A Fast Ecological Inference Algorithm for the R×C case

Annual Conference OR66

Pablo Ubilla Pavez, Charles Thraves



September 12, 2024

Motivation

	Candidate	
Ballot-Box	A	B
1	58	42
2	47	53
3	59	41
4	21	79
5	52	48

Motivation

	Candidate	
Ballot-Box	A	B
1	58	42
2	47	53
3	59	41
4	**21**	**79**
5	52	48

Motivation

Ballot-Box	Candidate A	B
1	58	42
2	47	53
3	59	41
4	**21**	**79**
5	52	48

We could have demographic groups

Ballot-Box	Voters 18-39	40+
1	40	60
2	39	61
3	42	58
4	**11**	**89**
5	45	55

Motivation

We could have demographic groups

Ballot-Box	Candidate A	B
1	58	42
2	47	53
3	59	41
4	**21**	**79**
5	52	48

Ballot-Box	Voters 18-39	40+
1	40	60
2	39	61
3	42	58
4	**11**	**89**
5	45	55

We could have more candidates

Election	# of Candidates
(2R) Presidential 2021	2
(1R) Presidential 2021	7
Const. Council 2023	~ 20

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Related Literature

Ecological Inference

- **Survey:** King, Tanner, and Rosen 2004, W. T. Cho and Manski 2008
- **Parametric models:** King 1997, King, Rosen, and Tanner 1999, Rosen et al. 2001, James Greiner and Quinn 2009, Glynn and Wakefield 2010
- **Entropy methods:** R. J. Johnston and Hay 1983, R. Johnston and Pattie 2000, Antweiler 2007, El Deeb 2023, and Judge, Miller, and W. K. Cho 2003
- **Integer Programming:** McCarthy and Ryan 1977, Tziafetas 1986, Corominas, Lusa, and Dolors Calvet 2015 and Pavía and Romero 2021

Sampling integer points from a polyhedron

- Mete and Zabinsky 2012; and Baumert et al. 2009

Approximation of sum of Multivariate distributions

- Roos 1999, Roos 2002, Novak 2019, and Lin, Wang, and Hong 2022

p-values for Multivariate Gaussian

- Maatouk and Bay 2016; and Ellis and Maitra 2007

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Setting

Ecological inference is a technique to estimate individual behaviour by using aggregated data

Group	Candidate 1	$\cdots$	C	
1	z_{b11}	$\cdots$	z_{b1C}	w_{b1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	z_{bG1}	$\cdots$	z_{bGC}	w_{bG}
	x_{b1}	$\cdots$	x_{bC}	

We observe the sum of rows and the sum of columns.

We do not observe what is inside the matrix.

Setting

Ecological inference is a technique to estimate individual behaviour by using aggregated data

		Candidate			
		1	$\cdots$	C	
Group	1	z_{b11}	$\cdots$	z_{b1C}	w_{b1}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	G	z_{bG1}	$\cdots$	z_{bGC}	w_{bG}
		x_{b1}	$\cdots$	x_{bC}	

We observe the sum of rows and the sum of columns.

We do not observe what is inside the matrix.

Setting

- $\mathcal{C}$: Set of C candidates.
- $\mathcal{G}$: Set of G groups.
- $\mathcal{B}$: Set of B ballot-boxes.
- $\mathcal{I}_b$: Set of I_b voters from ballot-box b.
- x_{bc}: Number of votes to candidate c in ballot box b
- w_{bg}: Number of voters of group g in ballot box b
- Y_{bic}: RV, 1 if voter i voted for candidate c in ballot box b, 0 o/w
- X_{bc}: RV, votes obtained by candidate c in ballot box b.
- p_{gc}: probability that a voter from group g votes for c
- q_{bgc}:probability that a voter from group g voted for c in ballot-box b

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Setting

- $\mathcal{C}$: Set of C candidates.
- $\mathcal{G}$: Set of G groups.
- $\mathcal{B}$: Set of B ballot-boxes.
- $\mathcal{I}_b$: Set of I_b voters from ballot-box b.
- x_{bc}: Number of votes to candidate c in ballot box b
- w_{bg}: Number of voters of group g in ballot box b
- Y_{bic}: RV, 1 if voter i voted for candidate c in ballot box b, 0 o/w
- X_{bc}: RV, votes obtained by candidate c in ballot box b.
- p_{gc}: probability that a voter from group g votes for c
- q_{bgc}:probability that a voter from group g voted for c in ballot-box b

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Setting

- $\mathcal{C}$: Set of C candidates.
- $\mathcal{G}$: Set of G groups.
- $\mathcal{B}$: Set of B ballot-boxes.
- $\mathcal{I}_b$: Set of I_b voters from ballot-box b.
- x_{bc}: Number of votes to candidate c in ballot box b
- w_{bg}: Number of voters of group g in ballot box b
- Y_{bic}: RV, 1 if voter i voted for candidate c in ballot box b, 0 o/w
- X_{bc}: RV, votes obtained by candidate c in ballot box b.
- p_{gc}: probability that a voter from group g votes for c
- q_{bgc}:probability that a voter from group g voted for c in ballot-box b

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Setting

- $\mathcal{C}$: Set of C candidates.
- $\mathcal{G}$: Set of G groups.
- $\mathcal{B}$: Set of B ballot-boxes.
- $\mathcal{I}_b$: Set of I_b voters from ballot-box b.
- x_{bc}: Number of votes to candidate c in ballot box b
- w_{bg}: Number of voters of group g in ballot box b
- Y_{bic}: RV, 1 if voter i voted for candidate c in ballot box b, 0 o/w
- X_{bc}: RV, votes obtained by candidate c in ballot box b.
- p_{gc}: probability that a voter from group g votes for c
- q_{bgc}:probability that a voter from group g voted for c in ballot-box b

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Expectation-Maximization Algorithm

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Use the Expectation-Maximization Algorithm (EM Algorithm)

1. **E-Step**: Find the distribution of Y_{bic} given the observed data ($\boldsymbol{X}_b = \boldsymbol{x}_b$) and a prior probability $\boldsymbol{p} = \boldsymbol{p}^{(old)}$

$$q_{bg(i)c} := P(Y_{bic} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p})$$

2. **M-Step**: Maximize the expected log-likelihood using the distribution of Y_{bgc} from the **E-Step**

$$\max_{\boldsymbol{p}} \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Iterate until convergence

Expectation-Maximization Algorithm

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Use the Expectation-Maximization Algorithm (EM Algorithm)

1. **E-Step**: Find the distribution of Y_{bic} given the observed data $(\boldsymbol{X}_b = \boldsymbol{x}_b)$ and a prior probability $\boldsymbol{p} = \boldsymbol{p}^{(old)}$

$$q_{bg(i)c} := P(Y_{bic} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p})$$

2. **M-Step**: Maximize the expected log-likelihood using the distribution of Y_{bgc} from the **E-Step**

$$\max_{\boldsymbol{p}} \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Iterate until convergence

Expectation-Maximization Algorithm

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Use the Expectation-Maximization Algorithm (EM Algorithm)

1. **E-Step**: Find the distribution of Y_{bic} given the observed data ($\boldsymbol{X}_b = \boldsymbol{x}_b$) and a prior probability $\boldsymbol{p} = \boldsymbol{p}^{(old)}$

$$q_{bg(i)c} := P(Y_{bic} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p})$$

2. **M-Step**: Maximize the expected log-likelihood using the distribution of Y_{bgc} from the **E-Step**

$$\max_{\boldsymbol{p}} \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Iterate until convergence

Expectation-Maximization Algorithm

$$l(\boldsymbol{p}) = \sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c})$$

Use the Expectation-Maximization Algorithm (EM Algorithm)

1. **E-Step**: Find the distribution of Y_{bic} given the observed data ($\boldsymbol{X}_b = \boldsymbol{x}_b$) and a prior probability $\boldsymbol{p} = \boldsymbol{p}^{(old)}$

$$q_{bg(i)c} := P(Y_{bic} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p})$$

2. **M-Step**: Maximize the expected log-likelihood using the distribution of Y_{bgc} from the **E-Step**

$$\max_{\boldsymbol{p}} \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B}, i \in \mathcal{I}_b, c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Iterate until convergence

M-Step

$$\max_{\mathbf{p}} \; \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B} i \in \mathcal{I}_b c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Equivalently

$$\max_{\mathbf{p}} \quad \sum_{b \in \mathcal{B}, g \in \mathcal{G}} w_{bg} \sum_{c \in \mathcal{C}} q_{bgc} \ln(p_{gc})$$

$$\text{s.t.} \quad \sum_{c \in \mathcal{C}} p_{gc} = 1 \qquad \forall g$$

$$p_{gc} \geq 0 \qquad \forall g, c$$

Obtaining

$$p_{gc} = \frac{\sum_{b \in \mathcal{B}} w_{bg} q_{bgc}}{\sum_{b \in \mathcal{B}} w_{bg}}$$

M-Step

$$\max_{\mathbf{p}} \; \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B} i \in \mathcal{I}_b c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Equivalently

$$\max_{\mathbf{p}} \quad \sum_{b \in \mathcal{B}, g \in \mathcal{G}} w_{bg} \sum_{c \in \mathcal{C}} q_{bgc} \ln(p_{gc})$$

$$\text{s.t.} \quad \sum_{c \in \mathcal{C}} p_{gc} = 1 \qquad \forall g$$

$$p_{gc} \geq 0 \qquad \forall g, c$$

Obtaining

$$p_{gc} = \frac{\sum_{b \in \mathcal{B}} w_{bg} q_{bgc}}{\sum_{b \in \mathcal{B}} w_{bg}}$$

M-Step

$$\max_{\mathbf{p}} \; \mathbb{E}_{Y_{bic} \sim q_{bg(i)c}} \left[\sum_{b \in \mathcal{B} i \in \mathcal{I}_b c \in \mathcal{C}} Y_{bic} \cdot \ln(p_{g(i)c}) \right]$$

Equivalently

$$\max_{\mathbf{p}} \quad \sum_{b \in \mathcal{B}, g \in \mathcal{G}} w_{bg} \sum_{c \in \mathcal{C}} q_{bgc} \ln(p_{gc})$$

$$\text{s.t.} \quad \sum_{c \in \mathcal{C}} p_{gc} = 1 \qquad \forall g$$

$$p_{gc} \geq 0 \qquad \forall g, c$$

Obtaining

$$p_{gc} = \frac{\sum_{b \in \mathcal{B}} w_{bg} q_{bgc}}{\sum_{b \in \mathcal{B}} w_{bg}}$$

E-Step

$$
\begin{aligned}
q_{bgc} &:= P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p}) \\
&= \frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c} = 1) \cdot p_{gc}}{P(\boldsymbol{X}_b = \boldsymbol{x}_b)}
\end{aligned}
$$

E-Step

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

$\mathcal{H}_{bf}$: Set of possible voting outcomes for group f

$T_{bf}(\boldsymbol{k})$: Probability of aggregated result $\boldsymbol{k} \in \mathbb{N}^C$ until group f

$T_{bf}(\boldsymbol{k}) = \sum_{\boldsymbol{h} \in \mathcal{H}_{bf} : \boldsymbol{h} \leq \boldsymbol{k}} T_{b,f-1}(\boldsymbol{k} - \boldsymbol{h}) \cdot \binom{w_{bf}}{h_1, \ldots, h_C} \prod_{c \in \mathcal{C}} p_{fc}^{h_c}$

$T_{b0}(\boldsymbol{k}) = \mathbf{1}\{\boldsymbol{k} = 0\}$

$P(\boldsymbol{X}_b = \boldsymbol{x}_b) = T_{bG}(\boldsymbol{x}_b)$

Needs $O(B \cdot G \cdot I_b^{2^C})$ computations.

E-Step

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

$U_{bfgc}(\boldsymbol{k}) \in \mathbb{N}^C$: Probability of aggregated result $\boldsymbol{k}$ until group f if one vote of group g already went to c

$$P(\boldsymbol{X}_b = \boldsymbol{x}_b) | Y_{bi(g)c} = 1) = U_{bGgc}(\boldsymbol{x}_b)$$

Needs $O(B \cdot G^2 \cdot C \cdot I_b^{2C})$ computations.

E-Step

$$
\begin{aligned}
q_{bgc} &:= P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p}) \\
&= \frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c} = 1) \cdot p_{gc}}{P(\boldsymbol{X}_b = \boldsymbol{x}_b)} \\
&= \frac{U_{bGgc}(\boldsymbol{x}_b) \cdot p_{gc}}{T_{bg}(\boldsymbol{x}_b)}
\end{aligned}
$$

We would like faster ways to compute q_{bgc}

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Simulation: Hit and Run

	1	$\cdots$	C	
1	z_{b11}	$\cdots$	z_{b1C}	w_{b1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	z_{bG1}	$\cdots$	z_{bGC}	w_{bG}
	x_{b1}	$\cdots$	x_{bC}	

$z_{bgc} \geq 0,\ z_{bgc} \in \mathbb{Z},\ \forall g, \forall c$

$z_b \in \Omega_b$, with Ω_b a polyhedron

Algorithm 1 Hit and Run

Input: w_b, x_b, s, S

Choose starting z_b, $\mathcal{S}_b = \phi$

for $i = 1$ to S **do**

  for $j = 1$ to s **do**

    Sample g, g' and c, c'

    if $z_{bgc} > 0$ and $z_{bg'c'} > 0$ **then**

      $z_{bgc} \leftarrow z_{bgc} - 1$

      $z_{bg'c'} \leftarrow z_{bg'c'} - 1$

      $z_{bgc'} \leftarrow z_{bgc'} + 1$

      $z_{bg'c} \leftarrow z_{bg'c} + 1$

  $\mathcal{S}_b = \mathcal{S}_b \cup \{z_b\}$

Output: $\mathcal{S}_b$

Simulation: Hit and Run

	1	$\cdots$	C	
1	z_{b11}	$\cdots$	z_{b1C}	w_{b1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	z_{bG1}	$\cdots$	z_{bGC}	w_{bG}
	x_{b1}	$\cdots$	x_{bC}	

$$z_{bgc} \geq 0,\ z_{bgc} \in \mathbb{Z},\ \forall g, \forall c$$

$\boldsymbol{z}_b \in \Omega_b$, with Ω_b a polyhedron

Algorithm 2 Hit and Run

Input: $\boldsymbol{w}_b, \boldsymbol{x}_b, s, S$
Choose starting $\boldsymbol{z}_b$, $\mathcal{S}_b = \phi$
for $i = 1$ to S **do**
  for $j = 1$ to s **do**
    Sample g, g' and c, c'
    if $z_{bgc} > 0$ and $z_{bg'c'} > 0$ **then**
      $z_{bgc} \leftarrow z_{bgc} - 1$
      $z_{bg'c'} \leftarrow z_{bg'c'} - 1$
      $z_{bgc'} \leftarrow z_{bgc'} + 1$
      $z_{bg'c} \leftarrow z_{bg'c} + 1$
  $\mathcal{S}_b = \mathcal{S}_b \cup \{\boldsymbol{z}_b\}$
Output: $\mathcal{S}_b$

Simulation: Hit and Run

We have a subset $\mathcal{S}_b$ of sampled scenarios. Now we can condition q_{bgc} to this subset.

$$
\begin{aligned}
q_{bgc} &= P(Y_{bi(g)c} = 1 | \mathbf{X_b} = \mathbf{x_b}, \mathcal{S}_b; \boldsymbol{p}) \\
&= P(Y_{bi(g)c} = 1 | \mathcal{S}_b; \boldsymbol{p}) \\
&= \sum_{\boldsymbol{z}_b \in \mathcal{S}_b} P(Y_{bi(g)c} = 1 | \boldsymbol{Z}_b = \boldsymbol{z}_b; \boldsymbol{p}) P(\boldsymbol{Z}_b = \boldsymbol{z}_b | \mathcal{S}_b; \boldsymbol{p}) \\
&= \frac{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \frac{z_{bgc}}{w_{bg}} \cdot \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}
\end{aligned}
$$

The E-step needs $O(B \cdot S \cdot G^2 \cdot C^2)$ computations.

