OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Estimating Vote Choice in U.S. Elections with Approximate Poisson-Binomial Logistic Regression

Nic Fishman NJWFISH@GMAIL.COM
Harvard University

Evan Rosenman ETRROSENMAN@GMAIL.COM
Claremont McKenna College

Abstract

We develop an approximate method for maximum likelihood estimation in Poisson-Binomial Logistic regression. The resulting approximate log-likelihood is generally non-convex but easy to optimize in practice. We investigate the geometry of the likelihood and propose simple but effective optimization procedures. We use these methods to fit logistic regressions in all statewide U.S. elections between 2016 and 2020, a total of 544 offices and over 1.75 billion votes.

1. Introduction

Understanding voter behavior is crucial to sustaining a healthy democracy. Elections are the primary mechanism through which citizens express their preferences and hold leaders accountable, so analyzing electoral behavior is necessary to uncover the electorate's priorities, assess campaign effectiveness, and ensure fair representation. There are basically two ways to understand who is voting for whom: those based primarily on polling and those based directly on electoral data. Survey-based approaches can produce excellent results [13], but as the quality of survey data has come under increasing scrutiny [3], approaches that circumvent survey methods become more attractive.

Ecological methods to study elections combine precinct-level vote results with voter covariates. Goodman's Regression [7] pioneered this approach, followed by a number of improvements leveraging aggregate covariate information to identify ecological correlations [4, 5, 10, 11, 18, 28, 29].

In the modern setting, we observe a set of covariates $\boldsymbol{x}_{ij}$ for each voter j within each precinct i, derived from the "voterfile," the roster of all registered voters within a given geography. Vote tallies are reported for each precinct i. One of the first models to use individual-level covariates, as opposed to aggregates, was Jackson et al. [8, 9] which modeled voters' individual vote choice as a Binomial. More recent approaches model individual vote choice explicitly via a Poisson-Binomial distribution, the distribution of the sum of independent but not identically distributed Bernoulli random variables [12, 16, 19, 21]. We follow this approach via an approximate likelihood which makes our methods more scalable and makes studying the geometry of the log-likelihood much more straightforward.

Using our approximate log-likelihood we fit ecological logistic regressions to all statewide elections in the United States between 2016 and 2020 using electoral outcomes documented by the Voting and Election Science Team [24–26] and the TargetSmart Voterfile. This totals 544 offices, over 1.75 billion votes, and hundreds of millions of unique voters. In addition to logistic regression predicting two-way vote choice, we extend our results to fit multinomial logistic regressions allowing us to additionally model votes for third-party candidates, as well as ballot roll-off.

In this paper we focus on the geometry of our approximation, developing primitives which allow us to understand when and why it is easy to optimize. We go into more detail on our application in App. A.1. We provide a more detailed review of existing approaches in App. A.3, along with an empirical evaluation showing our method outperforms existing methods. In App. B.2 we present several statistical results which give conditions for consistency and asymptotic normality. The consistency and asymptotic normality results assume identifiability and that our MLE is the global maximizer of the sample, and our goal in studying the geometry is determining when we can expect identifiability to hold and when we can expect the optimization problem to be simple.

2. Poisson-Binomial Logistic Regression

Our model is a logistic regression with missing outcomes. We assume there are n precincts each with m_i voters. Each voter has a "true" individual-level probability of voting for the Democrat $p^*_{ij} = \sigma(x_{ij}^\top \beta^*)$ determined by β^* pushed through the sigmoid function $\sigma(z) = 1/(1+\exp(-z))$, and an unobserved individual level vote sampled as $V_{ij} \sim \text{Bern}(p^*_{ij})$. The vote counts Y_i are just the sum over the votes $Y_i = \sum_j V_{ij}$. Under this model, the vote counts follow a Poisson Binomial distribution: $Y_i \sim \text{PoiBin}(\{p^*_{ij}\}_{j=1}^{m_i})$. We illustrate this data structure in Fig. 2(a) where we depict the voterfile containing voter covariates grouped into precincts with the precinct-level vote counts.

Assumption 1 ***(Realizability)*** *The votes V_{ij} and the vote counts Y_i are generated by the data generating process described above with a true, unknown, β^*.*

Our goal is to recover the parameter β^* via an estimate $\hat{\beta}$, fit using maximum likelihood estimation. We can write the likelihood we would like to optimize in terms of β:

$$\ell_{\text{PoiBin}}(\beta) = \left(\frac{1}{n \cdot \bar{m}}\right) \sum_{i=1}^{n} \log \left(\sum_{A \in \mathbb{P}_{Y_i}([m_i])} \left(\prod_{j \in A} \sigma(x_{ji}^\top \beta) \right) \left(\prod_{j \in A^c} (1 - \sigma(x_{ji}^\top \beta)) \right) \right) \tag{1}$$

Where $\mathbb{P}_{Y_i}([m_i])$ is the set of all partitions of $[m_i] = \{1, \ldots, m_i\}$ into sets of size Y_i and $m_i - Y_i$, and $\bar{m} = \frac{1}{n}\sum_{i=1}^n m_i$. This is essentially the logistic likelihood but we do not know who voted for the Democrat so we have to average over all partitions of voters into Democrats and Republicans.

3. Approximation via a Local Limit Theorem

The $\ell_{\text{PoiBin}}(\beta)$ likelihood is computationally intractable since we need to sum over all possible partitions of voters for every precinct. Since precincts are usually large we can use a local limit theorem to approximate the object by a Normal likelihood, giving us a tractable objective.

Theorem 1 ***(Poisson-Binomial Local Limit Theorem, [17])***
We define the mean $\mu_i(\beta) = \sum_{j=1}^{m_i} p_{ij}$ and variance $\varsigma_i(\beta)^2 = \sum_{j=1}^{m_i} p_{ij}(1-p_{ij})$. For any $\beta \in \mathbb{R}^d$:

$$\sup_{k \in [m_i]} \left| \sum_{A \in \mathbb{P}_k([m_i])} \left(\prod_{j \in A} p_{ij} \right) \left(\prod_{j \in A^c} (1 - p_{ij}) \right) - \frac{1}{\sqrt{2\pi\varsigma_i(\beta)^2}} \exp\left(-\frac{1}{2} \left(\frac{k - \mu_i(\beta)}{\varsigma_i(\beta)} \right)^2 \right) \right| = O\left(\frac{1}{\varsigma^2}\right)$$

Using the local limit theorem we can let $p_{ij} = \sigma(\boldsymbol{x}_{ij}^\top \beta)$ to define the approximate log-likelihood:

$$\ell_{\text{PoiBin}}(\beta) \approx \ell(\beta) = -\left(\frac{1}{n \cdot \bar{m}}\right) \sum_{i=1}^{n} \frac{1}{2} \left(\log(2\pi) + \log\left(\varsigma_i(\beta)^2\right) + \frac{(Y_i - \mu_i(\beta))^2}{\varsigma_i(\beta)^2} \right) \tag{2}$$

where constants have been dropped. From the local limit theorem a necessary condition for our approximation to be close to the true log-likelihood is clear:

Assumption 2 *The variance diverges for some precincts $\mathcal{S} \subseteq [n]$, so for all $i \in \mathcal{S}$:* $\varsigma_i^2 \xrightarrow{m_i \to \infty} \infty$.

Essentially we need to assume that (1) precincts are large, and (2) as the number of voters in a precinct grows the variance also grows. If $p_{ij} \in (\epsilon_i, 1-\epsilon_i)$ for some fixed ϵ_i then we satisfy the condition. In practice, our covariates are bounded so for any fixed β^* we will have $p_{ij} \in (\epsilon_i, 1-\epsilon_i)$.

We illustrate the relationship between this limit theorem and our data-generating process in Fig. 2(b, c). This is a common, easy-to-compute, approximation to the Poisson-Binomial likelihood [22], which also lends itself to a much more straightforward analysis than the full likelihood. The central remaining issue is that this approximate log-likelihood remains non-convex, so it is unclear how easy we should expect it to be to optimize in practice.

The approximate log-likelihood is simpler than the sum of its parts. Almost every individual precinct has a complex multimodal log-likelihood, but the sum of log-likelihoods across precincts tends to have simple geometries that are unimodal and often convex. Our analysis of the log-likelihood landscape will focus on the population problem, where we analyze the expectation of the log-likelihood under a "fixed design" where we re-sample Y_i conditional on the precincts:

$$\mathbb{E}_{Y_i \sim \text{PoiBin}\left(\{\sigma(\mathbf{x}_{ij}^\top \beta^*)\}_{j=1}^{m_i}\right)}\left[\frac{(Y_i - \mu_i(\beta))^2}{\varsigma_i(\beta)^2}\right] = \frac{\varsigma_i(\beta^*)^2}{\varsigma_i(\beta)^2} + \frac{(\mu_i(\beta^*) - \mu_i(\beta))^2}{\varsigma_i(\beta)^2}$$

This lets us compute the population log-likelihood and gradient (as in Appendix B.1):

$$\begin{aligned}
\mathbb{E}[\ell_i(\beta)] &= \frac{1}{2}\left(\log\left(\varsigma_i(\beta)^2\right) + \frac{\varsigma_i(\beta^*)^2}{\varsigma_i(\beta)^2} + \frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)^2}{\varsigma_i(\beta)^2}\right) \\
\mathbb{E}[\nabla_\beta \ell_i(\beta)] &= \frac{1}{2}\left(\frac{\varsigma_i(\beta^*)^2 - \varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}(2p_{ij}-1)p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right) \\
&\quad + \frac{1}{2}\left(\frac{(\mu_i(\beta^*) - \mu_i(\beta))^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}(2p_{ij}-1)p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right) - \left(\frac{\mu_i(\beta^*) - \mu_i(\beta)}{\varsigma_i(\beta)^2}\right)\left(\sum_{j=1}^{m_i} p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right)
\end{aligned}$$

We can see here that the population likelihood for a single precinct decomposes into a mean component and a variance component, and this decomposition carries through into the gradient. The coefficient on the first term of the gradient ensures that when $\varsigma_i(\beta)^2 = \varsigma_i(\beta^*)^2$ then that term is zero. The coefficients on the second and third terms guarantee a similar condition when $\mu_i(\beta) = \mu_i(\beta^*)$.

Many β will achieve either the correct mean or variance in a given precinct, as we can see in Fig 1(a). We will denote the sets $\mathcal{M}_i = \{\beta \mid \mu_i(\beta) = \mu_i(\beta^*)\}$ and $\Sigma_i = \{\beta \mid \varsigma_i(\beta)^2 = \varsigma_i(\beta^*)^2\}$ and their intersection $\mathcal{B}_i = \mathcal{M}_i \cap \Sigma_i$. Now since the approximate likelihood is a Normal likelihood any point achieving the correct mean and variance will be a global minimizer. Put another way, in any particular precinct every point in $\mathcal{B}_i$ is a global optimum. By realizability, we know $\beta^* \in \mathcal{B}_i$ but if there are any other points in $\mathcal{B}_i$ then we do not have identifiability in precinct i with respect to the approximate likelihood. Across precincts, we have identifiability if and only if $\cap_{i=1}^n \mathcal{B}_i = \beta^*$.

