Supplemental materials for Bayesian Population Size Estimation Using Dirichlet Process Mixtures

Daniel Manrique-Vallier[*]

December 17, 2015

This supplement includes a sensitivity analysis of the posterior distribution of N to the prior distribution of α; an example of the application of the NPLCM to a classical animal abundance estimation dataset; and additional tables.

1 Sensitivity of the posterior distribution of N to the prior distribution of α

In this section we empirically investigate the sensibility of the posterior distribution of N to the prior distribution of the Dirichlet process parameter $\alpha > 0$. Remember that if $\boldsymbol{\pi} = (\pi_1, \pi_2, ...) \sim SB(\alpha)$, then $\boldsymbol{\pi}$ takes values on the infinite-dimensional unit simplex Δ_∞, and parameter α acts a concentration parameter: the smallest the value of α, the fastest π_k decreases to zero with k. In the limit case, taking $\alpha \to 0$ concentrates all the probability mass on the first vertex of Δ_∞, resulting in a degenerate distribution at $\boldsymbol{\pi} = (1, 0, 0, ...)$. Thus α has complexity modulation effect in our model, with small values of α favoring sparse mixtures, and large values favoring more complex joint distributions. In an effort for letting the data themselves inform about the appropriate degree of sparsity required to model the

[*] Daniel Manrique-Vallier is Assistant Professor at the Department of Statistics, Indiana University, Bloomington, IN 47408 (e-mail: dmanriqu@indiana.edu).

Prior $\alpha \sim \text{Gamma}(a, b)$	Mean $\hat{N}$	Mean CI width	MSE	Empirical Coverage
$a = 0.25$, $b = 0.25$	1935.23	868.83	49038.44	0.92
$a = 1.0$, $b = 1.0$	1938.46	882.44	48663.07	0.94
$a = 10^{-5}$, $b = 10^{-3}$	1934.44	866.48	49680.04	0.92
$a = 10^{-3}$, $b = 10^{-5}$	1933.66	861.12	49534.73	0.91
$a = 1.0$, $b = 3.0$	1937.92	872.95	49136.01	0.93
$a = 3.0$, $b = 1.0$	1949.75	910.91	47837.93	0.94
$a = 0.6$, $b = 1.2 \cdot 10^{-3}$	1937.13	877.08	48438.76	0.93
$a = 0.0$, $b = 0.0$	1934.94	863.46	49117.99	0.92

Table 1: Summary of simulation results over 200 experiments with $N = 2000$ and $E[n] = 727$ using the NPLCM with $\alpha \sim \text{Gamma}(a, b)$ for different values of a and b.

joint distribution of capture patterns, we have followed the advice from Dunson and Xing (2009) recommendation of endowing α with the hyper-prior distribution Gamma(a, b).

In order to study the effect of the prior specification of α under controlled circumstances, we have expanded the repeated sampling simulated data experiment from the main manuscript (200 replications with $N = 2000$ and $J = 5$; refer to the main article for more details), adding several NPLCM models with different hyper-prior parameters a and b, including our default $a = b = 0.25$. Table 1 shows the results of our experiments. We see that most results are close to one another, resulting in a repeated sampling performance roughly equivalent. It is possible that this robustness to prior specification has to do with the relatively large sample size used in this example. To explore this issue, we have taken advantage of the example that we develop in the next section ($n = 68$). There we perform our calculations under several prior specifications (Table 3). Results also exhibit a considerable robustness to the prior specification.

2 An ecology example with small sample size

Different from the real-data examples already analyzed (casualties in conflicts in Kosovo and Colombia), that dealt with human populations, capture-recapture datasets in animal abundance studies typically detail a relatively large number of recapture occasions J, but a small observed sample size n. Here we illustrate the application of our methods to this type of data. To this end we have selected the now classic dataset detailing the multiple recapture of snowshoe hares, originally analyzed by Otis et al. (1978) and subsequently re-analyzed by several authors (e.g. Cormack, 1989; Agresti, 1994; Dorazio and Royle, 2003; Pledger, 2005). This dataset features a very small number of uniquely captured subjects, $n = 68$, trapped in $J = 6$ occasions. In Table 2 we reproduce the full capture history data, as obtained from Baillargeon and Rivest (2007).