Simulation: Hit and Run

We have a subset $\mathcal{S}_b$ of sampled scenarios. Now we can condition q_{bgc} to this subset.

$$
\begin{aligned}
q_{bgc} &= P(Y_{bi(g)c} = 1 | \mathbf{X_b} = \mathbf{x_b}, \mathcal{S}_b; \boldsymbol{p}) \\
&= P(Y_{bi(g)c} = 1 | \mathcal{S}_b; \boldsymbol{p}) \\
&= \sum_{\boldsymbol{z}_b \in \mathcal{S}_b} P(Y_{bi(g)c} = 1 | \boldsymbol{Z}_b = \boldsymbol{z}_b; \boldsymbol{p}) P(\boldsymbol{Z}_b = \boldsymbol{z}_b | \mathcal{S}_b; \boldsymbol{p}) \\
&= \frac{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \frac{z_{bgc}}{w_{bg}} \cdot \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}
\end{aligned}
$$

The E-step needs $O(B \cdot S \cdot G^2 \cdot C^2)$ computations.

Simulation: Hit and Run

We have a subset $\mathcal{S}_b$ of sampled scenarios. Now we can condition q_{bgc} to this subset.

$$
\begin{aligned}
q_{bgc} &= P(Y_{bi(g)c} = 1 | \mathbf{X_b} = \mathbf{x_b}, \mathcal{S}_b; \boldsymbol{p}) \\
&= P(Y_{bi(g)c} = 1 | \mathcal{S}_b; \boldsymbol{p}) \\
&= \sum_{\boldsymbol{z}_b \in \mathcal{S}_b} P(Y_{bi(g)c} = 1 | \boldsymbol{Z}_b = \boldsymbol{z}_b; \boldsymbol{p}) P(\boldsymbol{Z}_b = \boldsymbol{z}_b | \mathcal{S}_b; \boldsymbol{p}) \\
&= \frac{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \frac{z_{bgc}}{w_{bg}} \cdot \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}{\sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \prod_{g' \in \mathcal{G}} \left(\binom{w_{bg'}}{z_{bg'1}, \ldots, z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}} \right)}
\end{aligned}
$$

The E-step needs $O(B \cdot S \cdot G^2 \cdot C^2)$ computations.

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

MVN CDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \boldsymbol{\mu}_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

$$\mathcal{D}_{bc} = \prod_{c'} [x_{bc} - 0.5 - \mathbb{1}_{\{c'=c\}}, x_{bc} + 0.5 - \mathbb{1}_{\{c'=c\}}]$$

$$q_{bgc} = \frac{F_{bg}(\mathcal{D}_{bc}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} F_{bg}(\mathcal{D}_{bc'}) \cdot p_{gc'}}$$

$O(B \cdot G \cdot C^2 \cdot \epsilon^{-1/2})$ for M.C integration + $O(B \cdot G \cdot C^3)$ for Cholesky Decomposition

MVN CDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 \sim \mathcal{N}(\mu_b^g, \Sigma_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \mu_b^g = \mu_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \Sigma_b^g = \Sigma_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

$$\mathcal{D}_{bc} = \prod_{c'} [x_{bc} - 0.5 - \mathbb{1}_{\{c'=c\}}, x_{bc} + 0.5 - \mathbb{1}_{\{c'=c\}}]$$

$$q_{bgc} = \frac{F_{bg}(\mathcal{D}_{bc}) \cdot p_{gc}}{\sum_{c' \in C} F_{bg}(\mathcal{D}_{bc'}) \cdot p_{gc'}}$$

$O(B \cdot G \cdot C^2 \cdot \epsilon^{-1/2})$ for M.C integration + $O(B \cdot G \cdot C^3)$ for Cholesky Decomposition

MVN CDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \boldsymbol{\mu}_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

$$\mathcal{D}_{bc} = \prod_{c'} [x_{bc} - 0.5 - \mathbb{1}_{\{c'=c\}}, x_{bc} + 0.5 - \mathbb{1}_{\{c'=c\}}]$$

$$q_{bgc} = \frac{F_{bg}(\mathcal{D}_{bc}) \cdot p_{gc}}{\sum_{c' \in C} F_{bg}(\mathcal{D}_{bc'}) \cdot p_{gc'}}$$

$O(B \cdot G \cdot C^2 \cdot \epsilon^{-1/2})$ for M.C integration + $O(B \cdot G \cdot C^3)$ for Cholesky Decomposition

MVN CDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \boldsymbol{\mu}_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

$$\mathcal{D}_{bc} = \prod_{c'} [x_{bc} - 0.5 - \mathbb{1}_{\{c'=c\}}, x_{bc} + 0.5 - \mathbb{1}_{\{c'=c\}}]$$

$$q_{bgc} = \frac{F_{bg}(\mathcal{D}_{bc}) \cdot p_{gc}}{\sum_{c' \in C} F_{bg}(\mathcal{D}_{bc'}) \cdot p_{gc'}}$$

$O(B \cdot G \cdot C^2 \cdot \epsilon^{-1/2})$ for M.C integration + $O(B \cdot G \cdot C^3)$ for Cholesky Decomposition

MVN CDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \boldsymbol{\mu}_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

$$\mathcal{D}_{bc} = \prod_{c'} [x_{bc} - 0.5 - \mathbb{1}_{\{c'=c\}}, x_{bc} + 0.5 - \mathbb{1}_{\{c'=c\}}]$$

$$q_{bgc} = \frac{F_{bg}(\mathcal{D}_{bc}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} F_{bg}(\mathcal{D}_{bc'}) \cdot p_{gc'}}$$

$O(B \cdot G \cdot C^2 \cdot \epsilon^{-1/2})$ for M.C integration + $O(B \cdot G \cdot C^3)$ for Cholesky Decomposition

MVN PDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bi(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \mu_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

We could use the PDF instead:

$$q_{bgc} = \frac{f_{bg}(x_{b1}, ..., x_{bc} - 1, ..., x_{bC}) \cdot p_{gc}}{\sum_{c' \in C} f_{bg}(x_{b1}, ..., x_{bc'} - 1, ..., x_{bC}) \cdot p_{gc'}}$$

It need $O(B \cdot G \cdot C^3)$ computations.

MVN PDF

Remove the last candidate C, we get the following approximations:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \qquad \boldsymbol{X}_b^* | Y_{bi(g)c} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c$$

$$\boldsymbol{\mu}_b = \boldsymbol{p}^{*T} \boldsymbol{w}_b \qquad \boldsymbol{\mu}_b^g = \mu_b - \boldsymbol{p}_g^{*T}$$

$$\boldsymbol{\Sigma}_b = \text{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T} \text{diag}(\boldsymbol{w}_b) \boldsymbol{p}^* \qquad \boldsymbol{\Sigma}_b^g = \boldsymbol{\Sigma}_b - \text{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T} \boldsymbol{p}_g^*$$

We could use the PDF instead:

$$q_{bgc} = \frac{f_{bg}(x_{b1}, ..., x_{bc} - 1, ..., x_{bC}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} f_{bg}(x_{b1}, ..., x_{bc'} - 1, ..., x_{bC}) \cdot p_{gc'}}$$

It need $O(B \cdot G \cdot C^3)$ computations.

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Multinomial Approximation

Let's consider a single multinomial with weighted probabilities.

$$\boldsymbol{X}_b \sim \text{Mult}(I_b, \boldsymbol{r}_b) \qquad \boldsymbol{X}_b | Y_{bi(g)c} = 1 \sim \text{Mult}(I_b - 1, \boldsymbol{r}_b^{(g)}) + \boldsymbol{e}_c$$

$$r_{bc} = \frac{\sum_g w_{bg} p_{gc}}{I_b} \qquad r_{bc}^{(g)} = \frac{\sum_{g'} w_{bg'} p_{g'c} - p_{gc}}{I_b - 1}$$

$$q_{bgc} = \frac{x_{bc} \cdot p_{gc} \cdot \left(r_{bc}^{(g)}\right)^{-1}}{\sum_{c'} x_{bc'} \cdot p_{gc'} \cdot \left(r_{bc'}^{(g)}\right)^{-1}}$$

Needs $O(B \cdot G \cdot C)$ computations

Multinomial Approximation

Let's consider a single multinomial with weighted probabilities.

$$\boldsymbol{X}_b \sim \text{Mult}(I_b, \boldsymbol{r}_b) \qquad \boldsymbol{X}_b | Y_{bi(g)c} = 1 \sim \text{Mult}(I_b - 1, \boldsymbol{r}_b^{(g)}) + \boldsymbol{e}_c$$

$$r_{bc} = \frac{\sum_g w_{bg} p_{gc}}{I_b} \qquad r_{bc}^{(g)} = \frac{\sum_{g'} w_{bg'} p_{g'c} - p_{gc}}{I_b - 1}$$

$$q_{bgc} = \frac{x_{bc} \cdot p_{gc} \cdot \left(r_{bc}^{(g)}\right)^{-1}}{\sum_{c'} x_{bc'} \cdot p_{gc'} \cdot \left(r_{bc'}^{(g)}\right)^{-1}}$$

Needs $O(B \cdot G \cdot C)$ computations

E-Step Complexity

In summary:

Method	Complexity $O(\cdot)$
EXACT	$B \cdot G^2 \cdot C \cdot I_b^{2C}$
H&R	$B \cdot G \cdot C \cdot S \cdot s$
MVN CDF	$B \cdot G \cdot C^2 \cdot \epsilon^{-1/2} + B \cdot G \cdot C^3$
MVN PDF	$B \cdot G \cdot C^3$
MULT	$B \cdot G \cdot C$

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Time

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	2.8	5.8	38.2	0.13	0.02	0.0009
	3	4.7	5.9	37.4	0.22	0.02	0.0011
	4	6.4	5.8	34.8	0.32	0.03	0.0012
3	2	122.5	5.8	37.6	0.53	0.02	0.0009
	3	314.0	5.8	36.0	0.93	0.03	0.0011
	4	466.0	5.5	32.1	1.29	0.03	0.0013
4	2	6,168.8	5.6	36.2	1.06	0.03	0.0009
	3	13,589.1	5.5	33.7	1.75	0.05	0.0011
	4	22,599.2	5.3	29.8	2.29	0.04	0.0013
5	2	-	5.6	34.6	1.56	0.03	0.0009
	3	-	5.3	32.1	2.55	0.04	0.0011
	4	-	5.1	28.0	3.82	0.06	0.0012
10	2	-	5.3	30.6	5.64	0.05	0.0008
	3	-	5.0	26.7	10.66	0.08	0.0010
	4	-	4.7	24.0	13.58	0.12	0.0010

Table: Mean running time over 20 scenarios in seconds for the EM-algorithm for varying instances, with fixed values of $I_b = 100$ and $B = 50$.

Mean Absolute Error

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	0.012	0.013	0.012	0.013	0.013	0.011
	3	0.015	0.019	0.016	0.016	0.016	0.015
	4	0.018	0.029	0.020	0.019	0.019	0.018
3	2	0.011	0.014	0.011	0.013	0.011	0.010
	3	0.014	0.030	0.017	0.016	0.015	0.014
	4	0.017	0.051	0.027	0.020	0.018	0.018
4	2	0.011	0.016	0.011	0.013	0.012	0.011
	3	0.012	0.033	0.019	0.015	0.013	0.012
	4	0.015	0.051	0.032	0.019	0.017	0.015
5	2	-	0.013	0.009	0.010	0.009	0.008
	3	-	0.034	0.021	0.015	0.014	0.014
	4	-	0.043	0.030	0.015	0.014	0.014
10	2	-	0.016	0.010	0.009	0.009	0.006
	3	-	0.026	0.021	0.010	0.009	0.009
	4	-	0.032	0.028	0.013	0.012	0.012

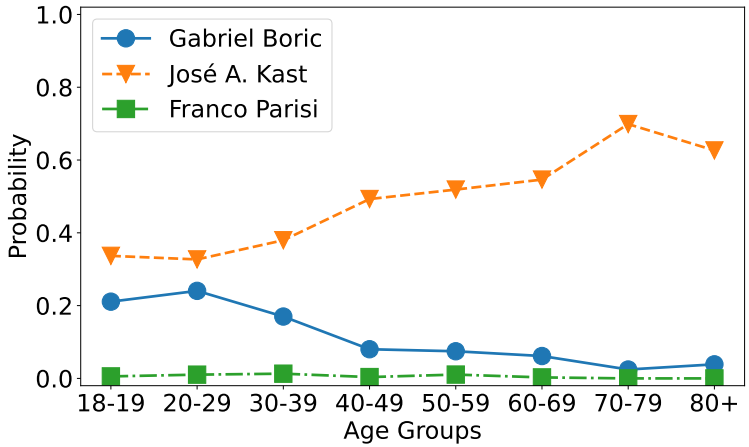
Table: Mean absolute prediction error over 20 scenarios for varying instances, with fixed values of $I_b = 100$ and $B = 50$.

