An EM Algorithm for Clustering Voter Types

Shiro Kuriwaki*

June 2020

*Ph.D. Candidate, Department of Government and Institute of Quantitative Social Science, Harvard University. Thanks to Shusei Eshima, Max Goplerud, Sooahn Shin, for their help. Thanks especially to Soichiro Yamauchi for his extensive help, including the suggestion of EM and deriving the original iteration of this algorithm for me.

1 Data Generating Process

1.1 Setup

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

Y_{ij} is recoded with respect to the top of the ticket. In the simplest case $L = 1$, Y_{ij} is an indicator for a split ticket. $Y_{ij} = 1$ 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 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.

1.2 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 $\boldsymbol{\pi}$, 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 $\boldsymbol{\pi}$ (the mixing proportion).

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

Here $\boldsymbol{\pi}$ is the same for every individual, indicating that before we observe outcomes, everyone has the same probabilty of being in a particular cluster. We later incorporate demographic covariates that changes the prior probability. In the code, therefore, we deal with the $N \times K$ matrix instead of a vector.

Let $\mu_{kj\ell} \in [0, 1]$ 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 $\{K \times J \times (L + 1)\}$ 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-J parameter $\boldsymbol{\mu}_k$. Therefore, we can express the joint density as follows.

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

The loop over ℓ simply represents the categorical distribution. In the binary case of $L = 1$, the

Categorical reduces to a Bernoulli:

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

2 Clustering as an Unobserved Variable Problem: EM

The common way to estimate mixture models is by using K-means, an iterative but deterministic algorithm. The usual MCMC sampler cannot reliably estimate clustering models like this one because of label-switching and multimodality. Moreover, the majority of existing clustering methods are dealt to model continuous or binary outcomes.

Instead, here I derive an Expectation Maximization (EM) algorithm, which is probabilistic and guaranteed to recover the (local) maximum likelihood estimates of the target parameters. Unlike off-the-shelf algorithms, this can handle extensions such as discrete and unordered multinomial outcomes, systematic missing data, and covariates. The rest of this paper outlines the EM algorithm.

2.1 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:

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

Therefore, the complete log-likelihood is:

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

To derive the EM algorithm, we first take expectations over the *posterior distribution* of the latent variable Z_i, therefore conditioning on the data and the initial parameter values:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] &= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=0}^{L}\mathbf{1}\left(Y_{ij}=\ell\right)\mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)|\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu}\right]\underbrace{\log\Pr(Y_{ij}=\ell|Z_i=k,\boldsymbol{\mu})}_{\equiv\log\mu_{kj\ell}} \\
&\quad + \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)|\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu}\right]\underbrace{\log\Pr(Z_i=k|\boldsymbol{\pi})}_{=\log\pi_k} \qquad (2.2) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=0}^{L}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}\log\mu_{kj\ell} + \sum_{i=1}^{N}\sum_{k=1}^{K}\zeta_{ik}\log\pi_k
\end{aligned}$$

where we represent the new unknown quantity in equation 2.2 as

$$\zeta_{ik} \equiv \mathbb{E}\left[\mathbf{1}\left(Z_i=k\right)\mid\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu}\right].$$

Going from 2.1 to 2.2 takes this form because in the second component of the sum, by the definition of expectation of a function of a discrete r.v.,

$$\begin{aligned}
\mathbb{E}_{Z_i\sim p(Z_i|\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu})}\left[\mathbf{1}\left(Z_i=k\right)\right] &= \sum_{z'=1}^{K}\Pr(Z_i=z'\mid\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu})\mathbf{1}\left(z_i'=k\right) \\
&= \Pr(Z_i=k\mid\mathbf{Y}_i,\boldsymbol{\pi},\boldsymbol{\mu})
\end{aligned}$$

where we get the last line because this is an expectation on a function of Z_i so we must go through each z' by plugging in all the potential values into Z_i. Only when $z'=k$ will $\mathbf{1}\left(Z_i=k\right)=1$, so we can simply reduce the sum to that case.

