A Comparative overview of some recent Bayesian nonparametric approaches for the size of a population

Una rassegna comparativa su alcuni recenti approcci nonparametrici bayesiani per la stima della numerosità di una popolazione

Luca Tardella and Danilo Alunni Fegatelli

Abstract We review some recent approaches that have been used to address the difficult problem of estimating the unknown size of a finite population. We start from illustrating what types of inferential difficulties one should expect when no parametric assumption is made on the class of distributions for the distribution of counts of the number of multiple occurrences of the same unit when the observed counts are modelled in terms of Poisson mixtures. We then consider the problem from the species sampling model perspective where each unit is represented by the distinct species and a sequence of exchangeable unit label observations are available. We discuss the implementation of the alternative approaches with real datasets and we compare their performance with simulated data.

Abstract *In questo lavoro passiamo in rassegna alcuni recenti approcci bayesiani nonparametrici per stimare la numerosità incognita di una popolazione finita. Iniziamo dall'illustrare le difficoltà inferenziali che si devono affrontare quando nessuna assunzione parametrica restringe la classe di distribuzioni che regola il tasso medio di conteggio non negativo delle occorrenze multiple delle unità distinte nel campione. Consideriamo quindi il problema dalla prospettiva dei modelli per species sampling dove le unità corrispondono alle specie distinte e si assume una successione scambiabile etichette di specie osservabili. Si effettuano analisi comparative sull'implementazione dei diversi approcci su dati reali e sulla performance con simulati*

Key words: Poisson Mixture counts, Finite Population Size estimation, Species Sampling, Bayesian Nonparametrics

Luca Tardella
Sapienza Università di Roma, Piazzale Aldo Moro,5 (00185) Roma, e-mail: luca.tardella@uniroma1.it

Danilo Alunni Fegatelli
Sapienza Università di Roma, Piazzale Aldo Moro,5 (00185) Roma, e-mail: danilo.alunnifegatelli@uniroma1.it

1 Introduction

In this paper we consider the estimation of the total size of a finite population based on a single sample of individual detections. There are many instances in which such problem is of interest starting from the complete enumeration of elusive populations (Böhning and van der Heijden, 2009), software debugging to get the total number of errors (Lloyd et al., 1999) and the species richness problem in ecology (Bunge and Fitzpatrick, 1993; Chao and Bunge, 2002; Wang and Lindsay, 2005; Chao and Chiu, 2016) to measure and preserve biodiversity. We assume there are N distinct units in the population labelled as $i = 1, ..., N$ which can be encountered (or detected) C_i times where each C_i is a non-negative integer. Indeed, only those units which are encountered at least once ($C_i > 0$) during the sampling stage are actually detected. Hence, if we denote with M the maximum number of multiple encounters (count) observed for a single unit, the positive frequencies of frequencies statistics $(f_1, ...f_k, ..., f_M)$, where f_k is the number of distinct units i which have been encountered exactly k times (i.e. for which $C_i = k$), provide a possibly incomplete enumeration $n = \sum_{k=1}^{M} f_k$ of the total population size N since $N = f_0 + n \geq n$. In fact, there are $f_0 > 0$ undetected units for which $C_i = 0$. There have been several attempts in the literature to address the problem of estimating N starting from modelling of the frequencies of frequencies distribution. In fact, this distribution is often determined by the nonparametric modelling of the individual encounter count data. One of the most flexible and general models commonly used in this setting is a mixture of Poisson distributions where the mixing distribution can be arbitrary. Indeed this setting has been dealt with both from the classical side with likelihood-based estimates Norris and Pollock (1998); Wang and Lindsay (2005) or Abundance-based Coverage Estimator (ACE), lower bounds and their variants (Chao and Lee, 1992; Mao, 2006; Mao et al., 2013) and from the Bayesian perspective Barger and Bunge (2010); Guindani et al. (2014). An alternative approach can be based on modelling the sequence of single encounters which is well known in the literature as species sampling model where it is assumed an exchangeable sequence of labels $X_1, ..., X_j, ..., X_s$ where a label uniquely and perfectly identifies each distinct species (to be understood as a distinct unit of the population) with $s = \sum_{k=1}^{M} k f_k$. In fact, in the species sampling terminology the word population size can be misleading since the total population size N of our original formulation corresponds to the total number of distinct species and is one of the main inferential objectives, whereas the total number of encounters of the different species corresponds to the number s of sequentially observed labels of the species sampling units. Exchangeability of labels ensures that the frequencies statistics $(f_1, ...f_k, ..., f_M)$ are the relevant statistics for inferring the population structure, including the probability of a new discovery and hence a possible assessment of the total number of distinct species N. Moreover, in most of the recent contribution in the Bayesian species sampling modelling (Lijoi et al., 2007; Arbel et al., 2017) the underlying population size, denoted with N in our setting, is indeed assumed to be infinite since the relative abundances of the population of species correspond to the random atoms of an almost surely discrete random probability measure belonging to the so-called Gibbs-type class. Notice-

