PyEI: A Python package for ecological inference

29 April 2021

Summary

An important question in some US voting rights cases and redistricting litigation is whether and to what degree voting is racially polarized. In the setting of voting rights cases, ecological inference involves using observed data about voting outcomes in each of a number of precincts and demographic information about each precint to infer voting patterns within each demographic group.

More generally, we can think of ecological inference as seeking to use knowledge about the margins of a set of tables (Figure 1) to infer associations between the row and column variables, by making (typically probablistic) assumptions about the underlying associations. In the setting of assessing racially polarized voting, each column in a table like the one in Figure 1 corresponds to a candidate or voting outcome, each row to a racial group, and each table to a precinct. Ecological inference methods then use the known counts of voting outcomes in each precinct and the known counts of people in demographic groups in each precinct to make inferences about the distribution of voting outcomes within each demographic group, thus addressing questions like: "What percentage of voters in Group 1 voted for candidate A?"). The "two by two" ecological inference problem in this example, where we have two groups and two voting outcomes, is a special case of the more general "R by C" ecological inference, in which we may have more than two groups or voting outcomes. Ecological inference is also applicable in other fields, such as epidemiology and sociology.

	GROUP A	GROUP B	
GROUP 1	?	?	Total in group 1
GROUP 2	?	?	Total in group 2
	Total in group A	Total in group B	

Figure 1: In "two by two" ecological inference we have information about the marginal counts for a set of tables like the one above and would like to make inferences about, for example the number or proportion of members of group 1 who were also in group A.

Statement of need

The results of ecological inference for inferring racially polarized voting are used in US voting rights cases; therefore, easy to use and high quality tools for performing ecological inference are of practical interest. There is a need for an ecological inference library that brings together a variety of ecological inference methods in one place and makes easy crucial tasks such as: quantifying the uncertainty associated with ecological inference results under a given model; making comparisons between methods; and bringing relevant diagnostic tools to bear on ecological inference methods. To address this need, we introduce `PyEI`, a Python package for ecological inference.

`PyEI` is meant to be useful to two main groups of researchers. First, it serves application-oriented researchers and practitioners who seek to run ecological inference on domain data (e.g. voting data), report results, and understand the uncertainty related to those results. Second, it facilitates exploration and benchmarking for researchers who are seeking to understand properties of existing ecological inference methods in different settings and/or develop new statistical methods for ecological inference.

`PyEI` brings together the following ecological inference methods in a common framework alongside plotting, reporting, and diagnostic tools:

- Goodman's ecological regression (Goodman 1953) and a Bayesian linear regression variant
- A truncated-normal based approach (King 1997)
- Binomial-Beta hierarchical models (King, Rosen, and Tanner 1999)
- Dirichlet-Multinomial hierarchical models (Rosen et al. 2001)
- A Bayesian hierarchical method for 2x2 EI following the approach of Wakefield (2004)

(Note in several of these cases, `PyEI` includes modifications to the models as originally proposed in the cited literature, such as reparametrizations or other changes to upper levels of the hierarchical models in order to ease sampling difficulties.)

`PyEI` is intended to be easily extensible, so that additional methods from the literature can continue to be incorporated (see e.g. James Greiner and Quinn (2009), currently implemented in the R package `RxCEcolInf` (Greiner, Baines, and Quinn 2019)), and newly developed statistical methods for ecological inference can be included and conveniently compared with existing methods.

Several R libraries implementing different ecological inference exist, such as `ei` (King and Roberts 2016), `eiCompare` (Collingwood et al. 2020), `eiPack` (Lau, Moore, and Kellermann 2020), and `RxCEcolInf` (Greiner, Baines, and Quinn 2019). `PyEI` presents a Python-based option that researchers who primarily use Python may appreciate. `PyEI` also incorporates the following key features and characteristics. First, the Bayesian hierarchical methods implemented in `PyEI` rest on modern probabilistic programming tooling (Salvatier, Wiecki, and

Fonnesbeck 2016) and MCMC methods such as the No U-Turn Sampler (NUTS) (Hoffman and Gelman 2014). Effective sample size is a measure of how the variance of the mean of drawn samples compare to the variance of i.i.d. samples from the posterior distribution (Gelman et al. 2013). In Metropolis-Hastings, the number of evaluations of the log-posterior required scales linearly with the effective sample size (after burn-in), while in Hamiltonian Monte Carlo approaches (e.g. NUTS), the scaling goes as only the fourth root of the number of evaluations of the gradient of the the log-posterior (Neal 2011). Second, integration with existing tools `PyMC` (Salvatier, Wiecki, and Fonnesbeck 2016) and ArViZ (Kumar et al. 2019) makes the results amenable to state of the art diagnostics (e.g. convergence diagostics) and some reasonable checks are automatically performed. Third, summary and plotting utilities for reporting, visualizing, and comparing results are included (see example plots below), with an emphasis on visualizations and reports that make clear the uncertainty of estimates under a model. Lastly, clear documentation is provided, including a set of introductory and example notebooks.

