A Bayesian model for estimating election results in Mexico

Gian Carlo Di-Luvi
Department of Statistics
University of British Columbia
gian.diluvi@stat.ubc.ca

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1 Introduction

In Mexico, presidential and gubernatorial elections take place every six years. Mexico has a multi-party electoral system where the candidate that receives the largest amount of votes wins the election. Background on Mexican democracy.

Because election results take a week to be certified, in recent years there has been a widespread effort by government authorities to determine the winner of each election the day it takes place. For this purpose, a sample of polling stations is usually selected before the election and statistical estimates based on those polling stations are computed as the results become available on election day—a process called a *quick count*.

Multiple Bayesian models have been used in Mexico's quick counts. Mendoza and Nieto-Barajas [5] proposed a simple Bayesian model with a tractable posterior to estimate the proportion of votes received by each candidate; the model was successfully used to predict the winners of the 2006 and 2012 elections. Di-Luvi et al. [3] modified that same model and used it to predict the results of the 2018 election. Anzarut et al. [2] proposed a Bayesian multilevel regression model and used it in the 2018 election as well.

Data from the polling stations arrive as field personnel finish counting the votes. On election day, every 15 minutes new data are supplied and models are retrained. Thus, models should be rich enough to do accurate inference with small sample sizes while allowing for fast online inference. The model in [2] accounts for data missing not at random but is computationally expensive—for the presidential election, it took well over 15 minutes to run even when parallelising over 48 CPU cores. On the other hand, although the model proposed by [5] is easy to train, they rely on a tractable posterior distribution that limits its flexibility. Indeed, Di-Luvi et al. [3] modified the prior distribution but did not account for the corresponding change in the posterior.

In this work, we *(i)* propose a new model that accounts for the correlation structure between candidates—something not done before—and compare the results with other models, specifically in the setting with small sample sizes; and *(ii)* design and implement an MCMC procedure to sample from the correct posterior distribution in Di-Luvi et al. [3]'s model, and compare it with the "wrong" posterior originally used. Paper outline.

2 Background

Quick counts are organized by the Mexican Electoral Institute but carried out by a committee of statisticians. Each member of the committe proposes and implements a different model. On election day, each model is trained every 5 minutes when new information becomes available, and the results from all models are combined into a single interval per candidate. When the committee decides that the sample is sufficiently large, the final combined intervals are reported to the government authorities.

To determine the sample of polling stations, a stratified sampling design is developed by the committee. The observational unit corresponds to the polling station and the strata—which vary from election to election—are defined based on geographical variables. Suppose that the population of polling stations is divided into N strata, and that stratum i has K_i polling stations. Although the number of voters is unknown before the election, the number of *registered voters* per polling station is known in advance. Let n_i^k be the number of registered voters in polling station k of stratum i.

The observables in the model correspond to the number of votes, which for each candidate j, stratum i, and polling station k we denote by X_{ij}^k. The latent variables correspond to the proportion of votes, which for candidate j and in stratum i are denoted by θ_{ij}. In these definitions, we use the term candidate loosely to refer to actual candidates as well as the set of null votes and the set of votes that were not cast. Note that we assume that all polling stations within a stratum have the same proportion of votes per candidate, i.e.

$$\theta_{ij} = \frac{\sum_{k=1}^{K_i} X_{ij}^k}{\sum_{k=1}^{K_i} n_i^k}$$

does not depend on k.

Although the model operates at a stratum level, national estimates are required. The proportion of votes in favor of candidate j is defined as

$$\theta_j = \frac{1}{n} \sum_{i=1}^{N} n_i \theta_{ij}, \tag{1}$$

where $n = \sum_{i=1}^N n_i$ is the total number of nation-wide registered voters. Observe, however, that θ_j refers to the proportion of votes out of registered voters, rather than out of cast votes. For this purpose we define, for each candidate j,

$$\lambda_j = \frac{\theta_j}{\sum_{\tilde{j}} \theta_{\tilde{j}}}, \tag{2}$$

where the sum is over all proper candidates (i.e. excluding θ_j corresponding to uncast ballots).

Mendoza and Nieto-Barajas [5] propose modeling the number of votes in polling station k of stratum i and in favor of candidate j via

$$X_{ij}^k \sim \mathcal{N}\left(n_i^k \theta_{ij}, \frac{\tau_{ij}}{n_i^k}\right), \tag{3}$$

where τ_{ij} is a precision parameter.[1] Observe that the number of votes are assumed to be independent between candidates. This is inaccurate because, at each polling station,

[1] Although $\frac{\tau_{ij}}{n_i^k}$ corresponds to the variance of the distribution.

the total votes should add up to the known number of registered voters, n_i^k. (Hence the importance of including uncast votes within j.) Mendoza and Nieto-Barajas [5] acknowledge this issue but argue that including a correlation structure, although correct in theory, produces small improvements in accuracy.

