A Fast Ecological Inference Algorithm for the R×C Case

Pablo Ubilla Pavez, Charles Thraves
Instituto de Sistemas Complejos de Ingeniería, Universidad de Chile

ISCI INSTITUTO SISTEMAS COMPLEJOS DE INGENIERÍA

INGENIERÍA INDUSTRIAL UNIVERSIDAD DE CHILE

Motivation

Ecological Inference: estimate individual behavior using aggregate data.

Demographic Group	Candidate 1	⋯	Candidate C
1	?		w_{b1}
⋮			⋮
G			w_{bG}
	x_{b1}	⋯	x_{bC}

Age Group	Candidate	*Ana*	*Bob*	
≤ 40		?	?	100
> 40		?	?	55
		65	90	

x_{bc}: Votes of candidate c in ballot-box b.
w_{bg}: Voters of demographic group g in ballot-box b.

Goal: Estimate the voting probability p_{gc} for each group g and candidate c

EM-Algorithm

To solve the ecological inference problem we can use the EM-Algorithm.

- **E-step:**

$$q_{bgc} := P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p}) = \frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c} = 1; \boldsymbol{p}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c'} = 1; \boldsymbol{p}) \cdot p_{gc'}} \quad (1)$$

Terms in red are the PMF of the sum of G Multinomial$(w_{bg}, p_{g1}, \ldots, p_{gC})$, **hard** to compute.

- **M-step:** Optimizing the expected log-likelihood has a closed-form solution:

$$p_{gc} = \frac{\sum_{b \in \mathcal{B}} w_{bg} q_{bgc}}{\sum_{b \in \mathcal{B}} w_{bg}} \qquad \forall g \in \mathcal{G}, \forall c \in \mathcal{C}. \quad (2)$$

The **E-step** and **M-step** iterate until convergence.

Approximation methods for q

H&R: Sample S integral points from a polytope
MVN CDF: Integrate the unit-hypercube with a Multivariate Normal approx.
MVN PDF: Directly computes the ratio in Eq. (1) with PDF of the MVN approx.
MULT: Approximate to a single Multinomial

Instance		Run Time [seconds]					Mean Absolute Error				
C	G	EXACT	H&R $S = 10^3$	MVN CDF	MVN PDF	MULT	EXACT	H&R $S = 10^3$	MVN CDF	MVN PDF	MULT
2	2	5.3	123.3	0.13	0.02	0.0014	0.012	0.012	0.013	0.013	0.011
	3	5.5	70.2	0.22	0.02	0.0011	0.015	0.016	0.016	0.016	0.015
	4	6.4	72.3	0.32	0.03	0.0012	0.018	0.020	0.019	0.019	0.018
3	2	132.1	80.8	0.53	0.02	0.0009	0.011	0.011	0.013	0.011	0.010
	3	354.2	73.8	0.93	0.03	0.0011	0.014	0.016	0.016	0.015	0.014
	4	466.0	76.3	1.29	0.03	0.0013	0.017	0.026	0.020	0.018	0.018
4	2	6,168.8	74.1	1.06	0.03	0.0009	0.011	0.011	0.013	0.012	0.011
	3	13,589.1	77.7	1.75	0.05	0.0011	0.012	0.018	0.015	0.013	0.012
	4	22,599.2	65.0	2.29	0.04	0.0013	0.015	0.032	0.019	0.017	0.015

Table 1: Average run time in seconds and MAE for 20 instances

Case Study

MULT was applied to the **First Round Chilean Presidential Election of 2021**

- $C = 8$: 7 candidates plus the Null/Blank option.
- $G = 8$: Age ranges: 18-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+
- The EM-Algorithm is run in each of the **681 districts**.

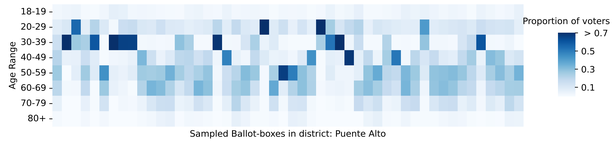


Figure 1: Distribution of demographic groups in a sample of ballot boxes from the Puente Alto district

Trade-off: Fit vs Error

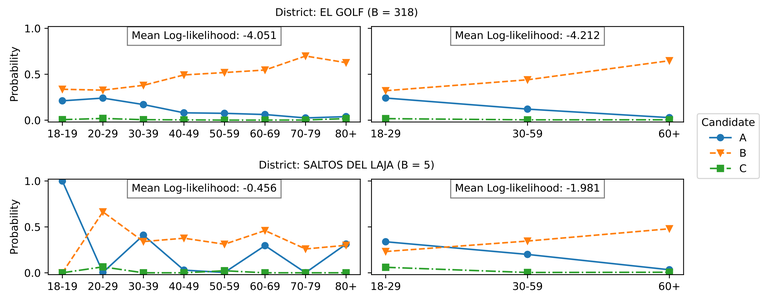


Figure 2: Estimated voting probabilities for groups (**left**) and macro-groups (**right**)

