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. When $L = 2$, we can consider three outcomes: abstention $Y_{ij} = 0$, split $Y_{ij} = 0$, or straight $Y_{ij} = 2$.

Parameters 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 $\mu_{k,j} \in [0, 1]$ be the parameter that governs the outcome of each type. Therefore $\boldsymbol{\mu}$ 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.

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

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

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

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

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.[1]

2 Clustering as an Unobserved Variable Problem: EM

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

Because the model is simple enough, we can derive an algorithm to obtain the global solution for the parameters.[2] An EM implementation makes it possible to handle extensions, such as systematic missing data, multinomial outcomes, and covariates.

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:

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

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

[1] That dependency can be expressed as $\mathbb{E}(\mathbf{y}) = \sum_{k=1}^{K} \theta_k \boldsymbol{\mu}_k$ and $\text{Cov}(\mathbf{y}) = \sum_k \theta_k (\boldsymbol{\Sigma}_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top) - \mathbb{E}(\mathbf{y})\mathbb{E}(\mathbf{y})^\top$, where $\Sigma_k = \text{diag}(\mu_{jk}(1 - \mu_{jk}))$.

[2] Thanks to Soichiro Yamauchi for deriving this algorithm in the original iteration.

$$\mathcal{L}_{\text{comp}}(\boldsymbol{\mu}, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell=0}^{2}\mathbf{1}\{Y_{ij}=\ell, Z_i=k\}\Big\{\log \Pr(Y_{ij}=\ell|Z_i=k,\boldsymbol{\mu})\Big\} + \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}\{Z_i=k\}\log\Pr(Z_i=k|\boldsymbol{\theta}) \tag{2.1}$$

We first take expectations over the latent variable Z_i,

$$\mathbb{E}\left[\mathcal{L}_{\text{comp}}\right] = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{K}\sum_{\ell=0}^{2}\mathbf{1}\{Y_{ij}=\ell\}\mathbb{E}\left[\mathbf{1}\{Z_i=k\}\right]\Big\{\underbrace{\log\Pr(Y_{ij}=\ell|Z_i=k,\boldsymbol{\mu})}_{\equiv\log\boldsymbol{\mu}_{kj\ell}}\Big\} + \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{Z_i=k\}\right]\underbrace{\log\Pr(Z_i=k|\boldsymbol{\theta})}_{=\log\boldsymbol{\theta}_k} \tag{2.2}$$

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,

$$\widehat{\zeta}_{ij} \propto \boldsymbol{\theta}_k \underbrace{\prod_{j=1}^{D}\prod_{\ell=0}^{2}\left(\boldsymbol{\mu}_{kj,\ell}\right)^{\mathbf{1}(Y_{ij}=\ell)}}_{\boldsymbol{\mu}_{kj,Y_{ij}}} \tag{2.3}$$

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

$$\widehat{\theta}_k = \frac{1}{N}\sum_{i=1}^{N}\zeta_{ik} \tag{2.4}$$

$$\widehat{\mu}_{jk\ell} = \frac{\sum_{i=1}^{N}\mathbf{1}\left(Y_{ij}=\ell\right)\zeta_{ik}}{\sum_{i=1}^{N}\zeta_{ik}} \tag{2.5}$$

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, $\mu_{jk}^{(0)} = \frac{\sum_{i=1}^{N}\mathbf{1}(Y_{ij}=1)\mathbf{1}\left(Z_i'=k\right)}{\sum_{i=1}^{N}\mathbf{1}\left(Z_i'=k\right)}$.