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Group Aggregation

Figure: El Golf. $B = 318, A = 8$



Consider A: number of selected age ranges

First Round - Chilean Presidential Election - 2021

Group Aggregation

Figure: El Golf. $B = 318, A = 8$

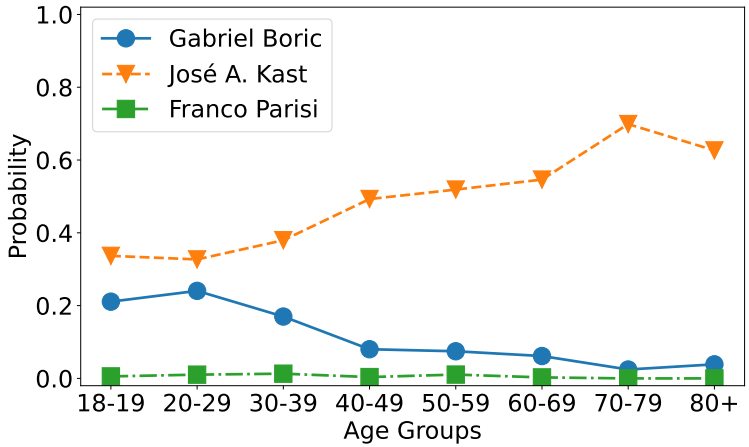
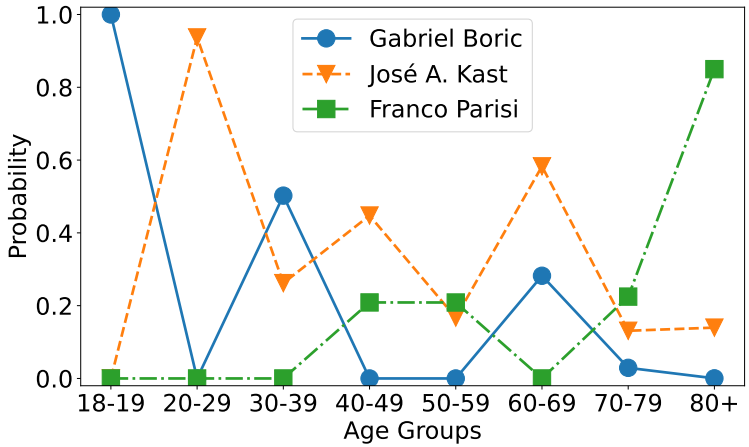


Figure: Saltos del Laja, $B = 5, A = 8$

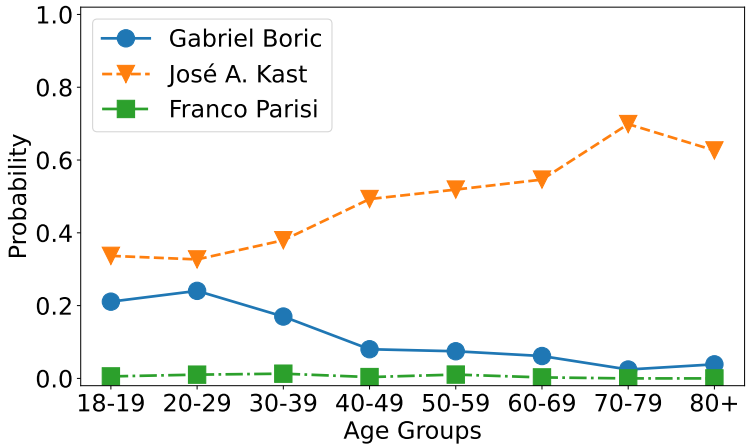


First Round - Chilean Presidential Election - 2021

Consider A: number of selected age ranges

Group Aggregation

Figure: El Golf. $B = 318, A = 8$



Consider A: number of selected age ranges

Figure: Saltos del Laja, $B = 5, A = 8$

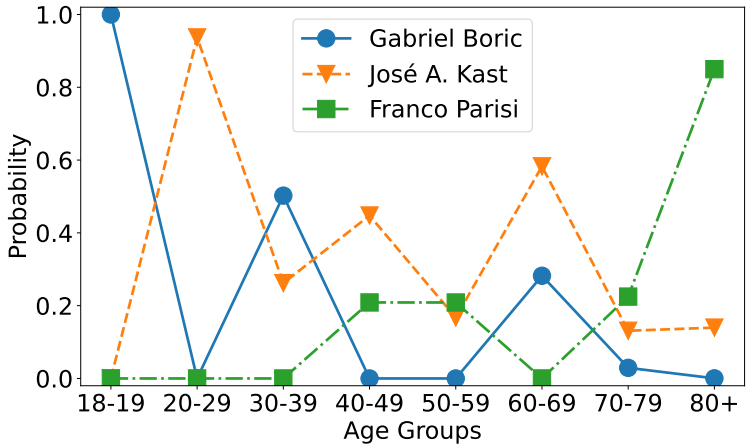
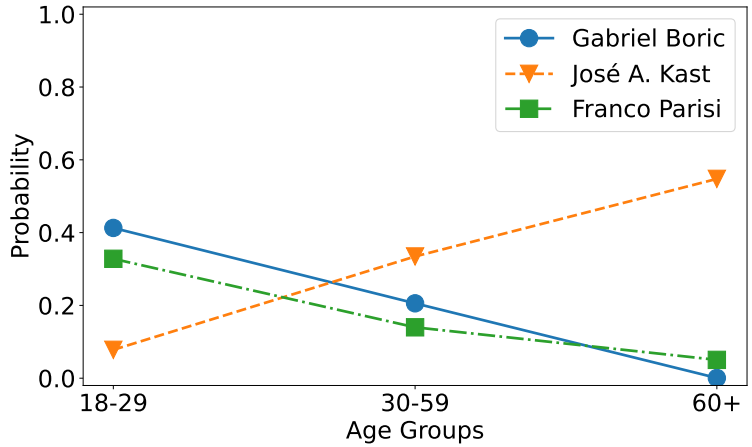


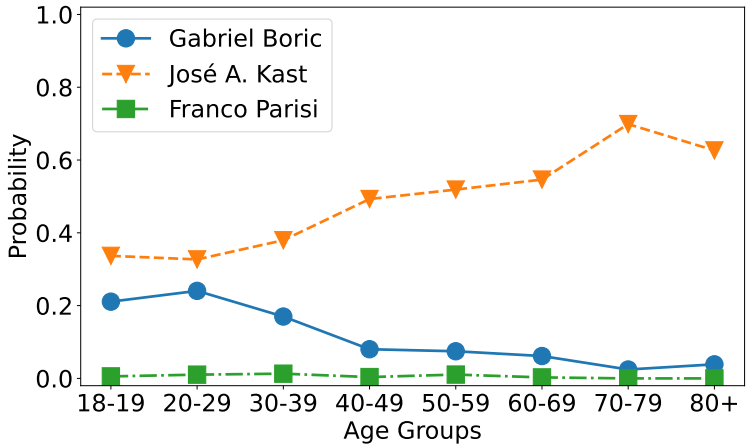
Figure: Saltos del Laja $B = 5, A = 3$



First Round - Chilean Presidential Election - 2021

Group Aggregation

Figure: El Golf. $B = 318, A = 8$



How to aggregate age ranges?

	# Agg. Groups (A) High	Low
Reliable	✗	✓
Good Fit	✓	✗

Figure: Saltos del Laja, $B = 5, A = 8$

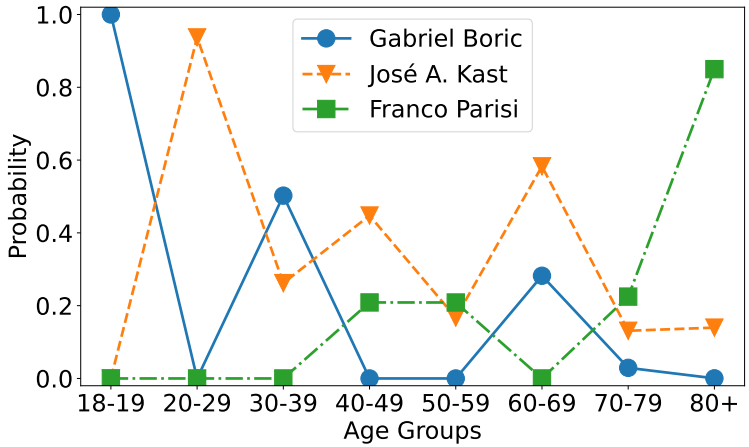
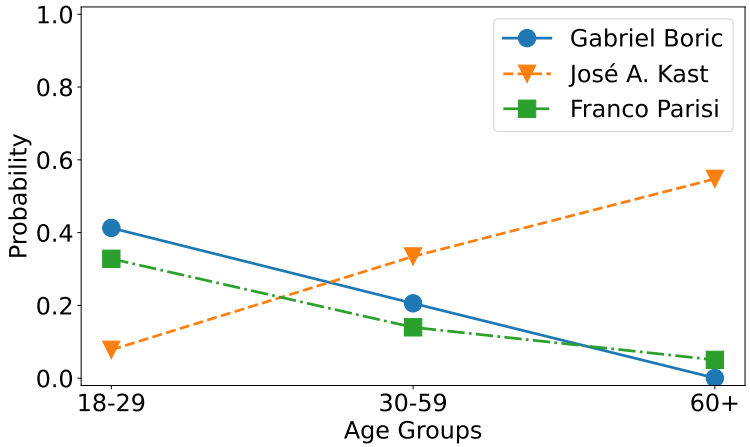


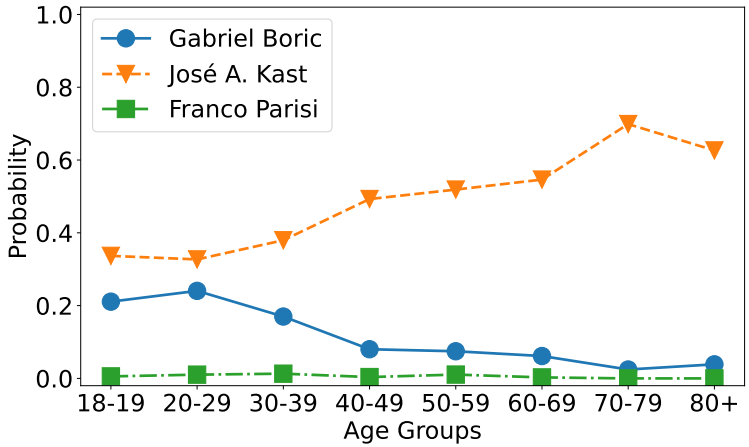
Figure: Saltos del Laja $B = 5, A = 3$



First Round - Chilean Presidential Election - 2021

Group Aggregation

Figure: El Golf. $B = 318, A = 8$



How to aggregate age ranges?

	# Agg. Groups (A) High	Low
Reliable	✗	✓
Good Fit	✓	✗

Figure: Saltos del Laja, $B = 5, A = 8$

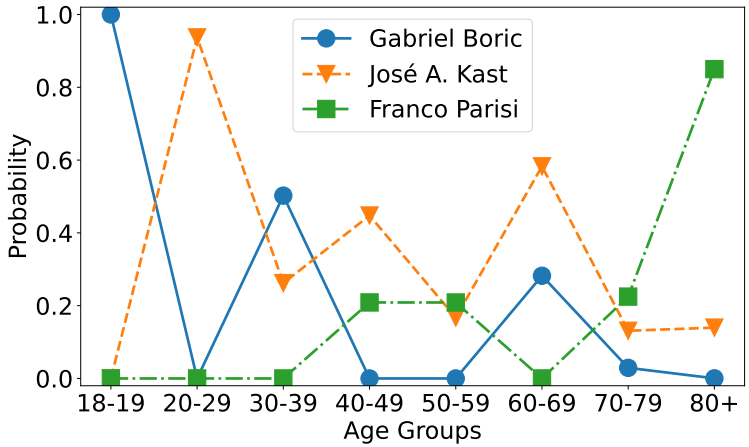
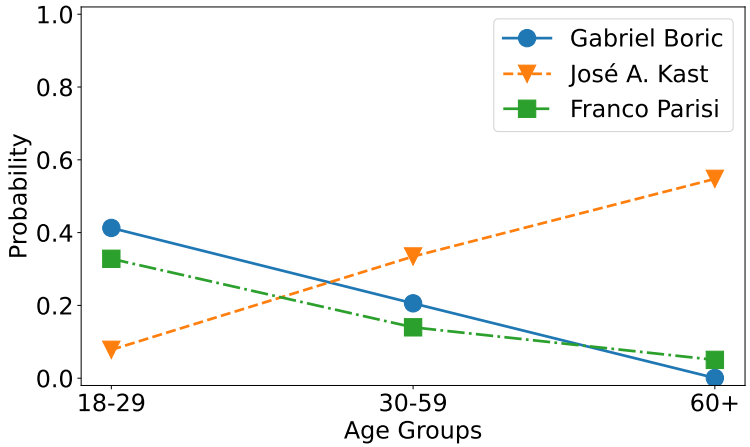


Figure: Saltos del Laja $B = 5, A = 3$



First Round - Chilean Presidential Election - 2021

Group Aggregation

- How do we measure reliability?
- **Maximum standard deviation** of the estimated voting probability ($\boldsymbol{p}$) using bootstrapping.

$$\text{std}(p_{gc}) := \sqrt{\frac{1}{S-1} \cdot \sum_{s=1}^{S} (p_{gc}^{(s)} - \bar{p}_{gc})^2}$$

$$\text{max-std}(\boldsymbol{p}) := \max_{g,c} \text{std}(p_{gc})$$

- How do we measure fitness?
- **Mean log-likelihood**:

$$I(\boldsymbol{p}; \mathcal{X}) = \frac{1}{B} \cdot \sum_{b,g} w_{bg} \sum_{c} q_{bgc} \ln \left(\frac{p_{gc}}{q_{bgc}} \right)$$

Group Aggregation

- How do we measure reliability?
- **Maximum standard deviation** of the estimated voting probability ($\boldsymbol{p}$) using bootstrapping.

$$\text{std}(p_{gc}) := \sqrt{\frac{1}{S-1} \cdot \sum_{s=1}^{S} (p_{gc}^{(s)} - \bar{p}_{gc})^2}$$

$$\text{max-std}(\boldsymbol{p}) := \max_{g,c} \text{std}(p_{gc})$$

- How do we measure fitness?
- **Mean log-likelihood**:

$$I(\boldsymbol{p}; \mathcal{X}) = \frac{1}{B} \cdot \sum_{b,g} w_{bg} \sum_{c} q_{bgc} \ln \left(\frac{p_{gc}}{q_{bgc}} \right)$$

Group Aggregation

- How do we measure reliability?
- **Maximum standard deviation** of the estimated voting probability ($\boldsymbol{p}$) using bootstrapping.

$$\text{std}(p_{gc}) := \sqrt{\frac{1}{S-1} \cdot \sum_{s=1}^{S} (p_{gc}^{(s)} - \bar{p}_{gc})^2}$$

$$\text{max-std}(\boldsymbol{p}) := \max_{g,c} \text{std}(p_{gc})$$

- How do we measure fitness?
- **Mean log-likelihood**:

$$I(\boldsymbol{p}; \mathcal{X}) = \frac{1}{B} \cdot \sum_{b,g} w_{bg} \sum_{c} q_{bgc} \ln \left(\frac{p_{gc}}{q_{bgc}} \right)$$

Group Aggregation

- How do we measure reliability?
- **Maximum standard deviation** of the estimated voting probability ($\boldsymbol{p}$) using bootstrapping.

