Clustering Voter Types with Multinomial Outcomes

Shiro Kuriwaki*

October 2019

1 Data Generating Process

Setup Index individuals by $i \in \{1, ..., N\}$ and the universe of races excluding the top of the ticket as $j \in \{1, ..., D\}$. The data we observe is a length-D vector of votes $\mathbf{Y}_i$. Y_{ij} is a discrete response value, $Y_{ij} \in \{0, ..., L\}$.

In the simplest case $L = 1$, we code each vote Y_{ij} an indicator for splitting their ticket or not. $Y_{ij} = 1$ would mean voter i splitting their ticket in some office j, with reference to a top of the ticket office like the President or Governor. In the multinomial. case of $L = 2$, which will be our default setting, we can consider three outcomes: $Y_{ij} = 0$ indicates *abstention*, $Y_{ij} = 1$ indicates ticket *splitting* and $Y_{ij} = 2$ indicates *straight* (co-party) voting.

Parameters There are two sets of parameters: $\boldsymbol{\mu}$, the propensity for a given outcome for a given type of voter in a given office; and $\boldsymbol{\theta}$, the mixing proportions of each type. Individuals are endowed with a cluster (or type) $k \in \{1, ..., K\}$, which is drawn from a distribution governed by length-K simplex $\boldsymbol{\theta}$ (the mixing proportion).

$$Z_i \sim \text{Cat}(\boldsymbol{\theta}),$$

Set $\mu_{kj\ell} \in [0, 1]$ be the probability parameter that governs the probability of a given outcome for a given office, by a given type of voter. That is, $\boldsymbol{\mu}$ is a $\{K \times D \times (L + 1)\}$ array, where

$$\Pr\left(Y_{ij} = \ell \mid Z_i = k\right) = \mu_{kj\ell}.$$

In other words, for each individual (who is type k), their observed vector $\mathbf{Y}_i$ is governed by a length-D parameter $\boldsymbol{\mu}_k$. Therefore, we can express the joint density as follows.

$$\Pr\left(\mathbf{Y}_i \mid Z_i = k, \boldsymbol{\theta}\right) = \prod_{j=1}^{D} \text{Cat}(Y_{ij}|\boldsymbol{\mu}_k) = \prod_{j=1}^{D} \prod_{\ell=0}^{L} \mu_{kj\ell}^{\mathbf{1}(Y_{ij}=\ell)} \tag{1.1}$$

*Thanks to Shusei Eshima, Max Goplerud, Sooahn Shin, and Soichiro Yamauchi for their generous time and help.

The loop over ℓ simply represents the categorical distribution. In the binary case of $L = 1$, the Categorical reduces to a Bernoulli:

$$\Pr\left(\mathbf{Y}_i \mid Z_i = k, \boldsymbol{\theta}\right) = \prod_{j=1}^{D} \mu_{kj}^{Y_{ij}} (1 - \mu_{kj})^{1 - Y_{ij}}$$

The benefit of this modeling exercise over that from a naive sample of $N \times D$ Bernoullis is that we have captured the correlations between variables.

2 Clustering as an Unobserved Variable Problem: EM

This mixture model lends itself to clustering analysis such as K-means clustering. Although we can estimate this model in a Bayesian fashion by setting a prior for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$, the MCMC sampler Stan cannot reliably estimate clustering models like this one because of label-switching and multimodality.

Instead, we can derive the EM algorithm that is guarateed to recover the (local) maximum likelihood estimates for the target parameters. Unlike off-the-shelf algorithms like K-means, an EM approach can handle extensions such as discrete and unordered multinomial outcomes, systematic missing data, and covariates. The rest of this paper outlines the EM algorithm.[1]

Complete Likelihood If we knew the cluster assignment, we would be able to write the complete log-likelihood ($\mathcal{L}_{\text{comp}}$). First start with the joint probability of the outcome data and the cluster assignment:

$$\begin{aligned}
\Pr(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\theta}) &= \Pr(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\theta}) \Pr(\mathbf{Z} \mid \boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{D} \Pr(Y_{ij} \mid \mathbf{Z}, \boldsymbol{\mu}) \prod_{i=1}^{N} \Pr(Z_i \mid \boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{D} \prod_{k=1}^{K} \left\{ \prod_{\ell=0}^{L} \Pr\left(Y_{ij} = \ell | Z_i = k\right)^{\mathbf{1}(Y_{ij}=\ell)} \right\}^{\mathbf{1}(Z_i=k)} \prod_{i=1}^{N} \prod_{k=1}^{K} \Pr\left(Z_i = k \mid \boldsymbol{\theta}\right)^{\mathbf{1}(Z_i=k)}
\end{aligned}$$

