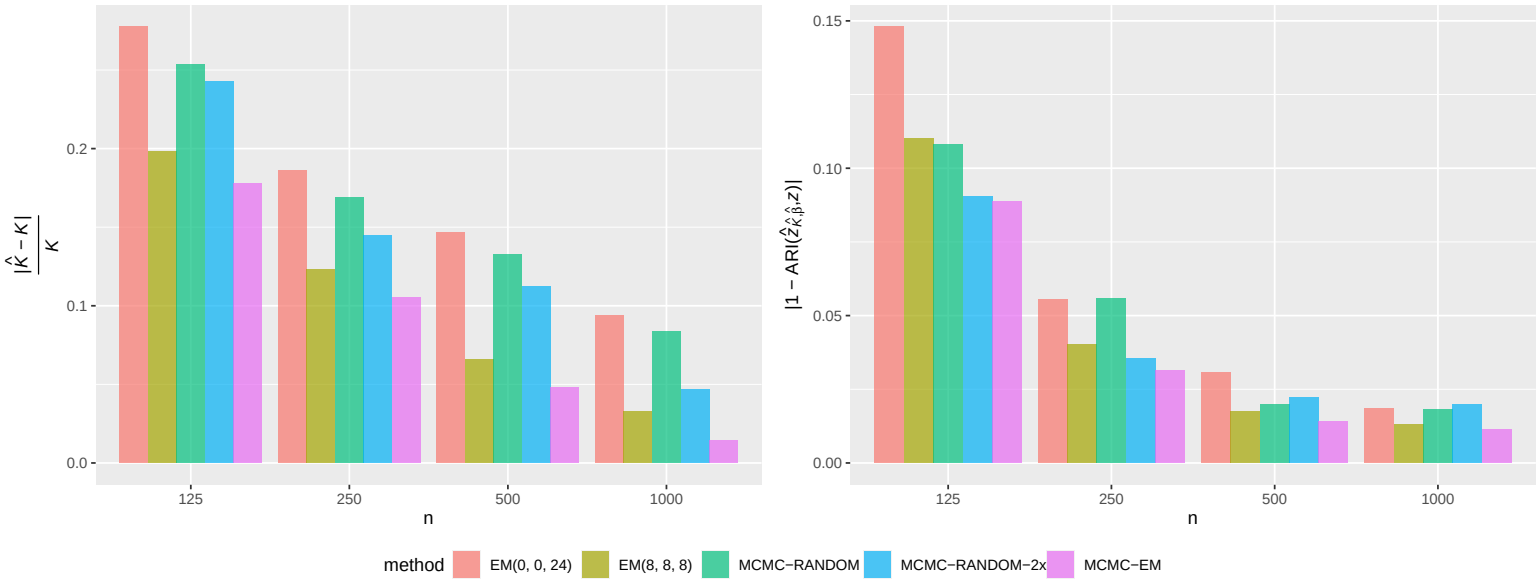
Simulations

- Setup

n	K	P	$J+1$
$\{125, 250, 500, 1000\}$	$\{1, \ldots, 8\}$	$\{2, 4, 6\}$	$\{6, 9, 12\}$

Table: Values for sample size (n), number of clusters (K), covariates (P) and number of categories ($J+1$) in the simulation study.

