Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Published by Pearson

www.pearson.com

ISBN 9788891910776

Contents

Specialized sessions

Solicited Sessions

Contributed papers and Posters

Preface

The COVID-19 pandemic is putting our society under incredible health, emotional, and economic stress. Facing its harmful effects and their uncertainty, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee, to ensure the highest level of safety for members and delegates, deliberated to cancel the 50th Meeting of the Italian Statistical Society originally planned to be held in Pisa in June 2020 and to postpone the conference to June 2021. The Executive Board and the Local Organizing Committee continue to monitor closely the pandemic evolving situation, and keep the members of SIS and the researchers informed about the potential new dates for the next meeting. To give value to the work of those who prepared their presentation for the conference, the Program Committee decided to publish the volume *Book of short papers - SIS 2020* despite the conference cancellation.

The conference program included 4 plenary sessions, 16 specialized sessions, 24 solicited sessions, 32 contributed sessions and the poster exhibition. Plenary sessions concerned with robust statistics, human longevity, statistical models for climate changes and small area estimation for educational poverty. The meeting had to host also 2 round tables on data privacy and innovation in statistics. Activities focused on topics of interest for a wider audience included two round tables on Teaching Statistics and on the SIS journal Statistical Methods & Applications, and the Stats Under the Stars (SUS6) competition for young statisticians. The SUS6 event attracted many sponsors from statistical, financial and editorial firms as well as numerous students. The conference committee had registered 345 accepted submissions, including 143 to be presented in invited plenary, specialized and solicited sessions, and 202 spontaneously submitted for oral and poster sessions.

This book includes most of the scientific contributions that had to be presented at the 50th Meeting of the Italian Statistical Society. It is organized into 49 chapters corresponding to 15 specialized, 23 solicited sessions, and to 11 general topics for contributed papers and posters. All 268 contributions provide a wide overview of the state-of-the-art of the subjects, from methodological and theoretical contributions, to applied works and case studies. The result is a very lively picture of the Italian statistical community with its international connections.

We would like to thank all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Although the Conference did not take place, the organization went on until cancellation was decided for safety reasons. It would have been impossible without the joint effort of Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members these three institutions took part actively in the Local Organizing Committee. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

This book is our contribution to encourage the scientific community and the network of the Italian Statistical Society to go on and transform this difficult period into an opportunity of scientific debate for better statistics in a better world.

Alessio Pollice
Università degli Studi di Bari Aldo Moro
Chair of the Program Committee

Nicola Salvati
Università di Pisa
Chair of the Local Organizing Committee

Francesco Schirripa Spagnolo
Università di Pisa

Program Committee: Alessio Pollice (Chair), Serena Arima, Marilena Barbieri, Alessandra Brazzale, Eugenio Brentari, Alessia Caponera, Antonio Lepore, Antonella Plaia, Tommaso Proietti, MarcoRiani, Nicola Salvati, Pasquale Sarnacchiaro, Mauro Scanu, Manuela Stranges, Valentina Tocchioni, Simone Vantini, Massimo Ventrucci, Paola Vicard, Donatella Vicari.

Local Organizing Committee: Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Alessandra Coli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

Organizers of Specialized and Solicited Sessions: Giada Adelfio, Bruno Arpino, Emanuele Aliverti, Nicoletta Balbo, Mara Bernardi, Silvia Bozza, Pierpaolo Brutti. Annalisa Busetta, Michela Cameletti, Carlo Cavicchia, Fabrizio Durante, Leonardo Egidi, Pietro D. Falorsi, Francesco Finazzi, Livio Finos, Stefania Galimberti, Michele Gallo, Caterina Giusti, Francesca Greselin, Alessandra Guglielmi, Francesca Ieva, Tiziana Laureti, Achille Lemmi, Brunero Liseo, Fabio Massimo Lo Verde, Daria Mendola, Roberta Pappadà, Lea Petrella, Alessandra Petrucci, Alessia Pini, Sabrina Prati, Maria Giovanna Ranalli, Davide Risso, Fabrizio Ruggeri, Silvana Salvini, Monica Scannapieco, Francesco Stingo, Luca Tardella, Grazia Vicario, Susanna Zaccarin, Maroussa Zagoraiou.