Therefore, the complete log-likelihood is computed by taking the log of this:

$$\begin{aligned}
\mathcal{L}_{\text{comp}}(\boldsymbol{\mu}, \boldsymbol{\theta} | \mathbf{Y}, \mathbf{Z}) = &\sum_{i=1}^{N} \sum_{j=1}^{D} \sum_{k=1}^{K} \sum_{\ell=0}^{L} \mathbf{1}\{Y_{ij} = \ell, Z_i = k\} \log \Pr(Y_{ij} = \ell | Z_i = k, \boldsymbol{\mu}) \\
&+ \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{Z_i = k\} \log \Pr(Z_i = k | \boldsymbol{\theta})
\end{aligned} \tag{2.1}$$

[1] Thanks to Soichiro Yamauchi for deriving the original iteration of this algorithm for me.

We first take expectations over the latent variable Z_i,

$$\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell=0}^{L}\mathbf{1}\left(Y_{ij}=\ell\right)\mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)\right]\underbrace{\log \Pr(Y_{ij}=\ell|Z_i=k,\boldsymbol{\mu})}_{\equiv\log \boldsymbol{\mu}_{kj\ell}}$$
$$+\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)\right]\underbrace{\log \Pr(Z_i=k|\boldsymbol{\theta})}_{=\log \boldsymbol{\theta}_k} \tag{2.2}$$

Let's define this unknown quantity as

$$\zeta_{ik} \equiv \mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)\right].$$

Then the E-step can be the normalized version of the posterior probability marginalized by the mixing proportion,

$$\widehat{\zeta}_{ik} \propto \theta_k \underbrace{\prod_{j=1}^{D}\prod_{\ell=0}^{L}\left(\mu_{kj\ell}\right)^{\mathbf{1}(Y_{ij}=\ell)}}_{\equiv\boldsymbol{\mu}_{kj,Y_{ij}}} \tag{2.3}$$

The M-step is derived by taking the derivatives of $\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right]$ with respect to the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$. This leads to a MLE-like M-step, which is shown in the next section (derivation in appendix).

EM Implementation We first need to set initial values for $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$. I do this by letting $\boldsymbol{\theta}^{(0)} = \left(\frac{1}{K},\ldots\frac{1}{K}\right)$, randomly assigning an initial cluster assignment $Z_i' \sim \text{Cat}(\boldsymbol{\theta}^{(0)})$, and setting the initial $\boldsymbol{\mu}$ by the sample means of the data within those initial assignments, $\mu_{kj}^{(0)} = \frac{\sum_{i=1}^{N}\mathbf{1}(Y_{ij}=1)\mathbf{1}\left(Z_i'=k\right)}{\sum_{i=1}^{N}\mathbf{1}\left(Z_i'=k\right)}$.

Then we iterate as follows. For each voter i, compute the probability that they belong in cluster k (E-step):

$$\zeta_{ik} \leftarrow \frac{\boldsymbol{\theta}_k\prod_{j=1}^{D}\boldsymbol{\mu}_{kj,Y_{ij}}}{\sum_{k'=1}^{K}\boldsymbol{\theta}_{k'}\prod_{j=1}^{D}\boldsymbol{\mu}_{k'j,Y_{ij}}} \tag{2.4}$$

Given those type probabilities, we update the parameters in the M-step. That will show that for updating θ_k, we should take the simple average of $\widehat{\zeta}_{ik}$ across all i. For updating $\widehat{\mu}_{kj\ell}$, we should take for each k and ℓ the sample proportion of the occurrence of $Y_{ij}=\ell$, but weighted

by $\widehat{\zeta}_{ik}$:

$$\text{for each } k, \text{ update:} \quad \widehat{\theta}_k \leftarrow \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_{ik} \tag{2.5}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{kj\ell} \leftarrow \frac{\sum_{i=1}^{N}\mathbf{1}\left(Y_{ij}=\ell\right)\widehat{\zeta}_{ik}}{\sum_{i=1}^{N}\widehat{\zeta}_{ik}}, \tag{2.6}$$

repeated until convergence.

