PyEI: A Python package for ecological inference

10 May 2021

Summary

An important question in some voting rights and redistricting litigation in the U.S. is whether and to what degree voting is racially polarized. In the setting of voting rights cases, there is a family of methods called “ecological inference” (see especially King 1997) that uses observed data, pairing voting outcomes with demographic information for each precinct in a given polity, to infer voting patterns for each demographic group.

More generally, we can think of ecological inference as seeking to use knowledge about the margins of a set of tables (Table 1) to infer associations between the row and column variables, by making (typically probabilistic) assumptions. In the context of assessing racially polarized voting, a table like the one in Table 1 will correspond to a precinct, where each column corresponds to a candidate or voting outcome and each row to a racial group. Ecological inference methods then use the vote counts and demographic data for each precinct to make inferences about the overall voting preferences by demographic group, thus addressing questions like: “What percentage of East Asian voters voted for Hardy?” This example is an instance of what is referred to in the literature as “R by C” ecological inference, where here we have R = 2 groups and C = 3 voting outcomes. `PyEI` was created to support performing ecological inference with voting data; however, ecological inference methods also applicable in other fields, such as epidemiology (Elliot et al. 2000) and sociology (Goodman 1953).

Table 1: In ecological inference we have information about the marginal counts for a set of tables like the one here and would like to make inferences about, for example, the number or proportion of East Asian voters who voted for Hardy. The system is underdetermined and ecological inference methods proceed by making statistical assumptions.

	Hardy	Kolstad	Nadeem	
East Asian	?	?	?	Total East Asian
non- East Asian	?	?	?	Total non- East Asian
	Total for Hardy	Total for Kolstad	Total for Nadeem	

Statement of need

The results of ecological inference for inferring racially polarized voting are routinely used in US voting rights cases (King 1997); 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 to facilitate 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 the 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 2×2 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)). 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 eiPack (Lau, Moore, and Kellermann 2020), RxCEcolInf (Greiner, Baines, and Quinn 2019), ei (King and Roberts 2016), and eiCompare (Collingwood et al. 2020). 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; Betancourt 2018). Using NUTS where possible should allow for faster convergence than existing implementations that rest primarily on Metropolis-Hastings and Gibbs sampling steps. Consider effective sample size, which is a measure of how the variance of the mean of drawn samples compare to the variance of independent samples from the posterior distribution (or, very roughly, how "effective" the samples are for computing the posterior mean, compared to independent samples) (Gelman et al. 2013). Under certain assumptions on the target posterior distribution, 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 number of required evaluations of the gradient of the log-posterior scales only as the fourth root of the dimension (Neal 2011). Reasonable scaling with the dimensionality of the parameter space is important in ecological inference, as that dimensionality is large when there are many precincts.

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 diagnostics) and some reasonable checks are automatically performed.

Third, summary and plotting utilities for reporting, visualizing, and comparing results are included (e.g. Figure 1, Figure 2), with an emphasis on visualizations and reports that clarify the uncertainty of estimates under a model.

Lastly, clear documentation is provided, including a set of introductory and example notebooks.

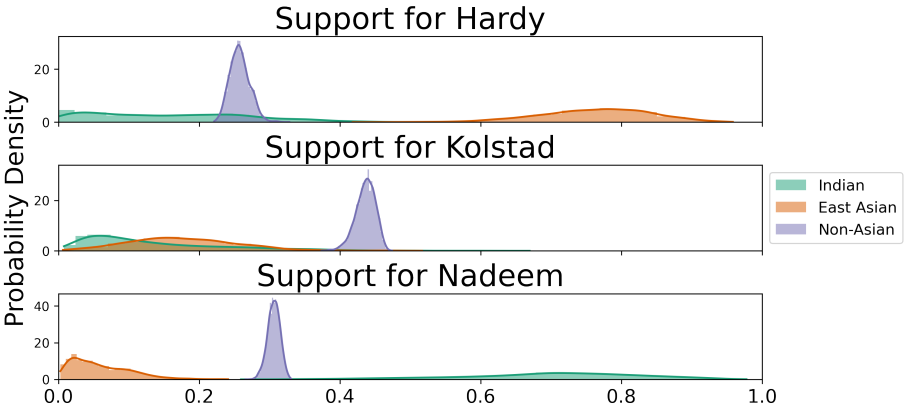

Figure 1: Kernel density estimation plots for visualizing uncertainty of support for candidates within each group.

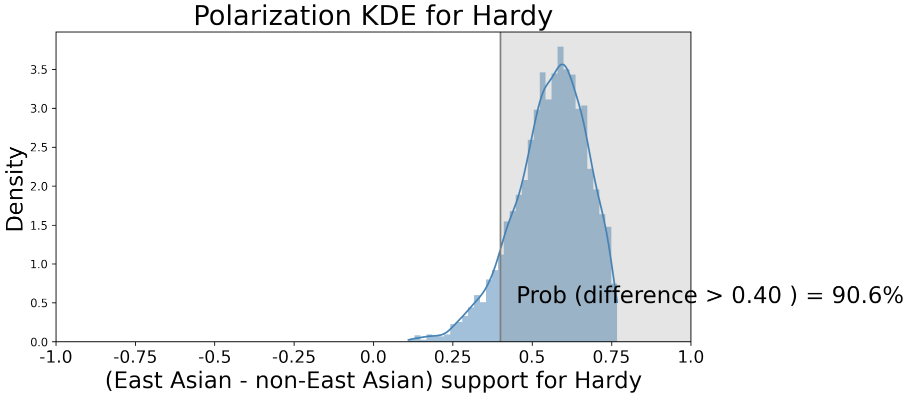

Figure 2: Visualizing and quantifying degree of polarization.

Acknowledgments

This software development is part of a research project comparing methods, joint with Moon Duchin and Thomas Weighill. We thank Colin Carroll, JN Matthews, and Matthew Sun for their helpful contributions to PyEI.

References

Betancourt, Michael. 2018. "A Conceptual Introduction to Hamiltonian Monte Carlo." http://arxiv.org/abs/1701.02434.

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.

Elliot, Paul, Jon C Wakefield, Nicola G Best, David John Briggs, and others. 2000. *Spatial Epidemiology: Methods and Applications*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198515326.001.0001.

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*. https://doi.org/10.2307/2088121.

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. https://doi.org/10.1201/b10905-7.

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.