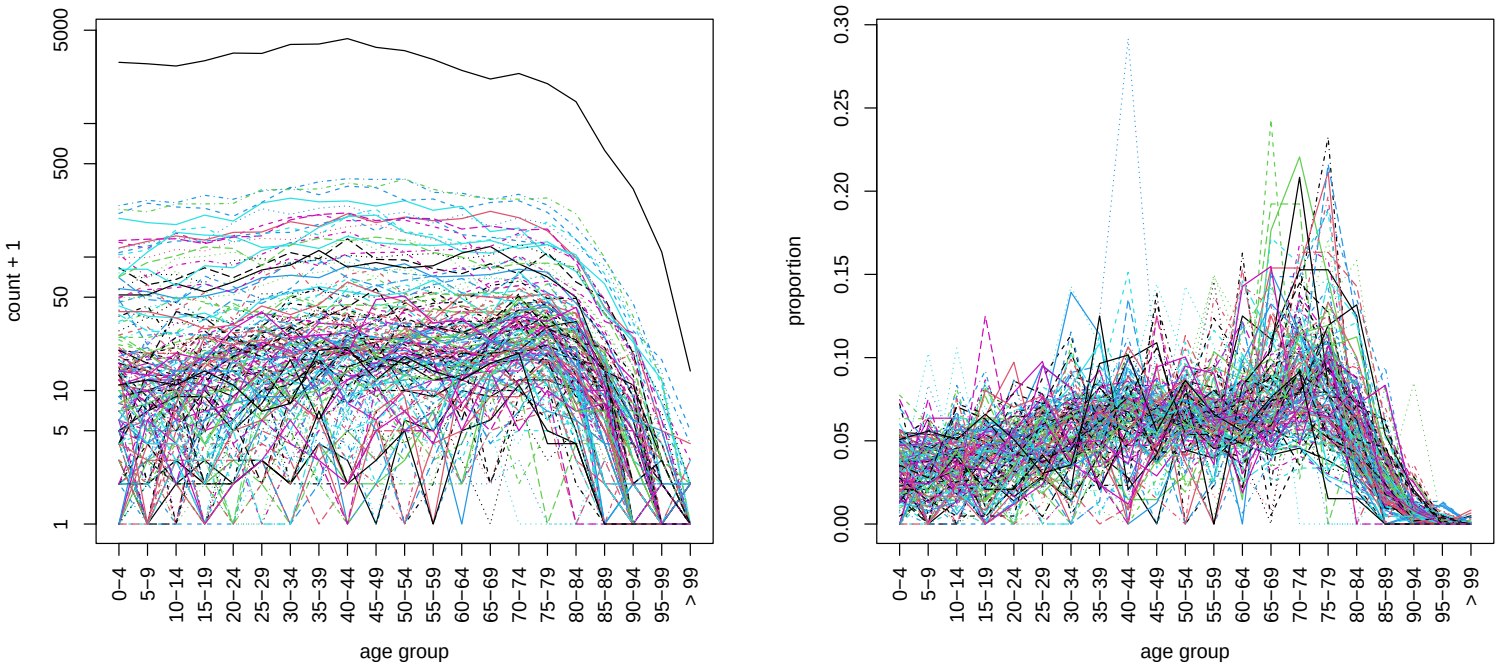
- Clustering accuracy



Phthiotis population dataset

- Population characteristics in the Phthiotis regional unit (Greece)
- Retrieved from http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census
- 21 age groups: $0-4, 5-9, \ldots, 95-99, >99$ y.o.
- Data consists of number of people per age group for $n = 187$ places
- Group the places based on their population profiles per age group
- Clustering without covariates
  - `mixtools`: 8-9 clusters
  - very small differences
- It makes sense to use covariates here such as
  - Distance from Lamia (capital city of the regional unit)
  - Altitude

Data



(a) (b)

Figure: Age profiles for $n = 187$ settlements in the Phthiotis regional unit according to the 2011 census of Eurostat. (a): Population counts (increased by 1) displayed in log-scale in the y axis and (b): relative frequency of population counts.

Model

- y_{ij}: number of people in age group j for place i
- $\boldsymbol{Y}_i|\text{cluster } k \sim \mathcal{M}_{21}(S_i, \boldsymbol{\theta}_k^{(i)})$
  - S_i: total population for place i
  - $\boldsymbol{\theta}_k^{(i)} = (\theta_{k1}^{(i)}, \ldots, \theta_{k21}^{(i)})$
  - $\theta_{kj}^{(i)}$: proportion of population being in age group j for cluster k in settlement i
- For cluster k and population group j

$$\log \frac{\theta_{kj}^{(i)}}{\theta_{k,1}^{(i)}} = \beta_{kj0} + \beta_{kj1} x_{i1} + \beta_{kj} x_{i2}, \quad j = 2, \ldots, 21$$

  x_{i1}: distance (in Km) from *Lamia* (capital city of the regional unit)
  x_{i2}: logarithm of the altitude (elevation, in m).

- the number of clusters (K) is unknown

Clustering results

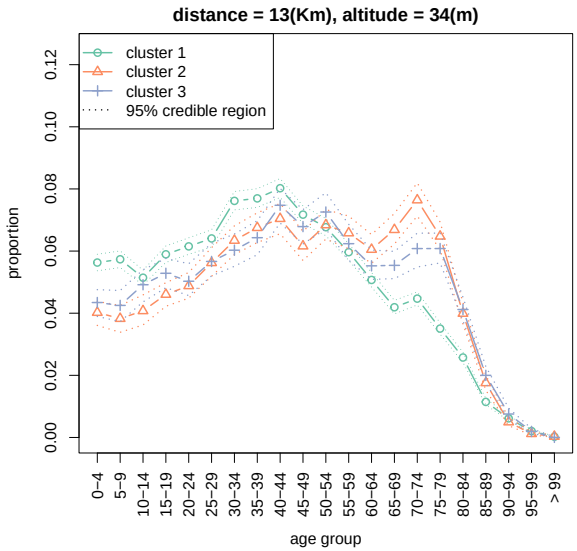
- Both EM and MCMC algorithm select 3 clusters
- Confusion matrix

