eiPack: R × C Ecological Inference and Higher-Dimension Data Management

by Olivia Lau, Ryan T. Moore, and Michael Kellermann

Introduction

Ecological inference (EI) models allow researchers to infer individual-level behavior from aggregate data when individual-level data is unavailable. Table 1 shows a typical unit of ecological analysis: a contingency table with observed row and column marginals and unobserved interior cells.

	col_1	col_2	...	col_C	
row_1	N_{11i}	N_{12i}	...	N_{1Ci}	$N_{1\cdot i}$
row_2	N_{21i}	N_{22i}	...	N_{2Ci}	$N_{2\cdot i}$
...	...	...		...	...
row_R	N_{R1i}	N_{R2i}	...	N_{RCi}	$N_{R\cdot i}$
	$N_{\cdot 1i}$	$N_{\cdot 2i}$	...	$N_{\cdot Ci}$	N_i

Table 1: A typical $R \times C$ unit in ecological inference; red quantities are typically unobserved.

In ecological inference, challenges arise because information is lost when aggregating across individuals, a problem that cannot be solved by collecting more aggregate-level data. Thus, EI models are unusually sensitive to modeling assumptions. Testing these assumptions is difficult without access to individual-level data, and recent years have witnessed a lively discussion of the relative merits of various models (Wakefield, 2004).

Nevertheless, there are many applied problems in which ecological inferences are necessary, either because individual-level data is unavailable or because the aggregate-level data is considered more authoritative. The latter is true in the voting rights context in the United States, where federal courts often base decisions on evidence derived from one or more EI models (Cho and Yoon, 2001). While packages such as **MCMCpack** (Martin and Quinn, 2006) and **eco** (Imai and Lu, 2005), provide tools for 2×2 inference, this is insufficient in many applications. In **eiPack**, we implement three existing methods for the general case in which the ecological units are $R \times C$ tables.

Methods and Data in eiPack

The methods currently implemented in **eiPack** are the method of bounds (Duncan and Davis, 1953), ecological regression (Goodman, 1953), and the Multinomial-Dirichlet model (Rosen et al., 2001).

The functions that implement these models share several attributes. The ecological tables are defined using a common formula of the form `cbind(col1, ..., colC) ~ cbind(row1, ...,rowR)`. The row and column marginals can be expressed as either proportions or counts. Auxiliary functions renormalize the results for some subset of columns taken from the original ecological table, and appropriate `print`, `summary`, and `plot` functions conveniently summarize the model output.

In the following section, we demonstrate the features of **eiPack** using the (included) `senc` dataset, which contains individual-level party affiliation data for Black, White, and Native American voters in 8 counties in southeastern North Carolina. These counties include 212 precincts, which form the ecological units in this dataset. Because the data are observed at the individual level, the interior cell counts are known, allowing us to benchmark the estimates generated by each method.

Method of Bounds

The method of bounds (Duncan and Davis, 1953) uses the observed row and column marginals to calculate upper and lower bounds for functions of the interior cells of each ecological unit. The method of bounds is not a statistical procedure in the traditional sense; the bounds implied by the row and column marginals are deterministic and there is no probabilistic model for the data-generating process.

As implemented in **eiPack**, the method of bounds allows the user to calculate for a specified column $k' \in k = \{1, \ldots, C\}$ the deterministic bounds on the proportion of individuals in each row who belong in that column. For each unit being considered, let j be the row of interest, k index columns, k' be the column of interest, k'' be the set of other columns considered, and $\tilde{k}$ be the set of columns excluded. For example, if we want the bounds on the proportion of Native American two-party registrants who are Democrats, j is Native American, k' is Democrat, k'' is Republican, and $\tilde{k}$ is No Party. The unit-level quantity of interest is

$$\frac{N_{jk'i}}{N_{jk'i} + \sum_{k \in k''} N_{jki}}$$

The lower and upper bounds on this quantity given by the observed marginals are, respectively:

$$\frac{\max(0, N_{ji} - \sum_{k \neq k'} N_{ki})}{\max(0, N_{ji} - \sum_{k \neq k'} N_{ki}) + \min(N_{ji}, \sum_{k \in k''} N_{ki})}$$

and

$$\frac{\min(N_{ji}, N_{k'i})}{\min(N_{ji}, N_{k'i}) + \max(0, N_{ji} - N_{k'i} - \sum_{k \in \tilde{k}} N_{ki})}$$

The intervals generated by the method of bounds can be analyzed in a variety of ways. Grofman (2000) suggests calculating the intersection of the unit-level bounds. If this intersection (calculated by **eiPack**) is non-empty, it represents the range of values that are consistent with the observed marginals in each of the ecological units.