Examples of plotting functionality

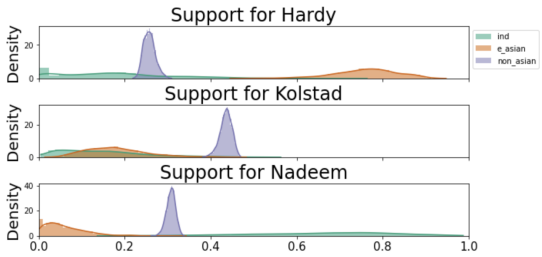

Figure 2: KDE plots for visualing uncertainty of support for candidates within each group.

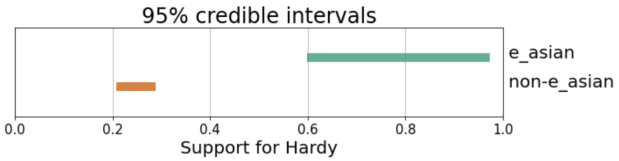

Figure 3: Bayesian credible intervals for support of candidates within groups.

Notes for draft

seeks to address some limitations of the existing R libraries, namely that:

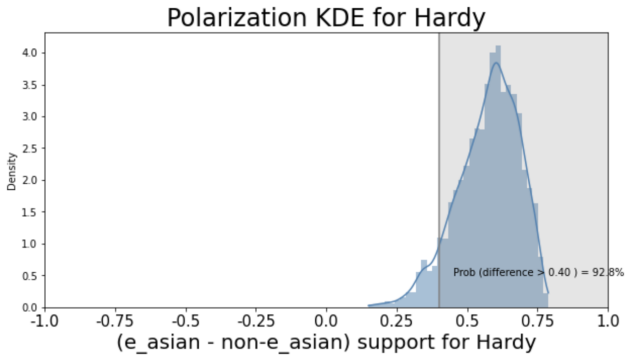

Figure 4: Visualizing and quantifying degree of polarization.

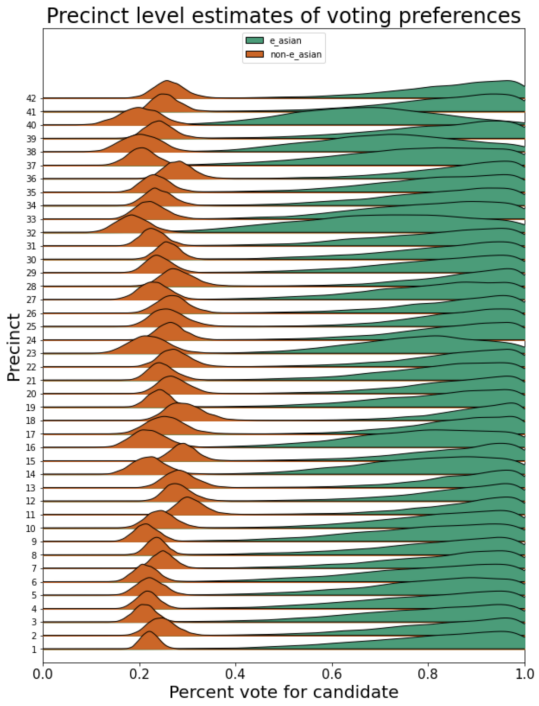

Figure 5: Visualizing estimates and uncertainty for precinct-level estimates.

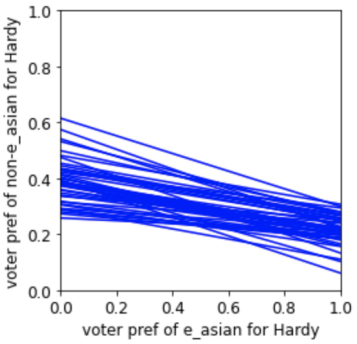

Figure 6: "Tomography" plots for two-by-two ecological inference.