		MCMC		
		1	2	3
	1	29	2	0
EM	2	1	73	5
	3	0	12	65

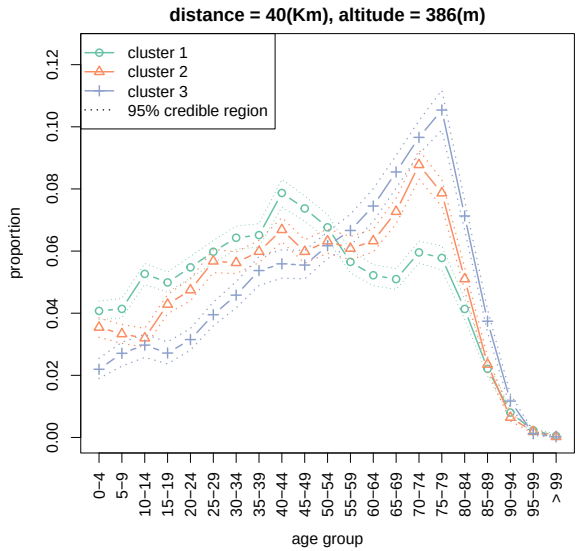
Table: Confusion matrix between the single best clustering of the Phthiotis Population Dataset arising from the EM and MCMC algorithms (after post-processing the MCMC output for correcting label switching).

- Adjusted Rand Index 0.67

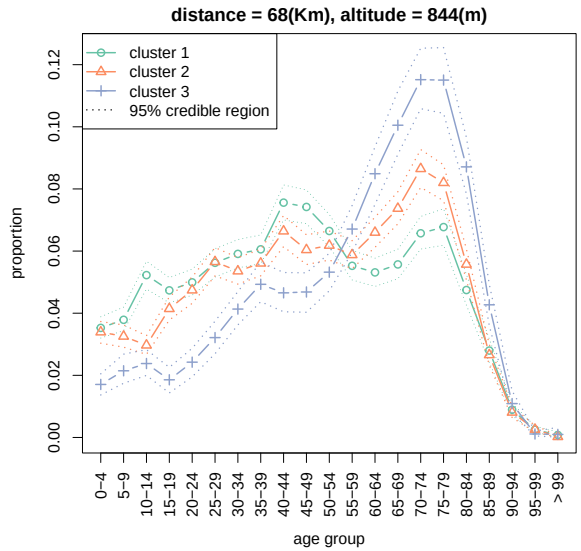
Cluster profiles



(a):



(b)



(c)

Figure: Posterior mean and 95% credible region of age profiles per cluster for the Phthiotis population data. The two covariates (distance from Lamia and altitude) are set equal to the corresponding 0.1 (a), 0.5 (b) and 0.9 (c) percentiles.

