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

Martin Elff[1], Thomas Gschwend[1], and Ron Johnston[2]

[1]University of Mannheim

[2]University of Bristol

useR 2006, R User Conference, Wirtschaftsuniversität Wien, 15-17 Juni 2006, Wien

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 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

Formulation of Johnston-Hay Model:

- First stage probability model of unknown data x_{ijk}:

$$f_{\mathrm{Mt}}(\boldsymbol{x}) = \frac{n!}{\prod_{i,j,k} x_{ijk}!} \prod_{i,j,k} p_{ijk}^{x_{ijk}},$$

- Expectations:

$$\mathrm{E}(x_{ijk}) = np_{ijk} = ne^{\alpha_{ij}+\beta_{ik}+\gamma_{jk}} e^{\tau-1} = n\frac{e^{\alpha_{ij}+\beta_{ik}+\gamma_{jk}}}{\sum_{r,s,t} e^{\alpha_{rs}+\beta_{rt}+\gamma_{st}}}$$

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:

$$L(\boldsymbol{p}) = -n\sum_{i,j,k} p_{ijk}\log p_{ijk} + \sum_{i,j}\alpha_{ij}\left(n\sum_k p_{ijk} - n_{ij.}\right) + \sum_{i,k}\beta_{ik}\left(n\sum_j p_{ijk} - n_{i.k}\right) + \sum_{j,k}\gamma_{jk}\left(n\sum_i p_{ijk} - n_{.jk}\right) + \tau\left(n\sum_{i,j,k} p_{ijk} - n\right)$$

Maximizing Entropy at the Second Stage – Extending the Johnston-Hay Model by a Infinite Mixture of the p_{ijk}

- Mixing distribution: Dirichlet

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

- Maximize $H_{\text{Dt}} := -\int f_{\text{Dt}}(\boldsymbol{p})\ln f_{\text{Dt}}(\boldsymbol{p})d\boldsymbol{p}$ for all θ_{ijk} subject to $\pi_{ijk} := \text{E}(p_{ijk}) = \frac{\theta_{ijk}}{\sum_{r,s,t}\theta_{rst}} \stackrel{!}{=} \hat{p}_{ijk}$, that is, maximize

$$\sum_{i,j,k}\ln\Gamma(\theta_0\hat{p}_{ijk}) - \ln\Gamma(\theta_0) + (\theta_0 - IJK)\Psi(\theta_0) - \sum_{i,j,k}(\theta_0\hat{p}_{ijk} - 1)\Psi(\theta_0\hat{p}_{ijk})$$

for θ_0 and set $\theta_{ijk} = \theta_0\hat{p}_{ijk}$. ($\Psi(x) := \text{d}\ln\Gamma(x)/\text{d}x$)

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.

`DirichletParms()` Produces entropy-maximizing parameters $\tilde{\theta}_{ijk}$ 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 $\tilde{\theta}_{ijk}$ and marginal Beta distribution of p_{ijk}.

A Simulation Study – Check of the Two-Stage Maximum Entropy Approach

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.

	Population size	
Number of cells	100,000	10,000,000
3×3×50	0.565	0.564
3×3×200	0.579	0.574
7×7×50	0.827	0.817
7×7×200	0.867	0.829

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

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

	Population size	
Number of cells	100,000	10,000,000
3×3×50	94.7	95.0
3×3×200	93.2	95.0
7×7×50	92.7	94.4
7×7×200	86.9	94.3

Possible Causes of Undercoverage

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

Application to Split-Ticket Voting: See poster!