But here we have a problem: $\mathcal{B}_i$ almost always contains more than just β^*. We illustrate this in Fig. 1. The first precinct is the best-case scenario: the mean and variance surfaces only intersect at the true optima. But it turns out this is only because the covariates are reflectively symmetric about β^*. In the center precinct, we see that when we shift the covariates to the right, the mean surface

tilts to the left, and the variance tilts the surface to the right, leading to two intersections which are both global optima. Shifting the covariates to the left has the opposite effect.

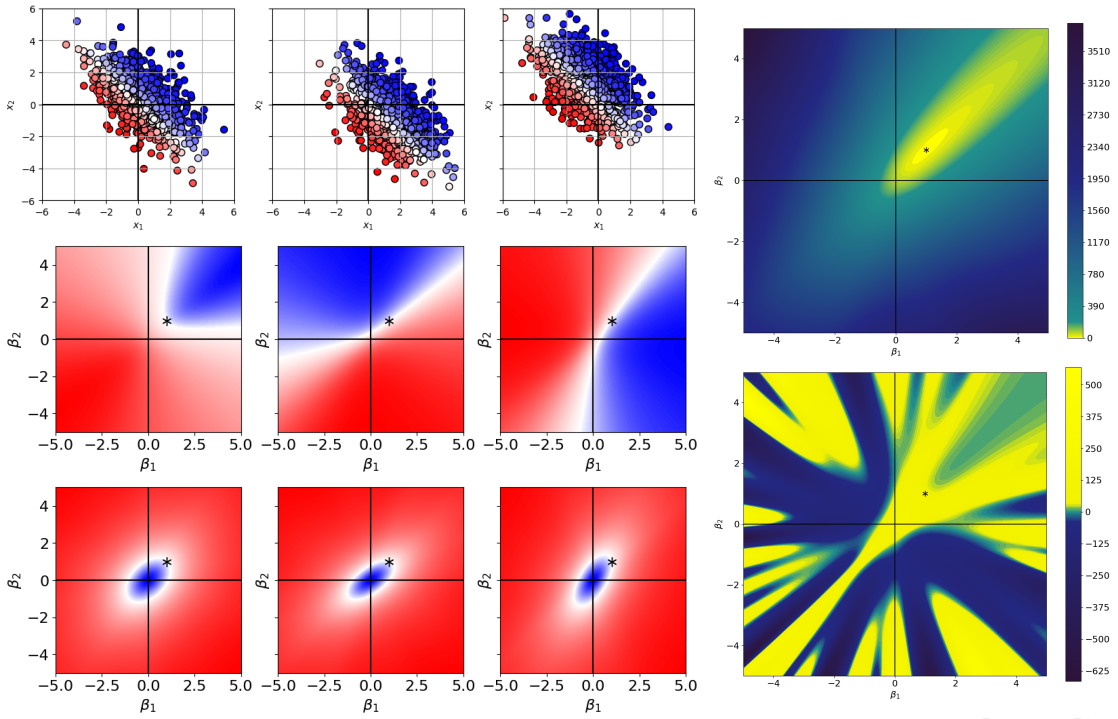
The key intuition to generalize this to higher dimensions is that the mean and variance sets $\mathcal{M}_i$ and Σ_i will be $d-1$ dimensional surfaces that must intersect at β^*. The variance surface Σ_i will describe a compact manifold, whereas $\mathcal{M}_i$ will be non-compact. In two dimensions these correspond to the one-dimensional curves in Fig. 1. Since these surfaces intersect by realizability, unless they touch at a tangent point, they must intersect at multiple points. In two dimensions this means we will have (at least) two global optima, but in higher dimensions, this will result in a $d-2$ dimensional surface: in three dimensions we will have a "ring" of global optima, and in four dimensions we will have a "sphere" of solutions, etc. The only way to ensure these surfaces touch at a tangent point is to require the mean and variance surfaces to be rotationally symmetric about β^*, which in turn means the covariates have to be rotationally symmetric about β^*. In two dimensions this rotational symmetry reduces to reflective symmetry. In $d \geq 3$, barring rotational symmetry, $\mathcal{B}_i$ will contain an infinite surface of points achieving the global optimum. We can formalize this:

Lemma 2 *Assuming the model is well-specified with β^*, a necessary condition for a precinct to have a single global optimum is that the covariates, $\{\boldsymbol{x}_{ij}\}_{j=1}^{m_i}$, are rotationally symmetric about β^*.*

Rotational symmetry is an extremely strong condition, which will virtually never be satisfied in any single precinct, much less in every single precinct in a real dataset. These results clarify the nature of the non-convexity in the approximate log-likelihood. Not only is the problem not *convex*, but almost every individual precinct will have multiple *global* optima to say nothing of local optima. And yet, as we demonstrate in Fig. 1, when we have even a few precincts with sufficiently distinct covariate distributions all of these complexities disappear and we are left with extremely simple log-likelihood landscapes. Certainly, there are pathological cases where $|\cap_{i=1}^n \mathcal{B}_i| > 1$ which guarantees a complex landscape. There are even more cases where $\cap_{i=1}^n \mathcal{B}_i = \beta^*$ but where we have local optima, especially when the covariate distributions are similar. But, setting aside pathological cases, if we have enough precincts with sufficient variation in covariate distributions we can expect the approximate log-likelihood landscape to have a unique global minima and no local optima. Here it is worth noting that this is not a mere artefact of our approximation: for large m_i any minimum in the approximate log-likelihood will be a minimum in the exact log-likelihood. although this does not pose a threat to identifiability it does make computing the MLE intractable.

Beyond the local convexity we expect from Theorem 6 and the unimodality we might hope for based on Fig. 1(b, top) we find that in most of the simulated settings and our applications the convex neighborhood containing the sample optimum actually stretches all the way to $\beta = 0$. In Fig. 1(b) we plot the minimum eigenvalues of the Hessian as functions of β, demonstrating convexity holds for the population log-likelihood when we have even a few sufficiently differentiated precincts.

When the convex neighborhood covers both zero and $\hat{\beta}$ optimization of the approximate log-likelihood is straightforward with Newton's method. Gradient descent and other non-linear optimizers like (L-)BFGS, on the other hand, struggle with the very sharp curvature around the optimum, usually failing to converge. The Hessian is not always positive definite even initializing at zero, so we developed a modified version of Newton's method which runs a line search over $\epsilon > 0$ projecting the Hessian onto the set of matrices with minimum eigenvalue ϵ using the Armijo condition as an acceptance criterion [1, 2]. We fully describe this algorithm in Appendix B.3, Algorithm 1.



(a) Precinct-level covariates, $\mu(\beta)$, and $\varsigma(\beta)^2$.

(b) $\mathbb{E}\left[\ell(\beta)\right]$ and $\lambda_{\min}\left(\mathbb{E}\left[\nabla^2_\beta \ell(\beta)\right]\right)$.

Figure 1: Here we present a simple three-precinct setting. (a) The first row depicts three different sets of covariates in two dimensions, colored by $p^*_{ij} = \sigma(\boldsymbol{x}_{ij}^\top \beta^*)$ where $\beta^* = (1, 1)$. The second and third rows plot the corresponding mean and variance landscapes for each set of covariates as functions of β, $\mu_i(\beta)$ and $\varsigma_i(\beta)^2$. Red indicates the mean/variance is too low while blue indicates too high, and the white lines indicate the $\mathcal{M}_i$ and Σ_i surfaces. (b) Depicts the sum of the expected log-likelihood landscape $\mathbb{E}\left[\ell(\beta)\right]$ (top) over all three precincts and the minimum eigenvalue of the expected Hessian $\lambda_{\min}(\mathbb{E}[\nabla^2_\beta \ell(\beta)])$ (bottom) as functions of β. The black lines denote the axes and the $*$ denotes β^*. Even though the second and third precincts each have two global optima the log-likelihood landscape is unimodal and in fact convex in a neighborhood containing zero and β^*.

These procedures are fast and robust. The optima computed in simulations recover $\hat{\beta}$ close to the β^* and the estimates behave asymptotically normally, as we would expect of the MLE based on Theorem 7. We present the details of this simulation study in Appendix A.2.

4. Conclusion

We have presented an efficient algorithm to approximate the Poisson-Binomial logistic regression. We have also documented the complexity of this non-convex optimization problem, a complexity that is inherent to the geometry of the problem, not some mere artifact of our approximation. However, we found that in practice this problem is easy to optimize. We have developed some intuition for why this is the case. In our appendices, we give more details on our application, along with detailed simulations and a comparison to existing methods. The appendices also detail technical results building up to theorems for consistency and asymptotic Normality of the (approximate) MLE.

References

[1] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700, 2021.

[4] Seth Flaxman, Dougal Sutherland, Yu-Xiang Wang, and Yee Whye Teh. Understanding the 2016 US presidential election using ecological inference and distribution regression with census microdata. *arXiv preprint arXiv:1611.03787*, 2016.

[5] Seth R. Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported Obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298. ACM, 2015.

[6] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[7] Leo A Goodman. Ecological regressions and behavior of individuals. *American sociological review*, 1953.

[8] C. H. Jackson, N. G. Best, and S. Richardson. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society, Series A: Statistics In Society*, 171(1):159–178, 2008.

[9] Christopher Jackson, Nicky Best, and Sylvia Richardson. Improving ecological inference using individual-level data. *Statistics in medicine*, 25(12):2136–2159, 2006.

[10] Gary King. A solution to the ecological inference problem, 1997.

[11] Gary King, Ori Rosen, and Martin A Tanner. Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research*, 28(1):61–90, 1999.

[12] H. Kück and N. de Freitas. Learning about individuals from group statistics. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, page 332, 2005.

[13] Shiro Kuriwaki, Stephen Ansolabehere, Angelo Dagonel, and Soichiro Yamauchi. The geography of racially polarized voting: Calibrating surveys at the district level. *OSF Preprints. December*, 4, 2021.

[14] Shiro Kuriwaki, Mason Reece, Samuel Baltz, Aleksandra Conevska, Joseph R. Loffredo, Taran Samarth, Can E. Mutlu, Kevin E. Acevedo Jetter, Zachary Djanogly Garai, Kate

Murray, Shigeo Hirano, Jeffrey B. Lewis, James M. Jr. Snyder, and Charles H. III Stewart. Cast Vote Records: A Database of Ballots from the 2020 U.S. Election, 2024. URL https://doi.org/10.7910/DVN/PQQ3KV.

[15] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

[16] G. Patrini, R. Nock, P. Rivera, and T. Caetano. (Almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

[17] VV Petrov. of independent random variables. *Yu. V. Prokhorov. V. StatuleviCius (Eds.)*, 1972.

[18] Ross L Prentice and Lianne Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82(1):113–125, 1995.

[19] N. Quadrianto, A. J. Smola, T. S Caetano, and Q. V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(Oct):2349–2374, 2009.

[20] Stefan Rueping. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 911–918, 2010.

[21] Tao Sun, Dan Sheldon, and Brendan O?Connor. A probabilistic approach for learning with label proportions applied to the us presidential election. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 445–454. IEEE, 2017.

[22] Wenpin Tang and Fengmin Tang. The poisson binomial distribution—old & new. *Statistical Science*, 38(1):108–119, 2023.

[23] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[24] Voting and Election Science Team. 2016 precinct-level election results. https://doi.org/10.7910/DVN/NH5S2I, 2018.