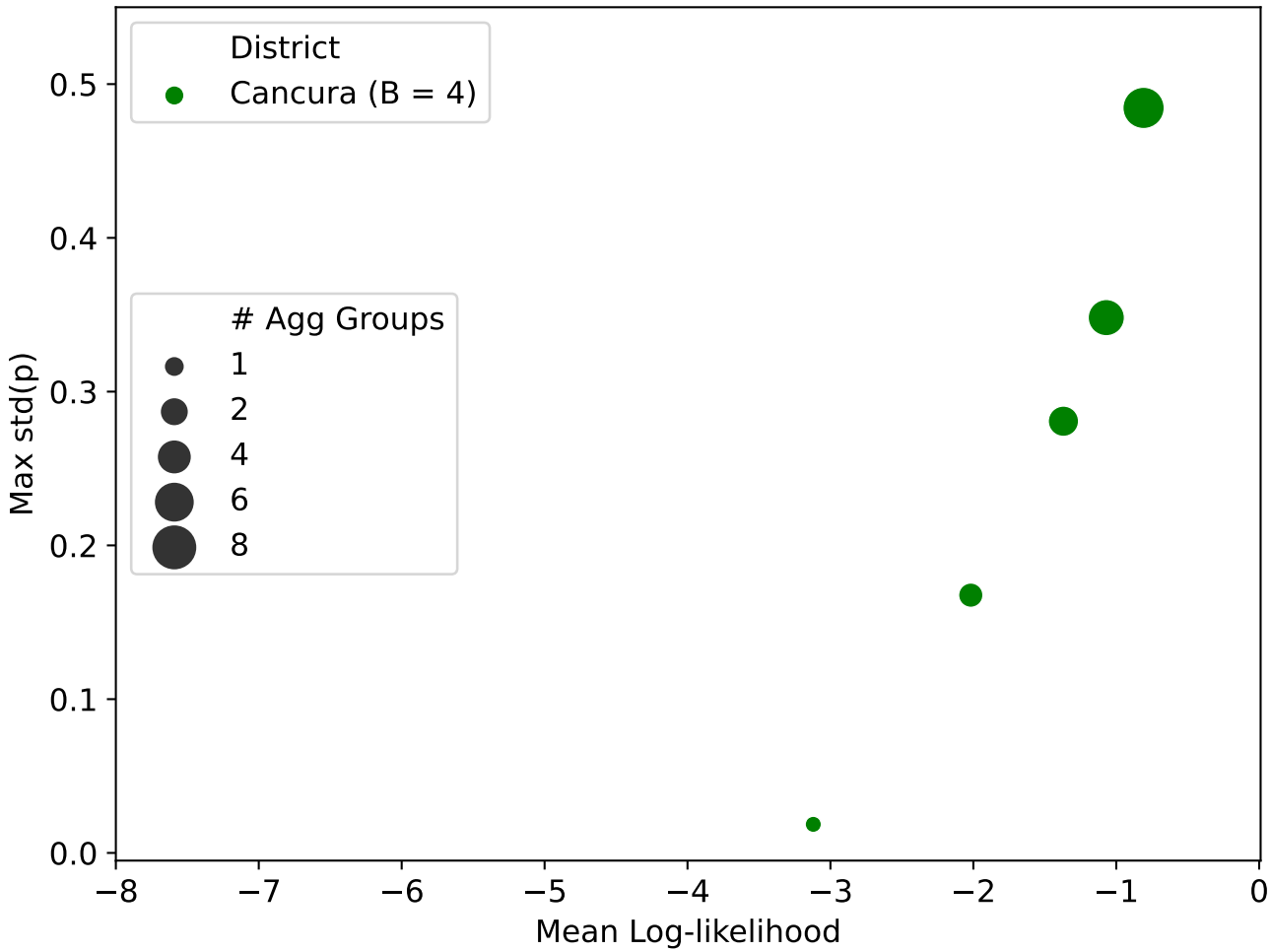
$$\text{std}(p_{gc}) := \sqrt{\frac{1}{S-1} \cdot \sum_{s=1}^{S} (p_{gc}^{(s)} - \bar{p}_{gc})^2}$$

$$\text{max-std}(\boldsymbol{p}) := \max_{g,c} \text{std}(p_{gc})$$

- How do we measure fitness?
- **Mean log-likelihood**:

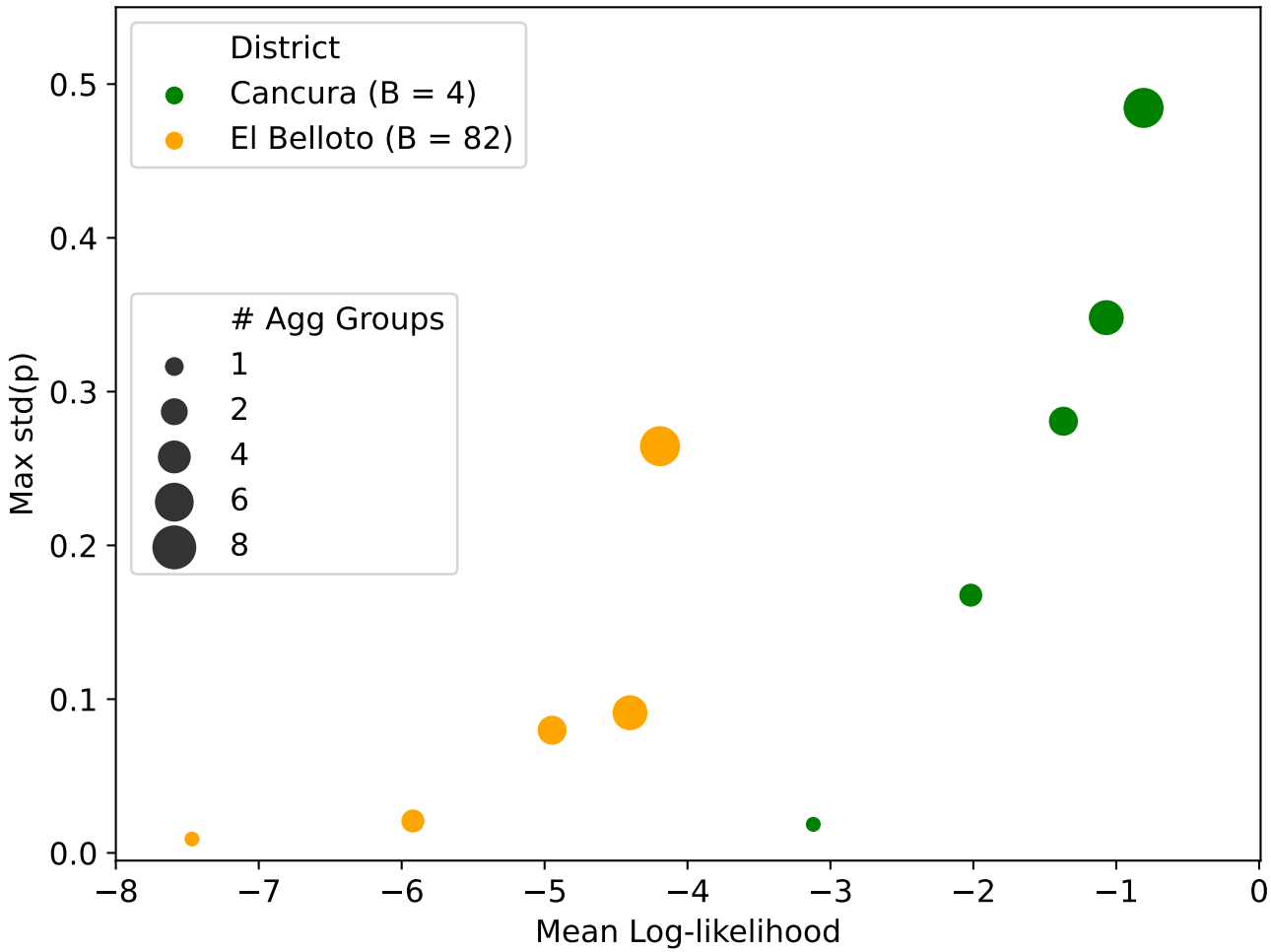
$$I(\boldsymbol{p}; \mathcal{X}) = \frac{1}{B} \cdot \sum_{b,g} w_{bg} \sum_{c} q_{bgc} \ln \left(\frac{p_{gc}}{q_{bgc}} \right)$$

Trade-off: log-likelihood vs standard deviation



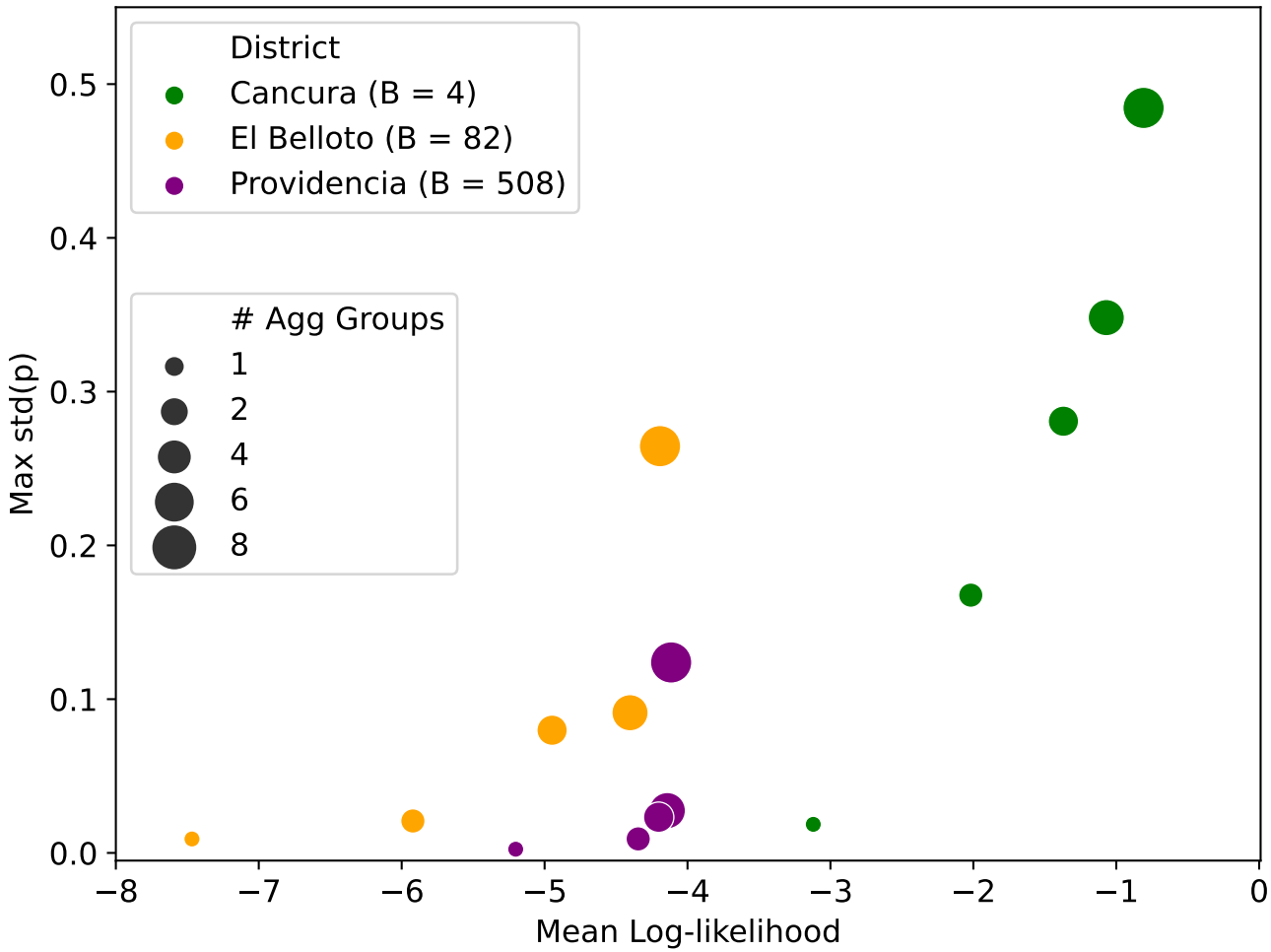
First Round - Chilean Presidential Election - 2021

Trade-off: log-likelihood vs standard deviation



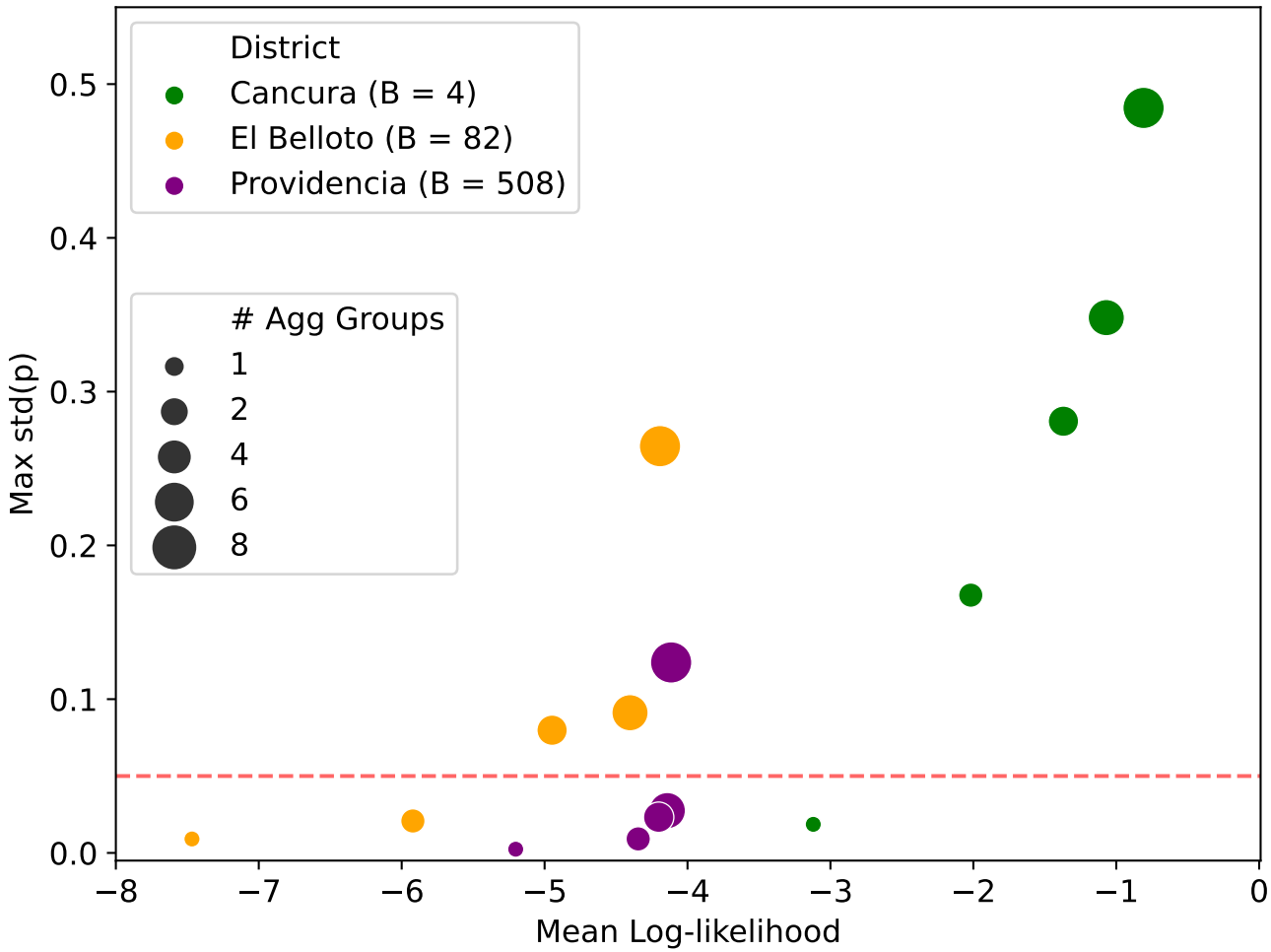
First Round - Chilean Presidential Election - 2021