able exceptions of species sampling models assuming a finite population structure are considered in Gnedin (2010), Cerquetti (2011), Bissiri et al. (2013) and Zhou et al. (2017).

2 Alternative Bayesian Nonparametric approaches

In this section we briefly review the main features of some recent alternative approaches used to infer on the characteristics of a population which have individuals that have varying probability to be encountered in a sampling stage. More specifically we will consider the Dirichlet process mixture approach of Guindani et al. (2014) and the moment based approach in Alunni Fegatelli (2013). We also find it interesting to provide a comparative analysis with nonparametric Bayesian species sampling approach based on a two-parameter (α,β) Poisson-Dirichlet random measure as in Lijoi et al. (2007) with $0 \leq \alpha < 1$ and $\beta \geq -\alpha$. Indeed the comparability with the latter approach should take into account the structurally different underlying assumption on the population size although, in practice, the alternative approaches can be used to analyze the same real datasets. However, we will also consider a more appropriate comparison with a structurally different two-parameter (α,β) Poisson-Dirichlet random measure with $\alpha < 0$ and $\beta = -N\alpha$ for which the random measure has a finite support on N distinct units.

2.1 Dirichlet process mixture of Poisson counts

Guindani et al. (2014) propose to analyse observed positive counts of unique proteomic and genomic units with a semiparametric mixture of Poisson distributions in the presence of overdispersion and uncertainty on the true number of unique proteins or genes in a specific tissue (population). They assume the following hierarchical model: for a fixed population size N, any population unit $i = 1,...,N$ is possibly detected according to $C_i|\lambda_i \sim Pois(\lambda_i)$, $\lambda_i|F \sim F$ and $F \sim DP(F_0,\tau)$ i.e. a Dirichlet process prior on an almost surely random discrete distribution F on the individual Poisson rate parameter λ_i. The Dirichlet process prior requires the specification of an expected distribution F_0 for λ and a positive total mass parameter τ regulating the concentration of the expected relative abundances corresponding to each unit of the population. They propose the use of a $Gamma(a,b)$ distribution for F_0. Indeed N is the main unknown parameter of interest and a prior distribution is needed. They acknowledge that the choice of the prior on N has a relevant impact and requires careful consideration. They starts arguing that in lack of genuine expert prior information a prior centered around the number of observed sequences n can provide a reasonable default choice. However, for simulation study purposes they implement their model with a uniform prior over a compact support.