[25] Voting and Election Science Team. 2018 precinct-level election results. https://doi.org/10.7910/DVN/UBKYRU, 2019.

[26] Voting and Election Science Team. 2020 precinct-level election results. https://doi.org/10.7910/DVN/K7760H, 2020.

[27] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[28] Jon Wakefield. Ecological inference for $2\times$ 2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–425, 2004.

[29] Jonathan Wakefield and Ruth Salway. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):119–137, 2001.

Appendix A. Application

A.1. Predicting vote choice in U.S. Elections

Here we present some of the details of our central application: prediction of all 544 statewide elections in the United States taking place between 2016 and 2020. Our first figure illustrates the data-generating process from the voterfile, along with our core Normality assumption.

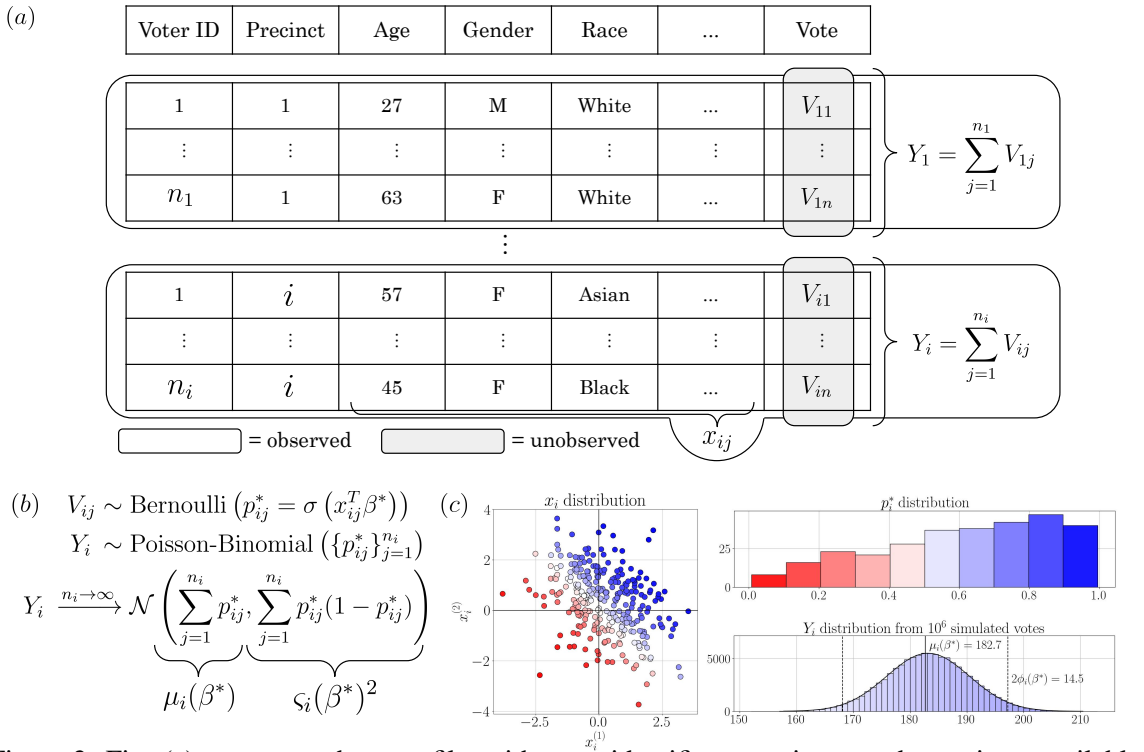
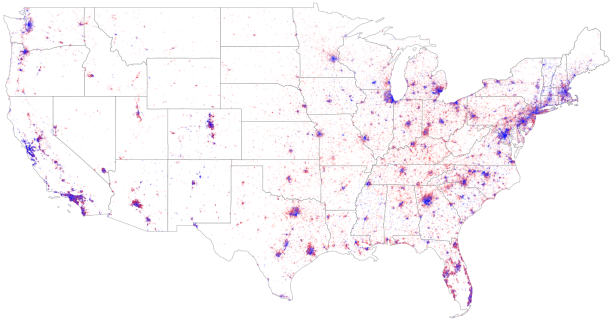
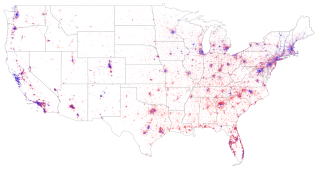


Figure 2: Fig. (a) represents the voterfile, with voter identifiers, precincts, and covariates available for every registered voter in the United States. The votes V_{ij} are not observed, but we observe the precinct-level vote counts Y_i. Fig. (b) represents our data-generating process. We assume some "true" parameter β^* determines each voter's probability of voting for the Democrat p^*_{ij}. This implies Y_i is Poisson-Binomial distributed and as long as m_i is large the Poisson-Binomial is approximately normally distributed with mean $\mu_i(\beta^*)$ and variance $\varsigma_i(\beta^*)^2$. Fig. (c) gives a visual representation of the data-generating process in (b): first we plot the covariates, then the distribution of the p^* corresponding to those covariates, and finally we show that resampling Y_i leads to the normal density. In the left panel, we plot the first two principal components of the voterfile for a single precinct in North Carolina, where each $x_{ij} = \left(x_{ij}^{(1)}, x_{ij}^{(2)}\right)$. Next we compute the probabilities for each voter by setting $\beta^* = (1, 1)$, so $p^*_{ij} = \sigma\left(x_{ij}^{(1)} + x_{ij}^{(2)}\right)$. We plot a histogram of these probabilities in the top right panel. Finally in the bottom right we simulate Y_i by sampling votes according to p^*_{ij} and summing. We simulate this process 10^6 times and plot the histogram, which demonstrates the empirical probability mass function matches the (appropriately scaled) Normal density as in (b).

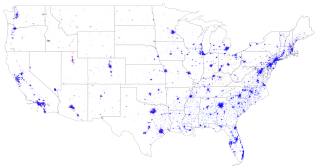
Our estimates allow researchers to ask fundamental questions about individual-level preferences in federal and state-level elections during a notoriously difficult-to-understand moment in American politics. We can produce individual-level estimates of party-line voting on a given ballot. We can also generate voter trajectories over time, showing how voters flow from one candidate to another. We can estimate candidate support by any crosstab – race, gender, age, and any possible interaction thereof, in any level of geographic specificity. We illustrate the power of this approach in Fig. 3, where we plot individual-level presidential vote choice estimates for all U.S. voters in 2020, chart the estimated flow of voters between presidential candidates from 2016 to 2020, and directly estimate the prevalence of party-line voting with validation using county-level cast vote records aggregated to the county level [14].



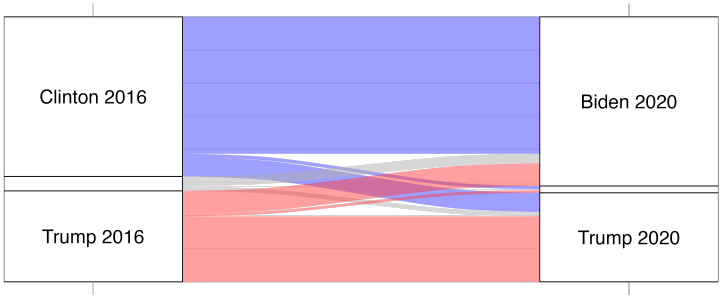
(a) US voters' individual-level estimated probability of supporting Joe Biden in the 2020 Presidential Election, plotted by registered address.



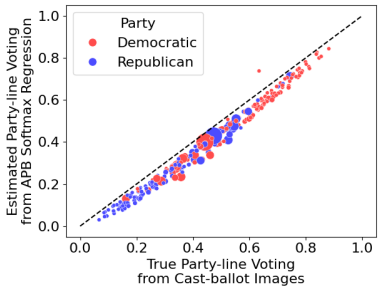
(b) White voters



(c) Black voters



(d) Estimated flow of candidate support between the 2016 and 2020 Presidential Elections for people who voted in both elections.



(e) Actual party-line voting vs estimated party-line voting in 2020.

Figure 3: Fig. (a) shows individual-level estimates of support. Figs. (b) and (c) show we can decompose these estimates by race. Fig. (d) shows we can study how support changes across elections in different populations of voters. Figs. (e) that we can estimate party-line voting, e.g. how often voters who support the Democrat for president support the Democrat for Senate and other down-ballot offices. Using county-level ballot data from several states in the 2020 election [14] we see that our estimates almost exactly recover the true rates of party-line voting.

Our main estimates are based on a simple specification using just a few core covariates, capturing party registration, race, gender, urbanicity, and voting history in the past two elections, as well as voting method:

```
party_dem , party_gop , party_other ,
race_black , race_hispanic , race_asian , race_white , race_native , race_missing ,
gender_male , gender_female ,
usr_r1 , usr_r2 , usr_s3 , usr_s4 , usr_u5 , usr_u6 , usr_missing ,
p{year}_voted_gop , p{year}_voted_dem ,
g{year - 4}_early , g{year - 4}_absentee , g{year - 4}_in_person ,
g{year - 2}_early , g{year - 2}_absentee , g{year - 2}_in_person ,
g{year}_early , g{year}_absentee , g{year}_in_person ,
gen_z , gen_millenial , gen_x , gen_boomer , gen_silent , gen_greatest , gen_missing
```

All of these variables are standard voterfile variables except for race, which is modeled using a commercial race model when race is not available on the voterfile. In addition to these base variables for 2018 and 2020 we also included all previous individual-level predictions for voters who were registered to vote in the previous elections with an interaction for whether they voted in that election. For new registrants, we filled in 0.5 and included an indicator.

A.2. Simulation Studies

Our central simulation setting involves iid Gaussian covariates which we randomly select into precincts. We fix the dimension d, the number of precincts n, and the total number of voters k. We start by sampling our population of voters, by drawing k random normal vectors $\boldsymbol{x}_{ij} \sim \mathcal{N}(0, I_d)$. Next we randomly sample the $\beta^* \sim \mathcal{N}(0, I_d)$ as above and a $\gamma^* \sim \mathcal{N}(0, I_{n\times d})$ which will select voters into precincts. Then we sample a precinct for each voter by letting $g_j \sim \text{Categorical}(\text{Softmax}(x_j^T \gamma^*))$. We then simulate votes and vote counts as $V_j \sim \text{Bernoulli}(\sigma(x_j^T \beta^*))$ and $Y_i = \sum_{j=1}^{k} \mathbb{I}\{g_j = i\} V_j$, replicating a Poisson-Binomial sampling process. We provide a colab notebook here which runs this simulation.

A.3. Comparison Against Other Ecological Inference Methods

Our goal with these models is to predict the candidate selected by each voter— an unknown outcome. This poses a challenge for evaluating the performance of our models and comparing them against other ecological inference methods.

To address this issue, we use a related task: modeling the probability that an individual casts a ballot (rather than modeling the candidate he or she supports). We train only on aggregated ballot counts from each precinct. This task is not a perfect proxy for modeling candidate selections, but it is an attractive option because it allows us to leverage the same set of covariates and the same aggregation structure, and we also have access to the individual-level outcomes for performance evaluation.