Because the quick counts are organized by governmental authorities, prior distributions should assign the same likelihood of winning to each candidate. Mendoza and Nieto-Barajas [5] use uniform distributions for θ_{ij} and an independent improper prior for τ_{ij}. With these, the posterior distribution after observing the information from c_i polling stations in stratum i has a closed form, namely

$$\begin{aligned} p(\theta_{ij}, \tau_{ij} \mid X_{ij}^1, ..., X_{ij}^{c_i}) \propto \mathcal{N}\left(\theta_{ij} \,\middle|\, \frac{\sum_{k=1}^{c_i} x_{ij}^k}{\sum_{k=1}^{c_i} n_i^k},\ \tau_{ij} \sum_{k=1}^{c_i} n_i^k \right) \mathbb{1}(0 < \theta_{ij} < 1) \\ \times \mathsf{Gamma}\left(\tau_{ij} \,\middle|\, \frac{c_i - 1}{2},\ \frac{1}{2}\left\{ \sum_{k=1}^{c_i} \frac{(x_{ij}^k)^2}{n_i^k} - \frac{\left(\sum_{k=1}^{c_i} x_{ij}^k\right)^2}{\sum_{k=1}^{c_i} n_i^k} \right\}\right). \end{aligned} \tag{4}$$

Sampling from Eqn. 4 can be done expeditiously for each stratum—there is no need to run an MCMC scheme. Samples from all strata are then combined via Eqns. 1 and 2 to obtain the national estimates for each candidate.

Because data arrive as field officers finish counting the votes, oftentimes entire strata in the sample have no information available. In those strata, Mendoza and Nieto-Barajas [5] sample from the prior distribution—which is the probability distribution that describes the uncertainty in the latent variables. However, because all candidates are assumed independent and $(\theta_{i1}, ..., \theta_{iJ})^\top$ is not constrained to add up to 1, it is possible that the estimated voter turnout exceeds 100% in some strata. This in turn affects the national voter turnout estimte, which overestimates the true voter turnout.

To remedy this, Di-Luvi et al. [3] proposed new prior distributions for θ_{ij}—namely, Beta distributions with a mean equal to the historic voter turnout divided by the number of candidates. Although this seemingly takes care of the issue in practice, it does not technically resolve it because it fails to acknowledge that the proportion of votes, as a vector, lives in the probabilistic simplex. Furthermore, Di-Luvi et al. [3] only use the modified prior when no information is available from a given stratum. If information is available, samples are obtained from the tractable posterior, 4, which does not correspond to the correct posterior associated with the modified prior distributions.

3 Modeling proportions in the simplex

To design a Bayesian model correctly addresses the estimation of voter turnout, we first write the likelihood of votes in each stratum in Eqn. 3 as a multivariate Normal distribution. Specifically, letting $X_i^k = (X_{i1}^k, ..., X_{iJ}^k)^\top$ and $\theta_i = (\theta_{i1}, ..., \theta_{iJ})^\top$, Eqn. 3 is equivalent to

$$X_i^k \sim \mathcal{N}_J\left(n_i^k \theta_i, \Sigma_i^k\right),$$

where $\Sigma_i^k = \mathrm{diag}(\tau_{i1}, ..., \tau_{iJ})/n_i^k$. Observe that the number of votes is, clearly, discrete. A multinomial distribution would be technically better suited to model X_i^k, and it would also ensure that the votes add up to the number of registered voters. However, it would also limit the flexibility of the model because, while possible to control the mean, the correlation structure of a multinomial distribution is determined by the means of each marginal Binomial distribution.

Given that the number of registered voters is usually high,[2] using a Normal approximation to a Binomial distribution is well justified. Furthermore, by adding variance parameters

[2] In 2018, the average n_i^k was 550

independent of the mean, Di-Luvi et al. [3], Mendoza and Nieto-Barajas [5] are able to model more complex variability structures. There are two issues with this approach. The first is that it assumes that the vector θ_i lives in $[0,1]^J$, when in reality θ_i lives in the probabilistic simplex $\Delta^{J-1} = \{\theta \in \mathbb{R}^J \mid \theta_j > 0, \sum_{j=1}^J \theta_j = 1\}$. This is the cause of the voter turnout overestimation. The second issue is that it assumes a trivial covariance structure.