Researchers and practitioners may also choose to restrict their attention to units in which one group dominates, since the bounds will typically be more informative in those units. **eiPack** allows users to set row thresholds to conduct this *extreme case analysis* (known as *homogeneous precinct analysis* in the voting context). For example, suppose the user is interested in the proportion of two-party White registrants registered as Democrats in precincts that are at least 90% White. **eiPack** calculates the desired bounds:

```
> out <- bounds(cbind(dem, rep, non) ~ cbind(black,
+   white, natam), data = senc, rows = "white",
+   column = "dem", excluded = "non",
+   threshold = 0.9, total = NULL)
```

These calculated bounds can then be represented graphically. Segments cover the range of possible values (the true value for each precinct is the red dot, not included in the standard bounds plot). In this example, the intersection of the precinct-level bounds is empty.

```
> plot(out, row = "white", column = "dem")
# add true values to plot
> idx <- as.numeric(rownames(out$bounds$white.dem))
> truth <- senc$whdem[idx]/(senc$white[idx]
+   -senc$non[idx])
> plot((1:length(idx)) / (length(idx) + 1), truth)
```

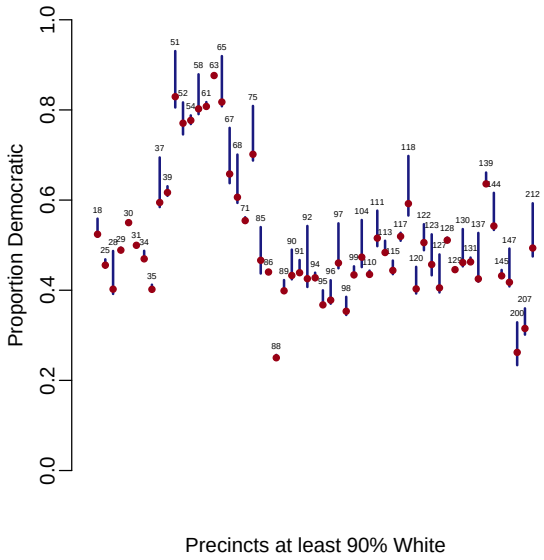

Figure 1: A plot of deterministic bounds.

Ecological Regression

In ecological regression (Goodman, 1953), observed row and column marginals are expressed as proportions and each column is regressed separately on the row proportions, thus performing C regressions. Regression coefficients then estimate the population internal cell proportions. For a given unit i, define

- X_{ri}, the proportion of individuals in row r,
- T_{ci}, the proportion of individuals in column c, and
- β_{rci}, the proportion of row r individuals in column c

The following identities hold:

$$T_{ci} = \sum_{r=1}^{R} \beta_{rci} X_{ri} \quad \text{and} \quad \sum_{c=1}^{C} \beta_{rci} = 1$$

Defining the population cell fractions β_{rc} such that $\sum_{c=1}^{C} \beta_{rc} = 1$ for every r, ecological regression assumes that $\beta_{rc} = \beta_{rci}$ for all i, and estimates the regression equations $T_{ci} = \beta_{rc} X_{ri} + \epsilon_{ci}$. Under the standard linear regression assumptions, including $E[\epsilon_{ci}] = 0$ and $Var[\epsilon_{ci}] = \sigma_c^2$ for all i, these regressions recover the population parameters β_{rc}. **eiPack** implements frequentist and Bayesian regression models (via `ei.reg` and `ei.reg.bayes`, respectively).

In the Bayesian implementation, we offer two options for the prior on β_{rc}. As a default, `truncate = FALSE` uses an uninformative flat prior that provides point estimates approaching the frequentist estimates (even when those estimates are outside the feasible range) as the number of draws $m \to \infty$. In cases where the cell estimates are near the boundaries, choosing `truncate = TRUE` imposes a uniform prior over the unit hypercube such that all cell fractions are restricted to the range $[0, 1]$.