We use a data set comprising all voters from Morris County, New Jersey, an affluent and historically Republican-leaning county of about half a million residents. The voter file contains 316,724 registered voters and includes limited demographic information as well as information about whether each voter cast a ballot in general elections and primaries stretching back to the year 2000. There are 396 voting precincts in the county.

		Coverage Probability			MSE		
	d	3	5	10	3	5	10
$\bar{m}$	n						
10	100	0.89	0.804	0.642	0.035	0.102	0.946
	500	0.882	0.794	0.446	0.00541	0.0109	0.0434
	1000	0.868	0.752	0.362	0.00248	0.00667	0.0204
50	100	0.946	0.948	0.956	0.00396	0.00702	0.0188
	500	0.942	0.946	0.884	0.000837	0.00116	0.00386
	1000	0.954	0.928	0.888	0.000416	0.00067	0.00169
100	100	0.96	0.968	0.98	0.00216	0.00289	0.00687
	500	0.956	0.938	0.918	0.000376	0.000591	0.00127
	1000	0.934	0.946	0.934	0.000204	0.000291	0.000568
500	100	0.952	0.948	0.962	0.000433	0.000637	0.00133
	500	0.962	0.934	0.962	7.04e-05	0.000115	0.000203
	1000	0.954	0.95	0.964	4.18e-05	5.28e-05	9.87e-05
1000	100	0.946	0.942	0.966	0.000229	0.000342	0.000734
	500	0.946	0.96	0.948	4.23e-05	5.61e-05	0.0001
	1000	0.958	0.95	0.96	1.95e-05	2.47e-05	4.84e-05

Table 1: Simulation results for various $\bar{m}$, d, and n. Each result is computed based on 500 simulations. We present both coverage and mean squared error results. When $\bar{m} \geqslant 50$ and as n grows we see the joint confidence ellipse achieves nominal coverage and decreasing mean squared error.

We fit models to eight data sets in total. To explore performance in different outcome regimes, we predict whether voters participated in each election from 2014 to 2017, in which 34%, 19%, 76%, and 45% of all voters in our data set cast a ballot, respectively. For each year, we fit two models: a parsimonious "demographics-only" model containing just four covariates (age, party, gender, and whether the voter lives in an apartment); and a "demographics and voter history" model that also contains nine variables corresponding to the voter's participation and voting method in the given year's primary and the primaries and general elections of the prior four years.

We compare the performance of a number of methods:

- To obtain an upper bound on performance, we fit a logistic regression and a Gradient Boosted Machine (GBM) to the data set while giving them access to the individual-level outcomes [6]. Because these models "see" individual-level data, they should outperform methods that only have access to aggregated data.
- We fit three variants of our logistic regression formulation.
  - In the first ("Logit with Gaussian Gradient"), the coefficients are fit via gradient ascent exclusively using the Gaussian approximation. We run for 120 iterations using a learning rate of 2×10^{-5}.
  - In the second ("Logit with Gaussian Gradient, PoiBin Backtracking"), we run ten iterations using the approximated gradient and fixed step size; for the remaining 110 iterations, we use the normal-approximation gradient to choose an ascent direction but use backtracking line search based on the *true* likelihood to choose a step size.
  - The third algorithm ("Logit with Gaussian Gradient, PoiBin Backtracking, True Gradient") is identical to the second, except we run only 100 iterations using backtracking line search. The final ten iterations are then instead run using the true gradient derived in Appendix B.1.

  These three variants are used to explore the practical effect of the Gaussian approximation on our model's accuracy.
- Following the approach in [20], we baseline against the simplest ecological inference method: assigning each unit in a given aggregation block the average of the outcomes in that aggregation block. In our setting, this means each voter in a precinct is assigned the voter turnout proportion in that precinct as a pseudo-outcome, and a logistic regression model is fit to these data.
- We baseline against ecological regression as implemented in the `ecoreg` package in R [8].
- We baseline against Rueping's Inverse Calibration method [20]. Aggregate accuracy on the 40 precincts in the development set is again used, this time for tuning the C and ϵ parameters.
- Lastly, we baseline against the Mean Map [19], Laplacian Mean Map, and Alternating Mean Map [16]. Hyperparameters are again tuned using squared error on the development set.

Results for the demographics-only model are provided in Table 2 and results from the expanded data set are provided in Table 3. The relative strength of the logistic regression formulation is immediately obvious: these models achieve the highest ROC AUC values in all but one of the

Table 2: ROC AUC scores for models predicting voter turnout, fit to demographics-only data sets. The highest values among ecological inference models are underlined.

	Demographics Only			
	2017	**2016**	**2015**	**2014**
Standard Methods (non-ecological)				
Logistic Regression	72.0%	71.2%	75.2%	76.9%
GBM	73.0%	72.7%	75.5%	77.2%
Proposed Methods				
Logit with Gaussian Gradient	<u>69.3%</u>	68.3%	<u>73.6%</u>	74.7%
Logit with Gaussian Gradient, PoiBin Backtracking	69.3%	<u>68.4%</u>	73.6%	74.7%
Logit with Gaussian Gradient, PoiBin Backtracking, True Gradient	69.3%	68.4%	72.2%	74.5%
Comparison Methods				
Logistic Regression on Aggregates	65.9%	60.0%	71.7%	69.0%
Ecological Regression	67.6%	66.7%	72.8%	<u>75.0%</u>
Inverse Calibration	61.1%	61.9%	72.9%	41.5%
Mean Map	51.5%	60.1%	33.3%	31.9%
Laplacian Mean Map	51.0%	46.1%	37.9%	51.1%
Alternating Mean Map	58.9%	62.6%	58.0%	58.4%

eight conditions and frequently come very close to the performance of methods with access to the individual outcomes. Also evident is the fact that little to no predictive power is gained by making use of the real likelihood rather than the approximation. Backtracking on the true Poisson Binomial likelihood or using the true gradient actually slightly degrades performance in most cases, while also slowing training.

Ecological regression performs well in all conditions and outperforms our proposed methods in the demographics-only model for 2014. The other tested methods are generally not competitive. The logistic regression on aggregates technique performs surprisingly well given its extreme simplicity, but it still underperforms the proposed logistic methods. Inverse calibration sees a noticeable performance bump with the inclusion of additional covariates. The Mean Map, LMM, and AMM methods typically do poorly, with only AMM consistently beating random guessing in the demographics-only case. Each of these methods is somewhat sensitive to hyperparameter values, and tuning is extremely challenging in the absence of labeled data in the development set. We are using squared error across development precincts as a proxy measure, and it's highly plausible that alternative proxies would yield better hyperparameter values. Nonetheless, a strength of our proposed methods is that they require very little tuning to get good performance.

Table 3: ROC AUC scores for models predicting voter turnout, fit to demographics and voter history data sets. The highest values among ecological inference models are underlined.

	Demographics and Voting History			
	2017	**2016**	**2015**	**2014**
Standard Methods (non-ecological)				
Logistic Regression	85.9%	84.5%	88.6%	89.5%
GBM	86.2%	85.5%	88.8%	89.6%
Proposed Methods				
Logit with Gaussian Gradient	<u>83.9%</u>	<u>82.0%</u>	<u>81.0%</u>	86.3%
Logit with Gaussian Gradient, PoiBin Backtracking	83.8%	82.0%	81.0%	<u>86.4%</u>
Logit with Gaussian Gradient, PoiBin Backtracking, True Gradient	83.8%	81.9%	80.6%	86.3%
Comparison Methods				
Logistic Regression on Aggregates	75.0%	72.4%	77.2%	76.8%
Ecological Regression	67.5%	68.7%	71.8%	76.1%
Inverse Calibration	64.2%	77.6%	78.4%	66.9%
Mean Map	45.4%	54.4%	48.4%	51.8%
Laplacian Mean Map	49.5%	51.5%	57.6%	49.4%
Alternating Mean Map	51.9%	52.9%	44.4%	46.2%

Appendix B. Methods

B.1. Gradients and Hessians

We start by writing out the gradients and Hessians as well as their expected forms, which will be useful throughout our theoretical discussion.

B.1.1. EXACT LIKELIHOOD

We can write the exact likelihood and its gradient and hessian:

$$\ell_i(\beta) = \log\left(\sum_{A\in\mathbb{P}_{Y_i}([m_i])} \exp\left(\sum_{j\in A} \boldsymbol{x}_{ij}^\top\beta\right)\right) - \sum_{j=1}^{m_i}\log\left(1+\exp(\boldsymbol{x}_{ij}^\top\beta)\right)$$

$$\nabla_\beta \ell_i(\beta) = \sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\exp\left(\sum_{j\in A}\boldsymbol{x}_{ij}^\top\beta\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\exp\left(\sum_{j\in A'}\boldsymbol{x}_{ij}^\top\beta\right)}\cdot\sum_{j\in A}\boldsymbol{x}_{ij} - \sum_{j=1}^{m_i}\frac{\exp(\boldsymbol{x}_{ij}^\top\beta)}{1+\exp(\boldsymbol{x}_{ij}^\top\beta)}\boldsymbol{x}_{ij}$$

$$\begin{aligned}\nabla^2_\beta \ell_i(\beta) = &\sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\exp\left(\sum_{j\in A}\boldsymbol{x}_{ij}^\top\beta\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\exp\left(\sum_{j\in A'}\boldsymbol{x}_{ij}^\top\beta\right)}\cdot\left(\sum_{j\in A}\boldsymbol{x}_{ij}\right)\left(\sum_{j\in A}\boldsymbol{x}_{ij}\right)^\top \\ &- \left(\sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\exp\left(\sum_{j\in A}\boldsymbol{x}_{ij}^\top\beta\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\exp\left(\sum_{j\in A'}\boldsymbol{x}_{ij}^\top\beta\right)}\cdot\sum_{j\in A}\boldsymbol{x}_{ij}\right)\left(\sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\exp\left(\sum_{j\in A}\boldsymbol{x}_{ij}^\top\beta\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\exp\left(\sum_{j\in A'}\boldsymbol{x}_{ij}^\top\beta\right)}\cdot\sum_{j\in A}\boldsymbol{x}_{ij}\right)^\top \\ &- \sum_{j=1}^{m_i}\frac{\exp(\boldsymbol{x}_{ij}^\top\beta)}{(1+\exp(\boldsymbol{x}_{ij}^\top\beta))^2}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\end{aligned}$$

The gradient and Hessian here actually have a lot of structure which we can use to analyze their behavior. Using the probabilistic structure we will show how we can re-write the gradient in terms very comparable to the standard logistic regression.

