Nonparametric Density Estimation for Capture-Recapture

Capture-Recapture = the science of estimating the size of a population from multiple incomplete lists

Zach Kurtz
Department of Statistics
Carnegie Mellon University

Advisers:
William F. Eddy
Stephen E. Fienberg
Cosma Shalizi
Rebecca Steorts

NIPS Conference at Lake Tahoe
December 7, 2012

How many fish are in this lake?

(1) Catch and weigh 1000 fish, tag each one with a unique identifier, and release.

(2) Again capture and weigh 1000 fish, and observe that (say) 50 of them have a tag.

		List L_1	
		yes	no
List L_2	yes	50	950
	no	950	$c_0 = ?$

List independence $\rightarrow$ the odds ratio is 1

$$c_0 \approx (950)(950)/50 = 18050$$

But independence fails if big fish are easier to catch than small fish!

Imputation with heterogeneity

Counts:

		List L_1: yes	List L_1: no
List L_2	yes	c_{11}	c_{01}
	no	c_{10}	$c_{00} = ?$

Probabilities:

		List L_1: yes	List L_1: no
List L_2	yes	$p(11, x)$	$p(01, x)$
	no	$p(10, x)$	$p(00, x) = ?$

Conditional Probabilities:

$$\pi(\mathbf{y}, x) := \frac{p(\mathbf{y}, x)}{1 - p(\mathbf{0}, x)}$$

		List L_1: yes	List L_1: no
List L_2	yes	$\pi(11, x)$	$\pi(01, x)$
	no	$\pi(10, x)$	$\pi(00, x) = ?$

$$\hat{c}_{\mathbf{0}} = \sum_{i=1}^{n_c} \hat{\pi}(\mathbf{0}, x_i)$$

Two Frontiers in Capture-Recapture

(1) Estimation of conditional probabilities

		List L_1	
		yes	no
List L_2	yes	$\pi(11,x)$	$\pi(01,x)$
	no	$\pi(10,x)$	

Use nonparametric conditional density estimation!

(2) Imputing the missing cell

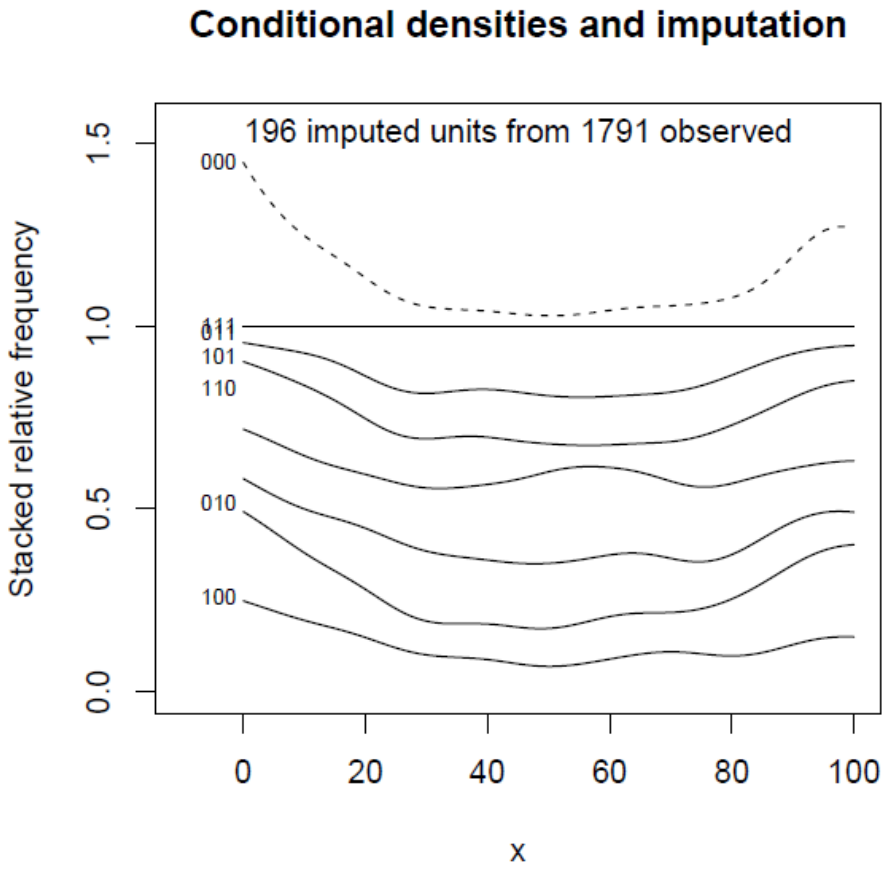
Independence model: $\pi(00,x) = \frac{\pi(01,x)\pi(10,x)}{\pi(11,x)}$

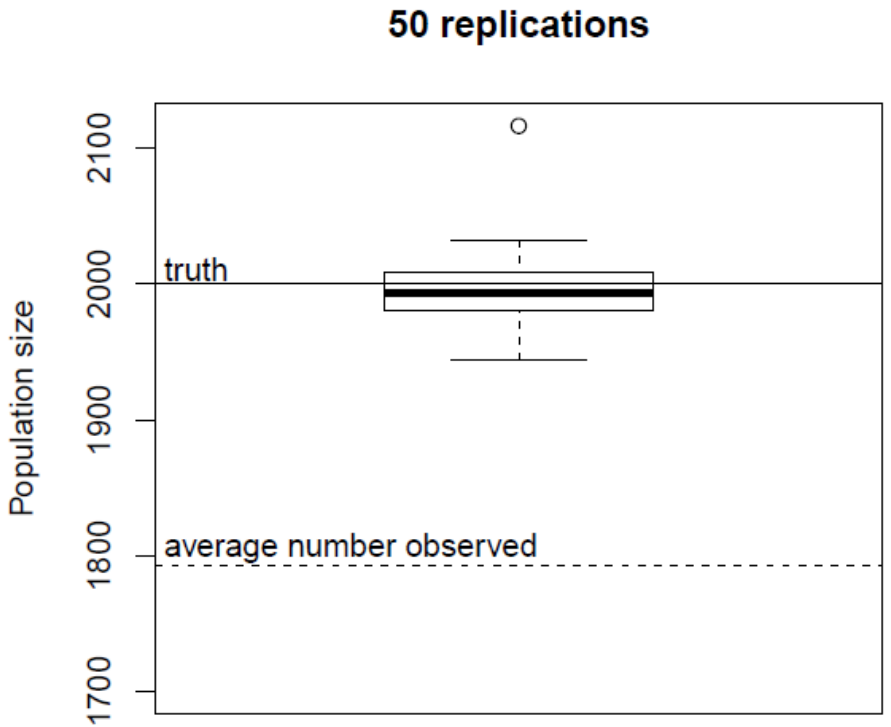
For three lists, a saturated model:

$$\log \pi(\mathbf{y},x) = u + u_1\mathbf{y}_1 + u_2\mathbf{y}_2 + u_3\mathbf{y}_3 + u_{12}\mathbf{y}_1\mathbf{y}_2 + u_{13}\mathbf{y}_1\mathbf{y}_3 + u_{23}\mathbf{y}_2\mathbf{y}_3$$

where $u = u(x)$.

Extra slide … Q&A





Challenges

- Local model selection
- Balancing model complexity over x versus complexity over y
- Dealing with heterogeneity not explained by covariates