Table 3 shows our estimates using the NPLCM for a variety of prior specifications of the α parameter. We see that most estimates are close to one another, with a point estimate around $\hat{N} = 77$ and a posterior 95% intervals close to $(70, 90)$. These estimates are close to those obtained by Agresti (1994) and Dorazio and Royle (2003) with “latent class” M_h and M_{th} models using random effects distributions with finite support, and are among the most conservative estimates obtained from models that assume heterogeneity. We note that, as these authors discuss (see also Pledger, 2005), the small sample size in this example makes it essentially impossible to assert if the true heterogeneity structure implies the existence of hidden sub-populations with low probability of capture. If this were the case, at small sample-size levels these hypothetical sub-populations would be, for all practical purposes, invisible. Our method reflects this fact by producing estimates consistent with the most conservative estimates obtained under heterogeneity models.

Pattern($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern($\mathbf{x}$)	$n_{\mathbf{x}}$
010000	6	010101	3	101100	1	001011	1
001000	5	100100	2	110010	1	011011	1
000010	4	100010	2	101010	1	000111	1
000001	4	001001	2	000110	1	100111	1
000101	4	011001	2	100001	1	010111	1
100000	3	011101	2	010001	1	011111	1
000100	3	000011	2	110001	1	$\ldots$	0
010100	3	111111	2	001101	1	$\ldots$	0
010010	3	101000	1	010011	1	000000	??

Table 2: Snowshoe hare data ($J = 6, n = 68$). Showing all the 33 capture patterns observed in the sample.

Prior	$\hat{N}$	95%-CI
$a = 0.25$, $b = 0.25$	76	(70, 90)
$a = 10^{-5}$, $b = 10^{-3}$	75	(70, 85)
$a = 10^{-3}$, $b = 10^{-5}$	76	(70, 88)
$a = 1$, $b = 1$	77	(70, 91)
$a = 10$, $b = 10$	77	(70, 91)
$a = 1$, $b = 3$	77	(70, 91)
$a = 3$, $b = 1$	76	(70, 90)

Table 3: Summary of results using the snowshoe hare data ($J = 6, n = 68$) for NPLCM model with different priors $\alpha \sim \text{Gamma}(a, b)$.

Table 4: Casanare data ($J = 15$, $n = 2629$). Showing the all 70 capture patterns actually observed in the sample. Note that 99.79% of the 2^{15} cells of the contingency table are empty.

Pattern ($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern ($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern ($\mathbf{x}$)	$n_{\mathbf{x}}$	Pattern ($\mathbf{x}$)	$n_{\mathbf{x}}$
000000001000000	1215	100001001000000	5	101001000000010	2	101011010101000	1
000010001000000	403	100011001100000	5	100010000000000	1	001000011101000	1
000000000100000	284	100001001010000	5	101001000000000	1	000001000000100	1
000010000000000	221	000011000000000	4	100001100000000	1	100001000000100	1
000010001100000	119	000001001000000	4	101000001000000	1	010001000000100	1
000001000000000	49	000011001100000	4	000110001000000	1	000101000000100	1
100000000000000	48	100010001000000	3	000011001000000	1	000101100000100	1
000010000100000	46	100011001000000	3	001000011000000	1	100101100000100	1
100000001000000	37	000010001010000	3	100001000100000	1	000100001000100	1
000000000010000	30	000001001010000	3	000011000100000	1	000110001000100	1
000000001100000	20	000000001100100	3	100010001100000	1	100000101000100	1
100001000000000	19	001000001000010	3	001010001100000	1	000100000010100	1
100001000010000	10	100000000100000	2	000001001100000	1	101000001000010	1
001000000000010	10	100010000100000	2	100000001010000	1	001010001100010	1
000000001010000	9	000001000100000	2	000000000110000	1	001000011000110	1
000001000010000	6	100000001100000	2	100001001110000	1	101001000101011	1
000000001000100	6	000010000000100	2	001010001001000	1	...	...
001000001000000	5	100001000010100	2	101000010101000	1	000000000000000	??

3 Additional tables

In Table 4 we reproduce the Casanare data used in Section 5.3.

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

Baillargeon, S. and Rivest, L.-P. (2007), "Rcapture: Loglinear Models for Capture-Recapture in R," *Journal of Statistical Software*, 19.

Cormack, R. M. (1989), "Log-Linear Models for Capture-Recapture," *Biometrics*, 45, pp. 395–413.

Dorazio, R. M. and Royle, A. J. (2003), "Mixture models for estimating the size of a closed population when capture rates vary among individuals," *Biometrics*, 59, 351–364.

Dunson, D. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), "Statistical inference from capture data on closed animal populations," *Wildlife monographs*, 3–135.

Pledger, S. (2005), "The performance of mixture models in heterogeneous closed population capture-recapture," *Biometrics*, 61, 868–876.