We will show the first term of the gradient is equal to

$$\mathbb{E}\left(\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}V_{ij} \mid \sum_{j=1}^{m_i}V_{ij} = Y_i\right).$$

We can interpret this in the sampling setting as the expected sample sum of the covariate vectors, conditional on the total number of units sampled among those in $[m_i]$. Careful expansion yields

$$\begin{aligned}\mathbb{E}\left(\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}V_{ij} \mid \sum_{j=1}^{m_i}V_{ij} = Y_i\right) &= \sum_{A\in\mathbb{P}_{Y_i}([m_i])} P(\text{sample } A \text{ is drawn} \mid \text{sample of size } Y \text{ is drawn})\cdot\sum_{j\in A}\boldsymbol{x}_{ij} \\ &= \sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{P(\text{sample } A \text{ is drawn})}{P(\text{sample of size } Y \text{ is drawn})}\cdot\sum_{j\in A}\boldsymbol{x}_{ij} \\ &= \sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\prod_{i\in A}p_{ij}\prod_{i\in A^c}(1-p_{ij})}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\prod_{i\in A'}p_{ij}\prod_{i\in A'^c}(1-p_{ij})}\cdot\sum_{j\in A}\boldsymbol{x}_{ij} \\ &= \sum_{A\in\mathbb{P}_{Y_i}([m_i])} \frac{\exp\left(\sum_{j\in A}\boldsymbol{x}_{ij}^\top\beta\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m_i])}\exp\left(\sum_{j\in A'}\boldsymbol{x}_{ij}^\top\beta\right)}\cdot\sum_{j\in A}\boldsymbol{x}_{ij}\end{aligned}$$

which we recognize as the first term in our gradient from the prior section. Analogous computations apply for the Hessian, giving us the following equivalent forms:

$$\nabla_\beta \ell_i(\beta) = \mathbb{E}\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) - \mathbb{E}\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij}\right)$$
$$\nabla^2_\beta \ell_i(\beta) = \text{Cov}\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) - \text{Cov}\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij}\right)$$

Pushing the expectation into the sum we can rewrite this:

$$\begin{aligned}
\mathbb{E}_\beta\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) &= \sum_{j=1}^{m_i} \boldsymbol{x}_{ij} \mathbb{E}_\beta\left(V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) \\
&= \sum_{j=1}^{m_i} \boldsymbol{x}_{ij} P_\beta\left(V_{ij} = 1 \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) \\
&= \sum_{j=1}^{m_i} \frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} p_{ij} \boldsymbol{x}_{ij}
\end{aligned}$$

We can similarly rewrite the covariance:

$$\begin{aligned}
\text{Cov}_\beta\left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) &= \sum_{j=1}^{m_i} \text{Var}_\beta\left(V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^\top \\
&= \sum_{j=1}^{m_i} \mathbb{E}_\beta\left(\left(V_{ij} - \mathbb{E}\left(V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right)\right)^2 \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^\top \\
&= \sum_{j=1}^{m_i} \mathbb{E}_\beta\left(V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right) - \mathbb{E}\left(V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i\right)^2 \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^\top \\
&= \sum_{j=1}^{m_i} \left(\frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} p_{ij}\right)\left(1 - \frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} p_{ij}\right) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^\top
\end{aligned}$$

This gives us a form of the gradient that very closely resembles the gradient for the logistic regression and a form of the Hessian that involves the Hessian for the logistic regression and an additional reweighting term:

$$\nabla_\beta \ell_i(\beta) = \sum_{j=1}^{m_i} \left(\frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} - 1\right) p_{ij} \boldsymbol{x}_{ij}$$
$$\nabla^2_\beta \ell_i(\beta) = \sum_{j=1}^{m_i} \left(\left(\frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} p_{ij}\right)\left(1 - \frac{P_\beta\left(\sum_{k\neq j} V_{ij} = Y_i - 1\right)}{P_\beta\left(\sum_{j=1}^{m_i} V_{ij} = Y_i\right)} p_{ij}\right) - p_{ij}(1 - p_{ij})\right) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^\top$$

We can also write these objects in expectation:

$$\mathbb{E}[\nabla_\beta \ell_i(\beta)] = \sum_{j=1}^{m_i}\left(\sum_{k=1}^{m_i}\left(\frac{P_*\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)}{P_\beta\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)}\right) P_\beta\left(\sum_{j'\neq j} V_{ij'}=k-1\right)-1\right) p_{ij}\boldsymbol{x}_{ij}$$

$$\begin{aligned}\mathbb{E}[\nabla^2_\beta \ell_i(\beta)] = &\sum_{j=1}^{m_i}\left(\sum_{k=1}^{m_i}\left(\frac{P_*\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)}{P_\beta\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)}\right) P_\beta\left(\sum_{j'\neq j} V_{ij'}=k-1\right)-1\right) p_{ij}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top \\ &+\sum_{j=1}^{m_i}\left(\sum_{k=1}^{m_i}\left(\frac{P_*\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)}{P_\beta\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)^2}\right) P_\beta\left(\sum_{j'\neq j} V_{ij'}=k-1\right)^2-1\right) p_{ij}^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\end{aligned}$$

It is worth noting that when $P_*\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right) = P_\beta\left(\sum_{j'=1}^{m_i} V_{ij'}=k\right)$ we will have cancellation of the terms in the ratio, and the sum $\sum_{k=1}^{m_i} P_\beta\left(\sum_{j'\neq j} V_{ij'}=k-1\right)=1$. So under this condition, the gradient will be zero.

B.1.2. APPROXIMATE LIKELIHOOD

To study the geometry of the likelihood it will be useful to write out the likelihood, gradient, and hessian:

$$\begin{aligned}\ell_i(\beta) = &\ \frac{1}{2}\left(\log\left(\varsigma_i(\beta)^2\right)+\frac{(Y_i-\mu_i(\beta))^2}{\varsigma_i(\beta)^2}\right)\\
\nabla_\beta \ell_i(\beta) = &\ \frac{1}{2}\left(\frac{(Y_i-\mu_i(\beta))^2-\varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}(2p_{ij}-1)p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right)-\left(\frac{Y_i-\mu_i(\beta)}{\varsigma_i(\beta)^2}\right)\left(\sum_{j=1}^{m_i}p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right)\\
\nabla^2_\beta \ell_i(\beta) = &\ \frac{1}{2}\left(\frac{(Y_i-\mu_i(\beta))^2-\varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}p_{ij}(1-p_{ij})(1-6p_{ij}(1-p_{ij}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right)\\
&-\frac{(Y_i-\mu_i(\beta))}{\varsigma_i(\beta)^2}\left(\sum_{j=1}^{m_i}p_{ij}(1-p_{ij})(1-2p_{ij})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right)\\
&-\frac{1}{2\varsigma_i(\beta)^4}z_i z_i^\top+\frac{1}{\varsigma_i(\beta)^2}\left(y_i+\frac{(Y_i-\mu_i(\beta))}{\varsigma_i(\beta)^2}z_i\right)\left(y_i+\frac{(Y_i-\mu_i(\beta))}{\varsigma_i(\beta)^2}z_i\right)^\top\end{aligned}$$

where

$$y_i=\left(\sum_{j=1}^{m_i}p_{ij}(1-p_{ij})\boldsymbol{x}_{ij}\right)\quad z_i=\left(\sum_{j=1}^{m_i}p_{ij}(1-p_{ij})(1-2p_{ij})\boldsymbol{x}_{ij}\right).$$

We will also consider these objects under the expectation:

$$\mathbb{E}\left[\frac{Y_i-\mu_i(\beta)}{\varsigma_i(\beta)^2}\right]=\frac{\mu_i(\beta^*)-\mu_i(\beta)}{\varsigma_i(\beta)^2},\ \ \mathbb{E}\left[\frac{(Y_i-\mu_i(\beta))^2}{\varsigma_i(\beta)^2}\right]=\frac{\varsigma_i(\beta^*)^2}{\varsigma_i(\beta)^2}+\frac{(\mu_i(\beta^*)-\mu_i(\beta))^2}{\varsigma_i(\beta)^2}$$

$$\mathbb{E}[\ell_i(\beta)] = \frac{1}{2}\left(\log\left(\varsigma_i(\beta)^2\right) + \frac{\varsigma_i(\beta^*)^2}{\varsigma_i(\beta)^2} + \frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)^2}{\varsigma_i(\beta)^2}\right)$$

$$\nabla_\beta \mathbb{E}[\ell_i(\beta)] = \frac{1}{2}\left(\frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)^2 + \varsigma_i(\beta^*)^2 - \varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}(2p_{ij}-1)p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right)$$
$$- \left(\frac{\mu_i(\beta^*) - \mu_i(\beta)}{\varsigma_i(\beta)^2}\right)\left(\sum_{j=1}^{m_i} p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right)$$

$$\mathbb{E}[\nabla^2 \ell_i(\beta)] = \frac{1}{2}\left(\frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)^2 + \varsigma_i(\beta^*)^2 - \varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i} p_{ij}(1-p_{ij})(1-6p_{ij}(1-p_{ij}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right)$$
$$- \frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)}{\varsigma_i(\beta)^2}\left(\sum_{j=1}^{m_i} p_{ij}(1-p_{ij})(1-2p_{ij})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right)$$
$$+ \frac{1}{\varsigma_i(\beta)^4}\left(\frac{\varsigma_i(\beta^*)^2}{\varsigma_i(\beta)^2} - \frac{1}{2}\right)zz^\top + \frac{1}{\varsigma_i(\beta)^2}\left(y_i + \frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)}{\varsigma_i(\beta)^2}z_i\right)\left(y_i + \frac{\left(\mu_i(\beta^*) - \mu_i(\beta)\right)}{\varsigma_i(\beta)^2}z_i\right)^\top\Bigg)$$

Taking this expectation does not immediately help us; all of the problems and dynamics of the eigenvalues of the Hessian are similarly true for this expected objective. But it is significantly more tractable to analyze.

B.2. Statistical Results

Here we outline several statistical theorems in the exact and approximate settings focusing on identifiability and asymptotic normality. We present all the proofs in the following appendices.

B.2.1. EXACT IDENTIFIABILITY

We can start analyzing the likelihood by trying to understand its symmetries and the necessary conditions for identifiability.

Lemma 3 *For a single precinct let $p_{i()}$ denote the vector of ordered probabilities $p_{i()} = (p_{j_1} < \cdots < p_{j_{m_i}})$. The Poisson-Binomial distribution is identifiable up to $p_{i()}$, which is up to permutations.*

This result tells us that what matters for the identifiability of the Poisson-Binomial Logistic regression is the set of probabilities in each precinct with respect to β^*, relative to other β. Essentially if β^* induces a unique set of probabilities across all precincts then we will have identifiability. If there is some other β that induces the same sets of probabilities in every precinct as β^* then the regression will be non-identifiable.

More formally, we can think about the sets of ordered probabilities induced by β, which we index by j_k^β (breaking ties arbitrarily) so that:

$$\left(x_{ij_1^\beta}\right)^\top \beta \leq \cdots \leq \left(x_{ij_{m_i}^\beta}\right)^\top \beta.$$

Now we define the set of β which induce the same probabilities as β^* in precinct i:

$$\mathcal{B}_i^{(PB)} = \left\{\beta \;\middle|\; \left(\left(x_{ij_1^\beta}\right)^\top \beta, \ldots, \left(x_{ij_{m_i}^\beta}\right)^\top \beta\right) = \left(\left(x_{ij_1^{\beta^*}}\right)^\top \beta^*, \ldots, \left(x_{ij_{m_i}^{\beta^*}}\right)^\top \beta^*\right)\right\}$$

Using these sets we can start to understand the identifiability of the Poisson-Binomial logistic regression and the geometry of the likelihood.