2.2 Gibbs-type prior and nonparametric Bayesian species sampling

Lijoi et al. (2007) use a Bayesian nonparametric approach to evaluate the probability of discovering new species in the population conditionally on the number s of species recorded in a sample. The discovery probability represents a natural tool for a quantitative assessment of concepts such as species richness and sample coverage that is the proportion of distinct species present in the observed sample. In particular, they provide a way of estimating the proportion of yet unobserved species which is the complementary sample coverage fraction. However, we must point out from the outset that the species sampling setting and terminology should be carefully rephrased and understood within the original context described in Section 1. Indeed, in the species sampling model of Lijoi et al. (2007) an exchangeable sequence of s observable labels $X_1,...,X_j,...,X_s$ are sampled and the corresponding number n of distinct labels $X_1^*,...,X_n^*$ allows to compute the counts $C_{i,s} = \sum_{j=1}^{s} I(X_j = X_i^*)$ and those counts are sufficient statistics for inferring the sample coverage $1 - U_s = \sum_i \pi_i I(C_{i,s} = 0)$ conditionally on the observed labels. π_i's are the probability masses attached to each distinct label X_i^* which are in turn assumed to be random according to a Gibbs-type prior which selects a.s. discrete distributions with a countable support of distinct points corresponding to a countable subset of labels. In Favaro et al. (2012) and Arbel et al. (2017) an empirical Bayes approach is used to infer on the underlying parameters of the Gibbs-type prior and derive point estimate and interval estimate of the discovery probability of a new species in the Poisson-Dirichlet case. In fact, one could try to relate this discovery probability to the fraction of yet unseen species which can then be turned into an estimate of the total population size N using the relation $E[n] = N(1 - U_s)$. However, this could be rigorously justified only if the number of point masses, i.e. N is assumed to be finite almost surely which happens in the presence of Gibbs-type prior of fixed type $\alpha < 0$ according to Gnedin and Pitman (2005). However in this case one can more directly derive a fully Bayesian inference based on the conditional (on a fixed N) probability of the observed counts and the underlying mixing measure for the finite number of species N. To our knowledge such approach has not been considered in the literature. Indeed a recent attempt in the same direction has been put forward by Zhou et al. (2017) although with no emphasis on the estimation of N.

2.3 Moment-based mixtures of truncated Poisson counts

In Alunni Fegatelli (2013) and Alunni Fegatelli and Tardella (2018) a Bayesian nonparametric approach is proposed. It starts from highlighting that when a finite sample of counts are observed from a mixture of Poisson distributions with unconstrained mixing F for the Poisson rate parameter the sample basically carries information on the mixing F only for a finite number of features. More precisely, if M is the maximum number of observed counts, it depends only on the first M moments of a suitable finite measure Q representing a one-to-one reparameterization

of F with the remaining moments of Q being completely unidentified by the sampling distribution (see details in Alunni Fegatelli and Tardella (2018)). In fact, the corresponding likelihood for N and the first M moments of Q, $(m_1(Q),....,m_M(Q))$ is

$$L(N,F;\mathbf{n}) = L(N,Q;\mathbf{n}) = L(N,m_1(Q),....,m_M(Q)) \propto \binom{N}{n} \prod_{k=0}^{T} \left[\frac{m_k(Q)}{k!}\right]^{n_k}$$

This yields the idea of working around a suitable moment-based approximation of the former likelihood which relies on a suitable truncation of the support of the rate parameter in a bounded interval $[0,u]$ and can then be used to derive a suitable default prior in terms of the Jeffreys rule for $(m_1,...,m_M)$ conditionally on N and u. A suitable Rissanen prior on the integer valued population size parameter N and a possible ad hoc choice of the prior on the truncation u complete the specification of the Bayesian model. Indeed it must be remarked that in Barger and Bunge (2010) alternative improper priors are derived as default priors from the reference and the Jeffreys prior approaches. The authors also provide justification for independent prior distributions for the parameter of interest N and the nuisance parameters of the stochastic abundance distribution.

