Ecological inference with distribution regression

Seth Flaxman

10 May 2017
Department of Politics and International Relations
Ecological inference

- How to draw conclusions about individuals from aggregate-level data?

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>White</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>2,200</td>
<td>1,300</td>
</tr>
</tbody>
</table>
Ecological inference


<table>
<thead>
<tr>
<th>Year</th>
<th>District</th>
<th>Voting for the Democratic Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1986</td>
<td>12</td>
<td>95.65%</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>100.06</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>103.47</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>98.92</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>108.41</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>93.58</td>
</tr>
<tr>
<td>1988</td>
<td>12</td>
<td>95.67</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>102.64</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>105.00</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>100.20</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>111.05</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>97.49</td>
</tr>
</tbody>
</table>

Table 1.3 Sample Ecological Inferences: All Ohio State House Districts Where an African American Democrat Ran Against a White Republican, 1986–1990. Source: “Statement of Gordon G. Henderson,” presented as part of an exhibit in federal court. Figures above 100% are logically impossible.

Source: King, 1997
Ecological inference: a modern approach

We (often) have unlabeled individual level data!

- Learning from label proportions [e.g. Kueck and de Freitas 2005, Quadrianto et al 2009, Sheldon and Dietterich 2011, Patrini et al 2014]
- Distribution regression [e.g., Szabo et al 2015]
- Data privacy literature / de-anonymization [e.g. Narayanan and Shmatikov, 2009]
- Bayesian approaches [Jackson, Best, and Richardson 2006, Wakefield 2004 and 2008]
Electoral data

- Voting results per district, e.g. Obama 52%, Romney 48%
- Individual-level demographic data from the American Community Survey: race, education, income, occupation, time to work, etc for 5% of US residents
- Goal: infer how demographic subgroups voted (by district!)
Figure: Electoral data for 67 counties (left), census data for 127 public use microdata areas (not shown). Merging results in 37 electoral regions (right).
“Ground truth”: exit polls

- Costly, only carried out in 18 states in 2012
- Limited number of demographic variables
- Not representative at substate level
- Unreliable
- Other sources of ground truth: public polls, voter file, post-election surveys
My new approach

- Insight: ecological inference with individual level data can be framed as distribution regression
- Goal: scalable spatially varying Bayesian model
- Approach: new Bayesian distribution regression method with GPs and fast kernel methods

---

Distribution regression

Individual-level data with group-level labels:

\[ \left( \{ x_1^j \}_{j=1}^{N_1}, y_1 \right), \left( \{ x_2^j \}_{j=1}^{N_2}, y_2 \right), \ldots, \left( \{ x_n^j \}_{j=1}^{N_n}, y_n \right) \]
Distribution regression

Individual-level data with group-level labels:

\[
(\{x^j_1\}_{j=1}^{N_1}, y_1), (\{x^j_2\}_{j=1}^{N_2}, y_2), \ldots (\{x^j_n\}_{j=1}^{N_n}, y_n)
\]

Learn a function:

\[
f : \{x^j\}_{j=1}^N \rightarrow y
\]
Distribution regression

Individual-level data with group-level labels:

\[(\{x^j_1\}_{j=1}^{N_1}, y_1), (\{x^j_2\}_{j=1}^{N_2}, y_2), \ldots, (\{x^j_n\}_{j=1}^{N_n}, y_n)\]

Learn a function:

\[f : \{x^j\}_{j=1}^{N} \rightarrow y\]

Adapt it for ecological inference: make predictions for subgroups.

\[\hat{f}(\{x^j_{\text{women}}\}_{j=1}^{N_1})\]
Learning from distributions


- Distribution regression / distribution classification relies on the kernel mean embedding [see Muandet et al 2017’s survey]

- Given kernel $k(x, \cdot)$, RKHS $\mathcal{H}_k$, and corresponding embedding $\phi(x) \in \mathcal{H}_k$, consider a measure with $X \sim \mathcal{P}$. Then define:

$$\mu_\mathcal{P} := E[\phi(X)] = \int_X \phi(x) d\mathcal{P}(x) \quad (1)$$

Obvious empirical estimator for samples $x_1, \ldots, x_n$:

$$\hat{\mu}_\mathcal{P} := \frac{1}{n} \sum_i \phi(x_i) \quad (2)$$

- Learning: use any supervised learning method to learn a function $f(\mu_\mathcal{P})$. 