Output from ecological regression can be summarized numerically just as in `lm`, or graphically using density plots. We also include functions to calculate estimates and standard errors of shares of a subset of columns in order to address questions such as, "What is the Democratic share of 2-party registration for each group?" For the Bayesian model, densities represent functions of the posterior draws of the β_{rc}; for the frequentist model, densities reflect functions of regression point estimates and standard errors calculated using the δ-method.

```
> out.reg <- ei.reg(cbind(dem, rep, non)
+   ~ cbind(black, white, natam), data = senc)
> lreg <- lambda.reg(out.reg,
    columns = c("dem", "rep"))
> density.plot(lreg)
```

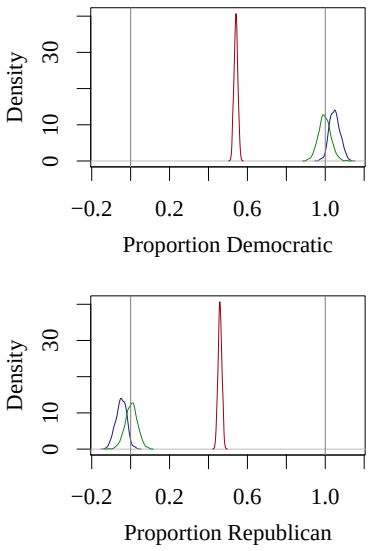

Figure 2: Density plots of ecological regression output.

Multinomial-Dirichlet (MD) model

In the Multinomial-Dirichlet model proposed by Rosen et al. (2001), the data is expressed as counts and a hierarchical Bayesian model is fit using a Metropolis-within-Gibbs algorithm implemented in C. Level 1 models the observed column marginals as multinomial (and independent across units); the choice of the multinomial corresponds to sampling with replacement from the population. Level 2 models the unobserved row cell fractions as Dirichlet (and independent across rows and units); Level 3 models the Dirichlet parameters as i.i.d. Gamma. More formally, without a covariate, the model is

$$\begin{aligned}
(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) &\overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \\
&\qquad \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri}) \\
(\beta_{r1i}, \ldots, \beta_{rCi}) &\overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}(\alpha_{r1}, \ldots, \alpha_{rC}) \\
\alpha_{rc} &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)
\end{aligned}$$

With a unit-level covariate Z_i in the second level, the model becomes

$$\begin{aligned}
(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) &\overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \\
&\qquad \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri}) \\
(\beta_{r1i}, \ldots, \beta_{rCi}) &\overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}(d_r e^{(\gamma_{rc} + \delta_{rc} Z_i)}, \ldots, \\
&\qquad d_r e^{(\gamma_{r(C-1)} + \delta_{r(C-1)} Z_i)}, d_r) \\
d_r &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)
\end{aligned}$$

In the model with a covariate, users have two options for the priors on γ_{rc} and δ_{rc} . They may may assume an improper uniform prior, as was suggested by Rosen et al. (2001), or they may specify normal priors for each γ_{rc} and δ_{rc} as follows:

$$\begin{aligned}
\gamma_{rc} &\sim \text{N}(\mu_{\gamma_{rc}}, \sigma^2_{\gamma_{rc}}) \\
\delta_{rc} &\sim \text{N}(\mu_{\delta_{rc}}, \sigma^2_{\delta_{rc}})
\end{aligned}$$

As Wakefield (2004) notes, the weak identification that characterizes hierarchical models in the EI context is likely to make the results sensitive to the choice of prior. Users should experiment with different assumptions about the prior distribution of the upper-level parameters in order to gauge the robustness of their inferences.

The parameterization of the prior on each $(\beta_{r1i}, \ldots, \beta_{rCi})$ implies that the following log-odds ratio of expected fractions is linear with respect to the covariate Z_i:

$$\log\left(\frac{E(\beta_{rci})}{E(\beta_{rCi})}\right) = \gamma_{rc} + \delta_{rc} Z_i$$

Conducting an analysis using the MD model requires two steps. First, `tuneMD` calibrates the tuning parameters used for Metropolis-Hastings sampling:

```
> tune.nocov <- tuneMD(cbind(dem, rep, non)
+   ~ cbind(black, white, natam), data = senc,
+   ntunes = 10, totaldraws = 100000)
```

Second, `ei.MD.bayes` fits the model by calling C code to generate MCMC draws:

```
> out.nocov <- ei.MD.bayes(cbind(dem, rep, non)
+   ~ cbind(black, white, natam),
+   covariate = NULL, data = senc,
+   tune.list = tune.nocov)
```

The output of this function can be returned as `mcmc` objects or arrays; in the former case, the standard diagnostic tools in **coda** (Plummer et al., 2006) can be applied directly. The MD implementation includes `lambda` and `density.plot` functions, usage for which is analogous to ecological regression:

```
> lmd <- lambda.MD(out.nocov,
+   columns = c("dem", "rep"))
> density.plot(lmd)
```