Theorem 4 ***(Identifiability)***

(i) The Poisson-Binomial logistic regression is identifiable if and only if $\cap_{i=1}^{n} \mathcal{B}_i^{(PB)} = \beta^*$

(ii) Assuming $\boldsymbol{x}_{ij}$ *are bounded, the Hessian is asymptotically positive semi-definite and so the log-likelihood is asymptotically concave at any* $\beta \in \cap_{i=1}^{n} \mathcal{B}_i^{(PB)}$*. Furthermore, we can consider the precinct-level vectors:*

$$\nabla_\beta \ell_i(\beta) = \sum_{j=1}^{m_i} \left(\frac{P_\beta \left(\sum_{k \neq j} V_{ij} = Y_i - 1 \right)}{P_\beta \left(\sum_{j=1}^{m_i} V_{ij} = Y_i \right)} - 1 \right) p_{ij} \boldsymbol{x}_{ij},$$

when these vectors $\nabla_\beta \ell_i(\beta)$ *span* $\mathbb{R}^d$ *the log-likelihood will be strictly concave at* β^**.*

The likelihood will be asymptotically log-concave in a neighborhood around the true parameter value if there is sufficient differentiation in the covariates $\boldsymbol{x}_{ij}$ across precincts. In practical examples, this condition holds as long n is sufficiently large compared to p, and the covariates differ sufficiently across precincts.

Assumption 3 ***(Identifiability)*** *In addition to Asmp. 1 we assume the set of precincts are identifiability with respect to* β^* *and that the Hessian is full rank at* β^**.*

Our next theorem shows that under this assumption we will have consistency of the maximum likelihood estimator and asymptotic normality:

Theorem 5 ***(Consistency and Asymptotic Normality of the Likelihood)*** *We will assume that Assumption 3 holds. Let* $\hat{\beta}$ *denote the maximum likelihood estimator for the likelihood* $\ell_{PoiBin}(\beta)$*. Then as* $n \to \infty$*, we will have:*

$$\hat{\beta}_n \xrightarrow{n \to \infty} \beta^* \text{ and } \sqrt{n} \left(\hat{\beta}_n - \beta^* \right) \xrightarrow{n \to \infty} \mathcal{N} \left(0, \mathbb{E} \left[\nabla_\beta^2 \ell_{PoiBin} \left(\beta^* \right) \right] \right). \tag{3}$$

In the next section, we will explore what this result means, both in the well-specified setting we have considered up to this point and in the more challenging mis-specified setting.

B.2.2. APPROXIMATE IDENTIFIABILITY

We can formalize this into an identifiability theorem, similar to the theorem in Section 4:

Theorem 6 ***(Identifiability)***

(i) The Poisson-Binomial logistic regression is identifiable with respect to the approximate likelihood if and only if $\cap_{i=1}^{n} \mathcal{B}_i = \beta^*$

(ii) The Hessian of the approximate likelihood is asymptotically negative semi-definite and so the log-likelihood is asymptotically concave at β^*. Furthermore, we can consider the precinct-level vectors:

$$\nabla_\beta \ell_i(\beta) = \frac{1}{2}\left(\frac{(Y_i - \mu_i(\beta))^2 - \varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)\left(\sum_{j=1}^{m_i}(2p_{ij}-1)p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right) - \left(\frac{Y_i - \mu_i(\beta)}{\varsigma_i(\beta)^2}\right)\left(\sum_{j=1}^{m_i} p_{ij}(1-p_{ij})\mathbf{x}_{ij}\right),$$

when these vectors $\nabla_\beta \ell_i(\beta)$ span $\mathbb{R}^d$ the log-likelihood will be strictly concave at β^*.

Now we can again assume these conditions hold which allows us to state consistency and asymptotic normality results for the approximate log-likelihood:

Assumption 4 ***(Identifiability with respect to the Approximate Log-Likelihood)*** *Extending the requirements of Assumptions 3 and 2 we further assume that $\mathcal{P}$ is identifiable with respect to the approximate log-likelihood with the approximate Hessian full rank at β^*.*

Theorem 7 ***(Consistency and Asymptotic Normality of the Approximate Likelihood)*** *We will assume that the approximate Poisson-Binomial logistic regression is identifiable with respect to $\mathcal{P}$ and that Assumptions 3- 4 hold. Let $\hat{\beta}$ denote the maximum likelihood estimator for the approximate likelihood $\ell(\beta)$. Then as $n \to \infty$, we will have:*

$$\hat{\beta}_n \xrightarrow{n\to\infty} \beta^* \text{ and } \sqrt{n}(\hat{\beta}_n - \beta^*) \xrightarrow{n\to\infty} \mathcal{N}\left(0, \mathbb{E}\left[\nabla^2_\beta \ell(\beta)\right]\right) \tag{4}$$

B.2.3. IDENTIFIABILITY RESULTS

We will start with our main identifiability results for the Poisson-Binomial log-likelihood and the approximate log-likelihood under the Poisson-Binomial model.

Proof of Lemma 3 **Proof** The Poisson-Binomial is identifiable up to the ordered probability vector $p_{()}$ if

$$\left\{\text{for all } k\text{: } \sum_{A\in P_k([n])}\left(\prod_{i\in A} p_i\right)\left(\prod_{i\in A^c}(1-p_i)\right) = \sum_{A\in P_k([n])}\left(\prod_{i\in A} q_i\right)\left(\prod_{i\in A^c}(1-q_i)\right)\right\} \Rightarrow p_{()} = q_{()}.$$

We will prove this condition holds by constructing a set of implied equivalences and then using those as coefficients in a pair of polynomials, which will imply $p_{()} = q_{()}$. We will start by analyzing the case $k = n$ which gives us:

$$\prod_i p_i = \prod_i q_i$$

Then we can examine $k = n - 1$:

$$\sum_j \prod_{i\neq j} p_i(1-p_j) = \sum_j \prod_{i\neq j} q_i(1-q_j) \Rightarrow \sum_j \prod_{i\neq j} p_i = \sum_j \prod_{i\neq j} q_i$$

since the remaining $\prod_i p_i$ and $\prod_i q_i$ terms cancel.
Likewise we can consider the case $k = n - 2$:

$$\sum_k \sum_{j\neq k} \prod_{i\neq k,j} p_i(1-p_j)(1-p_k) = \sum_k \sum_{j\neq k} \prod_{i\neq k,j} q_i(1-q_j)(1-q_k) \Rightarrow \sum_k \sum_{j\neq k} \prod_{i\neq k,j} p_i = \sum_k \sum_{j\neq k} \prod_{i\neq k,j} q_i$$

where again we have cancellation by the $k = n$ and $k = n - 1$ cases above. Proceeding in this fashion we will have that:

$$\forall k = 1, \ldots, n : \sum_{A \in P_{n-k}([n])} \prod_{i \in A} p_i = \sum_{A \in P_{n-k}([n])} \prod_{i \in A} q_i$$

Now we can rewrite these relations as the coefficients on pair of degree n polynomials:

$$\sum_{k=0}^{n-1} (-1)^n \left(\sum_{A \in P_{n-k}([n])} \prod_{i \in A} p_i \right) x^k + x^n = 0, \quad \sum_{k=0}^{n-1} (-1)^n \left(\sum_{A \in P_{n-k}([n])} \prod_{i \in A} q_i \right) x^k + x^n = 0$$

It turns out these sets of coefficients are precisely the elementary symmetric polynomials of p and q, and so by Vieta's formulas we know that p and q exactly describe the roots of the respective polynomials. Since we have already shown that these polynomials in p and q have the same coefficients we have that $p_{()} = q_{()}$ whenever $\mathcal{L}_{\text{PoiBin}}(k; p_{()}) = \mathcal{L}_{\text{PoiBin}}(k; q_{()})$ for all k. ■

Proof of Theorem 4 **Proof** These two results are largely independent, and we prove them separately:

(i) If $|\cap_{i=1}^{\infty} \mathcal{B}_i^{(PB)}| = 1$ then $\cap_{i=1}^{\infty} \mathcal{B}_i^{(PB)} = \beta^*$ and by Lemma 3 we have identifiability.

(ii) Per the results in Appendix B.1, we can write the scaled Hessian as:

$$\frac{1}{n} \nabla^2 \ell(\beta) = \frac{1}{n} \sum_{i=1}^{n} \text{Cov} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i \right) - \text{Cov} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \right)$$

where n is the number of precincts. By Kolmogorov's Strong Law [23], we see

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \text{Cov} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i \right) &\xrightarrow{a.s.} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left(\text{Cov} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \text{Cov} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \right) - \text{Cov} \left(\mathbb{E} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i \right) \right) \end{aligned}$$

where the second line is due to the Law of Total Covariance. Thus

$$\frac{1}{n} \nabla^2 \ell(\beta) \xrightarrow{a.s.} -\frac{1}{n} \sum_{i=1}^{n} \text{Cov} \left(\mathbb{E} \left(\sum_{j=1}^{m_i} \boldsymbol{x}_{ij} V_{ij} \mid \sum_{j=1}^{m_i} V_{ij} = Y_i \right) \right) .$$

and the result follows from the fact that any covariance matrix must be positive semidefinite. By an identical SLLN argument the Hessian and the outer product of the gradient will behave identically, so the full rank condition is rather immediate. Since the likelihood, gradient, and Hessian will be the same for all $\beta \in \cap_{i=1}^{n} \mathcal{B}_i^{(PB)}$ we can conclude that the Hessian will be positive semidefinite for all $\beta \in \cap_{i=1}^{n} \mathcal{B}_i^{(PB)}$.

■

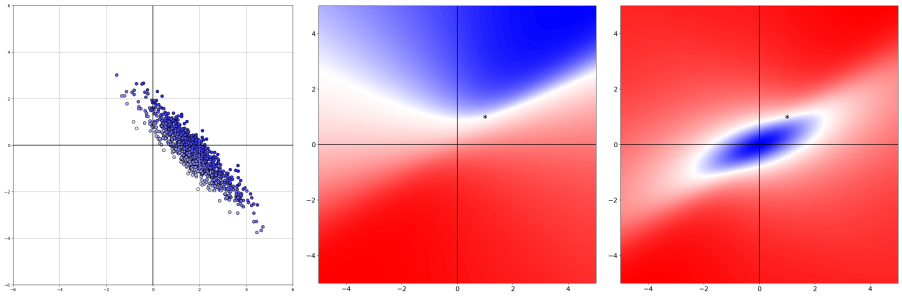
Proof of Theorem 6 **Proof** The proof follows the exact structure of the proof of Theorem 4 replacing $\mathcal{B}_i^{(PB)}$ with $\mathcal{B}_i$ and $\nabla \ell_{\text{PoiBin}}(\beta^*)$ with $\nabla \ell_i(\beta^*)$. ■

Proof of Lemma 2 **Proof** The necessity is immediate from the counterexamples in Fig. 1.