2.2 EM Implementation

The E-step in the algorithm is to compute this posterior membership probability. We can see that this should be proportional to the likelihood of observing the voting pattern we do see, with weights for the mixing proportion:

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

This is true because the posterior probability is the prior of the cluster multiplied by the likelihood of the data given the parameter under that cluster. i.e. by Bayes rule,

$$\begin{aligned}\zeta_{ik} &= \mathbb{E}\left[\mathbf{1}\left(Z_i = k\right)|\mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\mu}\right] \\ &= \Pr(Z_i = k|\mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\mu}) \\ &\propto \Pr(Z_i = k|\boldsymbol{\pi}, \boldsymbol{\mu}) p(\mathbf{Y}_i|Z_i = k, \boldsymbol{\pi}, \boldsymbol{\mu}) \\ &= \pi_k \prod_{j=1}^{J}\prod_{\ell=0}^{L}(\mu_{kj\ell})^{\mathbf{1}(Y_{ij}=\ell)}.\end{aligned}$$

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 $\boldsymbol{\pi}$. This leads to a MLE-like M-step, which is shown in the next section (derivation in Appendix A), and the results are shown here.

E-step For each voter i, compute the probability that they belong in cluster k:

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

M-step Given those type probabilities, we update the parameters in the M-step. That will show that for updating π_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 $\widehat{\zeta}_{ik}$:

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

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

repeated until convergence.

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

2.3 Evaluating Convergence

We evaluate convergence by the observed log likelihood,

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

So the observed log-likelihood is

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

At each iteration, I check the relative change in the observed log likelihood ($|\mathcal{L}_{\text{obs}}^{(t)} - \mathcal{L}_{\text{obs}}^{(t-1)}| / |\mathcal{L}_{\text{obs}}^{(t-1)}|$) and declare convergence once that is smaller than a small threshold (e.g. 10^{-5}).

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. $\max \left\{ |\widehat{\pi}_1^{(t+1)} - \widehat{\pi}_1^{(t)}|, ..., |\widehat{\mu}_{K,J}^{(t+1)} - \widehat{\mu}_{K,J}^{(t)}| \right\}$

3 Speed-Up by Collapsing to Unique Profiles

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.

This is only the case when there are no demographic covariates — When we allow for demographic covariates that vary at the individual level, this no longer holds.

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:

$$\text{for each } u, k, \text{ update:} \quad \widehat{\zeta}_{uk} \leftarrow \frac{\boldsymbol{\pi}_k \prod_{j=1}^{J} \boldsymbol{\mu}_{kj,Y_{uj}}}{\sum_{k'=1}^{K} \boldsymbol{\pi}_{k'} \prod_{j=1}^{J} \boldsymbol{\mu}_{k'j,Y_{uj}}} \tag{3.1}$$

Then given those type probabilities, update with

$$\text{for each } k, \text{ update:} \quad \widehat{\pi}_k \leftarrow \frac{1}{N}\sum_{u=1}^{U} n_u \widehat{\zeta}_{uk} \tag{3.2}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{kj\ell} \leftarrow \frac{\sum_{u=1}^{U} n_u \mathbf{1}\left(Y_{uj} = \ell\right) \widehat{\zeta}_{uk}}{\sum_{u=1}^{U} n_u \widehat{\zeta}_{uk}} \tag{3.3}$$

$$\tag{3.4}$$

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

$$\mathcal{L}_{\text{obs}} = \sum_{u=1}^{U} \log n_u + \sum_{u=1}^{U} \log \left\{ \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \boldsymbol{\mu}_{kj,Y_{ij}} \right\} \tag{3.5}$$

4 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*.

4.1 Categories of uncontestedness

In uncontested races, some options are not available to choose from. To show this, we introduce a new layer, following the notation in Yamamoto (2014):[1] 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 $\ell \in S_m$.

[1] Yamamoto, Teppei. 2014. A Multinomial Response Model for Varying Choice Sets, with Application to Partially Contested Multiparty Elections.