3 Numerical Illustration

A simulation study was used to investigate on the frequentist performance of the three Bayesian nonparametric procedure. We considered the same 12 simulation settings proposed in Wang (2010) where the distribution of the species abundance varied from gamma, gamma mixture, lognormal and lognormal mixture, to discrete distributions, with the expected coverage of the sampled species ranging from 0.20 to 0.90. The corresponding results labelled s_1 through s_{12} are shown in Figure 1 where on the top row shows the average point estimates resulting from 100 simulated datasets for each simulation setting. In the other two rows the root mean square error and the coverage of equal tail 0.95 interval estimates are reported. Differently from the original work, where the estimates were evaluated only for $N = 1000$, we considered also a true population size of 10000. For a fairer comparison with respect to the Poisson-Dirichlet species sampling distribution we have also considered simulation settings s_{13}, s_{14} and s_{15} using a Poisson-Dirichlet structure with parameters $\alpha = -2$ and $\beta = -N\alpha$ and with observed sample coverage equal to 0.3, 0.5 and 0.8 respectively. For both values of N there wasn't a procedure resulting better than others for each simulation setting. However, overall, point and interval estimates for the moment-based method seemed to be the most stable, robust and with a smaller average (over all simulation settings) mean square error. Poor adaptivity of the Poisson-Dirichlet model might be explained by the fact that it indeed incorporates a smaller number of free parameters. We again stress on the fact that simulations were based on finite values of N. Hence, comparisons with the model proposed in Lijoi et al.

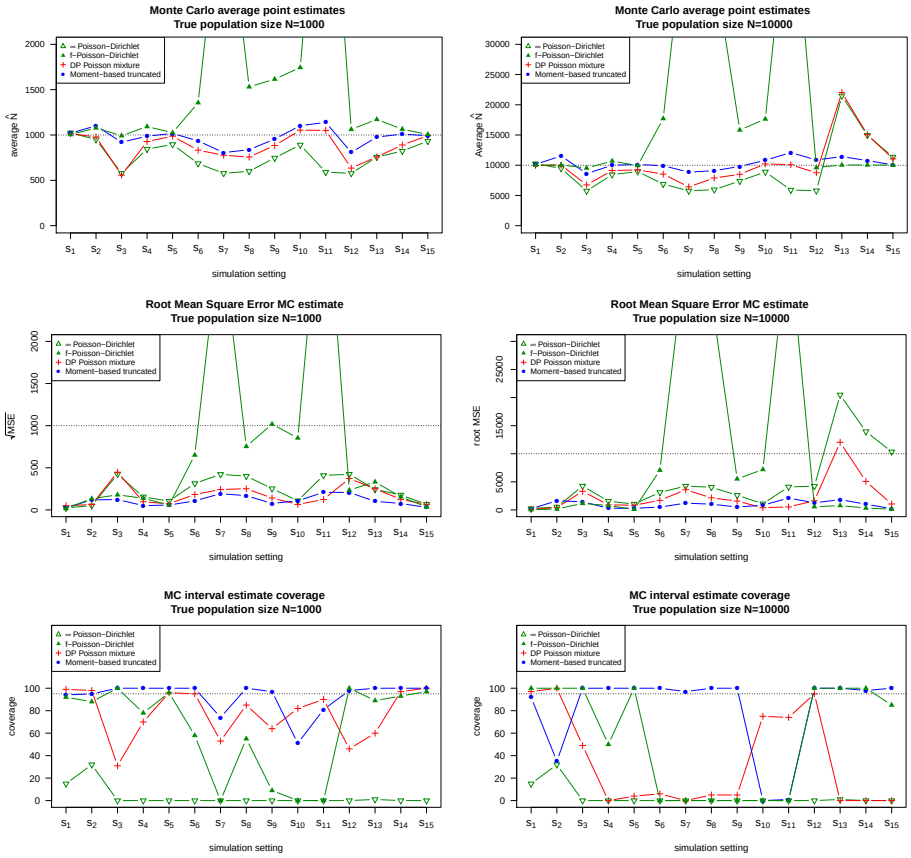

Fig. 1 Comparative performance of 4 alternative Bayesian methods

(2007) are admittedly unfair since in their approach they consider an infinite number of species. However one must also take into account that the comparison is of interest since that approach can be used in real applications where the population size cannot be indeed assumed to unbounded.