Now we can state a stronger version of Lemma 2 with the necessary and sufficient conditions:

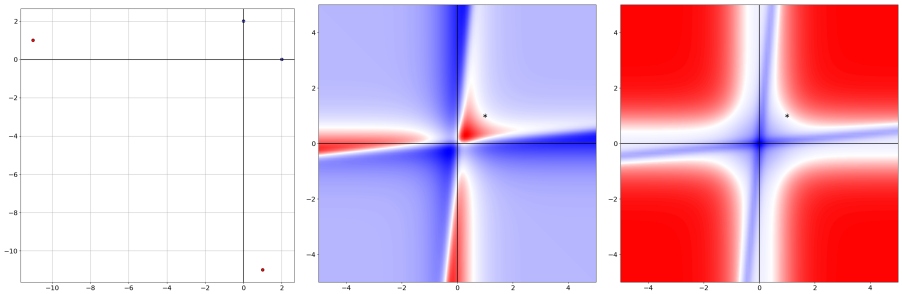
Lemma 8 *Assuming the model is well-specified with β^*, the necessary and sufficient conditions for a precinct to have a single global optimum are that the covariates, $\{\boldsymbol{x}_{ij}\}_{j=1}^{m_i}$:*

1. *span $\mathbb{R}^d$*
2. *are rotationally symmetric about β^**
3. *satisfy partisan monotonicity with respect to β^**



(a) x_i (b) $\mu_i(\beta)^2$ (c) $\varsigma_i(\beta)^2$

(d) A precinct which satisfies partisan monotonicity but has multiple global minima since the covariates are not rotationally symmetric.



(e) x_i (f) $\mu_i(\beta)^2$ (g) $\varsigma_i(\beta)^2$

(h) A precinct with rotationally symmetric covariates but which has multiple global minima since the covaraites do not satisfy partisan monotonicity with respect to β^*.

Figure 4: Two counter examples demostrating the necessity of both rotational symmetry and partisan monotonicity in the two dimensional setting. Left is the covariate distribution, in the middle we have the mean surface, and on the right we have the variance surface. Note that without partisan monotonicity the mean and variance surfaces can essentially be arbitrarily badly behaved, as seen in (b).

Clearly, if the covariates do not span $\mathbb{R}^d$ we will have multiple global optima. We prove the necessity of rotational symmetry and partisan monotonicity by counter-example in Fig. 4.

Now we can move on to sufficiency. Under partisan monotonicity and the full rank condition, we know that Σ_i will form a compact surface, symmetric about the separating hyperplane normal to β^*. We similarly know that $\mathcal{M}_i$ will be a non-compact surface. Since the covariates are radially symmetric about β^* we know that these surfaces must also be radially symmetric about β^*. Furthermore, we know that the surfaces must touch at β^*. Now by the radial symmetry of the covariates and partisan monotonicity we also know that β^* will be the minimum norm $\beta \in \beta^*$ and the maximum norm $\beta \in \Sigma_i$, so β^* will be the only point at which the two surfaces touch.

■

B.2.4. RADEMACHER COMPLEXITY, CONSISTENCY, AND ASYMPTOTIC NORMALITY

We will start by showing the Poisson-Binomial logistic log-likelihood and the approximate log-likelihood are Lipshitz assuming bounded covariates and a compact parameter space. Using the Lipschitz constant we will bound the Rademacher complexity. This will give us a Uniform Strong Law of Large Numbers, which combined with identifiability and compactness will give us consistency. We will then argue for asymptotic normality leveraging the fact we can take a second-order Taylor exapansion of both of our log-likelihoods. All of our arguments will hold under misspecification of the model for p^*, as long as our strong identifiability conditions hold. For simplicity here we treat the case where every precinct i has the same number of voters: $m_i = m$.

Lemma 9 *The Poisson-Binomial logistic log-likelihood is Lipschitz with constant $\frac{2}{\sqrt{m}}$, and including the linear connection the log-likelihood is Lipschitz with constant $2\sqrt{\frac{1}{m}\sum_{j=1}^{m}||x_{ij}||_2^2}$.*

Proof *Taking the gradient with respect to z_k we can take a straightforward bound:*

$$\left(\frac{\partial}{\partial z_k}\ell_i(z)\right)^2 = \frac{1}{m^2}\left(\sum_{A\in\mathbb{P}_{Y_i}([m])}\frac{\mathbb{1}\{k\in A\}\exp\left(\sum_{j\in A}z_{ij}\right)}{\sum_{A'\in\mathbb{P}_{Y_i}([m])}\exp\left(\sum_{j\in A'}z_{ij}\right)} - \frac{\exp(z_{ik})}{1+\exp(z_{ik})}\right)^2 \leq \frac{4}{m^2}$$

Then summing over each $k = 1, \ldots, m$ we have $||\nabla_z \ell_i(z)||_2 \leq \frac{2}{\sqrt{m}}$. This bound is tight up to the factor of 2 since setting $k = 0$ we have:

$$||\nabla_z \ell_i(z)||^2 = \sqrt{\sum_{k=1}^{m}\left(\frac{\exp(z_{ik})}{1+\exp(z_{ik})}\right)^2} \leq \lim_{(z_{i1},\ldots,z_{im})\to\infty}\sqrt{\sum_{k=1}^{m}\left(\frac{\exp(z_{ik})}{1+\exp(z_{ik})}\right)^2} = \frac{1}{\sqrt{m}}$$

The second claim is almost immediate since a linear function $X_i^\top \beta$ is Lipschitz with $\sqrt{\sum_{j=1}^{m}||\boldsymbol{x}_{ij}||_2^2}$ and the composition of Lipschitz functions is Lipschitz with constant the products of the constants.
■

Lemma 10 *The approximate log-likelihood under the Poisson-Binomial model is Lipschitz with constant*

$$L = \sqrt{\frac{31K^2m^2+4}{1728m^3}}$$

and including the linear connection the log-likelihood is Lipschitz with constant

$$L = \sqrt{\left(\frac{31K^2}{1728} + \frac{1}{432m^2}\right)\frac{1}{m}\sum_{i=1}^{m}||\boldsymbol{x}_{ij}||_2^2}$$

Proof *Now we can write the gradient of the log-likelihood in terms of z:*

$$\begin{aligned}\left(\frac{\partial}{\partial z_j}\ell_i(z)\right)^2 &= \left(\frac{1}{2m}\left(\frac{(Y_i-\mu_i(\beta))^2-\varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)(2p_{ij}-1)p_{ij}(1-p_{ij}) - \frac{1}{m}\left(\frac{Y_i-\mu_i(\beta)}{\varsigma_i(\beta)^2}\right)p_{ij}(1-p_{ij})\right)^2 \\ &\leq \frac{1}{4m^2}\left(\frac{(Y_i-\mu_i(\beta))^2-\varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)^2((2p_{ij}-1)p_{ij}(1-p_{ij}))^2 + \frac{1}{m^2}\left(\frac{Y_i-\mu_i(\beta)}{\varsigma_i(\beta)^2}\right)^2(p_{ij}(1-p_{ij}))^2\end{aligned}$$

Now $((2p_{ij}-1)p_{ij}(1-p_{ij}))^2 \leq \frac{1}{108}$ and $(p_{ij}(1-p_{ij}))^2 \leq \frac{1}{16}$. All we need to do is bound the coefficients:

$$\left(\frac{(Y_i-\mu_i(\beta))^2-\varsigma_i(\beta)^2}{\varsigma_i(\beta)^4}\right)^2 \leq \left(\frac{(Y_i-\mu_i(\beta))^2}{\varsigma_i(\beta)^4}\right)^2 + \frac{1}{\varsigma_i(\beta)^4}$$

Now we can separately maximize the numerator and minimize the denominator. We can upper bound the numerator since $0 \leq Y_i \leq m$ so $|Y_i - \mu_i(\beta)| \leq m$. Further assuming the covariates are bounded such that $||\boldsymbol{x}_{ij}||_2 \leq C$ and $||\beta||_2 \leq B$ we can lower bound $\varsigma_i(\beta)^2 \geq n\sigma(BC)(1-\sigma(BC))$. We let $K = \frac{1}{\sigma(BC)(1-\sigma(BC))}$. Thus we have the bound $\left(\frac{\partial}{\partial z_j}\ell_i(z)\right)^2 \leq \frac{31K^2m^2+4}{1728m^4}$ which lets us bound the norm of the gradient:

$$||\nabla_z\ell_i(z)||_2 \leq \sqrt{\frac{31K^2m^2+4}{1728m^3}}$$

■

Lemma 11 *Under the Poisson-Binomial model we can upper bound the Rademacher compexity of the Poisson-Binomial logistic log-likelihood and the approximate log-likelihood:*

$$\mathcal{R}(\ell_{PoiBin} \odot \mathcal{H}) \leq \frac{2B}{\sqrt{n}}\sqrt{\frac{2}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}||\boldsymbol{x}_{ij}||_2^2\right)}$$

$$\mathcal{R}(\ell \odot \mathcal{H}) \leq \frac{B}{\sqrt{n}}\sqrt{\frac{2}{n}\left(\frac{31K^2}{1728} + \frac{1}{432m^2}\right)\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}||\boldsymbol{x}_{ij}||_2^2\right)}$$

Proof *We will prove the first bound first. Our proof will make use of a vector contraction inequality for Rademacher complexities [15, Theorem 3] to peel off the L-Lipschitz loss function, incurring $\sqrt{2}L$ and summing over all coordinates $j = 1, \ldots, m$:*

$$\mathcal{R}(\ell_{PoiBin} \odot \mathcal{H}) \leq \left(\frac{2}{n}\sqrt{\frac{2}{m}}\right)\mathbb{E}_\epsilon \sup_{||\beta||\leq B}\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{ij}\beta^\top\boldsymbol{x}_{ij}$$

So now we need to bound the linear transformation:

$$\mathbb{E}_\epsilon\left[\sup_{\beta\in B}\sum_{ij}\epsilon_{ij}\beta^\top \boldsymbol{x}_{ij}\right] = \mathbb{E}_\epsilon\left[\sup_{\beta\in B}\beta^\top\left(\sum_{ij}\epsilon_{ij}\boldsymbol{x}_{ij}\right)\right] \leq B\mathbb{E}_\epsilon\left|\left|\sum_{ij}\epsilon_{ij}\boldsymbol{x}_{ij}\right|\right|_2 \leq B\sqrt{\mathbb{E}_\epsilon\left|\left|\sum_{ij}\epsilon_{ij}\boldsymbol{x}_{ij}\right|\right|_2^2} \leq B\sqrt{\sum_{ij}||\boldsymbol{x}_{ij}||_2^2}$$

Combining this with the contraction inequality we have the desired bound on the Rademacher complexity. The proof for the approximate likelihood is identical with the appropriate Lipschitz constant. ■

Note that these bounds are effectively free of m, which is what we should expect. Changing the number of people in a precinct should not materially change the complexity of the learning problem since we have fixed our functionclass to the linear model. It is worth noting that although it may seem in the above proof that we could replace β with a separate β_j for each coordinate of j without incurring a complexity cost this is actually not the case. This would require $||\beta_j||_F \leq B$ and as $m \to \infty$ we would drive $||\beta_j||_2 \to 0$ very quickly. Using the same β for every voter ensures the magnitude of β can be free of m.