Cluster membership

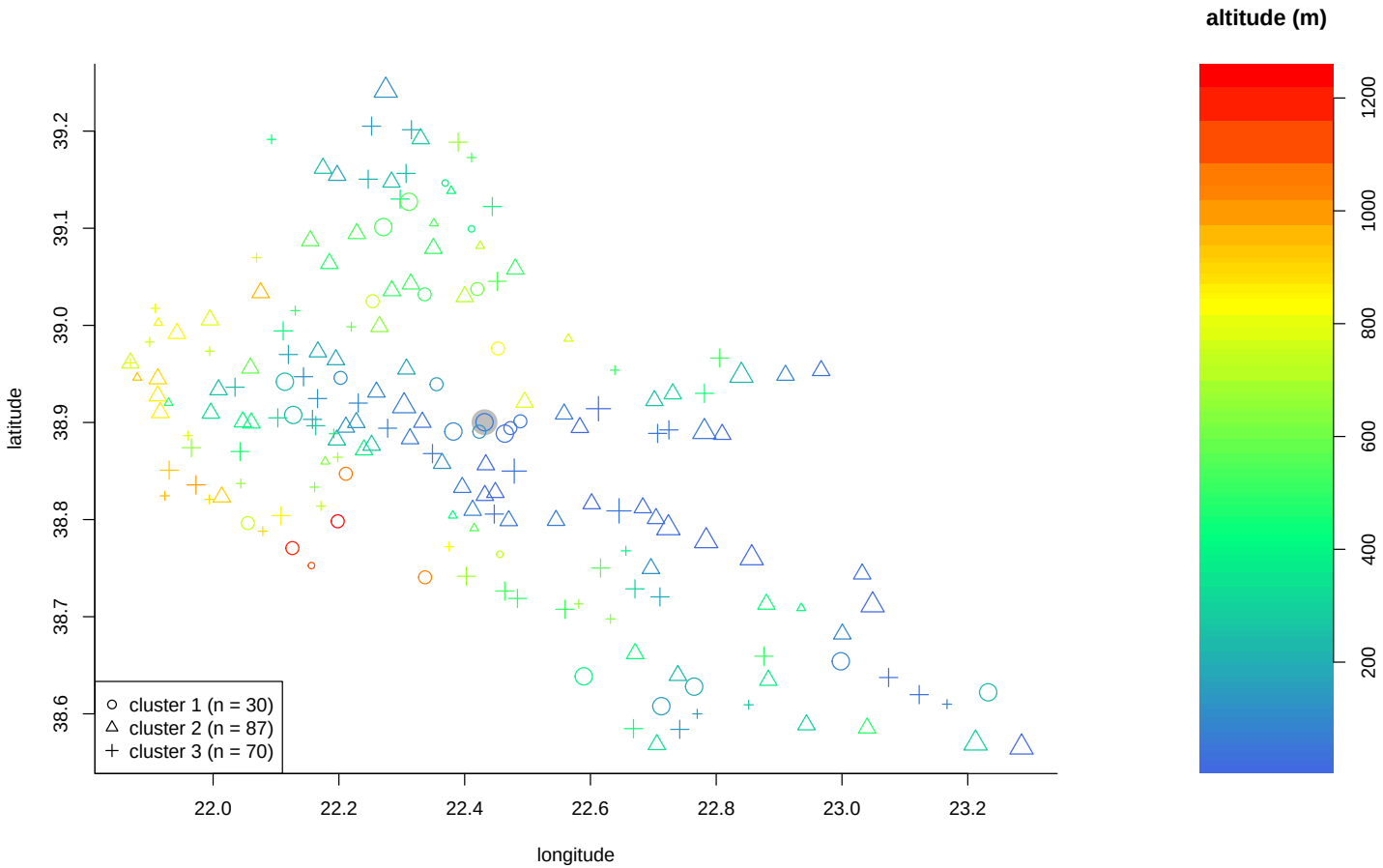
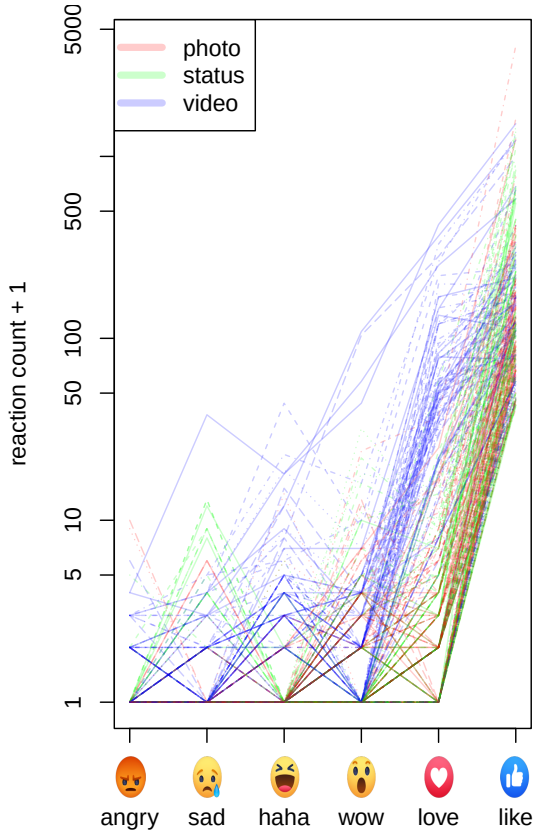


Figure: The gray circle indicates Lamia, that is, the central city of the Phthiotis region. Different point sizes are used according to the total population of each settlement: small ($S_i < 150$), medium $150 \leqslant S_i \leqslant 999$ and larger ($S_i > 999$).

UCI Facebook sellers dataset

- Engagement metrics of Facebook pages for Thai fashion and cosmetics retail sellers