Evaluating Convergence We evaluate convergence by the observed log likelihood,

$$\mathbf{L}_{\text{obs}} = \prod_{i=1}^{N}\sum_{k=1}^{K}\theta_k\prod_{j=1}^{D}\boldsymbol{\mu}_{kj,Y_{ij}}$$

So the observed log-likelihood is

$$\mathcal{L}_{\text{obs}} = \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\theta_k\prod_{j=1}^{D}\boldsymbol{\mu}_{kj,Y_{ij}}\right\} = \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\theta_k\prod_{j=1}^{D}\prod_{\ell=0}^{L}\left(\mu_{kj\ell}\right)^{\mathbf{1}(Y_{ij}=\ell)}\right\} \tag{2.7}$$

Calculating eq. 2.7 is computationally intensive, so a quick way to check convergence is to track the maximum of the change in parameters which are all on the probability scale, i.e. $\max\left\{|\widehat{\theta}_1^{(t+1)}-\widehat{\theta}_1^{(t)}|, ..., |\widehat{\mu}_{K,D}^{(t+1)}-\widehat{\mu}_{K,D}^{(t)}|\right\}$

Speed-Ups Because this EM algorithm deals with discrete data, the algorithm needs only sufficient statistics. In our setting the unique number of voting profiles is much smaller than the number of observations, because vote vectors follow a systematic pattern and most votes are straight-ticket votes. Therefore, we can re-format the dataset so that each row is a unique combination.

Let $u \in \{1, ..., U\}$ index the unique voting profiles, and n_u be the number of such profiles in the data. We re-cycle the objects $\mathbf{Y}$ and $\boldsymbol{\zeta}$ so that each row indexes profiles rather than voters.

We repeat the EM algorithm described earlier. For each profile u, compute the probability that it belong in type k:

$$\text{for each } u, k, \text{ update:} \quad \widehat{\zeta}_{uk} \leftarrow \frac{\boldsymbol{\theta}_k\prod_{j=1}^{D}\boldsymbol{\mu}_{kj,Y_{uj}}}{\sum_{k'=1}^{K}\boldsymbol{\theta}_{k'}\prod_{j=1}^{D}\boldsymbol{\mu}_{k'j,Y_{uj}}} \tag{2.8}$$

Then given those type probabilities, update with

$$\text{for each } k, \text{ update:} \quad \widehat{\theta}_k \leftarrow \frac{1}{N}\sum_{u=1}^{U} n_u \widehat{\zeta}_{uk} \tag{2.9}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{kj\ell} \leftarrow \frac{\sum_{u=1}^{U} n_u \mathbf{1}\left(Y_{uj} = \ell\right) \widehat{\zeta}_{uk}}{\sum_{u=1}^{U} n_u \widehat{\zeta}_{uk}} \tag{2.10}$$

And the observed log-likelihood will also only require looping through the profiles:

$$\mathcal{L}_{\text{obs}} = \sum_{u=1}^{U} \log n_u + \sum_{u=1}^{U} \log \left\{ \sum_{k=1}^{K} \theta_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj,Y_{ij}} \right\} \tag{2.11}$$

3 Modeling Uncontested Races

A majority of elections for state and local offices are uncontested, which means that a voter technically votes in a choice but does not have the option to vote for one of the candidates. These qualitatively different settings require us to model *varying choice sets*.

Categories of uncontestedness In uncontested races, some options are not available to choose from. To show this, we introduce a new layer: voter i for a given office j is in one of three settings, denoted by $M_{ij} \in \{1, 2, 3\}$. Unlike the cluster Z_i, that status is exactly observed in the data.