Then we iterate as follows. For each voter i, compute the probability that they belong in

cluster k (E-step):

$$\zeta_{ik} \leftarrow \frac{\boldsymbol{\theta}_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj,Y_{ij}}}{\sum_{k'=1}^{K} \boldsymbol{\theta}_{k'} \prod_{j=1}^{D} \boldsymbol{\mu}_{k'j,Y_{ij}}} \tag{2.6}$$

Then given those type probabilities, update the parameters with the MLE (M-step)

$$\text{for each } k, \text{ update:} \quad \widehat{\theta}_k \leftarrow \frac{1}{N} \sum_{i=1}^{N} \widehat{\zeta}_{ik} \tag{2.7}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{jk\ell} \leftarrow \frac{\sum_{i=1}^{N} \mathbf{1}\left(Y_{ij} = \ell\right) \widehat{\zeta}_{ik}}{\sum_{i=1}^{N} \widehat{\zeta}_{ik}} \tag{2.8}$$

We repeat this 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 log-likelihood is

$$\mathcal{L}_{\text{obs}} = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \theta_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj,Y_{ij}} \right\} \tag{2.9}$$

Calculating eq. (2.9) 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:

$$\text{for each } u, k, \text{ update:} \quad \widehat{\zeta}_{uk} \leftarrow \frac{\boldsymbol{\theta}_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj,Y_{uj}}}{\sum_{k'=1}^{K} \boldsymbol{\theta}_{k'} \prod_{j=1}^{D} \boldsymbol{\mu}_{k'j,Y_{uj}}} \tag{2.10}$$

Then given those type probabilities, update with

$$\text{for each } k, \text{ update:} \quad \widehat{\theta}_k \leftarrow \frac{1}{N} \sum_{u=1}^{U} n_u \widehat{\zeta}_{uk} \tag{2.11}$$

$$\text{for each } k, j, \ell, \text{ update:} \quad \widehat{\mu}_{jk\ell} \leftarrow \frac{\sum_{u=1}^{U} n_u \mathbf{1}\left(Y_{uj} = \ell\right) \widehat{\zeta}_{uk}}{\sum_{u=1}^{U} n_u \widehat{\zeta}_{uk}} \tag{2.12}$$

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

$$\mathcal{L}_{\text{obs}} = \sum_{u=1}^{U} \log n_u + \sum_{u=1}^{U} \log \left\{ \sum_{k=1}^{K} \theta_k \prod_{j=1}^{D} \boldsymbol{\mu}_{kj,Y_{ij}} \right\} \tag{2.13}$$

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 run for one of the candidates. These choices are qualitatively different from contested races in their data generating process.

Consider the trichotomous outcome $L = 2$ where $Y_{ij} = 0$ indicates abstention, $Y_{ij} = 1$ indicates ticket splitting (appearing like they are voting for their less preferred party as expressed in the top of the ticket) and $Y_{ij} = 2$ indicates co-partisan voting.

Voters for a given office now fall into one of these three categories, denoted by $M_{ij} \in \{1, 2, 3\}$. Let $M_{ij} = 3$ denote the contested case, so the voter has three options. Let $M_{ij} = 2$ denote the case when only the preferred party candidate is in the race, so the voter has options $Y_{ij} \in \{0, 2\}$. Finally let $M_{ij} = 1$ denote the case when only the opposed party candidate is in the race, so the voter only has the option to abstain or "reluctantly" (as it might seem) vote for the less favored option by splitting: $Y_i \in \{0, 1\}$.

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. $\psi_{jk,(\ell=0)} = 0 \; \forall \; j, k$. Then the log-likelihood can be expressed by

$$\mu_{jk\ell} = \mathbf{1}\left(M_{ij}=1\right)\frac{\exp\psi_{kj\ell}}{\sum_{\ell'\in\{0,1\}}\exp\psi_{kj\ell'}} + \mathbf{1}\left(M_{ij}=2\right)\frac{\exp\psi_{kj\ell}}{\sum_{\ell'\in\{0,2\}}\exp\psi_{kj\ell'}} + \mathbf{1}\left(M_{ij}=3\right)\frac{\exp\psi_{kj\ell}}{\sum_{\ell'\in\{0,1,2\}}\exp\psi_{kj\ell'}} \tag{3.1}$$

Remark: Because $\exp(\psi_{jk\ell}) = 1$ 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 $\boldsymbol{\psi}_{jk}$ 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. We can therefore think of equation 3.1 as representing that the relevant parameter μ for each individual takes one of the types depending on the (known) class of menu options, without introducing more parameters into the model.