Trade-off: log-likelihood vs standard deviation



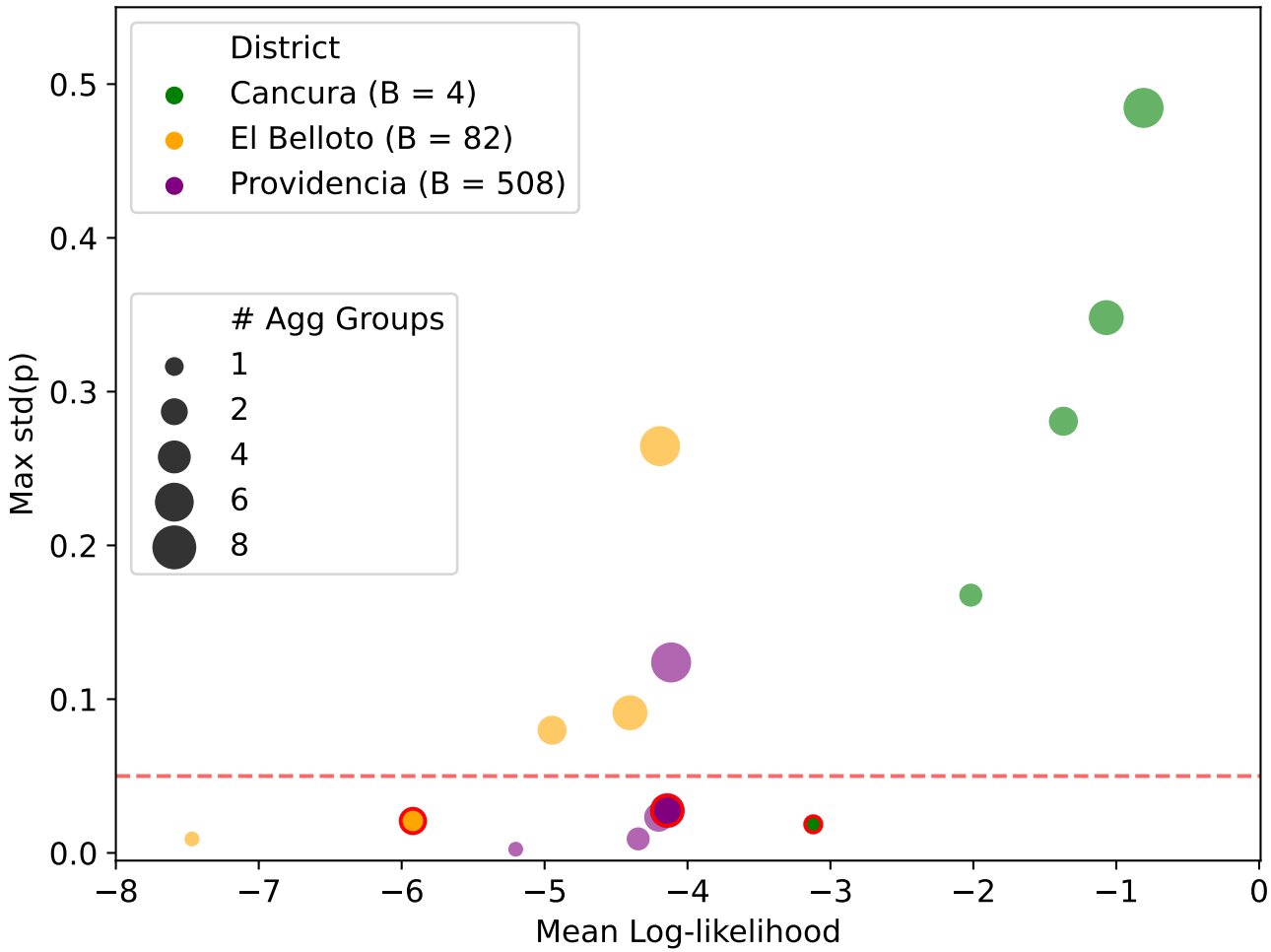
First Round - Chilean Presidential Election - 2021

Trade-off: log-likelihood vs standard deviation



First Round - Chilean Presidential Election - 2021

Trade-off: log-likelihood vs standard deviation



First Round - Chilean Presidential Election - 2021

Group Aggregation

Note that there are multiple group aggregations sets: $\mathcal{A}$. For example:

- 1 Group: $\mathcal{A} = \{18^+\}$
- 4 Groups: $\mathcal{A} = \{18\text{-}29, 30\text{-}49, 50\text{-}69, 70^+\}$
- 8 Groups: $\mathcal{A} = \{18\text{-}19, 20\text{-}29, 30\text{-}39, 40\text{-}49, 50\text{-}59, 60\text{-}69, 70\text{-}79, 80^+\}$

Even for fixed A groups we may have multiple sets. For example:

- 3 Groups: $\mathcal{A} = \begin{matrix} \{18\text{-}29, 30\text{-}69, 70^+\} \\ \{18\text{-}39, 40\text{-}69, 80^+\} \\ \{18\text{-}19, 20\text{-}39, 40^+\}, \text{ etc...} \end{matrix}$

There are $\binom{G-1}{A-1}$ possible aggregations sets of size A

How do we select the best one?

Group Aggregation

Note that there are multiple group aggregations sets: $\mathcal{A}$. For example:

- 1 Group: $\mathcal{A} = \{18^+\}$
- 4 Groups: $\mathcal{A} = \{18\text{-}29, 30\text{-}49, 50\text{-}69, 70^+\}$
- 8 Groups: $\mathcal{A} = \{18\text{-}19, 20\text{-}29, 30\text{-}39, 40\text{-}49, 50\text{-}59, 60\text{-}69, 70\text{-}79, 80^+\}$

Even for fixed A groups we may have multiple sets. For example:

- 3 Groups: $\mathcal{A} = \begin{matrix} \{18\text{-}29, 30\text{-}69, 70^+\} \\ \{18\text{-}39, 40\text{-}69, 80^+\} \\ \{18\text{-}19, 20\text{-}39, 40^+\}, \text{ etc...} \end{matrix}$

There are $\binom{G-1}{A-1}$ possible aggregations sets of size A

How do we select the best one?

Group Aggregation

Note that there are multiple group aggregations sets: $\mathcal{A}$. For example:

- 1 Group: $\mathcal{A} = \{18^+\}$
- 4 Groups: $\mathcal{A} = \{18\text{-}29, 30\text{-}49, 50\text{-}69, 70^+\}$
- 8 Groups: $\mathcal{A} = \{18\text{-}19, 20\text{-}29, 30\text{-}39, 40\text{-}49, 50\text{-}59, 60\text{-}69, 70\text{-}79, 80^+\}$

Even for fixed A groups we may have multiple sets. For example:

- 3 Groups: $\mathcal{A} = \begin{matrix} \{18\text{-}29, 30\text{-}69, 70^+\} \\ \{18\text{-}39, 40\text{-}69, 80^+\} \\ \{18\text{-}19, 20\text{-}39, 40^+\}, \text{ etc...} \end{matrix}$

There are $\binom{G-1}{A-1}$ possible aggregations sets of size A

How do we select the best one?

Group Aggregation

Note that there are multiple group aggregations sets: $\mathcal{A}$. For example:

- 1 Group: $\mathcal{A} = \{18^+\}$
- 4 Groups: $\mathcal{A} = \{18\text{-}29, 30\text{-}49, 50\text{-}69, 70^+\}$
- 8 Groups: $\mathcal{A} = \{18\text{-}19, 20\text{-}29, 30\text{-}39, 40\text{-}49, 50\text{-}59, 60\text{-}69, 70\text{-}79, 80^+\}$

Even for fixed A groups we may have multiple sets. For example:

- 3 Groups: $\mathcal{A} = \begin{matrix} \{18\text{-}29, 30\text{-}69, 70^+\} \\ \{18\text{-}39, 40\text{-}69, 80^+\} \\ \{18\text{-}19, 20\text{-}39, 40^+\}, \text{ etc...} \end{matrix}$

There are $\binom{G-1}{A-1}$ possible aggregations sets of size A

How do we select the best one?

MIP

$\mathcal{M}$: Set of all possible group aggregations m

σ_m: Standard deviation of group aggregation m proportion

Decision variables: $a_m = \begin{cases} 1 & \text{group aggregation } m \text{ is included} \\ 0 & \text{otherwise} \end{cases}$

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & \sum_{m \in \mathcal{M}} a_m \cdot \sigma_m && \text{(Maximize standard deviation)} \\ \text{s.t:} \quad & \sum_{m \in \mathcal{M} | g \in m} a_m = 1, \ \forall g && \text{(Each } g \text{ is selected only once)} \\ & \sum_{m \in \mathcal{M}} a_m = A && \text{(Choose } A \text{ aggregations)} \\ & a_m \in \{0, 1\} && \text{(If aggregation } m \text{ is included)} \end{aligned}$$

MIP

$\mathcal{M}$: Set of all possible group aggregations m

σ_m: Standard deviation of group aggregation m proportion

Decision variables: $a_m = \begin{cases} 1 & \text{group aggregation } m \text{ is included} \\ 0 & \text{otherwise} \end{cases}$

$$
\begin{aligned}
\max_{\boldsymbol{a}} \quad & \textstyle\sum_{m \in \mathcal{M}} a_m \cdot \sigma_m && \text{(Maximize standard deviation)} \\
\text{s.t:} \quad & \textstyle\sum_{m \in \mathcal{M} | g \in m} a_m = 1, \ \forall g && \text{(Each } g \text{ is selected only once)} \\
& \textstyle\sum_{m \in \mathcal{M}} a_m = A && \text{(Choose } A \text{ aggregations)} \\
& a_m \in \{0, 1\} && \text{(If aggregation } m \text{ is included)}
\end{aligned}
$$

Methodology

Algorithm 3 Group Aggregation Algorithm

Input: $\mathcal{I}$, $\mathcal{G}$, $\mathcal{C}$, $\mathcal{B}$, $\boldsymbol{x}$, $\boldsymbol{w}$, γ

Output: $\mathcal{A}$

for $A = G$ to 1 **do**

  Solve MIP and obtain: $\mathcal{A} = \{m \in \mathcal{M} | a_m = 1\}$

  $w'_{ba} = \sum_{g \in a} w_{bg}, \forall a \in \mathcal{A}$

  Apply bootstrapping to obtain $\text{std}(p_{ac})$

  if $\text{std}(p_{ac}) \leq \gamma, \forall a \in \mathcal{A}, \forall c \in \mathcal{C}$ **then**

    Stop

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Application to a Real Election

- We ran the models in the Fist Round Chilean Presidential Election of 2021.
- C=8: 7 candidates: Eduardo Artés, Franco Parisi, Gabriel Boric, José A. Kast, Marco E.-Ominami, Sebastián Sichel and Yasna Provoste; plus the Null/Blank option.
- G=8: 8 age ranges: 18-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80^{+}
- 7,080,276 votes emitted across 681 districts containing 2808 polling places with a total of 46,639 ballot-boxes.

Results for Group Aggregation Algorithm

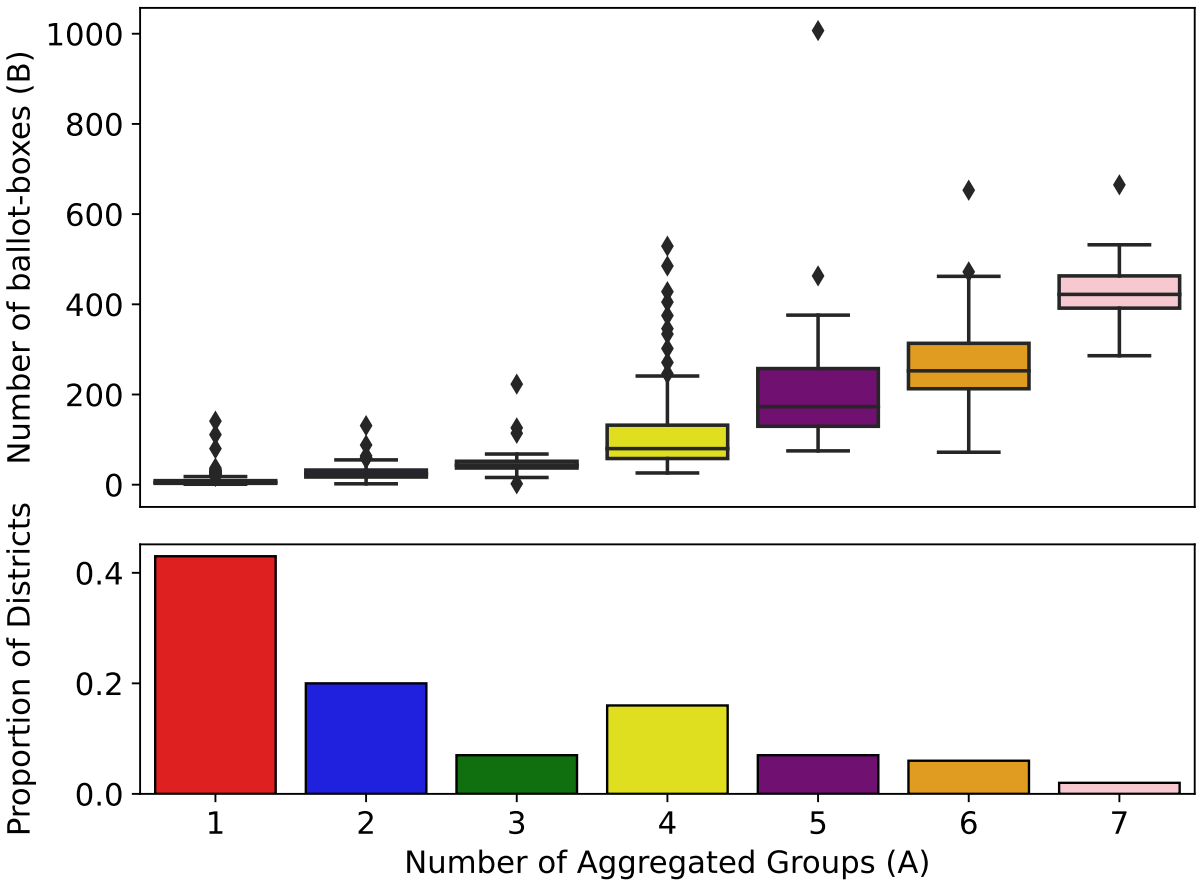


Figure: Statistics for aggregated groups

Estimated Probabilities

How do estimated probabilities look?