Denote $M_{ij} = 3$ to mean vote j for voter i falls in the *contested* case, so the voter has all three options on the "menu". Denote $M_{ij} = 2$ as the case when only the *preferred party's* candidate is in the contest, so the voter only has options $Y_{ij} \in \{0, 2\}$. Finally denote $M_{ij} = 1$ as the case when only the *opposed party's* candidate is in the contest, so the voter only has the option to abstain or reluctantly (perhaps) vote for the less favored option by splitting: $Y_i \in \{0, 1\}$. For shorthand, I use the notation $\boldsymbol{S}_m$ for the set of possible of values of Y_{ij} allowed for a given category of contestedness:

$$S_m = \begin{cases} \{0, 1\} & \text{if } m = 1 \\ \{0, 2\} & \text{if } m = 2 \\ \{0, 1, 2\} & \text{if } m = 3 \end{cases}$$

Therefore the complete likelihood is modified by replacing the loop $\ell = \{0, ..., L\}$ to $\ell in S_m$.

Moving from a Multinomial model to a logit model To express the choice probability for option ℓ for office j among voters of type k, let us introduce another parameter $\boldsymbol{\psi}$ which represents the intensity of preference for option $\ell \in \{1, 2\}$ relative to $\ell = 0$ (abstention). We set the baseline for abstention to be 0, i.e. $\psi_{kj,(\ell=0)} = 0 \; \forall \; k, j$.

Parameterization To show the essence of this modeling choice, the below shows the case when there are no clusters $K = 1$ and no latent heterogeneity across individuals for simplicity. Then the indices i, k drop and we get

$$\begin{aligned}\mu_{j\ell} &= \mathbf{1}(M_j = 1)\frac{\exp(\psi_{j\ell})}{\sum_{\ell' \in \{0,1\}} \exp(\psi_{j\ell'})} + \mathbf{1}(M_j = 2)\frac{\exp(\psi_{j\ell})}{\sum_{\ell' \in \{0,2\}} \exp(\psi_{j\ell'})} + \mathbf{1}(M_j = 3)\frac{\exp(\psi_{j\ell})}{\sum_{\ell' \in \{0,1,2\}} \exp(\psi_{j\ell'})} \\ &= \sum_{m=1}^{3} \mathbf{1}(M_j = m)\frac{\exp(\psi_{j\ell})}{\sum_{\ell' \in S_m} \exp(\psi_{kj\ell'})}\end{aligned}$$

Now we add the clustering and individual indices back in for the real data, effectively another layer to account for the fact that individuals are both of a type (Z_i) and each separate office is also of a missingness type M_{ij}.

$$\mu_{kj\ell} = \frac{1}{N\theta_k}\sum_{i=1}^{N} \mathbf{1}(Z_i = k, M_{ij} = m)\frac{\exp(\psi_{kj\ell})}{\sum_{\ell' \in \boldsymbol{S}_m} \exp(\psi_{kj\ell'})} \tag{3.1}$$

Analog to Multinomial Logit Because $\exp(\psi_{kj\ell}) = 1$ for $\ell = 0$, which exists in all three components, each component is analogous to a simple multinomial logit. In the first two cases, since we consider only two possibilities, it reduces to a simple intercept-only logit. Also notice that we use the same set of parameters $\boldsymbol{\psi}_{kj}$ regardless of M_{ij}. This represents the well-known independence of irrelevant alternatives (IIA) assumption in multinomial logit. The choice probabilities when one option is not on the "menu" is assumed to follow the same type of decision rule as the ratio between the existing options.

EM Estimation We can use this new representation of the parameter μ in the EM algorithm. The three components in equation 3.1 can be cast as a multinomial logit. R packages of multinomial logit typically presume IIA if an outcome value is missing and implicitly do the kind of three-way subsetting as in equation 3.1. The required data would be of the "long" form shown in Table 1.