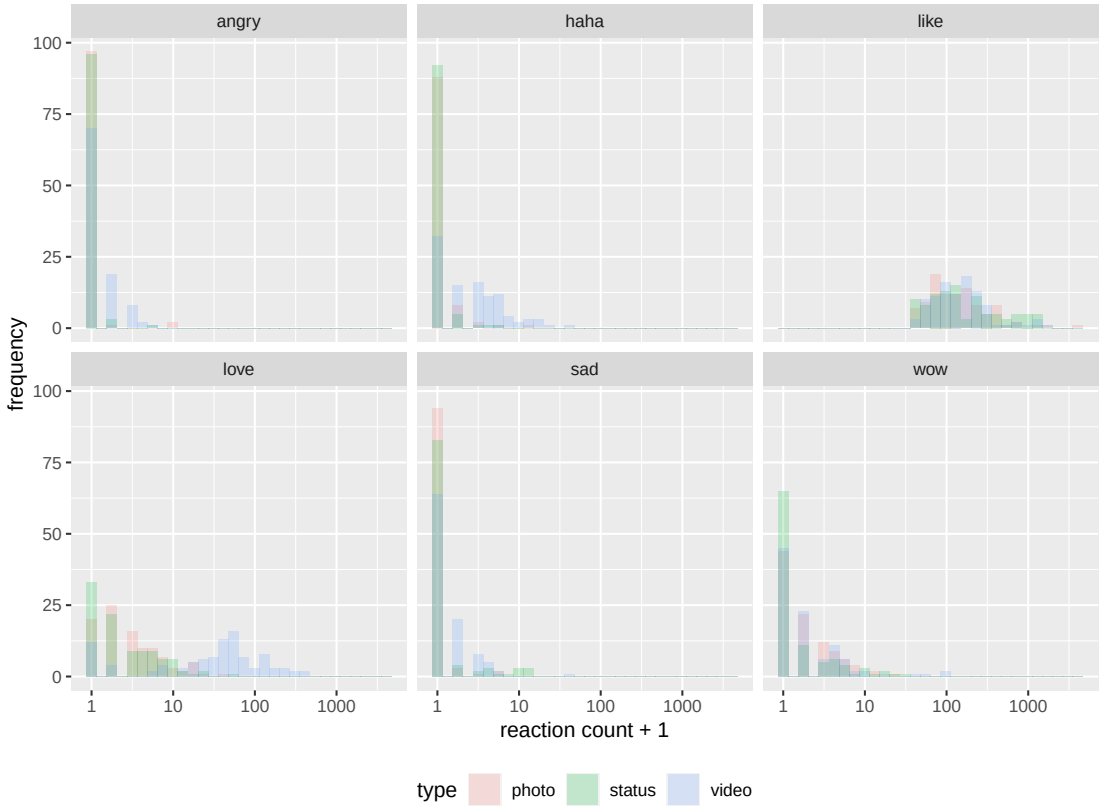


Figure: Reaction counts for 300 posts of the Facebook Live Sellers Dataset. A different colour displays the type of each post (100 video, 100 photos and 100 statuses). Note that the y axis of the left graph and x axis of the right graph is displayed in log-scale after increasing each observed count by one.

Model

- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})^\top$: reaction counts for post $i = 1, \ldots, n$ ($n = 300$)

Model

- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})^\top$: reaction counts for post $i = 1, \ldots, n$ ($n = 300$)
- The type of each post serves as a categorical predictor with three levels ("video", "photo" and "status").

Model

- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})^\top$: reaction counts for post $i = 1, \ldots, n$ ($n = 300$)
- The type of each post serves as a categorical predictor with three levels ("video", "photo" and "status").
- Selecting the probability of "like" as the reference category and conditional on cluster $k = 1, \ldots, K$, the multinomial logit model is written as

$$\log \frac{\theta_{kj}^{(i)}}{\theta_{k6}^{(i)}} = \beta_{kj0} + \beta_{kj1} x_i^{\text{status}} + \beta_{kj2} x_i^{\text{photo}}, \quad j = 1, 2, 3, 4, 5$$

where $\theta_{kj}^{(i)}$ denotes the probability of reaction j corresponding to "angry" ($j = 1$), "sad" ($j = 2$), "haha" ($j = 3$), "wow" ($j = 4$), "love" ($j = 5$) and "like" ($j = 6$).

Model

- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})^\top$: reaction counts for post $i = 1, \ldots, n$ ($n = 300$)
- The type of each post serves as a categorical predictor with three levels ("video", "photo" and "status").
- Selecting the probability of "like" as the reference category and conditional on cluster $k = 1, \ldots, K$, the multinomial logit model is written as

$$\log \frac{\theta_{kj}^{(i)}}{\theta_{k6}^{(i)}} = \beta_{kj0} + \beta_{kj1} x_i^{\text{status}} + \beta_{kj2} x_i^{\text{photo}}, \quad j = 1, 2, 3, 4, 5$$

  where $\theta_{kj}^{(i)}$ denotes the probability of reaction j corresponding to "angry" ($j = 1$), "sad" ($j = 2$), "haha" ($j = 3$), "wow" ($j = 4$), "love" ($j = 5$) and "like" ($j = 6$).
- Dummy variables

$$x_i^{\text{status}} = \begin{cases} 1, & \text{if post } i \text{ is "status"} \\ 0, & \text{otherwise} \end{cases}, x_i^{\text{photo}} = \begin{cases} 1, & \text{if post } i \text{ is "photo"} \\ 0, & \text{otherwise} \end{cases}$$

Clustering results

		MCMC					
		1	2	3	4	5	6
EM	1	153	0	1	0	0	0
	2	4	52	0	0	0	0
	3	1	0	40	0	0	0
	4	2	1	0	29	0	0
	5	0	0	0	0	10	0
	6	0	0	0	0	1	6

Table: Confusion matrix between the single best clustering of the Facebook Live Sellers Dataset arising from the EM and MCMC algorithms (after post-processing the MCMC output for correcting label switching).