Estimated Probabilities

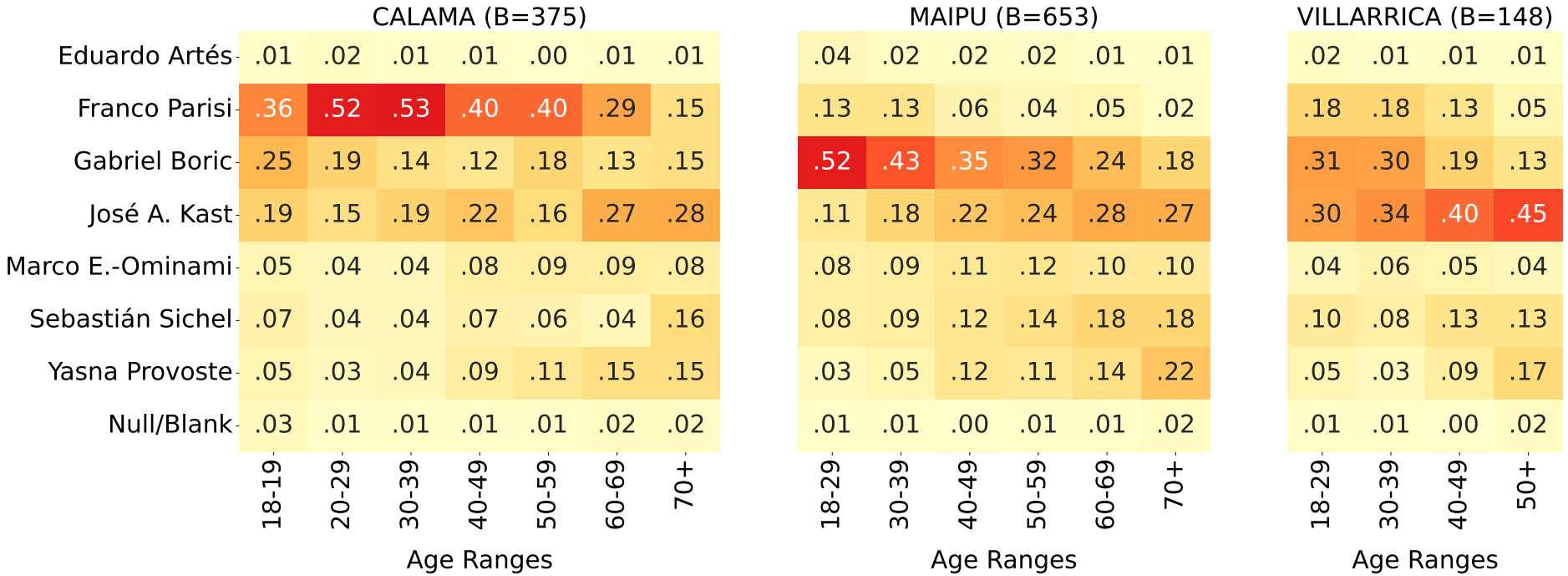


Figure: Group probability results for 3 selected polling places

Estimated Probabilities

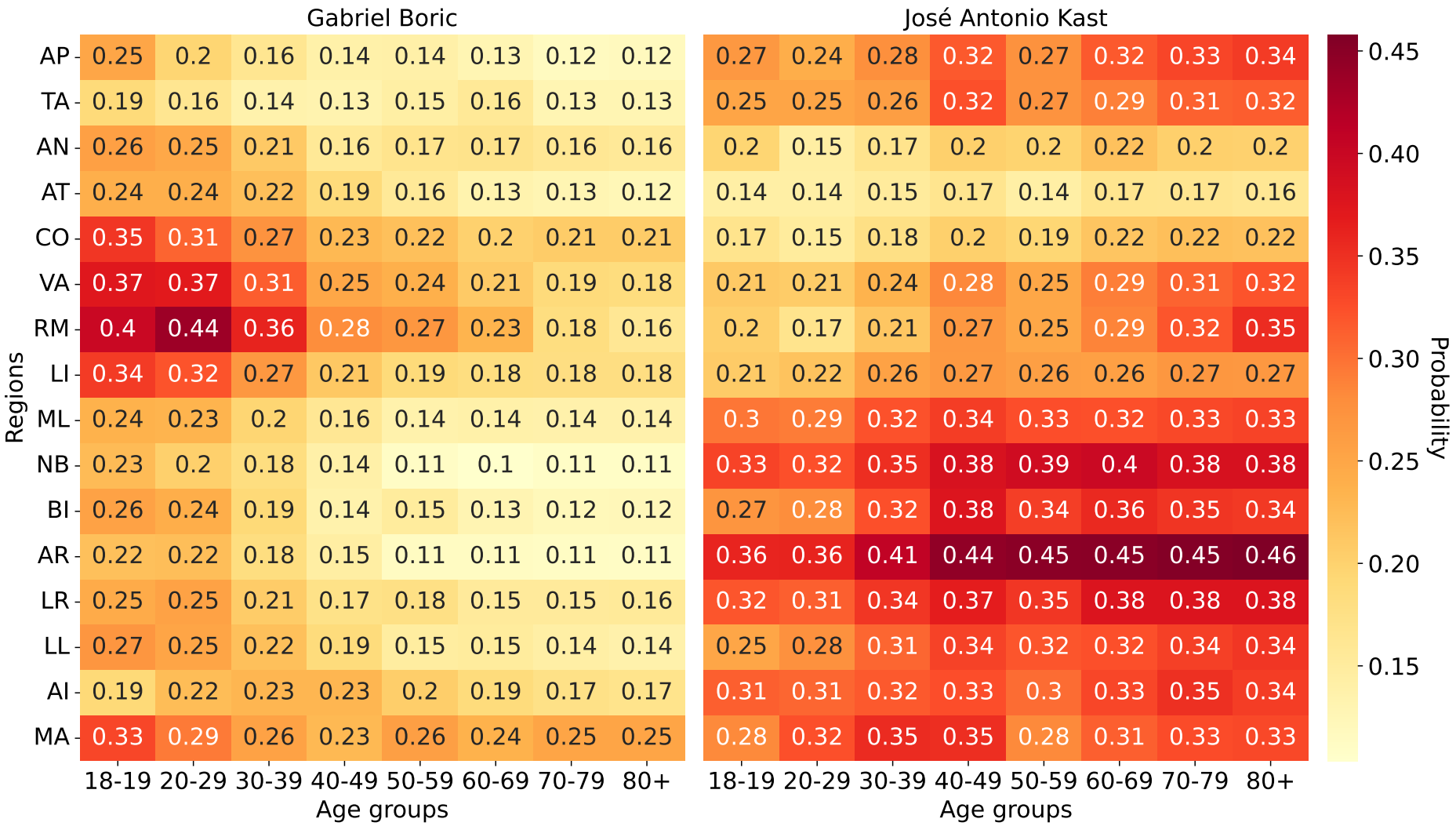


Figure: Aggregated group probabilities across regions

p-values

We would like to see how odd a ballot-box is using the estimated probabilities

p-values

p-value($\boldsymbol{x}_b$): Probability of obtaining an outcome equally or less likely than $\boldsymbol{x}_b$.

- **Exact Calculation**:

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

Similar complexity to the E-step: $O(B \cdot G \cdot C \cdot I_b^{2C})$

- **Simulation**: Consider $\boldsymbol{x}_b^{(1)}, ..., \boldsymbol{x}_b^{(H)}$ sampled from $\text{Mult}(I_b, \boldsymbol{r}_b)$.

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

p-values

p-value($\boldsymbol{x}_b$): Probability of obtaining an outcome equally or less likely than $\boldsymbol{x}_b$.

- **Exact Calculation**:

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

Similar complexity to the E-step: $O(B \cdot G \cdot C \cdot I_b^{2C})$

- **Simulation**: Consider $\boldsymbol{x}_b^{(1)}, ..., \boldsymbol{x}_b^{(H)}$ sampled from $\text{Mult}(I_b, \boldsymbol{r}_b)$.

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

p-values

p-value($\boldsymbol{x}_b$): Probability of obtaining an outcome equally or less likely than $\boldsymbol{x}_b$.

- **Exact Calculation**:

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

Similar complexity to the E-step: $O(B \cdot G \cdot C \cdot I_b^{2C})$

- **Simulation**: Consider $\boldsymbol{x}_b^{(1)}, ..., \boldsymbol{x}_b^{(H)}$ sampled from Mult($I_b, \boldsymbol{r}_b$).

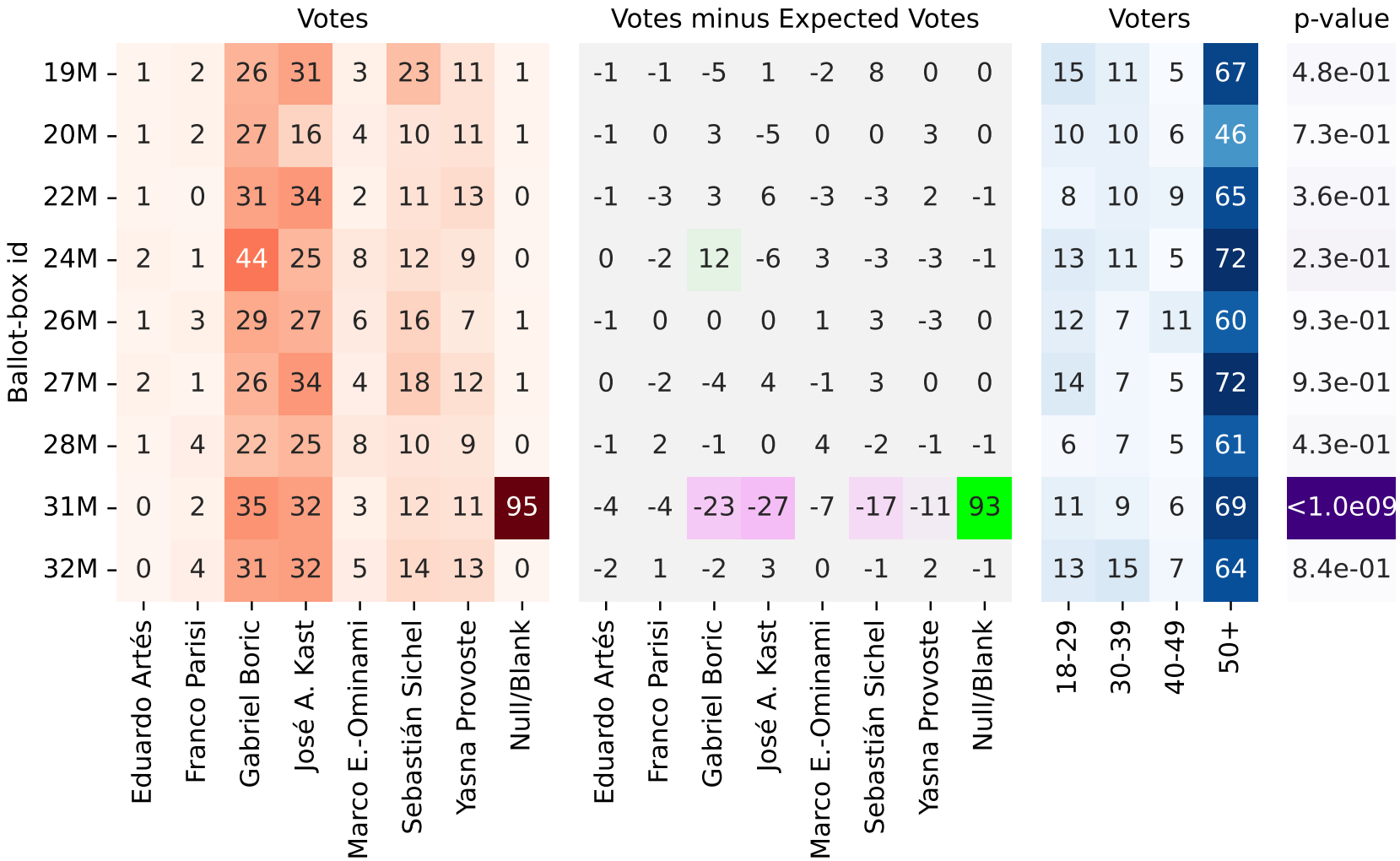
$$\text{p-value}(\boldsymbol{x}_b) = \frac{1}{H} \sum_{h=1}^{H} \mathbb{1}_{\{P(\boldsymbol{x}_b^{(h)}) \leq P(\boldsymbol{x}_b)\}}$$