Bayesian Model Averaging for Latent Class Models in Capture–Recapture

Model Averaging Bayesiano per Modelli a Classi Latenti in Cattura-Ricattura

Davide Di Cecco

Abstract Model selection appears to be crucial in capture-recapture problems as it is common that different models with an equally good level of adaptation to the observed data lead to rather different estimates of the undercounts. We consider log–linear Latent Class Models as our capture-recapture model and propose Bayesian model averaging to overcome the difficulties of model selection within this class. We show that, by focusing on graphical decomposable models, we can design a simple Gibbs–based MCMC to sample over the space of eligible models.
Abstract *In problemi di cattura–ricattura, la selezione del modello risulta essere un aspetto delicato e difficoltoso. Non è inusuale, infatti, trovare modelli con valori di bontà di adattamento molto simili tra di loro che conducono a delle stime del numero di unità non catturate molto diverse. In questo lavoro trattiamo una famiglia estesa di modelli a classi latenti che rilassa l'ipotesi di indipendenza condizionata modellando le interazioni tra le variabili attraverso un modello log–lineare con una variabile latente. Per superare la difficoltà di scelta del modello all'interno di questa classe ampliata, proponiamo un model averaging Bayesiano. Mostriamo come, se ci limitiamo a considerare i modelli log-lineari decomponibili, è possibile costruire un semplice Gibbs sampler per ottenere la distribuzione a posteriori della numerosità della popolazione d'interesse.*

Key words: Bayesian Model Averaging, Latent Class Models, Capture-Recapture

1 Introduction

As pointed out by many authors, see, e.g., [4], the problem of estimating the size of a population in a capture-recapture model is essentially a problem of forecasting. As a consequence, it is not unusual that different models with a comparable level of

Davide Di Cecco
Sapienza University, viale del Castro Laurenziano 9, e-mail: davide.dicecco@uniroma1.it

goodness of fit lead to rather different estimates of the total population count. Given the lack of specific criteria for model choice, and the impossibility to validate the estimates, it is not unusual in capture-recapture practice to simply rule out a model resulting in unrealistic estimates. We think that a reliable procedure to deal with model selection in a more automatic way would certainly be of interest.

We treat the case of Multiple Record System, that is, the data consists of a set of capturing lists, usually originating from different sources, reporting partial listing of the same target population. In this setting it is common to assume different capture probabilities for the various sources. As a consequence, log-linear models are the tool of choice in capture–recapture modeling, and Latent Class Models (LCM) represent the natural extension when one wants to include unobserved heterogeneity. The use of LCM in capture–recapture dates back at least to [1] with many developments thereafter. The simplest formulation of LCM envisages the conditional independence assumption (CIA) which appears to be too restrictive in many situations. There are many proposals in literature to relax the CIA resulting in more flexible models. We focus on log-linear LCM where the additional dependencies are directly modeled by interaction parameters. Previous works on this class include [3], [13], [12]. We propose a Bayesian approach to the class as previously introduced in [6] and [7]. To overcome the difficulty of model selection within this class, we propose Bayesian model averaging to analyze the posterior distribution of the population count over a set of eligible models. Usually a full Bayesian approach to model averaging requires the use of a Reversible Jump algorithm ([8]) which is in general hard to implement. See [11] for an example of use of the algorithm within the class of log-linear models (without a latent variable). We show that, if we restrict ourselves to the subclass of decomposable models, it is possible to implement a simple Gibbs–based MCMC. Some preliminary results on simulated data (not shown in this work for space limitation), seem to indicate that the restriction to that subclass does not affect the efficacy of the procedure.

2 The model

Consider k capturing variables $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_i = 1$ if a certain unit is listed in the i-th source and 0 otherwise, and let X be the latent variable taking values in $\{1, \ldots, m\}$ identifying the latent classes of our population. The LCM under the CIA can be equivalently expressed as the mixture model

$$P(\mathbf{Y} = \mathbf{y}) = p_{\mathbf{y}} = \sum_{x=1}^{m} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Y_i = y_i|X = x)$, or as the log-linear model

$$[XY_1][XY_2]\cdots[XY_k], \tag{2}$$

where we use the classic notation reporting only the higher order interactions (generators) of the model. The dependence graph of this model is a star–shaped graph, where the node representing X is connected to all other nodes, like the one in Figure 1 on the left. Any additional interaction term with respect to (2), (i.e. any additional arc in the graph), constitutes a relaxation of the CIA.

Denote the number of observed units as n_{obs}, and the number of units presenting the capture profile $\mathbf{y} \in \{0,1\}^k$ as $n_{\mathbf{y}}$. Let $n_{x,\mathbf{y}}$ be the number of units having that profile which belong to latent class x, so that $\sum_x n_{x,\mathbf{y}} = n_{\mathbf{y}}$, and $n_{\mathbf{x},\mathbf{y}}$ be the vector $(n_{x,\mathbf{y}})_{x=1,\ldots,m}$. Let $n_{\mathbf{0}}$ be the number of uncaptured units to be estimated and N the total population count such that $\sum_{\mathbf{y}} n_{\mathbf{y}} + n_{\mathbf{0}} = N$. Let M be the (random variable associated to) the model to be chosen in a pre determined set $\mathscr{M}$, and Θ_M the set of parameters associated to model M.

