Bayesian modeling of presence-only data

Fabio Divino, Giovanna Jona Lasinio, Natalia Golini

Abstract The prediction of species distribution in suitable regions is essential for planning conservation and management strategies. Unfortunately, quite often the only available information is the presence of the species at few locations while the associated environmental covariates can be observed over the complete area of interest. This kind of situation can be seen as a missing data problem with asymmetric and partial information, we say that data are presence-only data. In this paper we present a Bayesian approach to handle with presence-only data, we also consider the case when a spatial effect acts among the observations. MCMC computation has been implemented through a data augmentation algorithm allowing us to result consistent estimates for the regression parameters jointly with the unknown species prevalence.

1 Introduction

In ecology, the evaluation and the prediction of the spatial distributions of species and their interactions with environmental factors are of primary interest in order to better plan and manage strategies in habitat conservation. Binary responses, indicating the presence-absence process of a species, are usually related to the set of the explicative covariates through the use of logistic regressions. However, in many ecological analyses the complete collection of the binary responses could be quite expensive or even very difficult to be obtained, we could observe the true presence of the species only at few locations of the study area while the associated environmental covariates can be available over the whole region. In that case we define the data as presence-only data. In this work we present the hierarchical modeling introduced by Ward et al. (2009) and developed in a Bayesian framework by Divino et al. (2011). Through the use of a random approximation, it is possible to adapt the adjusted logistic model for

[1] Fabio Divino, Università del Molise; email: fabio.divio@unimol.it
Giovanna Jona Lasinio, Università di Roma "La Sapienza"
Natalia Golini, Università di Roma "La Sapienza"

case-controls studies also in the setting of the presence-only data, overcoming the need to know the prevalence of the species *a priori*. Estimation of the regression parameters jointly with the prevalence can be carried out through a data augmentation MCMC algorithm.

2 Model and Computation

With respect to a population U of spatially referenced sites i, let $\mathbf{Y}=(Y_i\ ;\ i\in U)$ be a binary process concerning the presence-absence of a species of interest, $\mathbf{X}=(X_i\ ;\ i\in U)$ a set of covariates and U_p the subset of U where the species is present (Y=1). When only presences are observed, samples from the process $\mathbf{Y}$ can be drawn only from the population U_p and the usual case-control approach through the logistic regression can not be adopted as the absences (Y=0) are not directly observed. Lancaster and Imbens (1996) and Ward et al. (2009) proposed to overcome this problem by considering a combination of two independent random samples: the first sample S_p is a sample of cases from the population of presences U_p while the second one S_u is a sample of "pseudo" controls collected from the whole population U. In this way the complete data sample S is composed by n_p presences (observed in S_p) and n_u unobserved values (collected in S_u). When the binary response Y is rare, this approach represents a naive approximation of the standard case-control design, here we present a different implementation of the design. Let Z be a stratum variable such that Z_i=0 if $i\in S_u$ and Z_i=1 if $i\in S_p$. Notice that Z_i=1 implies Y_i=1 while Z_i=0 implies that Y_i can assume a value in {0, 1}. The relation between Y and Z at the sample level can be represented in the following table.

Table 1: Sample composition with respect to Y and Z.

Y/Z	*Z=0*	*Z=1*	
Y=0	n_{0u}	*0*	n_0
Y=1	n_{1u}	n_{1p}	n_1
	n_u	n_p	n

The only quantities known are n_u and n_p (obviously also n_{1p}). We remark that all the unknowns can be considered random quantities due to the "censor" effect acting on the subsample S_u. In particular we can write n_{1u} as $\tilde{n}_i=\sum_{i\in S_u}\tilde{Y}_i$, where the symbol ~ just indicates the random nature of the quantity. Now let π=P(Y=1) be the prevalence of the species in U. Under the assumption that S_u is a random sample from the whole population U we have that $E[\tilde{n}_{1u}]=\pi n_u$. If we assume that the covariates $\mathbf{X}$ are available for all the sites of the population U, we can use the approach introduced by Ward et al. (2009) and developed in a Bayesian framework by Divino et al. (2011). In the case-control framework the logistic regression for a generic observation enclosed in the sample S with covariates X=x, is given by:

$$P(Y=1\,|\,s=1,\eta,x)=\frac{\exp\left\{\eta(x)+\ln\frac{\gamma_1}{\gamma_0}\right\}}{1+\exp\left\{\eta(x)+\ln\frac{\gamma_1}{\gamma_0}\right\}},$$