Optimal Group Aggregation

The following optimization problem can be formulated to choose the best group aggregation, where a macro-group m is a feasible group aggregation.

$$\max_{\boldsymbol{\alpha}} \sum_{m \in \mathcal{M}} \alpha_m \cdot \sigma_m \qquad \text{Maximize sum of proportional std} \quad (3a)$$

$$\text{subject to :} \sum_{m \in \mathcal{M} | g \subseteq m} \alpha_m = 1 \qquad \text{Each macro-group } m \text{ appears only once} \quad (3b)$$

$$\sum_{m \in \mathcal{M}} \alpha_m = A \qquad A \text{ macro-groups are selected} \quad (3c)$$

$$\alpha_m \in \{0, 1\} \qquad 1\text{: if macro-group } m \text{ is selected; } 0\text{: if not} \quad (3d)$$

Select group aggregation

for $A = G$ to 1 **do**
  Get $\{\alpha_m\}_{m \in \mathcal{M}}$ from Optimization Problem (3)
  $\mathcal{A} = \{m \in \mathcal{M} | \alpha_m = 1\}$
  $w'_{bm} = \sum_{g \subseteq m} w_{bg}, \forall m \in \mathcal{A}$
  Get $\{\text{std}(p_{mc})\}_{c \in \mathcal{C}, m \in \mathcal{A}}$ from bootstrapping
  if $\text{std}(p_{mc}) \leq 5\%, \forall m \in \mathcal{A}, \forall c \in \mathcal{C}$ **then**
    Stop
  end if
end for
Output: Group aggregation $\mathcal{A}$

Results

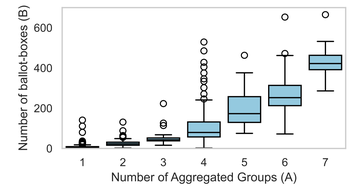


Figure 3: Number of ballot-boxes and number of aggregated groups

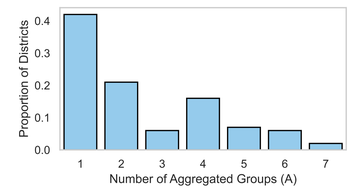


Figure 4: Proportion of districts for number of aggregated groups

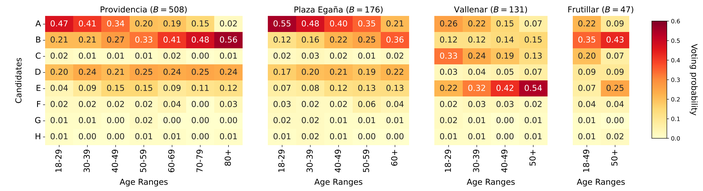


Figure 5: Probability estimation for 4 districts

Detecting Outliers Ballot-boxes

At each ballot-box b, compute the **p-value** as the probability of obtaining an outcome as or more unlikely than the observed outcome ($\boldsymbol{x}_b$). Namely:

- **Exact method:**

$$\text{p-val}(\boldsymbol{x}_b) = \sum_{\boldsymbol{x}'_b \in R_{\boldsymbol{X}_b} | P(\boldsymbol{x}'_b) \leq P(\boldsymbol{x}_b)} P(\boldsymbol{x}'_b)$$

- **Simulation method:**

$$\text{p-val}(\boldsymbol{x}_b) \approx \frac{1}{H} \sum_{h=1}^{H} \mathbb{1}_{\{P(\boldsymbol{x}_b^{(h)}) \leq P(\boldsymbol{x}_b)\}}$$

	Ballot boxes	
p-val Range	Number	Percentage
$\leq 10^0$	46,605	100.00
$\leq 10^{-1}$	9,174	19.68
$\leq 10^{-2}$	1,801	3.86
$\leq 10^{-3}$	567	1.22
$\leq 10^{-4}$	219	0.47
$\leq 10^{-5}$	131	0.28
$\leq 10^{-6}$	72	0.15
$\leq 10^{-7}$	45	0.10

Table 2: Count for p-value ranges

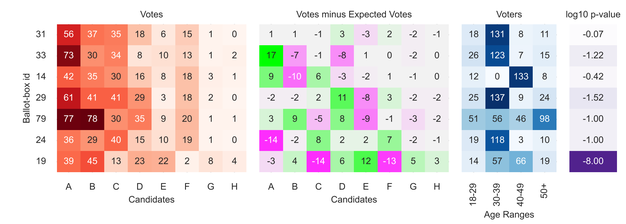


Figure 6: Sample of ballot-boxes showing candidates' votes, vote differences compared to expected results, voters' demographic groups, and p-values.

Conclusions

- MULT is the **fastest method** and has a similar error to the EXACT method
- Importance of doing **group aggregation** when having few ballot-boxes
- **Outlier ballot-boxes** can be identified to assess election transparency

LaTeX TikZposter