We will estimate this regression separately for each $k \in \{1, ..., K\}$, using the estimates of $\boldsymbol{\zeta}_k$, which represents the posterior probability of voter $i \in \{1, ..., K\}$ being in cluster k, as the weights of the logit:

$$\text{for each } k, \text{ update:} \quad \widehat{\theta}_k \leftarrow \frac{1}{N}\sum_{i=1}^{N} \widehat{\zeta}_{ik} \tag{3.2}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{kj\ell} \leftarrow \frac{\exp\left(\widehat{\psi}_{kj\ell}\right)}{1 + \exp\left(\widehat{\psi}_{kj1}\right) + \exp\left(\widehat{\psi}_{kj2}\right)} \tag{3.3}$$

Table 1: How uncontested races affect choice sets

Voter level				Voter-office level		Choice	
Voter	ζ_{i1}	$\cdots$	ζ_{iK}	Office	M_{ij} status	Option ℓ	$Y_{ij} = \ell$
$i = 1$	0.12	$\cdots$	0.05	$j = 1$	3 (contested)	0	FALSE
$i = 1$	0.12	$\cdots$	0.05	$j = 1$	3 (contested)	1	FALSE
$i = 1$	0.12	$\cdots$	0.05	$j = 1$	3 (contested)	2	TRUE
$i = 1$	0.12	$\cdots$	0.05	$j = 2$	2 (cannot split)	0	FALSE
$i = 1$	0.12	$\cdots$	0.05	$j = 2$	2 (cannot split)	1	NA
$i = 1$	0.12	$\cdots$	0.05	$j = 2$	2 (cannot split)	2	TRUE
$i = 1$	0.12	$\cdots$	0.05	$j = 3$	1 (cannot straight)	0	FALSE
$i = 1$	0.12	$\cdots$	0.05	$j = 3$	1 (cannot straight)	1	TRUE
$i = 1$	0.12	$\cdots$	0.05	$j = 3$	1 (cannot straight)	2	NA
$i = 2$	0.01	$\cdots$	0.80	$j = 1$	1 (cannot straight)	0	FALSE
$i = 2$	0.01	$\cdots$	0.80	$j = 1$	1 (cannot straight)	1	FALSE
$i = 2$	0.01	$\cdots$	0.80	$j = 1$	1 (cannot straight)	2	NA

Where the $\boldsymbol{\psi}_{kj}$ vector is estimated from the coefficients of a multinomial logit, of the form

```
mlogit(Y[[j]] ~ 1, data, weights = zeta_k).
```

Evaluating Convergence When following the EM algorithm on this data affected by uncontested choices, the observed log likelihood changes. Recall that in the no-missing case, we have equation 2.7. However, in cases of missingness, the contribution of a data point also depends on the contestedness class.

$$\mathcal{L}^{\star}_{\text{obs}} = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \theta_k \prod_{j=1}^{D} \prod_{\ell \in S_{M_{ij}}} \left\{ \left(\frac{\mu_{kj\ell}}{\sum_{\ell' \in S_{M_{ij}}} \mu_{kj\ell'}} \right)^{\mathbf{1}(Y_{ij}=\ell)} \right\} \right]$$

Appendix

A Deriving EM with complete data

Recall that the expectation of the likelihood from equation 2.2 is

$$\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell=0}^{L}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}\log\mu_{kj\ell} + \sum_{i=1}^{N}\sum_{k=1}^{K}\zeta_{ik}\log\theta_k$$

so to optimize we introduce Langrange multipliers λ and $\boldsymbol{\eta}$ for the constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{kj}$, respectively:

$$\widetilde{\mathcal{L}} = \mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] - \lambda\left(\sum_{k=1}^{K}\boldsymbol{\theta}_k - 1\right) - \sum_{k=1}^{K}\sum_{j=1}^{D}\eta_{kj}\left(\sum_{\ell=0}^{L}\mu_{kj\ell} - 1\right) \tag{A.1}$$

Then, for $\boldsymbol{\theta}$ we have that

$$\frac{\partial}{\partial\theta_k}\widetilde{\mathcal{L}} = \frac{\sum_{i=1}^{N}\zeta_{ik}}{\theta_k} - \lambda = 0$$

along with the constraint $\sum_{k=1}^{K}\theta_k = 1$. Notice that when we sum the FOC for $\boldsymbol{\theta}$ across k, the first condition becomes $\sum_{k=1}^{K}\theta_k = \frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=i}^{N}\zeta_{ik}$, and because the LHS sums to 1 due to the constraint and in the RHS $\sum_{i=1}^{N}\sum_{k'=1}^{K}\zeta_{ik'}$ sums to N, we have $\lambda = N$.