Cluster profiles

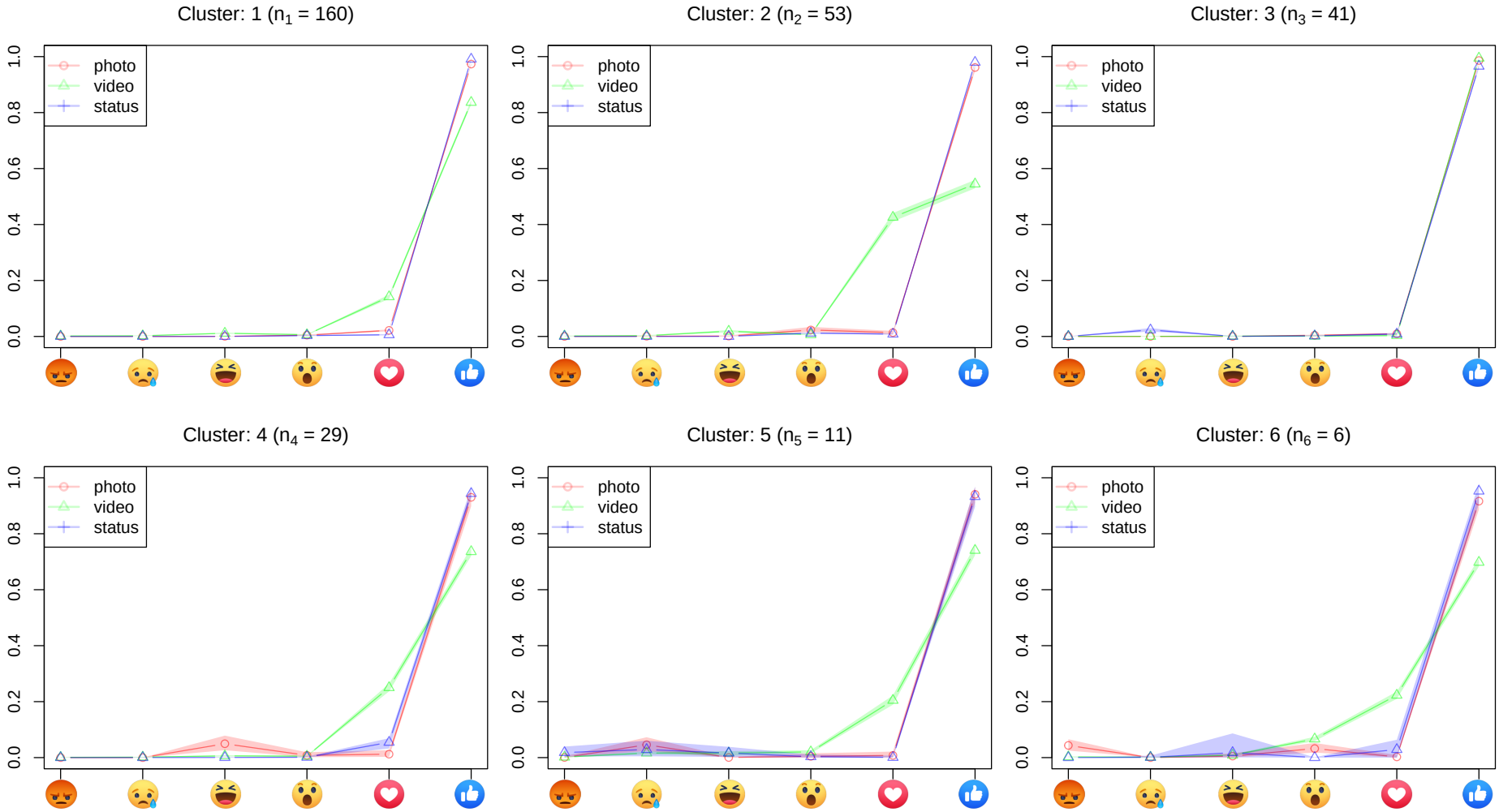
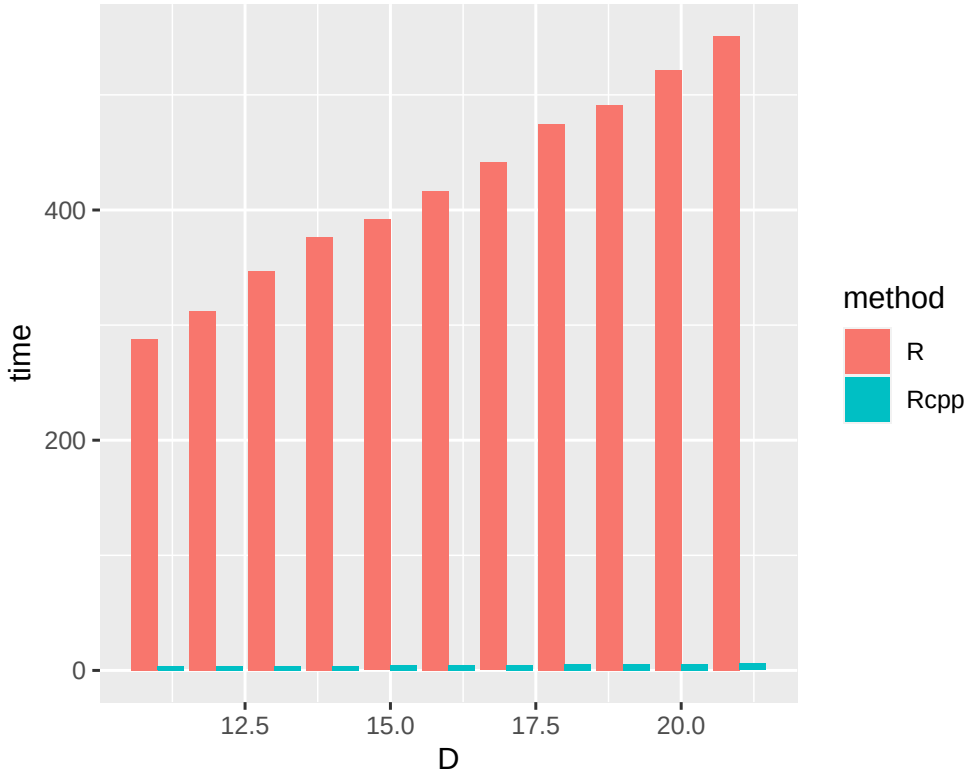


