ONE-INFLATED BAYESIAN MIXTURES FOR POPULATION SIZE ESTIMATION

Davide Di Cecco [1], Andrea Tancredi[1] and Tiziana Tuoto[2]

[1] Sapienza University of Rome, (e-mail: davide.dicecco@uniroma1.it, andrea.tancredi@uniroma1.it)

[2] ISTAT, (e-mail: tuoto@istat.it)

ABSTRACT:

The phenomenon of one-inflation frequently affects the estimates of population size when the available dare are represented by frequencies of counts. A particular behavioral effect preventing subsequent captures after the first one may be the reason for such an effect. We consider a Bayesian semi-parametric approach by fitting a truncated Dirichlet process mixture model as a base tool for modeling repeated count data and extend this class to include one–inflation. The proposed methodology is briefly illustrated via a real data application.

KEYWORDS: capture-recapture, Dirichlet process mixture, repeated count data

1 Introduction

Consider a closed population composed of N individuals. Suppose that N is unknown, n distinct units have been identified for a fixed amount of time and a given unit may be identified exactly once or observed twice, three times, or more. Under time–homogeneity, without individual covariates, the data can be simply summarized as counts of units captured j times, $j = 1,2,...,$ commonly called "repeated count data" The common parametric approach for estimating N is to define a counting distribution for the number of captures in the population. In the absence of any additional individual information, it is crucial to model the unobserved heterogeneity. A well–established approach to this end is represented by the use of mixtures of counting distributions, see, Böhning *et al.* (2005).

Mixtures of Poisson distributions are a standard choice both for repeated captures and species sampling problems but they present several issues related to the selection of the number of components and the instability of the N estimator. The choice of the number of mixture components has been usually addressed by the use of the nonparametric maximum likelihood estimation

(NPMLE) approach (Norris & Pollock (1996)) which maximizes the likelihood of an over-fitting finite mixture model. The frequentist properties of the NPMLE approach have been discussed in Wang & Lindsay (2005) where a penalized NPMLE estimator of N with better inferential performance is proposed. Another critical issue that has been recently addressed for the N estimation problem with repeated counts is that the collected data set frequently exhibit an elevated number of individuals captured exactly once. See, for example, Godwin & Böhning (2017). This excess of singletons is also termed as "one–inflation". Failing to identify and model in the analysis such a mechanism implies a (possibly severe) overestimation of the total population count.

Bayesian semi-parametric approaches underlying population size estimation have already been proposed. Guindani *et al.* (2014) handled the heterogeneity problem proposing a Dirichlet process mixture (DPM) of Poisson distributions for modeling gene expression sequence abundance and estimating the number of different unique sequences. The DPM approach, as the NPMLE, avoids to fix the number of components and, by averaging over mixtures of different order, has the advantage of properly accounting for the clustering process uncertainty in the final estimate of N. A DPM latent class model has been also proposed in the context of multiple systems estimation by Manrique-Vallier (2016) under capture heterogeneity and list dependence. In this paper, we present an application of the DPM approach handling the presence of one inflation with repeated count data.

2 One-inflated mixture distributions

Let Y_i, $i = 1,\ldots,N$, be the integer-valued random variable representing the number of times a given unit has been captured. We assume that

$$Y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i) \quad \lambda_i|G \overset{iid}{\sim} \Lambda \quad \Lambda \sim DP(\phi\Lambda_0) \tag{1}$$

where $\Lambda \sim DP(\phi\Lambda_0)$ denotes a distribution generated by a Dirichlet process with base measure $\phi\Lambda_0$, see Guindani *et al.* (2014). Note that we only observe the n individuals which are captured at least once. Let n_j denote the number of units captured j times, such that $\sum_{j>0} n_j = n$. We want to estimate the number of uncaptured units n_0, or, equivalently, $N = n + n_0$. Considering the *truncated* version of the DPM model (1) (see Ishwaran & James (2001)), Y_i is a finite mixture of Poisson distributions with mixing weights $\pi_1,\ldots,\pi_k$ following a finite stick-breaking prior, that is, $\pi_1 = V_1$ and

$$\pi_i = (1-V_1)(1-V_2)\cdots(1-V_{i-1})V_i \quad i = 2,\ldots k \tag{2}$$

where V_i for $i = 1, \ldots, k-1$ are independent Beta$(1, \phi)$ random variables and $V_k = 1$. Denote as $f(j|\lambda_i)$ the probability $\lambda_i^j e^{-\lambda_i}/j!$ of being captured j times in the i–th component defined by the parameter λ_i and denote as θ the set of all parameters. The truncated Poisson DPM model is defined as $P(Y = j) = f(j|\theta) = \sum_{i=1}^{k} \pi_i f(j|\lambda_i)$ for $j = 0, 1, \ldots$ with the mixing weights given by (2).