Distribution embedding illustration

Figure: Each distribution is mapped into the reproducing kernel Hilbert space via an expectation operation. (Source: Muandet et al 2017)
My new approach: distribution regression

% vote for Obama

% vote for Obama

feature space

region 1

region 2

region 3

women

men

both
Bayesian distribution regression

- Estimate $\hat{\mu}_1, \ldots, \hat{\mu}_n \in \mathcal{R}^n$ using kernel embeddings:

  $$\hat{\mu}_i = \frac{1}{N} \sum_j k(x^j_i, \cdot) = \frac{1}{N} \sum_j \phi(x^j_i)$$

- Use GP logistic regression

- Additive kernels with a spatial component:

  $$K_{ij} = \sigma^2_x \langle \hat{\mu}_i, \hat{\mu}_j \rangle + k_s(s_i, s_j)$$

  $$\mathbf{f} \sim \mathcal{GP}(0, \mathbf{K})$$

  $$k_i|\mathbf{f}_i \sim \text{Binomial}(n_i, \logit^{-1}(f_i))$$

  Obama received $k_i$ out of $n_i$ votes in region $i$.

- Make predictions for demographic subgroups:

  $$\hat{\mathbf{f}}(\mu_i^{\text{women}}, s_i)$$
Kernel details

- Demographic attributes (Gaussian RBF):
  - Standardize coordinates
  - Expand discrete attributes:
    (low, medium, high income) → ([1 0 0], [0 1 0], [0 0 1]).
  - Use random Fourier features for speed:
    \[ k(x, x') = \langle \phi(x), \phi(x') \rangle \approx \langle \hat{\phi}(x), \hat{\phi}(x') \rangle \text{ with } \hat{\phi}(x) \in \mathbb{R}^{2048}. \]

- Spatial attributes with Matérn-$$\frac{3}{2}$$:

  \[ k(s, s') = (1 + \rho \| s - s' \|) \exp(-\rho \| s - s' \|) \]

Millions of observations, but the covariance matrix is 843 × 843 for the 843 electoral regions.
Algorithm details

- One pass through census data to create mean embeddings:

\[
\hat{\mu}_1 = \frac{\sum_j w_1^j \phi(x_1^j)}{\sum_j w_1^j}, \ldots, \hat{\mu}_n = \frac{\sum_j w_n^j \phi(x_n^j)}{\sum_j w_n^j}
\]  

- Setup GP regression:

\[
f \sim \mathcal{GP}(0, \sigma_x^2 K_x + \sigma_s^2 K_s)
\]

\[
k_i | f_i \sim \text{Binomial}(n_i, \logit^{-1}(f_i))
\]

- Laplace approximation for hyperparameter learning

\[
\theta = [\sigma_x, \sigma_s, \rho] \text{ w/ marginal likelihood:}
\]

\[
\arg\max_\theta p(y|\theta)
\]

- Bayesian posterior inference to make predictions for latent \(f\) at new “locations”:

\[
p(f_{\text{men}}^*|y, \hat{\theta})
\]
Experiments

Exit poll women

Ecological regression women

Exit poll men

Ecological regression men
Experiments

[Graphs showing ecological regression for women and men with exit poll data.]
Experiments

The graph shows the Obama support gender gap (percentage points) across different geographic regions. The x-axis represents the Obama support gender gap in percentage points, ranging from -5% to 15%, while the y-axis represents the fraction of geographic regions. The distribution peaks around the 0% mark, indicating a central tendency in the gender gap across the regions.
Refinements for 2016 election\textsuperscript{2}