- different ecological inference methods are spread out across different libraries
- existing implementations may use slower-to-converge MCMC approaches (BENCHMARK)
- existing libraries do not all expose samples directly (CHECK THAT THIS IS STILL THE CASE) and do not incorporate convergence-related checks and warnings
- not all libraries are oriented towards careful uncertainty quantification (CHECK AND EXPLAIN/SAY THIS MORE CAREFULLY)

Notes on R packages

- `ei`
  - main function is inplementation of King's EI (truncated normal~1997)... functionality is mainly oriented toward 2x2. Does include RxC generalization and Goodman ER
  - Can get samples of psi, aggregated betas (at the polity level), and precinct-level betas. The first two are easily grabed by eiread() function. the latter doesn't seem to be designed to be grabbed, but the samples can be read.
  - convergence testing/diagnostics are not transparent (if they are done at all)... there is a 'resamp' parameter that appears to be some sort of diagnostic, but I'm not sure what it is exactly (not explained much in documentation)
  - standard errors and 80% "confidence intervals" (these appear to be just .1 and .9 percentiles of sample) for precinct-level betas. Also gives standard errors for psis and aggregate betas.
- `eiPack`
  - includes multinomial dirichlet (ala Rosen), Goodman ER, ER w/ Bayesian normal regression, and some plotting functionality (density and bounds)
  - outputs draws for alphas, betas, and cell counts as well as acceptance

ratios of these variable draws
  - no convergence tests/diagnostics included or described. Presumably the outputs can make use of the functionality of the coda package to perform these, but user would have to write this script.
  - Has plotting functionality for unit (precinct)-level credible intervals, though this appears to be just for visualization and less for actually grabbing/using these values... no functionality/attention to uncertainty for other measures of interest
- `eiCompare`
  - Appears to be primarily a wrapper around `ei` and `eiPack` packages. Some added funcationality (including capturing uncertainty, visualization, and comparison of results across methods). Some instances appear to be literally rewritten from these packages with minor/trivial tweaks, rather than significantly different implementations.
- `RxCEcolInf`
  - package really just includes the model from the 2009 Greiner/Quinn paper
  - outputs draws for: internal cell counts, thetas, mu, and the standard deviations and correlations in sigma.
  - convergence tests not directly incorporated, but returns mcmc object with intention of using functionality of R's coda package. Documentation examples show using coda's Geweke's as well as Heidelberger and Welch's convergence diagnostics, but ackwledges that chains created by `RxCEcolInf` will cause error in coda's Gelman-Rubin diagnostic.
  - the `RxCEcolInf` package itself does not include careful uncertainty quantification functionality, but digging through some of their replication code from their paper does show some examples of uncertainty calculations and credible intervals, though not particularly well documented/clear how to apply

Acknowledgements

We acknowledge contributions from ...

References

Collingwood, Loren, Ari Decter-Frain, Hikari Murayama, Pratik Sachdeva, and Juandalyn Burke. 2020. *eiCompare: Compares Ecological Inference, Goodman, Rows by Columns Estimates*. https://CRAN.R-project.org/package=eiCompare.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC press.

Goodman, Leo A. 1953. "Ecological Regressions and Behavior of Individuals." *American Sociological Review*.

Greiner, D. James, Paul Baines, and Kevin M. Quinn. 2019. *RxCEcolInf: 'R x c Ecological Inference with Optional Incorporation of Survey Information'*. https://CRAN.R-project.org/package=RxCEcolInf.

Hoffman, Matthew D, and Andrew Gelman. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–623.

James Greiner, D, and Kevin M Quinn. 2009. "R× c Ecological Inference: Bounds, Correlations, Flexibility and Transparency of Assumptions." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172 (1): 67–81.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

King, Gary, and Molly Roberts. 2016. *Ei: Ecological Inference*. https://CRAN.R-project.org/package=ei.

King, Gary, Ori Rosen, and Martin A Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research* 28 (1): 61–90.

Kumar, Ravin, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. 2019. "ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python." *Journal of Open Source Software* 4 (33): 1143. https://doi.org/10.21105/joss.01143.

Lau, Olivia, Ryan T. Moore, and Michael Kellermann. 2020. *eiPack: Ecological Inference and Higher-Dimension Data Management*. https://CRAN.R-project.org/package=eiPack.

Neal, Radford. 2011. "MCMC Using Hamiltonian Dynamics." *Handbook of Markov Chain Monte Carlo* 2 (11): 2.

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The r× c Case." *Statistica Neerlandica* 55 (2): 134–56.

Salvatier, John, Thomas V Wiecki, and Christopher Fonnesbeck. 2016. "Probabilistic Programming in Python Using PyMC3." *PeerJ Computer Science* 2: e55.

Wakefield, Jon. 2004. "Ecological Inference for 2× 2 Tables." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167 (3): 385–425.