If the precinct-level parameters are returned or saved, `cover.plot` plots the central credible intervals for each precinct. The segments represent the 95% central credible intervals and their medians for each unit (the true value for each precinct is the red dot, not included in the standard `cover.plot`).

```
> cover.plot(out.nocov, row = "white",
+   column = "dem")
# add true values to plot
> points(senc$white/senc$total,
+   senc$whdem/senc$white)
```

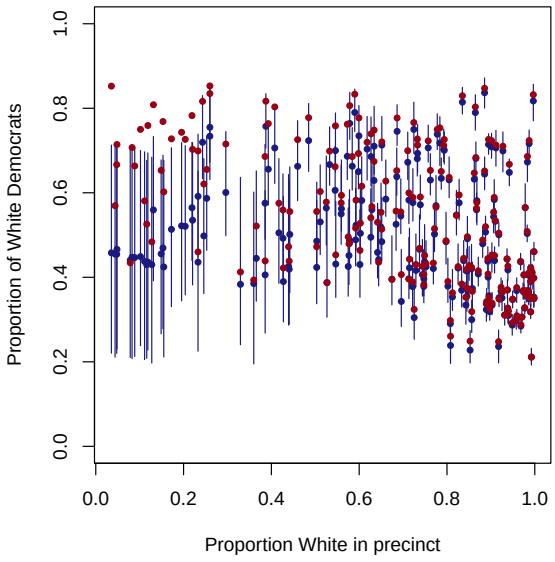

Figure 3: Coverage plot for MD model output.

Data Management

In the MD model, reasonable-sized problems produce unreasonable amounts of data. For example, a model for voting in Ohio includes 11000 precincts, 3 racial groups, and 4 parties. Implementing 1000 iterations yields about 130 million parameter draws. These draws occupy about 1GB of RAM, and this is almost certainly not enough iterations. We provide a few options to users in order to make this model tractable for large EI problems.

The unit-level parameters present the most significant data management problem. Rather than storing unit-level parameters in the workspace, users can save each chain as a `.tar.gz` file on disk using the option `ei.MD.bayes(..., ret.beta = "s")`, or discard the unit-level draws entirely using `ei.MD.bayes(..., ret.beta = "d")`. To reconstruct the chains, users can select the row marginals, column marginals, and units of interest, without reconstructing the entire matrix of unit-level draws:

```
> read.betas(rows = c("black", "white"),
+   columns = "dem", units = 1:150,
+   dir = getwd())
```

If users are interested in some function of the unit-level parameters, the implementation of the MD model allows them to define a function in R that will be called from within the `C` sampling algorithm, in which case the unit-level parameters need not be saved for post-processing.

Acknowledgments

eiPack was developed with the support of the Institute for Quantitative Social Science at Harvard University. Thanks to John Fox, Gary King, Kevin Quinn, D. James Greiner, and an anonymous referee for suggestions and Matt Cox and Bob Kinney for technical advice. For further information, see `http://www.people.fas.harvard.edu/~olau/software/eiPack.html`.

Bibliography

W. T. Cho and A. H. Yoon. Strange bedfellows: Politics, courts and statistics: Statistical expert testimony in voting rights cases. *Cornell Journal of Law and Public Policy*, 10:237–264, 2001.

O. D. Duncan and B. Davis. An alternative to ecological correlation. *American Sociological Review*, 18: 665–666, 1953.

L. Goodman. Ecological regressions and the behavior of individuals. *American Sociological Review*, 18: 663–664, 1953.

B. Grofman. A primer on racial bloc voting analysis. In N. Persily, editor, *The Real Y2K Problem: Census 2000 Data and Redistricting Technology*. Brennan Center for Justice, New York, 2000.

K. Imai and Y. Lu. *eco: R Package for Fitting Bayesian Models of Ecological Inference in 2x2 Tables*, 2005. URL `http://imai.princeton.edu/research/eco.html`.

A. D. Martin and K. M. Quinn. Applied Bayesian inference in R using MCMCpack. *R News*, 6:2–7, 2006.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence diagnostics and output analysis for MCMC. *R News*, 6:7–11, 2006.

O. Rosen, W. Jiang, G. King, and M. A. Tanner. Bayesian and frequentist inference for ecological inference: The $R \times C$ case. *Statistica Neerlandica*, 55(2):134–156, 2001.

J. Wakefield. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society*, 167:385–445, 2004.

Olivia Lau
`olau@fas.harvard.edu`
Ryan T. Moore
`rtmoore@fas.harvard.edu`
Michael Kellermann
`kellerm@fas.harvard.edu`

Institute for Quantitative Social Science
Harvard University, Cambridge, MA