- Explicitly model non-voters:

\[ y_i = [\text{Clinton votes, Trump votes, Non-votes and third party votes}]^\top \]

- Multinomial likelihood with softmax link, fit with penalized MLE with group lasso and $L_2$ penalty

- More interpretable / richer feature representation to allow for exploratory analysis / calculation of marginal effects:

\[ \phi(x_i^j) := [\phi_1(x_{r1}^j), \ldots, \phi_d(x_{rd}^j)]^\top \] (4)

- Incorporation of some exit polling data as extra set of labeled distributions

\textsuperscript{2}arXiv:1611.03787
### Results for 2016 Presidential Election

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Clinton</th>
<th>Trump</th>
<th>Frac. electorate</th>
<th>Participation rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>0.45</td>
<td>0.55</td>
<td>0.47</td>
<td>0.50</td>
</tr>
<tr>
<td>Women</td>
<td>0.56</td>
<td>0.44</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>18–29 year olds</td>
<td>0.62</td>
<td>0.38</td>
<td>0.17</td>
<td>0.42</td>
</tr>
<tr>
<td>30–44</td>
<td>0.54</td>
<td>0.46</td>
<td>0.25</td>
<td>0.54</td>
</tr>
<tr>
<td>45–64</td>
<td>0.46</td>
<td>0.54</td>
<td>0.39</td>
<td>0.58</td>
</tr>
<tr>
<td>65 and older</td>
<td>0.45</td>
<td>0.55</td>
<td>0.18</td>
<td>0.47</td>
</tr>
</tbody>
</table>
### Results for 2016 Presidential Election

<table>
<thead>
<tr>
<th></th>
<th>Clinton</th>
<th>Trump</th>
<th>Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language other than English spoken at home</td>
<td>0.74</td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>Mobility = lived here one year ago</td>
<td>0.45</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>Mobility = moved here from outside US and Puerto Rico</td>
<td>0.60</td>
<td>0.40</td>
<td>0.47</td>
</tr>
<tr>
<td>Mobility = moved here from inside US or Puerto Rico</td>
<td>0.56</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>Active duty military</td>
<td>0.45</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>Not enrolled in school</td>
<td>0.45</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>Enrolled in a public school or public college</td>
<td>0.61</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Enrolled in private school, private college, or home school</td>
<td>0.66</td>
<td>0.34</td>
<td>0.53</td>
</tr>
</tbody>
</table>
## Results for 2016 Presidential Election

<table>
<thead>
<tr>
<th>Income Level</th>
<th>Clinton</th>
<th>Trump</th>
<th>Frac</th>
<th>Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>personal income ≤ 50000 &amp; men</td>
<td>0.56</td>
<td>0.44</td>
<td>0.25</td>
<td>0.37</td>
</tr>
<tr>
<td>personal income ≤ 50000 &amp; women</td>
<td>0.63</td>
<td>0.37</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td>50000 &lt; personal income ≤ 100000 &amp; men</td>
<td>0.40</td>
<td>0.60</td>
<td>0.15</td>
<td>0.67</td>
</tr>
<tr>
<td>50000 &lt; personal income ≤ 100000 &amp; women</td>
<td>0.53</td>
<td>0.47</td>
<td>0.13</td>
<td>0.84</td>
</tr>
<tr>
<td>personal income &gt; 100000 &amp; men</td>
<td>0.49</td>
<td>0.51</td>
<td>0.08</td>
<td>0.70</td>
</tr>
<tr>
<td>personal income &gt; 100000 &amp; women</td>
<td>0.62</td>
<td>0.38</td>
<td>0.03</td>
<td>0.80</td>
</tr>
</tbody>
</table>
### Exploratory results