4 Concluding remarks

Particular attention to the performance of alternative methods in an inferential context where inference can be challenging and non standard asymptotics is expected. In this framework Bayesian posterior inference can be more sensitive to the prior input and this should be properly taken into account in the absence of genuine prior information. In this sense we believe that our simulation study conducted under al-

ternative simulation settings as those proposed in the frequentist analysis of Wang (2010) could provide some practical suggestion for practitioners. Indeed we have also highlighted some possible drawbacks in using Bayesian nonparametric methods for species sampling based on Gibbs-type prior relying on the assumption of infinitely many species. A more extensive simulation study should be carried out to understand at what extent Bayesian nonparametric methods are sensible and numerically robust when the size of the underlying population grows. To our knowledge there is lack of theoretical understanding of this asymptotic behaviour even in the classical frequentist framework Wang (2010).

References

D. Alunni Fegatelli. *New methods for capture-recapture modelling with behavioural response and individual heterogeneity*. PhD thesis, Dipartimento di Scienze Statistiche, Sapienza Università di Roma, 2013.

Danilo Alunni Fegatelli and Luca Tardella. Moment-based bayesian poisson mixtures for inferring unobserved units. *arXiv preprint arXiv:https://arxiv.org/submit/2299462*, 2018.

J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858, 2017.

Kathryn Barger and John Bunge. Objective Bayesian estimation for the number of species. *Bayesian Analysis*, 5(4):765–786, 2010.

P. G. Bissiri, A. Ongaro, and S. G. Walker. Species sampling models: consistency for the number of species. *Biometrika*, 100(3):771–777, 2013.

Dankmar Böhning and Peter G. M. van der Heijden. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Ann. Appl. Stat.*, 3(2):595–610, 2009.

J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88:364–373, 1993.

Annalisa Cerquetti. Reparametrizing the two-parameter gnedin-fisher partition model in a bayesian perspective. pages 4678–4683, 8 2011.

Anne Chao and John Bunge. Estimating the number of species in a stochastic abundance model. *Biometrics*, 58(3):531–539, 2002.

Anne Chao and Chun-Huo Chiu. *Species Richness: Estimation and Comparison*, pages 1–26. American Cancer Society, 2016.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.*, 87(417):210–217, 1992.

S. Favaro, A. Lijoi, and I. Prunster. A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196, Dec 2012.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12):83–102, 244–245, 2005.

Alexander Gnedin. A species sampling model with finitely many types. *Electron. Commun. Probab.*, 15:79–88, 2010.

Michele Guindani, Nuno Sepúlveda, Carlos Daniel Paulino, and Peter Müller. A Bayesian Semi-parametric Approach for the Differential Analysis of Sequence Counts Data. *Journal of the Royal Statistical Society Series C*, 63(3):385–404, 2014.

Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

C. J. Lloyd, P. S. F. Yip, and Kin Sun Chan. Estimating the number of faults: efficiency of removal, recapture, and seeding. *IEEE Transactions on Reliability*, 48(4):369–376, Dec 1999.

Chang Xuan Mao. Inference on the number of species through geometric lower bounds. *Journal of the American Statistical Association*, 101(476):1663–1670, 2006.

Chang Xuan Mao, Nan Yang, and Jinhua Zhong. On population size estimators in the Poisson mixture model. *Biometrics*, 69(3):758–765, 2013.

James L. Norris and Kenneth H. Pollock. Non-parametric mle for poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics*, 5(4):391–402, Dec 1998.

Ji-Ping Wang. Estimating species richness by a Poisson-compound gamma model. *Biometrika*, 97(3):727–740, 2010. With supplementary data available online.

Ji-Ping Z. Wang and Bruce G. Lindsay. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, 100(471):942–959, 2005.

Mingyuan Zhou, Stefano Favaro, and Stephen G Walker. Frequency of frequencies distributions and size-dependent exchangeable random partitions. *Journal of the American Statistical Association*, 112(520):1623–1635, 2017.