We overcome these issues by *(i)* forcing $\theta_i \in \Delta^{J-1}$ and *(ii)* letting Σ_i^k be an arbitrary covariance matrix. The only additional requirements to fit this more complex model are to carefully design a prior distribution for θ_i in Δ^{J-1} and a prior distribution for Σ_i^k in the space of covariance matrices.

3.1 Prior distributions

To design a prior distribution for θ_i, observe that the dimension of Δ^{J-1} is $J-1$; indeed, Δ^{J-1} and $\mathbb{R}^{J-1}$ are isomorphic. Consider the map $\mathcal{A} : \Delta^{J-1} \to \mathbb{R}^{J-1}$ given by

$$\mathcal{A}(\theta_1, ..., \theta_J) = \left(\log \frac{\theta_1}{\theta_J}, ..., \log \frac{\theta_{J-1}}{\theta_J}\right)^\top .$$

$\mathcal{A}$ is known as the *additive log-ratio* [1], and its inverse is given by

$$\mathcal{A}^{-1}(x_1, ..., x_{J-1}) = \left(\frac{\exp(x_1)}{1 + \sum_{i=1}^{J-1} \exp(x_i)}, ..., \frac{\exp(x_{J-1})}{1 + \sum_{i=1}^{J-1} \exp(x_i)}, \frac{1}{1 + \sum_{i=1}^{J-1} \exp(x_i)}\right)^\top .$$

The additive log-ratio transform is closely related to the center log-ratio transform, whose inverse is the popular softmax function. However, we choose to work with $\mathcal{A}$ because it is easier to interpret, and thus to define a prior on the transformed values.

In the transformed space $\mathbb{R}^{J-1}$, the coordinate jth corresponds, up to additive constants, to the log of the proportion of votes in favor of candidate j, for $j = 1, ..., J-1$. We assume that the Jth candidate corresponds to the set of uncast ballots, which is of less interest. For the rest of the candidates, we define a Normal prior and choose the mean so that, when exponentiated, it coincides with that of Di-Luvi et al. [3]: the historic voter turnout divided by the number of candidates. This is because the inverse map exponentiates each coordinate and then normalizes the result. By using a Normal prior, we can use the properties of the logNormal distribution to control the prior mean in the transformed space. Note that we assume vote proportions to be independent in transformed space. A dependence structure is introduced when mapping back into the simplex.

Perhaps the best known distribution in the space of covariance matrices is the Wishart distribution, which is the multivariate generalization of the Gamma distribution. The Wishart distribution was popularized because it is the conjugate prior for the covariance matrix of Normally-distributed populations. However, the purpose of the prior being to quantify the prior knowledge of the parameters, we consider it best to choose a distribution that reflects the vague knowledge we have of the correlation structure between candidates.

For this purpose, we decompose $\Sigma_i^k = \operatorname{diag}(\sigma_1, ..., \sigma_J) P \operatorname{diag}(\sigma_1, ..., \sigma_J)/n_i^k$, where P is the correlation matrix and σ_j is the variance of each candidate. Lewandowski et al. [4] propose a method to generate correlation matrices uniformly over the space of all correlation matrices. We use their method, through its Stan implementation, and couple it with uninformative priors for the variances. This results in an uninformative prior for the covariance matrix.

4 Experiments

We compare four different models: the one used by Di-Luvi et al. [3], which contains modified priors but samples from the posterior in Mendoza and Nieto-Barajas [5]; that same model but now sampling from the correct posterior via MCMC; the model proposed in Section 3 with the priors discussed therein; and the same model but assuming a trivial correlation structure, i.e., $P = I_J$. In this last model, the proportions of votes are forced to live in the probabilistic simplex but the correlation structure assumes a simple form, which results in faster sampling.

5 Conclusion

References

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

[2] M. Anzarut, L. F. González, and M. T. Ortiz. A heavy-tailed multilevel mixture model for the quick count in the mexican elections of 2018. In *National Statistics Forum (FNE) and Latin-American Congress of Statistical Societies (CLATSE)*, pages 1–13. Springer, 2018.

[3] G. C. Di-Luvi, G. Orantes-Jordan, and M. Mendoza. Statistics in the 2018 mexican general election quick counts. *Laberintos & Infinitoss*, 48:29–37, 2018.

[4] D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[5] M. Mendoza and L. E. Nieto-Barajas. Quick counts in the mexican presidential elections: A bayesian approach. *Electoral Studies*, 43:124–132, 2016.