Outlier ballot-boxes

Let's take a look at some outlier ballot-boxes

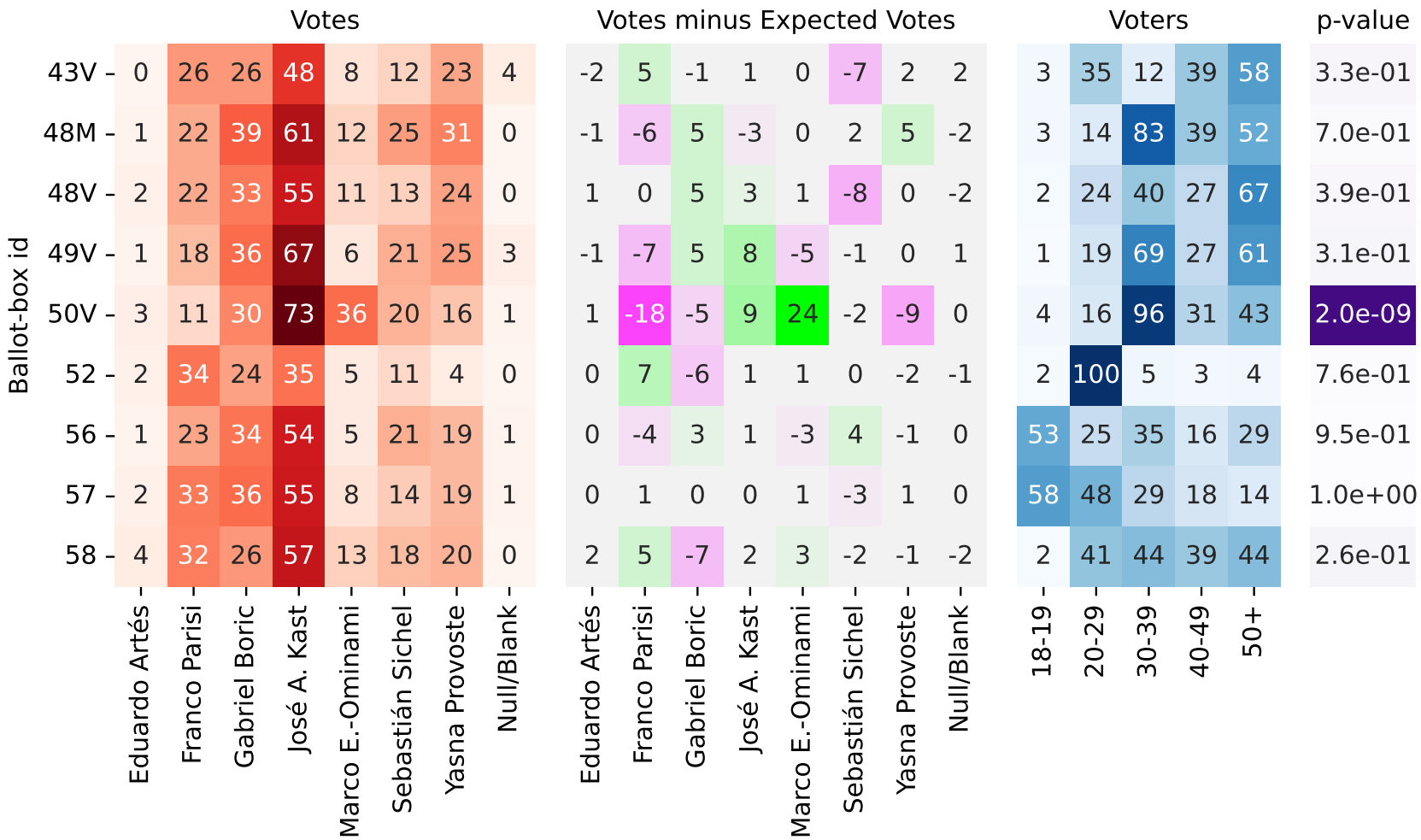
Outlier ballot-boxes

Figure: Liceo Javiera Carrera Local: 2



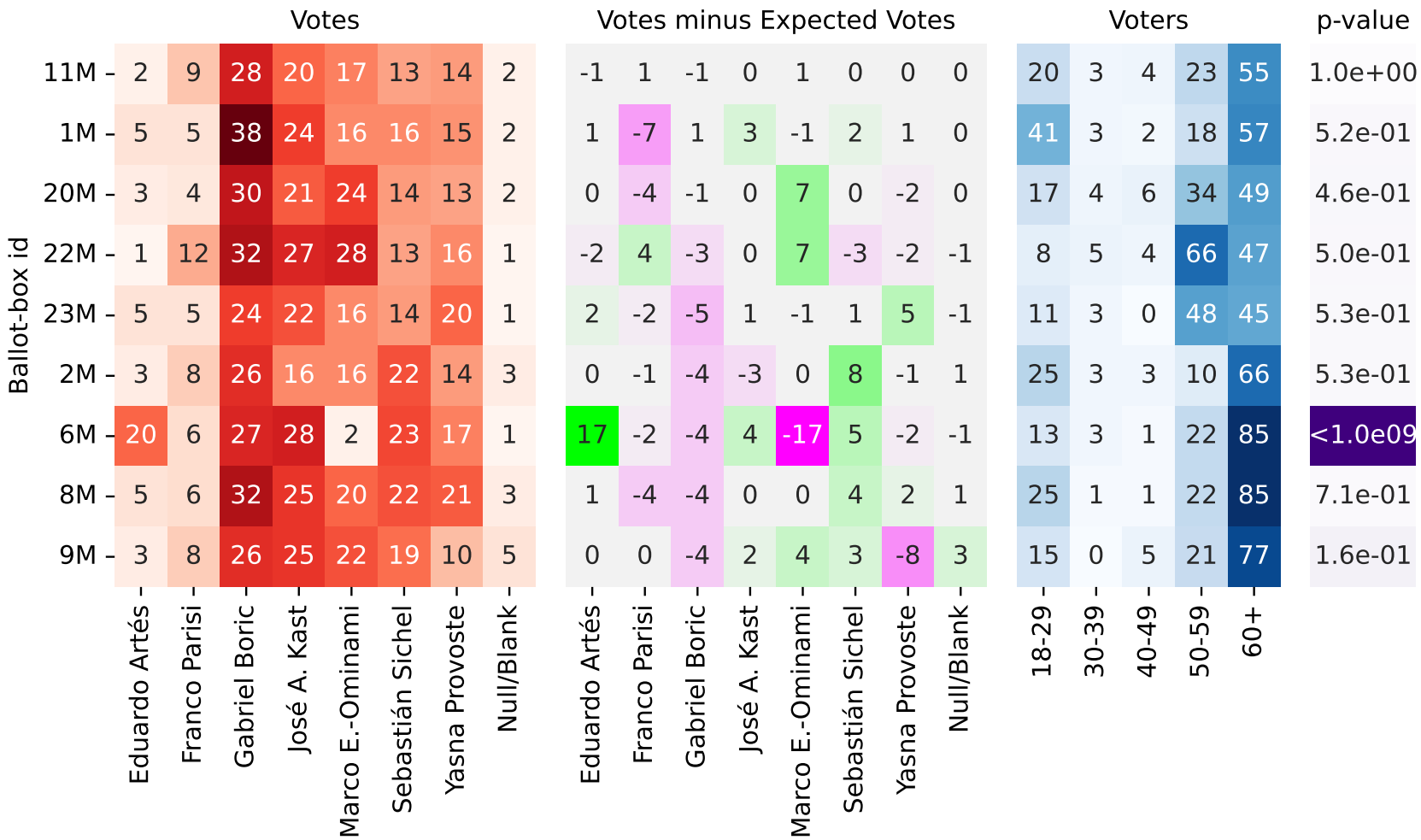
Outlier ballot-boxes

Figure: Escuela Gerónimo Lagos Lisboa



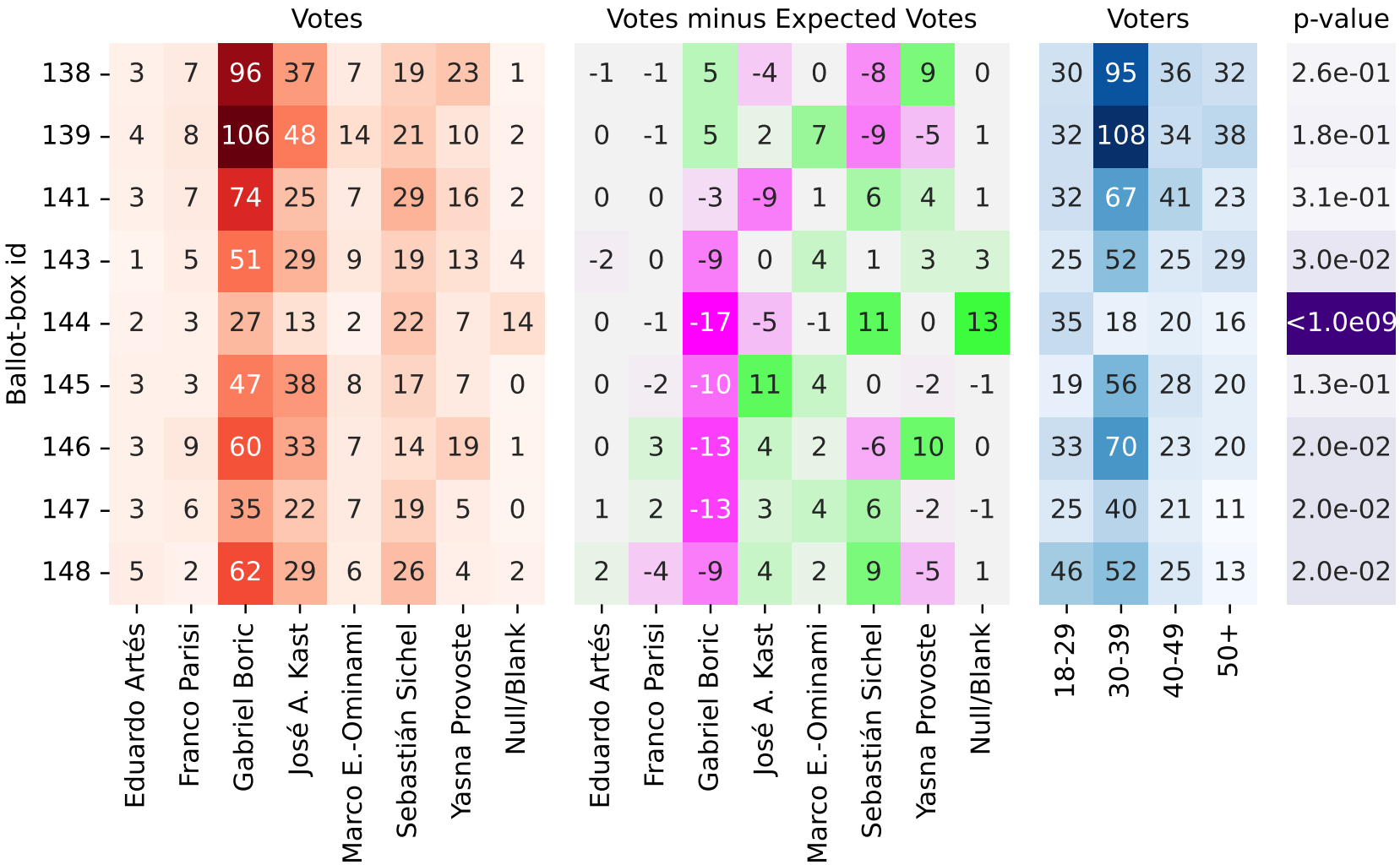
Outlier ballot-boxes

Figure: Escuela Monseñor Carlos Oviedo



Outlier ballot-boxes

Figure: Colegio Filipense



p-value Results

Table: p-values in the First Round Presidential election 2021

p-value	Total	%
$(10^{-1}, 1]$	36,584	78.50
$(10^{-2}, 10^{-1}]$	7,471	16.03
$(10^{-3}, 10^{-2}]$	2,055	4.41
$(10^{-4}, 10^{-3}]$	259	0.56
$(10^{-5}, 10^{-4}]$	111	0.24
$(10^{-6}, 10^{-5}]$	54	0.12
$(10^{-7}, 10^{-6}]$	28	0.06
$(10^{-8}, 10^{-7}]$	16	0.03
$(10^{-9}, 10^{-8}]$	12	0.03
$(0, 10^{-9}]$	15	0.03

p-value Results

Table: p-values in the First Round Presidential election 2021

p-value	Total	%
$(10^{-1}, 1]$	36,584	78.50
$(10^{-2}, 10^{-1}]$	7,471	16.03
$(10^{-3}, 10^{-2}]$	2,055	4.41
$(10^{-4}, 10^{-3}]$	259	0.56
$(10^{-5}, 10^{-4}]$	111	0.24
$(10^{-6}, 10^{-5}]$	54	0.12
$(10^{-7}, 10^{-6}]$	28	0.06
$(10^{-8}, 10^{-7}]$	16	0.03
$(10^{-9}, 10^{-8}]$	12	0.03
1 in a billion or less! $(0, 10^{-9}]$	15	0.03

Reasons for outliers

These models do not tell us why an outlier has happened. Possible explanations may be:

- The data provided by SERVEL has errors
- Model assumptions are missing relevant information
- Fraudulent behavior

Possible solutions

Possible actions SERVEL can take in the short and medium term:

- Verify that all protocols were followed
- Check that the official records match the data
- Delay the publication for the results of outlier ballot-boxes
- Change the ballot-box staff for future elections

Table of Contents

1. Model
   - Setting
   - EM Algorithm
2. Methods
   - Simulation: Hit and Run
   - Multivariate Normal Approximations
   - Multinomial Approximation
3. Numerical Results
4. Group Aggregation
5. Case Study
6. Conclusions

Conclusions

- Approximation methods reduces complexity from $O(B \cdot G^2 \cdot C \; I_b^{2C})$ to $O(B \cdot G \cdot C)$
- Estimated probabilities are almost identical to the ones obtained with the EXACT method
- Group aggregation: Methodology based on MIP and bootstrapping that balances reliability and fitness
- Applicable to real elections with multiple candidates and demographic groups. We ran the following:
  - 1st Round Presidential Election 2021
  - 2nd Round Presidential Election 2021
  - National Plebiscite 2022
  - Constitutional Council 2023

Any questions?

Bibliography I

- Antweiler, Werner (2007). "Estimating voter migration in Canada using generalized maximum entropy". In: *Electoral Studies* 26.4, pp. 756–771.
- Baumert, Stephen et al. (2009). "Discrete hit-and-run for sampling points from arbitrary distributions over subsets of integer hyperrectangles". In: *Operations Research* 57.3, pp. 727–739.
- Cho, Wendy Tam and Charles F Manski (2008). "Cross-Level/Ecological Inference". In.
- Corominas, Albert, Amaia Lusa, and M Dolors Calvet (2015). "Computing voter transitions: The elections for the Catalan Parliament, from 2010 to 2012". In: *Journal of Industrial Engineering and Management (JIEM)* 8.1, pp. 122–136.
- El Deeb, Omar (2023). "Entropic spatial auto-correlation of voter uncertainty and voter transitions in parliamentary elections". In: *Physica A: Statistical Mechanics and its Applications* 617, p. 128675.
- Ellis, Nick and Ranjan Maitra (2007). "Multivariate Gaussian simulation outside arbitrary ellipsoids". In: *Journal of Computational and Graphical Statistics* 16.3, pp. 692–708.

Bibliography II

- Glynn, Adam N and Jon Wakefield (2010). "Ecological inference in the social sciences". In: *Statistical methodology* 7.3, pp. 307–322.
- James Greiner, D and Kevin M Quinn (2009). "R× C ecological inference: bounds, correlations, flexibility and transparency of assumptions". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1, pp. 67–81.
- Johnston, Ron and Charles Pattie (2000). "Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting". In: *Political Analysis* 8.4, pp. 333–345.
- Johnston, Ron J and AM Hay (1983). "Voter transition probability estimates: an entropy-maximizing approach". In: *European Journal of Political Research* 11.1, pp. 93–98.
- Judge, George G, Douglas J Miller, and Wendy K Cho (2003). *An information theoretic approach to ecological estimation and inference*. Tech. rep.
- King, Gary (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press.

Bibliography III

- King, Gary, Ori Rosen, and Martin A Tanner (1999). "Binomial-beta hierarchical models for ecological inference". In: *Sociological Methods & Research* 28.1, pp. 61–90.
- King, Gary, Martin A Tanner, and Ori Rosen (2004). *Ecological inference: New methodological strategies*. Cambridge University Press.
- Lin, Zhengzhi, Yueyao Wang, and Yili Hong (2022). "The Poisson Multinomial Distribution and its Applications in Voting Theory, Ecological Inference, and Machine Learning". In: *arXiv preprint arXiv:2201.04237*.
- Maatouk, Hassan and Xavier Bay (2016). "A new rejection sampling method for truncated multivariate Gaussian random variables restricted to convex sets". In: *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*. Springer, pp. 521–530.
- McCarthy, Colm and Terence M Ryan (1977). "Estimates of voter transition probabilities from the British general elections of 1974". In: *Journal of the Royal Statistical Society: Series A (General)* 140.1, pp. 78–85.