Separately, for $\boldsymbol{\mu}_{kj}$ we have that

$$\frac{\partial}{\partial\mu_{kj\ell}}\widetilde{\mathcal{L}} = \frac{\sum_{i=1}^{N}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}}{\boldsymbol{\mu}_{kj\ell}} - \eta_{kj} = 0,$$

along with constraint $\sum_{\ell=0}^{L}\mu_{kj\ell} = 1$. Once we sum the FOC for $\boldsymbol{\mu}$ across ℓ the first condition becomes $\sum_{\ell=0}^{L}\mu_{kj\ell} = \frac{1}{\eta_{kj}}\sum_{i=1}\sum_{\ell=0}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}$, and because the LHS again sums to 1 and in the RHS $\sum_{i=1}^{N}\sum_{\ell=0}^{L}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}$ sums to the prevalence of the weights $\sum_{i=1}^{N}\zeta_{ik}$, we get $\eta_{kj} = \sum_{i=1}^{N}\zeta_{ik}$.

Together, the above imply that

$$\theta_k = \frac{1}{N}\sum_{i=1}^{N}\zeta_{ik} \quad \text{and} \quad \mu_{kj\ell} = \frac{\sum_{i=1}^{N}\mathbf{1}\{Y_{ij}=\ell\}\zeta_{ik}}{\sum_{i=1}^{N}\zeta_{ik}} \tag{A.2}$$

B Deriving EM with censored data

The modified log likelihood is

$$\mathcal{L}_{\text{comp}}(\boldsymbol{\mu}, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell \in S_{M_{ij}}} \mathbf{1}\left(Y_{ij} = \ell, Z_i = k\right) \log \Pr(Y_{ij} = \ell | Z_i = k, M_{ij} = m, \boldsymbol{\mu}) + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}\left(Z_i = k\right) \log \Pr(Z_i = k|\boldsymbol{\theta})$$

And the expected log likelihood, taking expectations over Z_i is

$$\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell \in M_{ij}}^{L} \mathbf{1}\left(Y_{ij} = \ell\right) \zeta_{ik} \underbrace{\log \Pr(Y_{ij} = \ell | Z_i = k, M_{ij} = m, \boldsymbol{\mu})}_{=\log\left(\frac{\mu_{kj\ell}}{\sum_{\ell' \in S_m} \mu_{kj\ell}}\right)} + \sum_{i=1}^{N}\sum_{k=1}^{K} \zeta_{ik} \log \boldsymbol{\theta}_k \tag{B.1}$$

E-step Then the E-step can be the normalized version of the posterior probability marginalized by the mixing proportion,

$$\widehat{\zeta}_{ik} \propto \theta_k \underbrace{\prod_{j=1}^{D}\prod_{\ell \in S_m}^{L}\left(\frac{\mu_{kj\ell}}{\sum_{\ell' \in S_m} \mu_{kj\ell'}}\right)^{\mathbf{1}(Y_{ij}=\ell)}}_{\equiv \boldsymbol{\mu}_{kj, Y_{ij}, M_{ij}}} \tag{B.2}$$

So in the E-step, we would be updating by this probability:

$$\zeta_{ik} \leftarrow \frac{\boldsymbol{\theta}_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj, Y_{ij}, M_{ij}}}{\sum_{k'=1}^{K} \boldsymbol{\theta}_{k'} \prod_{j=1}^{D} \boldsymbol{\mu}_{k'j, Y_{ij}, M_{ij}}} \tag{B.3}$$

M-step The M-step involves taking the derivative of one more layer of complication. Re-using notation we introduce Langrange multipliers λ and $\boldsymbol{\eta}$ for the constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_{kj}$, respectively and modify eq. A.1 as:

$$\widetilde{\mathcal{L}} = \mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] - \lambda\left(\sum_{k=1}^{K} \boldsymbol{\theta}_k - 1\right) - \sum_{k=1}^{K}\sum_{j=1}^{D} \eta_{kj}\left(\sum_{\ell \in S_m}\left(\frac{\mu_{kj\ell}}{\sum_{\ell' \in S_m} \mu_{kj\ell'}}\right) - 1\right)$$

We can deduce from the structure of $\boldsymbol{\mu}$ array, the equivalent thing to the E-step in the complete data case is to run a standard multinomial logit with a IIA assumption (also called conditional logit).