4.2 Shared Parameters across Varying Choice Sets

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. $\psi_{kj,(\ell=0)} = 0 \; \forall \; k, j$.

In the simplest case where clusters are completely homogenous, we paramterize our main variable of interest $\boldsymbol{\mu}$ as follows, while remembering that each individual is a member of type (Z_i) and each separate office is also of a missingness type M_{ij}.

$$\mu_{kj\ell} = \frac{1}{N\pi_k} \sum_{i=1}^{N} \mathbf{1}\left(Z_i = k, M_{ij} = m\right) \frac{\exp\left(\psi_{kj\ell}\right)}{\sum_{\ell' \in \boldsymbol{S}_m} \exp\left(\psi_{kj\ell'}\right)} \tag{4.1}$$

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.

4.3 EM Estimation of Varying Choice Sets

M-step We use this new representation of the parameter μ in the EM algorithm, replacing the weighted average M-step for μ with a weighted multinomial logit:

$$\text{for each } k, \text{ update:} \quad \widehat{\pi}_k \leftarrow \frac{1}{N} \sum_{i=1}^{N} \widehat{\zeta}_{ik} \tag{4.2}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{kj\ell} \leftarrow \frac{\exp\left(\widehat{\psi}_{kj\ell}\right)}{1 + \exp\left(\widehat{\psi}_{kj1}\right) + \exp\left(\widehat{\psi}_{kj2}\right)}, \tag{4.3}$$