Bibliography IV

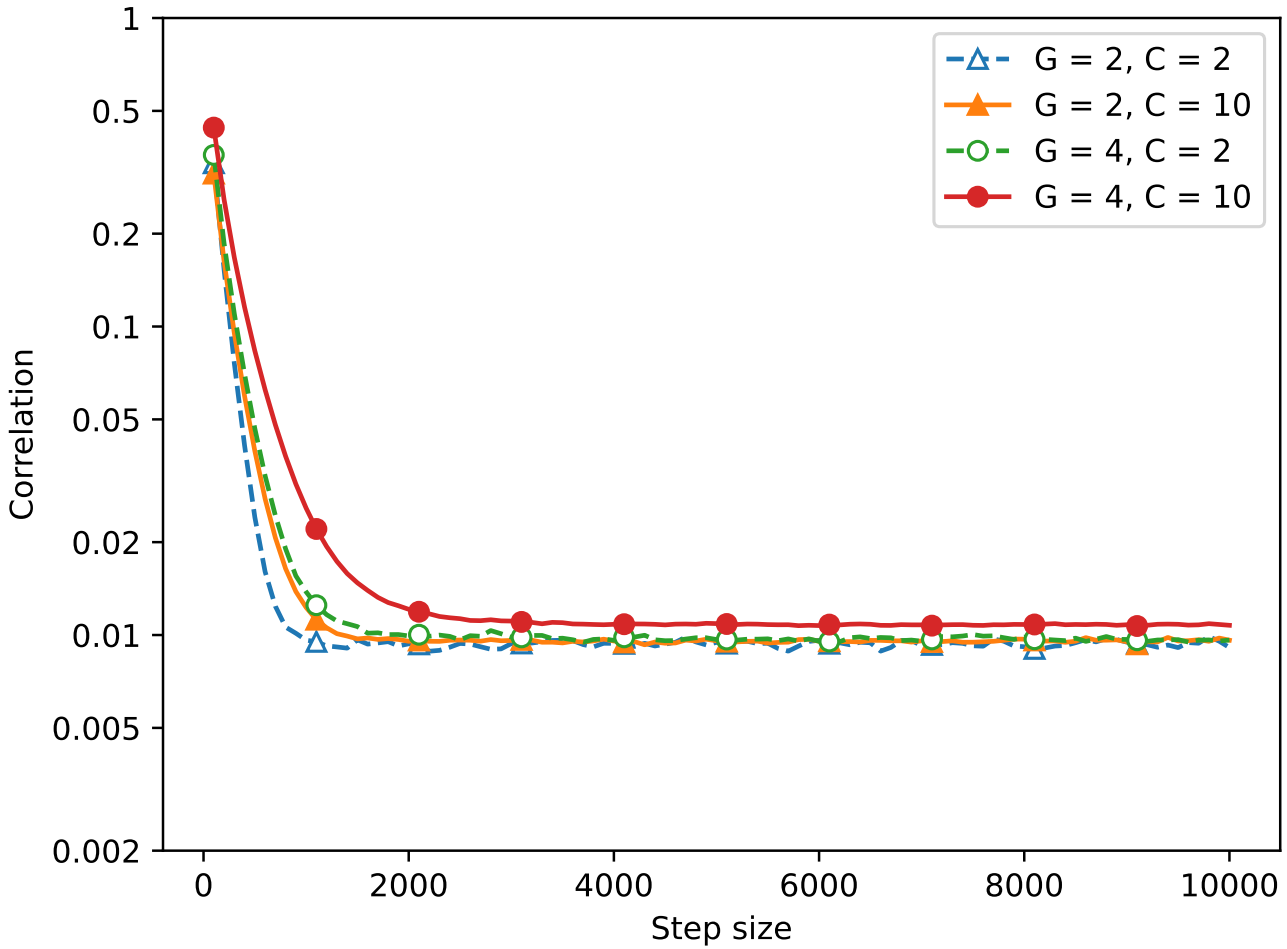
- Mete, Huseyin Onur and Zelda B Zabinsky (2012). "Pattern hit-and-run for sampling efficiently on polytopes". In: *Operations Research Letters* 40.1, pp. 6–11.
- Novak, SY (2019). "Poisson approximation". In.
- Pavía, Jose M and Rafael Romero (2021). "Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming". In: *Sociological Methods & Research*, p. 0049124121092725.
- Roos, Bero (1999). "Metric multivariate Poisson approximation of the generalized multinomial distribution". In: *Theory of Probability & Its Applications* 43.2, pp. 306–316.
- — (2002). "Multinomial and Krawtchouk approximations to the generalized multinomial distribution". In: *Theory of Probability & Its Applications* 46.1, pp. 103–117.
- Rosen, Ori et al. (2001). "Bayesian and frequentist inference for ecological inference: The R× C case". In: *Statistica Neerlandica* 55.2, pp. 134–156.
- Tziafetas, G (1986). "Estimation of the voter transition matrix". In: *optimization* 17.2, pp. 275–279.

Log-likelihood - Convergence

$$
\begin{aligned}
l(\boldsymbol{p};\mathcal{X}) &= \sum_{b\in\mathcal{B}} \ln P(\boldsymbol{X}_b = \boldsymbol{x}_b|\boldsymbol{p}) \\
&= \sum_{b\in\mathcal{B}} \ln \left[\sum_{i\in\mathcal{I}_b, c\in\mathcal{C}} P(\boldsymbol{X}_b = \boldsymbol{x}_b, Y_{bg(i)c} = 1|\boldsymbol{p}) \right] \\
&= \sum_{b\in\mathcal{B}} \ln \left[\sum_{i\in\mathcal{I}_b, c\in\mathcal{C}} P(Y_{bg(i)c} = 1|\boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p}^{(old)}) \cdot \frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b, Y_{bg(i)c} = 1|\boldsymbol{p})}{P(Y_{bg(i)c} = 1|\boldsymbol{X}_b = \boldsymbol{x}_b; \boldsymbol{p}^{(old)})} \right] \\
&= \ln \mathbb{E}\left[\left(\frac{P(\mathcal{X},\mathcal{Y}|\boldsymbol{p})}{P(\mathcal{Y}|\mathcal{X},\boldsymbol{p}^{(old)})} \right) |\mathcal{X},\boldsymbol{p}^{(old)} \right] \\
&\geq \mathbb{E}\left[\ln \left(\frac{P(\mathcal{X},\mathcal{Y}|\boldsymbol{p})}{P(\mathcal{Y}|\mathcal{X},\boldsymbol{p}^{(old)})} \right) |\mathcal{X},\boldsymbol{p}^{(old)} \right] \qquad \text{(Jansen's Inequality)} \\
&= \mathbb{E}\left[l(\boldsymbol{p};\mathcal{X},\mathcal{Y})|\mathcal{X},\boldsymbol{p}^{(old)} \right] - \mathbb{E}\left[\ln P(\mathcal{Y}|\mathcal{X};\boldsymbol{p}^{(old)})|\mathcal{X},\boldsymbol{p}^{(old)} \right] \\
&= Q(\boldsymbol{p};\boldsymbol{p}^{(old)}) - \mathbb{E}\left[\ln P(\mathcal{Y}|\mathcal{X};\boldsymbol{p}^{(old)})|\mathcal{X},\boldsymbol{p}^{(old)} \right] \\
&= \sum_{b\in\mathcal{B}, g\in\mathcal{G}} w_{bg} \sum_{c\in\mathcal{C}} q_{bgc} \ln \left(\frac{p_{gc}}{q_{bgc}} \right)
\end{aligned}
$$

Choosing Step-size

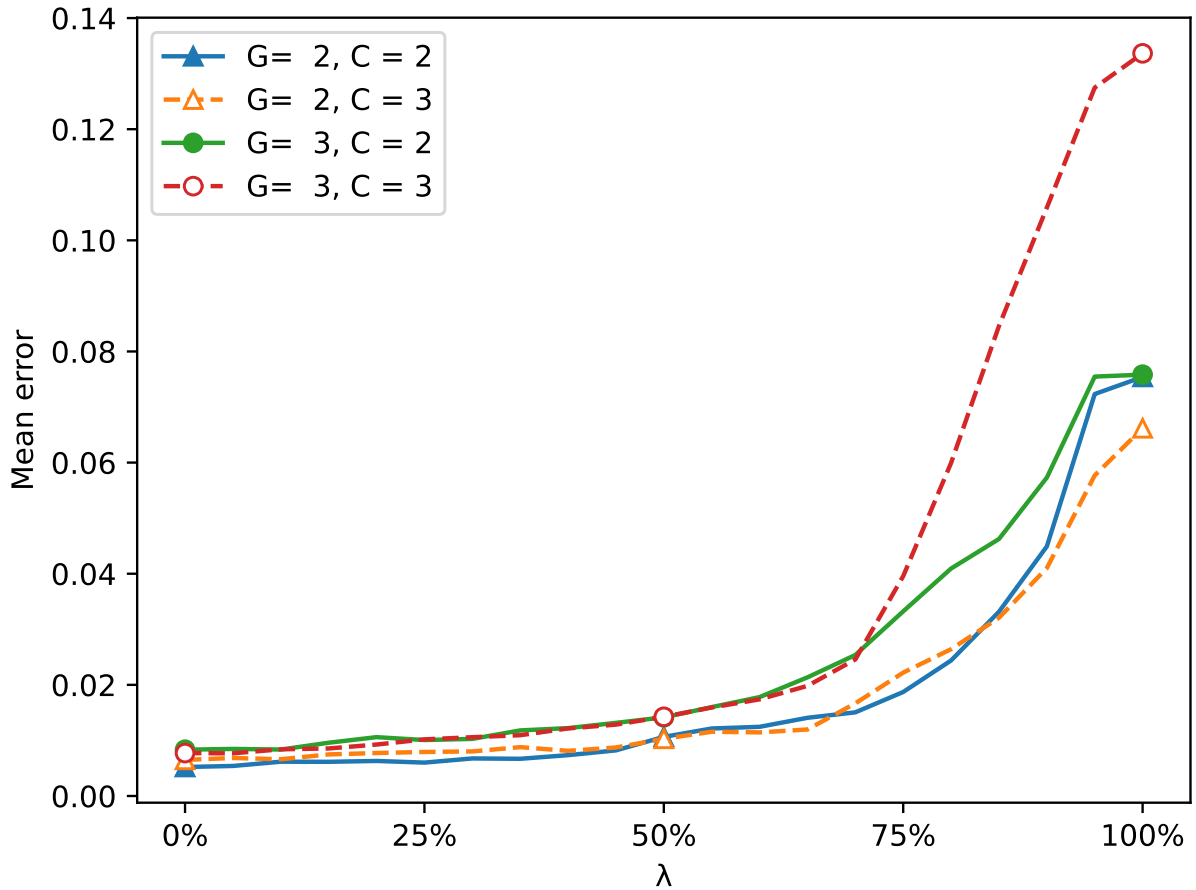
Figure: Mean correlation for various step-size, graph is log-scaled



Heterogeneity Effect on Predicted Probabilities

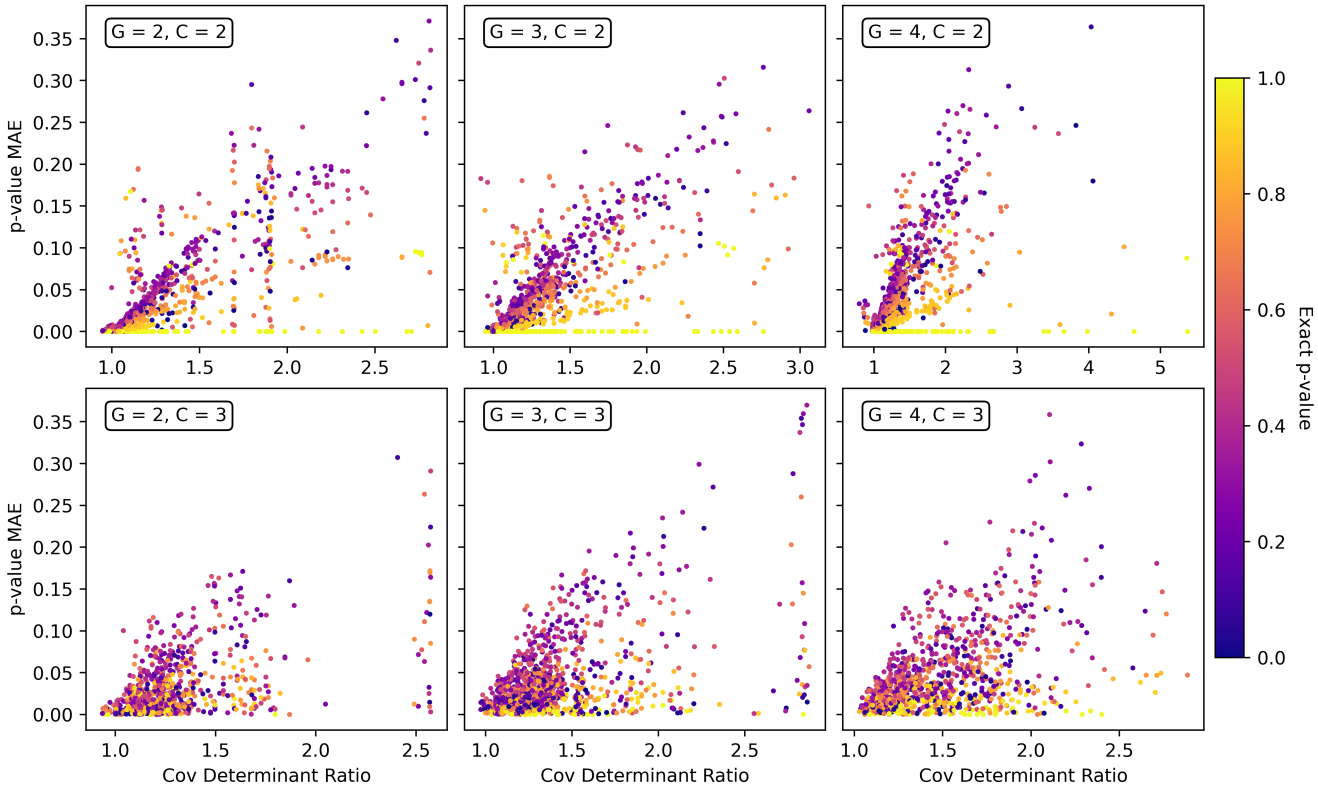
Consider λ: proportion of voters that get randomly mixed.

Figure: Mean absolute errors for different mixing parameters.



p-value Gap

Figure: Effect of covariance determinant increase in the mean absolute error of p-values



Group Aggregation

