How Much Can Be Inferred From Almost Nothing? A Two-Stage Maximum Entropy Approach to Uncertainty in Ecological Inference

Martin Elff¹, Thomas Gschwend¹, and Ron Johnston²

¹University of Mannheim
²University of Bristol


Problems of Ecological Inference

Ecological Inference

- Aim: estimation of individual-level behavior/properties from aggregate summaries
- If behavior/properties are categorical: estimation of a $I \times J \times K$-size data cube from $I \times K$, $J \times K$, and sometimes also $I \times J$-size marginal tables
- Big problem: more items of data to be estimated than items of data known
- Usual trick: use a model with less parameters

The Problem of Modelling Indeterminacy

- Restrictive model necessary to find estimates in ecological inference problem
- Assumptions of restrictive model cannot be tested – because of missing data
- Assumptions may be wrong – but a wrong model may lead to biased estimates
A Solution Template – A Two-Stage Approach

- Main Principle: Consider possible bias caused by model failure as a source of extra-variation of parameter estimates
- First stage: Use a “neutral” model: means maximize entropy subject to the constraints implied by known data
- Second stage: Use a entropy maximizing conjugate distribution of means derived from first-stage model
- Use means/expectations from first stage model to derive point estimates
- Use second stage model to derive confidence intervals

Maximizing Entropy at the First Stage – Example: The Johnston-Hay Model I

- Model for unknown counts in data cube with given marginal tables
- Entropy is maximized subject to the condition that sums of probabilities in each direction are equal to proportions in marginal tables

Maximizing Entropy at the First Stage – Example: The Johnston-Hay Model II

- First stage probability model of unknown data $x_{ijk}$:
  $$ f_{Mt}(x) = \frac{n!}{\prod_{i,j,k} x_{ijk}!} \prod_{i,j,k} p_{ijk}^{x_{ijk}} , $$

- Expectations:
  $$ E(x_{ijk}) = np_{ijk} = n e^{\alpha_{ij} + \beta_{ik} + \gamma_{jk}} e^{-1} = n \sum_{r,s,t} e^{\alpha_{rs} + \beta_{rt} + \gamma_{st}} $$
Maximizing Entropy at the First Stage – Example: The Johnston-Hay Model III

Entropy is maximized subject to constraints — that is, the following Lagrangian is maximized:

\[
L(p) = -n \sum_{i,j,k} p_{ijk} \log p_{ijk} + \sum_{i,j} \alpha_{ij} \left( n \sum_{k} p_{ijk} - n_{ij} \right) \\
+ \sum_{i,k} \beta_{ik} \left( n \sum_{j} p_{ijk} - n_{i,k} \right) + \sum_{j,k} \gamma_{jk} \left( n \sum_{i} p_{ijk} - n_{j,k} \right) \\
+ \tau \left( n \sum_{i,j,k} p_{ijk} - n \right)
\]

Mixing distribution: Dirichlet

\[
f_{\text{Dt}}(p) = \frac{\Gamma(\sum_{i,j,k} \theta_{ijk}) \prod_{i,j,k} \theta_{ijk}^{-1}}{\prod_{i,j,k} \Gamma(\theta_{ijk})}
\]

Maximize \( H_{\text{Dt}} := -\int f_{\text{Dt}}(p) \ln f_{\text{Dt}}(p) dp \) for all \( \theta_{ijk} \) subject to

\[
\pi_{ijk} := E(p_{ijk}) = \frac{\theta_{ijk}}{\sum_{r,s,t} \theta_{rst}} \hat{p}_{ijk}, \text{ that is, maximize}
\]

\[
\sum_{i,j,k} \ln(\theta_{0}\hat{p}_{ijk}) - \ln(\theta_{0}) + (\theta_{0} - IJK) \Psi(\theta_{0}) - \sum_{i,j,k} (\theta_{0}\hat{p}_{ijk} - 1) \Psi(\theta_{0}\hat{p}_{ijk})
\]

for \( \theta_{0} \) and set \( \theta_{ijk} = \theta_{0}\hat{p}_{ijk} \). \( \Psi(x) := d \ln(\Gamma(x))/dx \)

Implementation in R

**MaxEntMultinomial3()** Produces cell probability estimates \( p_{ijk} \) from marginal table counts \( n_{ij}, n_{ik}, \) and \( n_{jk} \) using iterative proportional scaling.

**DirichletParams()** Produces entropy-maximizing parameters \( \hat{\theta}_{ijk} \) of Dirichlet distribution subject to

\[
\theta_{ijk} / \sum_{r,s,t} \theta_{rst} = \hat{p}_{ijk}.
\]

**DirichletToBetaCI()** Produces confidence intervals for each of the \( \hat{p}_{ijk} \) based on \( \hat{\theta}_{ijk} \) and marginal Beta distribution of \( p_{ijk} \).
RMSE of First-Stage Point Estimates: Contrary to asymptotic theory, RMSE is unaffected by $n$.

Total root mean square error (TRMSE) of prediction after 2,000 replications with arbitrary configuration of “true” counts.

<table>
<thead>
<tr>
<th>Number of cells</th>
<th>Population size</th>
<th>100,000</th>
<th>10,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3 \times 3 \times 50$</td>
<td>0.565</td>
<td>0.564</td>
<td></td>
</tr>
<tr>
<td>$3 \times 3 \times 200$</td>
<td>0.579</td>
<td>0.574</td>
<td></td>
</tr>
<tr>
<td>$7 \times 7 \times 50$</td>
<td>0.827</td>
<td>0.817</td>
<td></td>
</tr>
<tr>
<td>$7 \times 7 \times 200$</td>
<td>0.867</td>
<td>0.829</td>
<td></td>
</tr>
</tbody>
</table>

Confidence Intervals from Second Stage Distribution: Nominal coverage $\approx$ real coverage if $n \to \infty$ (?)

Simulation Study of Extended Maximum Entropy Approach: Mean Effective Coverage (Percentage) of True Cell Counts after 2,000 replications

<table>
<thead>
<tr>
<th>Number of cells</th>
<th>Population size</th>
<th>100,000</th>
<th>10,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3 \times 3 \times 50$</td>
<td>94.7</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>$3 \times 3 \times 200$</td>
<td>93.2</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>$7 \times 7 \times 50$</td>
<td>92.7</td>
<td>94.4</td>
<td></td>
</tr>
<tr>
<td>$7 \times 7 \times 200$</td>
<td>86.9</td>
<td>94.3</td>
<td></td>
</tr>
</tbody>
</table>

Possible Causes of Undercoverage

- Proposed method rests on the approximation of the compound multinomial distribution by the Dirichlet distribution.
- If data cube is large and $n$ is “small,” the approximation is not so good.
- Confidence intervals based on compound multinomial distribution are difficult to construct (mixture of a discrete distribution with a continous distribution).

Application to Split-Ticket Voting: See poster!