Under the hypothesis of one–inflation caused by a specific behavioral effect, an individual that, without that effect, would face multiple captures, under this effect will be captured just once. The hypothesis can be modeled as follows: let B be the latent indicator variable identifying the units having this behavior. Each individual has a marginal probability ω of belonging to this subpopulation. Denote as Y^* the latent number of captures of a given unit that we would observe in absence of the behavioral mechanism and let $f^*(j|\theta) = P(Y^* = j \,|\, \theta)$ be its probability distribution. By assuming $P(Y = j|B = 0) = f^*(j|\theta)$ for all j and $P(Y = j|B = 1) = f^*(0|\theta)$ for $j = 0$ and $1 - f^*(0|\theta)$ for $j = 1$ the resulting distribution for Y is the one–inflated model defined as:

$$P(Y = j \,|\, \theta, \omega) = \begin{cases} f^*(0|\theta) & \text{if } j = 0; \\ (1-\omega) f^*(1|\theta) + \omega(1 - f^*(0|\theta)) & \text{if } j = 1; \\ (1-\omega) f^*(j|\theta) & \text{if } j > 1. \end{cases} \tag{3}$$

The one-inflated Poisson DPM model is then obtained by assuming for the baseline distribution $f^*(j|\theta)$ in (3) a Poisson DPM model.

3 Application

In this Section, we briefly illustrate the proposed methodology. We consider a data set that contains counts of treatment episodes by heroin users in Bangkok, see Godwin (2017). Upon visiting a treatment center, heroin users may find the treatment less pleasant than expected, and decide never to return, thus giving rise to one-inflation. Figure 1 shows the data set, the posterior distributions for N under the truncated DPM and the one-inflated version, the posterior distributions for the number of observed clusters, and the one-inflation parameters. These posterior distributions have been obtained via MCMC methods and a prior on the DPM parameter ϕ penalizing the overestimation of the number of clusters. As expected, the one-inflated model produces lower estimates of the population count by assigning a greater number of captures to a portion of singletons. The estimation for N and ω under the one-inflated DPM are comparable to those obtained by Godwin (2017) confirming that the DPM and its one-inflated counterpart represent valid competitors in this setting.

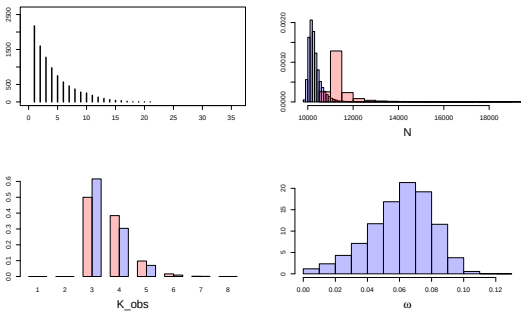

Figure 1. *Heroin users data set. Top left: count distribution. Top right: posterior distribution for N under the DPM (red) and one-inflated DPM (blue). Lower left: posterior distribution for the number of observed clusters. Lower right: posterior distribution for the DPM one-inflation parameter* ω

References

BÖHNING, D., DIETZ, E., KUHNERT, R., & SCHÖN, D. 2005. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, **14**, 29–43.

GODWIN, R.T. 2017. One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, **59**, 79–93.

GODWIN, R.T., & BÖHNING, D. 2017. Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 425–448.

GUINDANI, M., SEPÚLVEDA, N., PAULINO, C.D., & MÜLLER, P. 2014. A Bayesian semi-parametric approach for the Differential Analysis of Sequence Counts Data. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, **63**, 385.

ISHWARAN, H., & JAMES, L. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Stat. Association*, **96**, 161–173.

MANRIQUE-VALLIER, D. 2016. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, **72**, 1246–1254.

NORRIS, J., & POLLOCK, K. 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 639–649.

WANG, J. Z., & LINDSAY, B. G. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, **100**, 942–959.