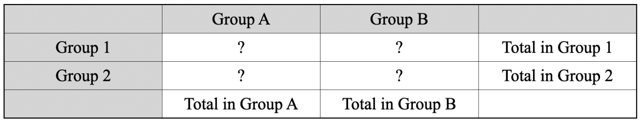
PyEI: A Python package for ecological inference

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 precinct 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 context 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)

(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 (for example, work is underway to add the method of James Greiner and Quinn (2009), currently implemented in the R package `RxCEcolInf` (Greiner, Baines, and Quinn 2019)), and so that newly developed statistical methods for ecological inference can be included and conveniently compared with existing methods.

Several R libraries implementing different ecological inference methods 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). In addition to presenting a Python-based option that researchers who primarily use Python may appreciate, `PyEI` 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 gradient-based MCMC methods such as the No U-Turn Sampler (NUTS) (Hoffman and Gelman 2014). Using NUTS where possible should allow for faster convergence than existing implementations that we are aware of that rest primarily on Metropolis-Hastings and Gibbs sampling steps. 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 for a given effective sample size scales linearly with the dimensionality of the parameter space, while in Hamiltonian Monte Carlo approaches such as NUTS, the the number of required evaluations of the gradient of the the log-posterior scales only as the fourth root of the dimension (Neal 2011).

Second, integration with the existing tools PyMC3 (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.

Acknowledgements

We thank Colin Carroll, Moon Duchin, JN Matthews, Matthew Sun, and Thomas Weighill for their helpful contributions.

Examples of plotting functionality

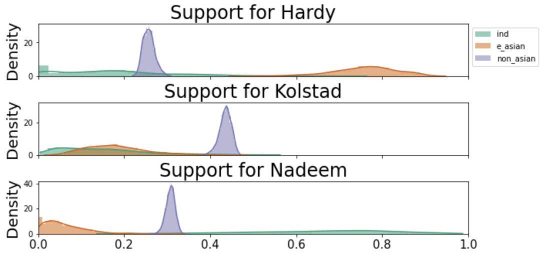


Figure 2: KDE plots for visualing uncertainty of support for candidates within each group.

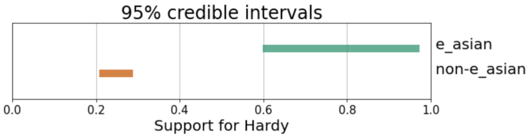


Figure 3: Bayesian credible intervals for support of candidates within groups.

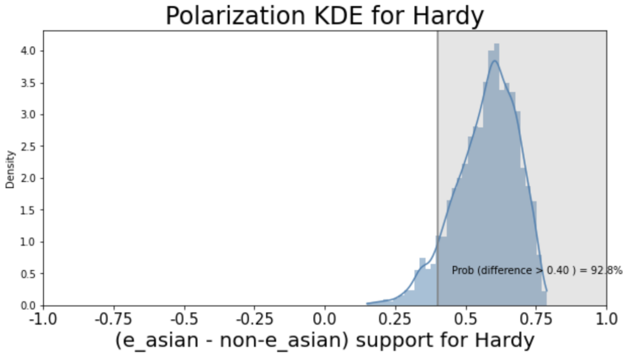


Figure 4: Visualizing and quantifying degree of polarization.

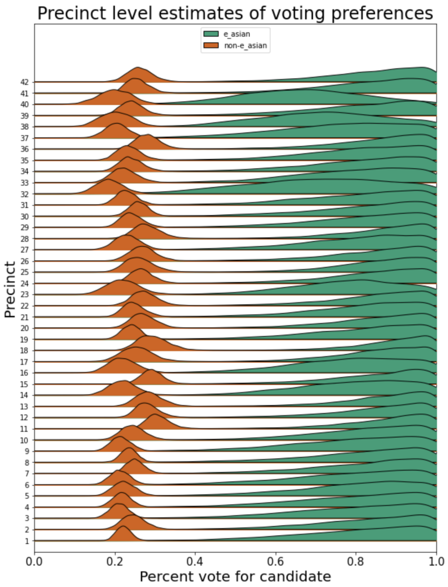


Figure 5: Visualizing estimates and uncertainty for precinct-level estimates.

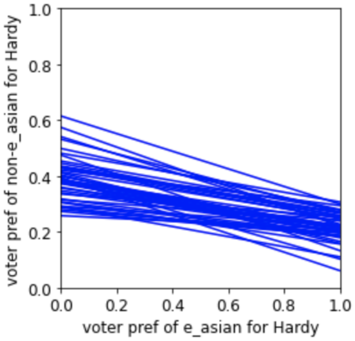


Figure 6: "Tomography" plots for two-by-two ecological inference.

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.