where $s=1$ denotes that the site is included in the sample S, $\eta(x)$ is the regression function, γ_0 and γ_1 are the unknown proportion of sampling respectively from the absences and the presences. The ratio γ_1/γ_0 adjusts the logistic model under the case-control design. Following Ward et al. (2009), we can manage presence-only data by considering the joint probability distribution of Y and Z and write the full likelihood model (see Ward et al. (2009) for details) or alternatively consider the observed likelihood, defined only with respect to the distribution of Z, that results in an average over Y. In both the likelihood models, the unknown ratio γ_1/γ_0 can be approximated as follow: $\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} \approx \frac{\tilde{n}_{1u} + n_p}{\tilde{n}_{1u}}$. That expression can be handled into the estimation algorithm through the use of a data augmentation step. In fact, given a current value for $\eta(x)$, it is possible to use the predictive probability distribution of Y to have consistent simulations of the unobserved variables related to the locations enclosed in S_u, then a simple summation over S_u will result an approximation for the quantity n_{1u} allowing us to get available a value for the ratio γ_1/γ_0. In this work we consider regression functions which are linear in the covariates while we present also a model with a spatially structured random effect $\boldsymbol{v}$ accounting for the geographical dependence eventually introduced by latent factors into the species distribution. We can now write the hierarchical Bayesian model. Let θ be the vector of hyperparameters with prior $p(\theta)$. Conditioned on θ, the regression parameters βs are assumed to be Normal distributed while the random effect $\boldsymbol{v}$ is a Gaussian Markov random field. Given β, $\boldsymbol{v}$ and the covariate x, the binary response is a Bernoulli random variable with conditional probability of occurrence given by $\pi_s(x) = P(Y = 1 \mid s = 1, \beta, x)$. At the lowest level of the hierarchy, the conditional distribution of Z given Y can be derived from the above Table 1. Notice that the spatial structure of the random effect $\boldsymbol{v}$ is given by the geographical neighborhood system among all sites in the population U. In the following scheme we describe the MCMC algorithm:

- step 1: initialize θ, β, $\boldsymbol{v}$, the unobserved of $\mathbf{Y}$ over U and set $n_{1u} = \sum_{i \in S_u} Y_i$
- step 2: sample θ from $P(\theta \mid \mathbf{Y}, \mathbf{Z}, \beta, \boldsymbol{v})$
- step 3: sample β from $P(\beta \mid \mathbf{Y}, \mathbf{Z}, \theta)$
- step 4: sample $\boldsymbol{v}$ from $P(\boldsymbol{v} \mid \mathbf{Y}, \mathbf{Z}, \theta)$ over U
- step 5: sample Y_i from $P(Y_i \mid \mathrm{Z}, \beta, v_i, x_i)$ over U

Remark that we need to perform the data augmentation (step 4 and step 5) over the entire population U for both $\boldsymbol{v}$ and the unobserved responses of $\mathbf{Y}$ in order to consider the spatial structure of the sites enclosed in both the samples S_u and S_p. An important feature of this estimation procedure is that we can easily obtain a consistent estimate of the π by $\hat{\pi}_u = \frac{\overline{n}_{1u}}{n_u}$, where $\overline{n}_{1u}$ is the MCMC average of the values in step 1.

3 Simulation study

In this section we report some simulation results. At first, we considered the situation without spatial effect and with only one explicative covariate X. We simulated the

binary response Y from the following logistic model: $\text{logit}\pi(x) = \beta x$, where β=-1, while the covariate X has been generated from a mixture of two Gaussian components with common variance and central values respectively μ=-2 e μ=2. A population U of N=10000 observations over a regular grid 100x100 has been considered, then we randomly sampled the presences in S_p and the "pseudo" absences in S_u in a rate of 1:4. We fitted the Bayesian model, considering the observed likelihood, for two different situations: with prevalence unknown (M1) and with prevalence known (M2). The second situation represents our benchmark, ideally we can not do better than that. Both the models were fitted assuming the Gaussian $N(0,1)$ as prior distribution for β. We ran 20000 iterations and discarded the first 10000 as burn-in. In Table 2 we report the 95% credibility intervals (CI) for β and the estimates for π with respect to different levels of dispersion in the covariate X that generated different true prevalence values in the population π, and with respect to different sizes of S.

Table 2: Simulation results.

Size n		π=0,37		π=0,38		π=0,40	
		β (CI)	$\hat{\pi}_u$	β (CI)	$\hat{\pi}_u$	β (CI)	$\hat{\pi}_u$
N=50	M1	-0,75:0,45	0,49	-1,21:0,12	0,46	-1,87:-0,12	0,46
	M2	-0,80:0,46	0,48	-1,27:0,10	0,46	-1,90:-0,12	0,46
n=500	M1	-1,12:-0,61	0,43	-1,21:-0,69	0,40	-1,10:-0,55	0,41
	M2	-1,11:-0,60	0,42	-1,23:-0,69	0,39	-1,11:-0,58	0,41
n=5000	M1	-1,03:-0,95	0,36	-1,16:-0,99	0,37	-1,13:-092	0,40
	M2	-1,02:-0,94	0,36	-1,15:0,97	0,37	-1,11:-0,91	0,40
Variance of X		0,25		1,0		4,0	

In the second study we introduced the spatial effect $\boldsymbol{u}$ as an intrinsic Gaussian Markov random field but we refer to the oral presentation for the details of that experiment.

4 Comments and Conclusion

Results are encouraging, especially in term of predictive capacity. In the first experiment, as the sample size increase the prevalence estimates become more consistent and closer to the true parameters with respect to all the different levels of dispersion in the covariate X. Further work concerns the investigation of the parameters identifiability and the overfitting that often results when the spatial effect is enclosed into the model.

References

1. Divino, F., Jona Lasinio, G., Golini, N., Pettinen, A.: Data augmentation approach in Bayesian modelling of presence-only data. Procedia Environmental Sciences, 7, 38-43 (2011)
2. Lancaster, T., Imbens, G.: Case-control studies with contaminated controls. Journal of Econometrics 71,145-160 (1996)
3. Ward, G., Hastie, T., Barry, S., Elith, J., Leathwick, A.: Presence-only data and the EM algorithm. Biometrics, 65, 554-563 (2009)