2.1 Prior distributions

Obviously, the choice of a prior on the set of models $\mathscr{M}$, as on any other parameter, is subjectively arbitrary. However, we usually just want to exclude some cases, and set a uniform prior on all remaining cases. As we have said, we focus on decomposable models and, as an elementary additional criterion, we rule out all unidentifiable models. In [14] we can find a necessary and sufficient condition for the identifiability of any graphical model (hence of any decomposable model too).

In practice, to utilize the proposed methodology, we have to list out all possible models for a given number of variables. That is, all identifiable models represented by decomposable graphs containing the star–shaped graph relative to the CIA. Consider the case $k = 4$: in Figure 1 we have all decomposable graphs grouped by isomorphism. This leaves us with 19 eligible models. When $k = 5$ the number of identifiable decomposable models goes up to 355.

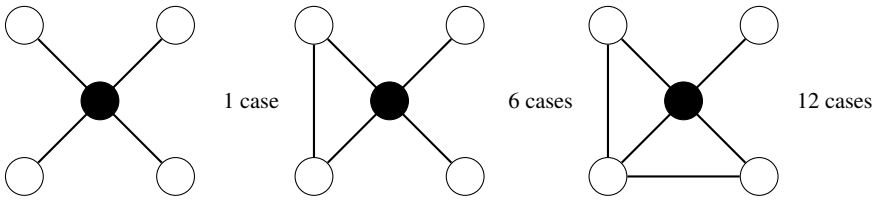

Fig. 1 Identifiable decomposable graph models with 4 manifests (empty nodes) and a latent (black node) grouped by classes of isomorphism

As for the prior distributions over the parameters Θ_M of each model M, we utilize the Hyper Dirichlet distribution described in [5]. Such a choice allows us to exploit

the result of [9] giving an analytical formula for the posterior probabilities of the models given the data.

3 Model averaging

We analyze the posterior distribution of N given the observed data $\{n_{\mathbf{y}}\}$ with $\mathbf{y} \neq \mathbf{0}$,

$$\pi(N \mid \{n_{\mathbf{y}}\}) = \sum_{M \in \mathscr{M}} \pi(N \mid M, \{n_{\mathbf{y}}\})\, \pi(M \mid \{n_{\mathbf{y}}\}).$$

The posterior probability of model M is given by:

$$\pi(M \mid \{n_{\mathbf{y}}\}) \propto \pi(M)\pi(\{n_{\mathbf{y}}\} \mid M) = \pi(M) \int \pi(\{n_{\mathbf{y}}\} \mid M, \Theta_M)\pi(\Theta_M \mid M) d\Theta_M.$$

In practice, given the computational complexity of calculating the last integral quantity, one can settle for the simplest (first order) approximation of the marginal likelihood of a model based on the Bayesian Information Criteria (BIC), that is, $\exp(-BIC/2)$, which, given equal prior probabilities for the models, leads to the following approximation of the weights $\pi(M \mid \{n_{\mathbf{y}}\})$:

$$\frac{\exp(-BIC_M/2)}{\sum_{M \in \mathscr{M}} \exp(-BIC_M/2)}, \tag{3}$$

where BIC_M is the BIC of model M. Then, one can use those weights in computing the averaged mean

$$E[N \mid \{n_{\mathbf{y}}\}] = \sum_{M \in \mathscr{M}} \widehat{N}_M\, \pi(M \mid \{n_{\mathbf{y}}\}), \tag{4}$$

where $\widehat{N}_M$ is the posterior mean of N under model M.

By using formula (4), one should keep in mind that the approximation quality can be poor in some cases, and, in any case, we limit ourselves to a point estimate of N. A full Bayesian approach to the problem, on the other hand, would result in an estimate of the whole posterior distribution of N marginalized over $\mathscr{M}$.