Figure: Posterior mean and 95% Credible Region of the reaction probabilities $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ per cluster for the Facebook Sellers data (clusters are displayed in decreasing order in terms of the assigned number of observations according to the MAP rule).

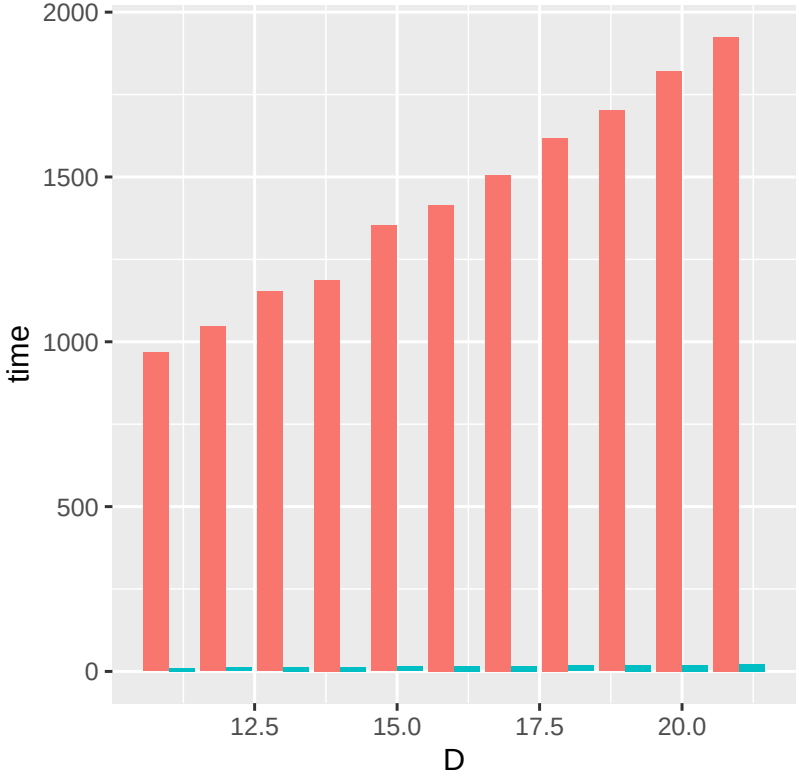
Useful extras

- Prior parallel tempering
  - Run different MCMC chains
  - Propose to exchange states
- Identifiability issues
  - ECR algorithm (Papastamoulis and Iliopoulos, 2010) in **`label.switching`** package (Papastamoulis, 2016)
  - Generic identifiability of multinomial mixtures Blischke, 1964; Grün and Leisch, 2008; Teicher, 1963; Titterington, Smith, and Makov, 1985
- Parallelization
  - **`foreach`** + **`doParallel`**
- `C++` and `R` integration
  - **`Rcpp`** + **`RcppArmadillo`** packages