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).
```

In other words, for each k, j, we estimate intercepts from regressing a vector of categorical votes for office $\mathbf{Y}_j$, using the estimates of $\boldsymbol{\zeta}_k$ as the weight `zeta_k`. 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 4.1.

4.4 MLE for varying choice multinomial logit

We can also solve the mlogit with varying choice sets by MLE.

We first introduce new notation $m_{ij\ell} \in \{0, 1\}$, for whether option ℓ is available for individual i in office j. Clearly, therefore, $m_{ij\ell}$ is a direct a mapping from M_{ij}.

$$m_{ij\ell} = \begin{array}{c} \\ \text{if } M_{ij} = 1 \\ \text{if } M_{ij} = 2 \\ \text{if } M_{ij} = 3 \end{array} \begin{array}{c} \begin{array}{ccc} \ell = 0 & \ell = 1 & \ell = 2 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

The log likelihood for the parameter of interest $\boldsymbol{\psi}_{jk} = \{\psi_{jk0}, \psi_{jk1}, ..., \psi_{jkL}\}$ is, for a fixed office j when considering the kth cluster:

$$\mathcal{L}(\boldsymbol{\psi}_{jk}) = \sum_{i=1}^{n} \sum_{\ell=0}^{L} m_{ij\ell} \zeta_{ik} \mathbf{1}\left(Y_{ij} = \ell\right) \log \left(\frac{\exp(\psi_{jk\ell})}{\sum_{\ell'=0}^{L} m_{ij\ell'} \exp(\psi_{jk\ell'})} \right) \tag{4.4}$$

Then we can solve the parameters numerically, i.e.,

$$\widehat{\boldsymbol{\psi}}_{jk}^{\text{MLE}} = \arg_{\boldsymbol{\psi}} \max \mathcal{L}(\boldsymbol{\psi}_{jk}) \tag{4.5}$$

by software like `optim`. Things will converge faster if we provide a gradient. We can take the partial derivative of the log likelihood, which returns a length-$(L+1)$ vector $\nabla\mathcal{L}(\boldsymbol{\psi}_{jk})$ where the $(\ell+1)$th element is derived in section C. All this significantly reduces time by reducing the overhead introduced in off-the-shelf packages like `mlogit`

4.5 Evaluating Convergence with missingness

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.

$$\mathcal{L}_{\text{obs}}^{\star} = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \prod_{\ell \in S_{M_{ij}}} \left\{ \left(\frac{\mu_{kj\ell}}{\sum_{\ell' \in S_{M_{ij}}} \mu_{kj\ell'}} \right)^{\mathbf{1}(Y_{ij}=\ell)} \right\} \right] \tag{4.6}$$

5 Incorporating Covariates

There are three types of auxilarly information one can include

1. Individual covariates (e.g. geography or demographics), that vary across i but not by j
2. Candidate covariates (e.g. like incumbency) that vary both across groups of i and appear in some j but not others
3. Office covariates (e.g. national vs. state offices) that vary across only j.

5.1 Individual-level covariates

Individual-level covariates affect who is in which cluster and the size of each cluster, but they will not change the dimensionality of μ. $\boldsymbol{\zeta}$ is already indexed by i, but $\boldsymbol{\pi}$ is not. Suppose we have a $N \times P_V$ numeric matrix $\mathbf{V}$. Then we parameterize:

$$\pi_{ik} = \frac{\exp \boldsymbol{V}_i^\top \boldsymbol{\gamma}_k}{\sum_{k'=1}^{K} \exp \boldsymbol{V}_i^\top \boldsymbol{\gamma}_k} \tag{5.1}$$

where $\boldsymbol{\gamma}_k$ is a length $P_V + 1$ vector of coefficients including one for the intercept. The general matrix γ is a $(P_V + 1) \times (K)$ matrix where the first column is the baseline and is all zeroes.

In the M-step for $\widehat{\pi}$, we now must adapt eq. 3.2 so that it not only upweights π_k for which ζ_k is high, but also systematically upweight voters whose covariates tend to correlate with high ζ_k. To do this, we regress `zeta` $\sim$ `V` i.e. zeta on Vs in `emlogit`[2]). Then to get predictions $\widehat{\pi}_i$ following the standard multinomial logit transformation as above (eq. 5.1). In the next iteration, we must use those $\widehat{\gamma}$ coefficients as starting values.

$$\text{for each } i, k, \text{ update:} \quad \widehat{\pi}_{ik} \leftarrow \frac{\exp \boldsymbol{V}_i^\top \widehat{\gamma}_k}{\sum_{k'=1}^{K} \exp \boldsymbol{V}_i^\top \widehat{\gamma}_k} \tag{5.2}$$

Even without any covariates, the dimensionality of $\boldsymbol{\pi}$ changes. Previously two voters with the same outcome $\boldsymbol{Y}$ were in the same cluster with probability 1. Now, with covariates, there is more variation. $\boldsymbol{\pi}$ is a $N \times K$ matrix instead of a $K \times 1$ vector. So we must rewrite:

$$\text{for each } i, k, \text{ update:} \quad \widehat{\pi}_{ik} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \widehat{\zeta}_{ik} \tag{5.3}$$