<table>
<thead>
<tr>
<th>feature</th>
<th>deviance</th>
<th>frac.deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAC3P - race coding</td>
<td>0.04</td>
<td>0.86</td>
</tr>
<tr>
<td>ethnicity interacted with has degree</td>
<td>0.04</td>
<td>0.74</td>
</tr>
<tr>
<td>schooling attainment</td>
<td>0.04</td>
<td>0.72</td>
</tr>
<tr>
<td>ANC2P - detailed ancestry</td>
<td>0.04</td>
<td>0.83</td>
</tr>
<tr>
<td>OCCP - occupation</td>
<td>0.04</td>
<td>0.75</td>
</tr>
<tr>
<td>COW - class of worker</td>
<td>0.04</td>
<td>0.67</td>
</tr>
<tr>
<td>ANC1P - detailed ancestry</td>
<td>0.05</td>
<td>0.77</td>
</tr>
<tr>
<td>NAICSP - industry code</td>
<td>0.05</td>
<td>0.71</td>
</tr>
<tr>
<td>RAC2P - race code</td>
<td>0.05</td>
<td>0.70</td>
</tr>
<tr>
<td>age interacted with usual hours worked per week (WKHP)</td>
<td>0.05</td>
<td>0.69</td>
</tr>
<tr>
<td>sex interacted with ethnicity</td>
<td>0.05</td>
<td>0.65</td>
</tr>
<tr>
<td>MSP - marital status</td>
<td>0.05</td>
<td>0.61</td>
</tr>
<tr>
<td>FOD1P - field of degree</td>
<td>0.05</td>
<td>0.61</td>
</tr>
<tr>
<td>ethnicity</td>
<td>0.06</td>
<td>0.57</td>
</tr>
<tr>
<td>RAC1P - recoded race</td>
<td>0.06</td>
<td>0.54</td>
</tr>
<tr>
<td>sex interacted with age</td>
<td>0.06</td>
<td>0.57</td>
</tr>
<tr>
<td>has degree interacted with age</td>
<td>0.06</td>
<td>0.55</td>
</tr>
<tr>
<td>age interacted with personal income</td>
<td>0.06</td>
<td>0.76</td>
</tr>
<tr>
<td>sex interacted with hours worked per week</td>
<td>0.06</td>
<td>0.62</td>
</tr>
<tr>
<td>personal income interacted with hours worked per week</td>
<td>0.06</td>
<td>0.69</td>
</tr>
<tr>
<td>personal income</td>
<td>0.06</td>
<td>0.59</td>
</tr>
<tr>
<td>RACSOR - single or multiple race</td>
<td>0.07</td>
<td>0.42</td>
</tr>
<tr>
<td>has degree interacted with hours worked per week</td>
<td>0.07</td>
<td>0.59</td>
</tr>
<tr>
<td>hispanic</td>
<td>0.07</td>
<td>0.56</td>
</tr>
<tr>
<td>sex interacted with personal income</td>
<td>0.07</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Marginal results

Clinton/Trump Vote Share

100% Clinton
50%/50%
100% Trump

Participation Rate

0% 20% 40% 60% 80% 100%

Hispanic, college degree
White, college degree
Black, college degree
Black, no degree
Amerindian, no degree
Other/Multiracial, no degree
Asian, college degree
Asian, no degree
Other/Multiracial, college degree

Hispanic, no degree
White, no degree
Amerindian, college degree
Conclusion: ecological inference

- New ecological inference method through Bayesian distribution regression
- Scalable to millions of observations through random features
- Good empirical results
- Realistic uncertainty intervals
- Simple method [off-the-shelf tools]
- Python package by Dougal Sutherland and replication code
- Next steps: fully Bayesian version of multinomial model, learning richer feature representations, validation on ground truth
Future work

- Fully Bayesian method (pre-print will be on arXiv later this week)
- Mean embedding as an “ecological” covariate in hierarchical modeling
- Poll design: how to define likely voters?
- Imputing demographic characteristics of voters from voter file
- Many other application areas (public health, etc)

Thanks! Papers and code: www.sethrf.com and on arXiv