Benchmarking time for MALA proposal



$K = 3$



$K = 10$

- D: number of categories
- K: number of clusters
- $n = 1000$ observations, $p = 5$ covariates

Final remarks

- Frequentint and Bayesian techiques for clustering multinomial counts under the presence of covariates

Final remarks

- Frequentint and Bayesian techiques for clustering multinomial counts under the presence of covariates
- EM algorithm
  - efficient initialization for improved results
  - implement ridge-stabilized version of Newton-Raphson updates at the M-step

Final remarks

- Frequentint and Bayesian techiques for clustering multinomial counts under the presence of covariates
- EM algorithm
  - efficient initialization for improved results
  - implement ridge-stabilized version of Newton-Raphson updates at the M-step
- MCMC scheme
  - outperforms the EM
  - allows greater flexibility in the resulting inference
  - of course, under the cost of increased computing time

Final remarks

- Frequentint and Bayesian techiques for clustering multinomial counts under the presence of covariates
- EM algorithm
  - efficient initialization for improved results
  - implement ridge-stabilized version of Newton-Raphson updates at the M-step
- MCMC scheme
  - outperforms the EM
  - allows greater flexibility in the resulting inference
  - of course, under the cost of increased computing time
- Possible extensions:
  - Variable selection
  - Further data augmentation schemes in the Bayesian model

References

Benaglia, T. et al. (2009). "mixtools: An R Package for Analyzing Finite Mixture Models". In: *Journal of Statistical Software* 32.6, pp. 1–29.

Biernacki, C., G. Celeux, and G. Govaert (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models". In: *Computational Statistics & Data Analysis* 41.3-4, pp. 561–575.

Blischke, W. R. (1964). "Estimating the Parameters of Mixtures of Binomial Distributions". In: *Journal of the American Statistical Association* 59.306, pp. 510–528. DOI: 10.1080/01621459.1964.10482176.

Crockett, J. B., H. Chernoff, et al. (1955). "Gradient methods of maximization.". In: *Pacific Journal of Mathematics* 5.1, pp. 33–50.

Fraley, C., A. Raftery, and R. Wehrens (2005). "Incremental model-based clustering for large datasets with small clusters". In: *Journal of Computational and Graphical Statistics* 14.3, pp. 529–546.

Girolami, M. and B. Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2, pp. 123–214.

Goldfeld, S. M., R. E. Quandt, and H. F. Trotter (1966). "Maximization by quadratic hill-climbing". In: *Econometrica: Journal of the Econometric Society*, pp. 541–551.

Grün, B. and F. Leisch (2008). "Identifiability of finite mixtures of multinomial logit models with varying and fixed effects". In: *Journal of classification* 25.2, pp. 225–247.

Papastamoulis, P. (2016). "label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs". In: *Journal of Statistical Software* 69.1, pp. 1–24.

Papastamoulis, P. and G. Iliopoulos (2010). "An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions". In: *Journal of Computational and Graphical Statistics* 19, pp. 313–331.

Papastamoulis, P., M.-L. Martin-Magniette, and C. Maugis-Rabusseau (2016). "On the estimation of mixtures of Poisson regression models with large number of components". In: *Computational Statistics & Data Analysis* 93, pp. 97–106.

Roberts, G. O. and J. S. Rosenthal (1998). "Optimal scaling of discrete approximations to Langevin diffusions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1, pp. 255–268. DOI: https://doi.org/10.1111/1467-9868.00123.

Roberts, G. O. and R. L. Tweedie (1996). "Exponential Convergence of Langevin Distributions and Their Discrete Approximations". In: *Bernoulli* 2.4, pp. 341–363. ISSN: 13507265.

Rousseau, J. and K. Mengersen (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, pp. 689–710.

Teicher, H. (Dec. 1963). "Identifiability of Finite Mixtures". In: *Ann. Math. Statist.* 34.4, pp. 1265–1269. DOI: 10.1214/aoms/1177703862.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley.

Further info

- Pre-print:

  Papastamoulis, P (2022): Model based clustering of multinomial count data, `arXiv:2207.13984v1`

- R package

  `https://CRAN.R-project.org/package=multinomialLogitMix`

- Reproducibility

  `https://github.com/mqbssppe/multinomialLogitMix`