Now assuming the covariates are bounded $||\boldsymbol{x}_{ij}||_2^2 \leq C$ we have that ℓ_{PoiBin} and ℓ are bounded by some constants K_{PoiBin}, K, so we can convert these Rademacher complexity bounds into the desired Uniform Strong Laws of Large Numbers by [27, Theorem 4.10], for any $\delta > 0$, $n \in \mathbb{N}$:

$$P\left(\sup_{\beta\in B}\left|\sum_{i=1}^n \ell_{\text{PoiBin}}(\beta) - \mathbb{E}_{\mathcal{P}}\left[\ell_{\text{PoiBin}}(\beta)\right]\right| \leq 2\mathcal{R}_n(\ell_{\text{PoiBin}} \odot \mathcal{H}) + \delta\right) \geq 1 - \exp\left(-\frac{n\delta^2}{2K_{\text{PoiBin}}^2}\right)$$

$$P\left(\sup_{\beta\in B}\left|\sum_{i=1}^n \ell(\beta) - \mathbb{E}_{\mathcal{P}}\left[\ell(\beta)\right]\right| \leq 2\mathcal{R}_n(\ell \odot \mathcal{H}) + \delta\right) \geq 1 - \exp\left(-\frac{n\delta^2}{2K^2}\right)$$

To show consistency by [23, Theorem 5.7] we require two conditions: first the a uniform law of large numbers and second the condition that the maximizer of the population problems $\mathbb{E}[\ell_{\text{PoiBin}}(\beta)]$ and $\mathbb{E}[\ell(\beta)]$ are well-separated.

We proved a USLLN via our Rademacher complexity arguments above. We have also assumed compactness, so the well-separatedness assumption reduces to the condition that the population problems are globally maximized at the unique points. Under Assumptions 3 and by [23, Lemma 5.35] this implies β^* is the unique maximizer of the population log-likelihood in the well specified case. In the misspecified case we will have a similar condition by a standard KL argument assuming our parameter space is convex. The approximate log-likelihood is more straightforward since we constructed our sets $\mathcal{B}_i$ in Assumption 4 directly based on their being global maximizers. For the misspecified case we will simply assume the population minimizer is unique. We will comment on this further below. Under this uniqueness condition, regardless of misspecification, we have that:

$$\hat{\beta}_{n,\text{PoiBin}} \xrightarrow{n\to\infty} \beta^*_{\text{PoiBin}} \quad \text{and} \quad \hat{\beta}_{n,\text{Normal}} \xrightarrow{n\to\infty} \beta^*$$

Now we will argue we have asymptotic normality by [23, Theorem 5.23]. We need a number of conditions: (i) measurability in $(Y_i, \{\boldsymbol{x}_{ij}\}_{j=1}^m)$ and differentiability in β, (ii) local Lipschitzness in a neighborhood of the population optimum, (iii) the estimates $\hat{\beta}_n$ are the maximizers of the sample

log-likelihoods, (iv) the consistency results above, and (v) a second-order Taylor expansion about the population optimum with non-singular Hessian.

Both measureability and differentiability are clear for both likelihoods. We have global Lipschitzness, proved for both likelihoods in Lemmas 9 and 10. Our $\hat{\beta}_n$ are indeed maximizers of the sample log-likelihoods and we assume the same necessary conditions for the above consistency results. Since the output space is simply $[1, \ldots, m]$ the expected log-likelihood, the expected gradient, and the expected Hessian all exist and are bounded at the population optimum. The same is true for the expected approximate log-likelihood and its gradient and Hessian. These conditions imply:

$$\sqrt{n}(\hat{\beta}_{n,\text{PoiBin}} - \beta^*_{\text{PoiBin}}) \xrightarrow{n\to\infty} \mathcal{N}(0, \mathcal{I}_{\text{PoiBin}}(\beta^*_{\text{PoiBin}}))$$
$$\sqrt{n}(\hat{\beta}_{n,\text{Normal}} - \beta^*) \xrightarrow{n\to\infty} \mathcal{N}(0, \mathcal{I}(\beta^*))$$

This completes our proof of consistency and asymptotic normality. These results rely crucially on our assumption that the expected log-likelihoods are unique maximizers. For the full log-likelihood this is a relatively mild condition since identifiability essentially relies on an absence of symmetry across precincts which should essentially always hold. For the approximate log-likelihood in the well specified case this is also a relatively mild issue. In the misspecified case since the same β^* should not minimize every precinct, but only minimize them on average, we might worry this is a strong condition, but in practice it is not.

B.3. Optimization

Although we can usually use Newton's method on its own for optimizing the binary likelihood, when the Hessian is close to singular or has negative eigenvalues this leads to catastrophic failure. These issues are much more acute in the categorical case where it is much less common to have a convex neighborhood that covers both zero and $\hat{\beta}$. To address this problem we develop a variant of Newton's Method where, when the Hessian is not positive definite, we run a line search over minimum eigenvalues projecting onto the set of matrices with that minimum eigenvalue. We give a full description of this procedure in Algorithm 1.

Algorithm 1 This is a variant of Newton's Method where we use a spectral projection with minimum eigenvalues of ϵ to ensure the estimated Hessian is positive definite. We run a line search over ϵ to ensure convergence. Note that if the Hessian is positive definite the clipping has no effect.

Require: $f(\boldsymbol{x})$ (objective function), $\nabla f(\boldsymbol{x})$ (gradient), $\nabla^2 f(\boldsymbol{x})$ (Hessian), $\boldsymbol{x}_0$ (initial point), α (step size), ϵ_{min} (initial and minimum clipping threshold), ϵ_{max} (maximum clipping threshold), $\gamma > 1$ (clipping increase factor), c (sufficient decrease parameter), tol (tolerance), maxiter (maximum iterations)

function SPECTRALPROJLINESEARCH($f, \boldsymbol{x}, \boldsymbol{g}, \boldsymbol{H}, \epsilon_{\text{min}}, \epsilon_{\text{max}}, \gamma, c$)
  $\epsilon \leftarrow \epsilon_{\text{min}}$
  $\boldsymbol{\lambda}, \boldsymbol{v} \leftarrow \text{eigh}(\boldsymbol{H})$ ▷ Compute eigendecomposition of the Hessian
  while $\epsilon \leqslant \epsilon_{\text{max}}$ **do**
    $\boldsymbol{\lambda}_\epsilon \leftarrow \max(\boldsymbol{\lambda}, \epsilon)$ ▷ Clip minimum eigenvalues to ϵ
    $\boldsymbol{H}_\epsilon^{-1} \leftarrow \boldsymbol{v}\ \text{Diag}(\boldsymbol{\lambda}_\epsilon^{-1})\ \boldsymbol{v}^\top$
    $\boldsymbol{p}_{\alpha,\epsilon} \leftarrow -\alpha \boldsymbol{H}_\epsilon^{-1} \boldsymbol{g}$
    if $f(\boldsymbol{x} + \boldsymbol{p}_{\alpha,\epsilon}) \leq f(\boldsymbol{x}) + c \nabla f(\boldsymbol{x})^\top \boldsymbol{p}_{\alpha,\epsilon}$ **then** ▷ Armijo Condition for sufficient decrease
      break
    end if
    $\epsilon \leftarrow \gamma \epsilon$
  end while
  return $\boldsymbol{p}_{\alpha,\epsilon}$
end function
function SPECTRALPROJNEWTONRAPHSON($f, \nabla f, \nabla^2 f, \boldsymbol{x}_0, \alpha, \epsilon_{\text{min}}, \epsilon_{\text{max}}, \gamma, c, \text{tol}, \text{maxiter}$)
  $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
  for $k = 1$ to maxiter **do**
    $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})$
    $\boldsymbol{H} \leftarrow \nabla^2 f(\boldsymbol{x})$
    $\boldsymbol{p}_{\alpha,\epsilon} \leftarrow$ SPECTRALLINESEARCH($f, \boldsymbol{x}, \boldsymbol{g}, \boldsymbol{H}, \alpha, \epsilon_{\text{min}}, \epsilon_{\text{max}}, \gamma, c$)
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{p}_{\alpha,\epsilon}$
    if $\|\boldsymbol{g}\| < \text{tol}$ **then**
      break
    end if
  end for
  return $\boldsymbol{x}$
end function
return SPECTRALPROJECTEDNEWTONRAPHSON($\ell, \nabla\ell, \nabla^2\ell, \boldsymbol{x}_0 = \boldsymbol{0}, \alpha = 1, \epsilon_{\text{min}} = 10^{-4}, \epsilon_{\text{max}} = 10^{-1}, \gamma = 1.5, c = 10^{-4}, \text{tol} = 10^{-5}, \text{maxiter} = 500$)

B.4. Extension to categorical outcomes

The core of the generalization is to substitute the logit link function for the softmax link function and to work with $\beta = (\beta^{(1)}, \ldots, \beta^{(K)})$ where for identifiability we assume that $\beta^{(K)} = 0$ so that $\beta \in \mathbb{R}^{d \times (K-1)}$.

With these details clear we know that we have a Poisson-Multinomial distribution:

$$\ell_{\text{PoiMult}}(\beta) = \sum_{i=1}^{n} \log \left(\sum_{(A_1 \ldots, A_k) \in \mathbb{P}_{\left(Y_i^{(1)}, \ldots, Y_i^{(K)}\right)}([m_i])} \left(\prod_{k=1}^{K} \prod_{j \in A_k} \frac{\exp(x_{ij}^\top \beta^{(k)})}{\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})} \right) \right) \tag{5}$$

Where we write $\mathbb{P}_{\left(Y_i^{(1)}, \ldots, Y_i^{(K)}\right)}([m_i])$ to denote the set of all possible partitions of m_i into K sets of sizes $(Y_i^{(1)}, \ldots, Y_i^{(K)})$ respectively. This is even more intractable than the binary likelihood.

The natural central limit theorem here would give us:

$$Y_i \sim \mathcal{N}\left(\begin{bmatrix} \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(1)})}{\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})} \\ \vdots \\ \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(K)})}{\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})} \end{bmatrix}, \begin{bmatrix} \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(1)}) \left(\sum_{k \neq 1} \exp(x_{ij}^\top \beta^{(k)})\right)}{\left(\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})\right)^2} & \cdots & \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(1)}) \exp(x_{ij}^\top \beta^{(K)})}{\left(\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})\right)^2} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(1)}) \exp(x_{ij}^\top \beta^{(K)})}{\left(\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})\right)^2} & \cdots & \sum_{j=1}^{m_i} \frac{\exp(x_{ij}^\top \beta^{(K)}) \left(\sum_{k \neq K} \exp(x_{ij}^\top \beta^{(k)})\right)}{\left(\sum_{k'=1}^{K} \exp(x_{ij}^\top \beta^{(k')})\right)^2} \end{bmatrix} \right)$$

But the covariance here is always singular. We can "fix" this problem by dropping one of the coordinates, but this solution does not work at $\beta = 0$ where the covariance matrix will always be zero. This causes irreparable issues in the loss landscape. So instead of using this covariance we just use its diagonal. This is exactly equivalent to summing the binary log-likelihood on each component of $Y^{(k)}$:

$$\ell_{\text{cat}}(\beta) = \left(\frac{1}{n \cdot \bar{m}} \right) \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \log \left(\varsigma_i \left(\beta^{(k)} \right)^2 \right) + \frac{\left(Y_i^{(k)} - \mu_i \left(\beta^{(k)} \right) \right)^2}{\varsigma_i \left(\beta^{(k)} \right)^2} \right) \tag{6}$$

This loss function remains easy to compute and inherits the nice geometry of the binary case we analyze elsewhere. This also inherits the bounds in App. B.2.4 incurring a multiplicative factor in K, again giving us a uniform strong law of large numbers for the approximate log-likelihood.