4 The Gibbs sampler

In this section we outline a Gibbs–based MCMC algorithm to sample from the joint distribution of $(N, N_{\mathbf{x},\mathbf{y}}, M, \Theta_M)$, conditioned on the observed data $\{n_{\mathbf{y}}\}$. Note that we cannot obtain the full conditionals for all terms: as pointed out in [2] and in [10], given $n_{\mathbf{x},\mathbf{0}}$, the value of N is deterministically defined. As a workaround, they propose to consider the conditional distribution of the couple $N, N_{\mathbf{x},\mathbf{0}}$ conditionally

on the rest. Similarly, we cannot obtain the conditional distribution of M given the parameters Θ_M. For these reasons, the algorithm loops over the following steps:

1) sample $n^{(t)}_{\mathbf{x},\mathbf{y}}$ from $\pi(N_{\mathbf{x},\mathbf{y}} \,|\, N, M, \Theta_M, \{n_{\mathbf{y}}\})$, for all $\mathbf{y} \neq \mathbf{0}$,

$$N_{\mathbf{x},\mathbf{y}} \sim Mult\left(n_{\mathbf{y}}, p_{\mathbf{x}|\mathbf{y}}\right),$$

where the $p_{\mathbf{x}|\mathbf{y}}$ are calculated according to the current value of M and Θ_M;

2) sample a couple $\left(N^{(t)}, n^{(t)}_{\mathbf{x},\mathbf{0}}\right)$ from

$$\pi(N, N_{\mathbf{x},\mathbf{0}} \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\})\, \pi(N_{\mathbf{x},\mathbf{0}} \,|\, N, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}).$$

Note that

$$\pi(N \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N \,|\, M, \Theta_M, n_{obs}) \propto \pi(N) \binom{N}{n_{obs}} p_{\mathbf{0}}^{N-n_{obs}} (1-p_{\mathbf{0}})^{n_{obs}},$$

then, if we choose the improper prior $\pi(N) \propto 1/N$, the conditional distribution of N results in a Negative Binomial distribution, and we simply have to

- sample $N^{(t)}$ from $NegBin\,(n_{obs}, 1-p_{\mathbf{0}})$;
- sample $n^{(t)}_{\mathbf{x},\mathbf{0}}$ from $Mult\big((N^{(t)} - n_{obs}), p_{\mathbf{x}|\mathbf{0}}\big)$.

where $p_{\mathbf{x}|\mathbf{0}}$ and $p_{\mathbf{0}}$ are calculated according to the current value of M and Θ_M;

3) sample from $\pi(M, \Theta_M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(\Theta_M \,|\, M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})\, \pi(M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$. That is,

- calculate the posterior probability of each eligible decomposable model and sample $M^{(t)}$ from $\pi(M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$. The posterior probability of each model M is defined as a product of Gamma functions (see [9]);
- then sample all parameters $\Theta^{(t)}_M$ from their posterior conditional distribution $\pi(\Theta_M \,|\, M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$, which is a product of Dirichlet distributions.

5 Conclusions

The proposed algorithm can be just used for model selection by simply inspecting the generated values of M, as the relative frequency of each model constitutes an estimate of its posterior probability, and select the best model accordingly. However, model averaging seems to be the best choice in capture-recapture problems. Compared to the usual approximation techniques, our estimates should be more accurate, and allow to inspect the whole posterior distribution of N at the cost of some additional computational effort which appears nonetheless reasonable. The restriction to decomposable models may seem a severe limiting factor, as they constitutes a minority fraction of the possible models, and many frequently used models, such as the one with all second order and no higher order interactions, are left out. How-

ever, some preliminary results based on simulations (not shown in this work for space limitation), appears to be encouraging. In fact, the proposed approach seems to work well even when used on data generated from non decomposable models.

References

[1] A. Agresti. Simple capture–recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2):494–500, 1994.

[2] S. Basu and N. Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.

[3] A. Biggeri, E. Stanghellini, F. Merletti, and M. Marchi. Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population. *Statistica Applicata*, 11(3):1–14, 1999.

[4] B.A. Coull and A. Agresti. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1):294–301, 1999.

[5] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272–1317, 1993.

[6] D. Di Cecco. Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data*, pages 169–196. Chapman and Hall/CRC, 2019.

[7] D. Di Cecco, M. Di Zio, and B. Liseo. Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal*, 2020.

[8] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[9] D. Madigan and J. C. York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.

[10] D. Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.

[11] A. M. Overstall and R. King. `conting`: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58(7):1–27, 2014.

[12] E. Stanghellini and M. G. Ranalli. Population size estimation using a categorical latent variable. In *Capture-Recapture Methods for the Social and Medical Sciences*, pages 277–290. Chapman and Hall/CRC, 2017.

[13] E. Stanghellini and P. G. M. van der Heijden. A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, 60(2):510–516, 2004.

[14] E. Stanghellini and B. Vantaggi. Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A):1920–1937, 2013.