For the E-step, similarly, we use the matrix for each i, k, and index $\boldsymbol{\pi}$ by i as well:

$$\text{for each } i, k, \text{ update:} \quad \widehat{\zeta}_{ik} \leftarrow \frac{\pi_{ik} \prod_{j=1}^{J} \boldsymbol{\mu}_{kj, Y_{ij}}}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{J} \boldsymbol{\mu}_{k'j, Y_{ij}}} \tag{5.4}$$

Because the variation of π_i across observations $i \in 1, ..., N$ can get very large, when we report

[2] Yamauchi, Soichiro. emlogit: An ECM Algorithm for the Multinomial Logit Model. https://github.com/soichiroy/emlogit

the size of the cluster or general mixing proportion we can provide the sample average

$$\widetilde{\boldsymbol{\pi}} = \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\pi}_{i1}, ..., \frac{1}{N} \sum_{i=1}^{N} \widehat{\pi}_{iK}, \right)$$

Appendix

A Deriving EM with complete data

Recall that the expectation of the likelihood from equation 2.2 is

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

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

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

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

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

along with the constraint $\sum_{k=1}^{K}\pi_k = 1$. Notice that when we sum the FOC for $\boldsymbol{\pi}$ across k, the first condition becomes $\sum_{k=1}^{K}\pi_k = \frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=i}^{N}\zeta_{ik}$, 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

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

B Deriving EM with censored data

The modified log likelihood is

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

And the expected log likelihood, taking expectations over Z_i is

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

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

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

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

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

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 $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_{kj}$, respectively and modify eq. A.1 as:

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

We can deduce from the structure of the $\boldsymbol{\mu}$ array that the equivalent thing to the M-step in the complete data case is to run a standard multinomial logit with a IIA assumption. Unfortunately, neither the M-step nor a multinomial logit has a closed-form solution.

C The gradient for varying multinomial logit

Our goal is to take the partial derivative of the likelihood in eq. 4.4 with respect to ψ_{jk1} and ψ_{jk2}:

$$\mathcal{L}(\boldsymbol{\psi}_{jk}) = \sum_{i=1}^{n}\sum_{\ell=0}^{L} m_{ij\ell}\zeta_{ik}\mathbf{1}\left(Y_{ij}=\ell\right)\log\left(\frac{\exp(\psi_{jk\ell})}{\sum_{\ell'=0}^{L} m_{ij\ell'}\exp(\psi_{jk\ell'})}\right) \tag{C.1}$$

It is easier to consider the gradient at i, because the rest will be the sum of the individual gradients.

$$\mathcal{L}(\boldsymbol{\psi}_{jk})_i = \zeta_{ik}\sum_{\ell=0}^{L}\left\{m_{i\ell}\mathbf{1}\left(Y_{ij}=\ell\right)\log\left(\frac{\exp(\psi_{jk\ell})}{\sum_{\ell'=0}^{L}\exp_{\psi_{jk\ell'}}}\right)\right\}$$

$$\text{Let } c_i = \sum_{\ell'=0}^{L}\exp_{\psi_{jk\ell'}} \text{ to abbreviate}$$

$$\begin{aligned}
\nabla\mathcal{L}(\psi_{jk1})_i =& \zeta_{ik}\Bigg\{m_{i0}\mathbf{1}\left(Y_{ij}=1\right)\frac{\partial}{\partial\psi_{jk1}}\log\left(\frac{1}{c_i}\right)+ \\
& m_{i1}\mathbf{1}\left(Y_{ij}=1\right)\frac{\partial}{\partial\psi_{jk1}}\log\left(\frac{\exp\psi_{jk1}}{c_i}\right)+m_{i2}\mathbf{1}\left(Y_{ij}=1\right)\frac{\partial}{\partial\psi_{jk1}}\log\left(\frac{\exp\psi_{jk2}}{c_i}\right)\Bigg\} \\
=& \zeta_{ik}\Bigg\{m_{i0}\mathbf{1}\left(Y_{ij}=1\right)\left(-c_i\left(\frac{1}{c_i}\right)^2\exp\psi_{jk1}\right)+ \\
& m_{i1}\mathbf{1}\left(Y_{ij}=1\right)\left(1-\frac{\exp(\psi_{jk1})}{c_i}\right)+m_{i2}\mathbf{1}\left(Y_{ij}=1\right)\left(-\frac{\exp\psi_{jk1}}{c_i}\right)\Bigg\} \\
=& \zeta_{ik}\left\{m_{i1}\mathbf{1}\left(Y_{ij}=1\right)\left(1-\frac{\exp\psi_{jk1}}{c_i}\right)+\sum_{\ell'\neq 1}m_{i\ell'}\mathbf{1}\left(Y_{ij}=1\right)\left(\frac{\exp\psi_{jk\ell}}{c_i}\right)\right\}
\end{aligned}$$

So generally, the $\ell+1$th gradient is

$$\nabla\mathcal{L}(\psi_{jk\ell}) = \sum_{i=1}^{n}\left[\zeta_{ik}\left\{m_{i\ell}\mathbf{1}\left(Y_{ij}=\ell\right)\left(1-\frac{\exp\psi_{jk\ell}}{c_i}\right)+\sum_{\ell'\neq\ell}m_{i\ell'}\mathbf{1}\left(Y_{ij}=\ell\right)\left(\frac{\exp\psi_{jk\ell}}{c_i}\right)\right\}\right]$$