Running head: OCCUPANCY MODEL EFFICIENCY

Something occupancy models

Lauren C. Ponisio[1,2], Nicholas Michaud[1], Perry de Valpine[1]

1. Department of Environmental Science, Policy, and Management
   University of California, Berkeley
   130 Mulford Hall
   Berkeley, California, USA
   94720

2. Department of Entomology
   University of California, Riverside
   417 Entomology Bldg.
   Riverside, California, USA
   92521

Abstract

1. occupancy models are everywhere, but model fitting and assessment are extremely computationally intensive
2. because models are so computationally intensive, users often forgo model assessment (determining if a model provides an adequate fit to a particular dataset) because if involves simulating from and refitting the model many times.
3. Using the NIMBLE package for R, we develop combined computational approaches including user-defined and automatic blocking of parameters for MCMC, filtering over latent states, and customized MCMC samplers for specific parameters to improve efficiency. We test these approaches using three representative occupancy models of varying levels of complexity including a single species model with spatial auto-correlation, a single species dynamic (multi-season) model, and a multi-species model. We also develop and implement methods for calculating calibrated predictive posterior *p*-values to assess model fit within the open source modeling software, NIMBLE.
4. These computation approaches lead to an improvement in MCMC sampling efficiency over, particularly with models including random effects.
5. Ours results highlight the need for more customizable approaches to MCMC to fit and assess hierarchical models in order to ensure occupancy models are accessible to practitioners. By implementing MCMC procedures and model assessment techniques in open source software, we have made progress toward this aim.
6. *Implications:*

NIMBLE, Markov chain Monte Carlo, latent states, block sampling, dynamic occupancy, mutli species occupancy, spatial occupancy, JAGS

Introduction

Estimating the proportion of sites occupied by a species is common challenge for many sub-disciplines ecology and evolution including meta-population, endangered species and invasion biology. Greater acceptance of the biases introduced by imperfect detection has lead to the development and proliferation of occupancy models — models where the occurrence of a species at a site as a latent state layered underneath a detection process (e.g., MacKenzie *et al.*, 2006; Royle & Kéry, 2007a). Now only a little over a decade after occupancy models were introduced to ecology, they are being used to model the occurrence of everything from bees (M'Gonigle *et al.*, 2015) to tigers (Hines *et al.*, 2010) in an endless variety of complexity.

Occupancy models are part of a larger class of models known as Hidden Markov Models. For discrete Hidden Markov Models like occupancy models where a species is either present or absent from a site, likelihood calculation involves summing over the distribution of latent states. Because estimating the effect of explanatory variables on site occupancy or shared variation of in occupancy across species is often of greatest interest to ecologists (e.g., Iknayan *et al.*, 2014), the Hidden Markov Models are embedded within a hierarchical model. In such cases, practitioners generally rely on Markov chain Monte Carlo (MCMC) to perform a Bayesian analysis. Standard MCMC software will including the latent state variables in MCMC sampling (e.g., Plummer *et al.*, 2003; win; ope). Such models are computationally intensive, and large models requiring hundreds or thousands of dimensions which require MCMC can be intractable.

In addition, fitting these models is such a challenge that users often forgo adding any additional computation to asses model fit. A common idea behind evaluating whether a model provides an adequate fit to a particular dataset is that if data is simulated from the model, the simulated data should resemble the observed data. This is the basis of posterior predictive p-values, which compare the distribution of summary statistics calculated from simulated datasets to the observed statistic. Posterior predictive p-values alone, however, often fail to reject poor-fitting models (Bayarri & Berger, 2000; Robins *et al.*, 2000; Hjort *et al.*, 2006). Methods for correcting posterior predictive p-values for better performance have been proposed (e.g., calibrated posterior predictive p-values, Hjort *et al.*, 2006), but refitting the model via MCMC iterativly. Given that with occupancy models fitting the model just once can be a time consuming task, efficient methods for MCMC are necessary to ensure methods for assessment are feasible for these models.

Beyond assessing the fit of a model, choosing between models is one of the most widely used applications of statistics by practitioners. Though many theoretically sound methods for Bayesian model selection such as cross-validation have been developed (Hooten & Hobbs, 2014), they, like model assessment, are computationally intensive — particularly for hierarchical models like occupancy models. A typical need for model selection arises when a practitioner is choosing whether to include a specific layer of hierarchy (i.e., random effect). This is often the case with so called "multi-species" occupancy models, where the occupancy of many species is estimated simultaneously in a model with a random effect of species (reviewed in, Iknayan *et al.*, 2014). Ecologists are often interested in whether there is some variability in the response of species to an explanatory variable such that a random effect of species accounts for that variability, or whether a fixed effect of that variable fits adequately (Pacifici *et al.*, 2014). Currently, the Deviance Information Criteria (DIC), originally derived to mimic AIC for Bayesian, non-hierarchical models, is now commonly used by scientists to evaluate hierarchical models. Though the limitations of DIC for hierarchical model selection are widely recognized by statisticians (Celeux *et al.*, 2006; Hooten & Hobbs, 2014), because it is built into open-source software such as WinBUGS (win), it is uncritically used by practitioners. Readily available and theoretically sound alternative methods are thus critically needed.

Materials & Methods

Computational approaches

Single species, single season occupancy model with spatial auto-correlation

The first model is a single species, single season occupancy model accounting for spatial auto-correlation. We let z_i denote the true occupancy of a species at site i. We then let $x_{i,j}$ indicate whether we detected ($x_{i,j} = 1$) or did not detect ($x_{i,j} = 0$) that species in the j^{th} visit to site i. We assumed that occupancy at the i^{th} site is a Bernoulli random variable $z_i \sim \text{Bern}(\psi_i)$ with probability ψ_i. We included the effect of an arbitrary covariate (e.g., elevation) on site occupancy. To model the spatial auto-correlation in occupancy between sites, we assume the co-variance between sites Y_i and Y_j is a function of distance between p_i and p_j. We computed the probability of occupancy at site i

$$\begin{aligned} \text{logit}(\psi_i) &= \alpha + \beta * elevation_i + \rho_i \\ \rho_i &\sim MVN(0, Cov(Y_i, Y_j)) \\ Cov(Y_i, Y_j) &= \sigma^2 exp^{(-\lambda \|p_i - p_j\|)} . \end{aligned} \tag{1}$$

Where λ is the exponential decay constant and σ^2 is SOMETHING....

To improve efficiency of this model...

Single species, multi season (dynamic) occupancy model

The second model is a relatively simple single species occupancy model over multiple seasons (Royle & Kéry, 2007b). We let $z_{i,t}$ denote the true occupancy of a species in year t at site i. We then let $x_{i,t,j}$ indicate whether we detected ($x_{i,t,j} = 1$) or did not detect ($x_{i,t,j} = 0$) that species in the j^{th} visit to site i in year t. We assumed that occupancy at the i^{th} site in the t^{th} year is a Bernoulli random variable $z_{i,t} \sim \text{Bern}(\psi_{i,t})$ with probability $\psi_{i,t}$.

Letting $\phi_{i,t}$ denote the probability the species persists at site i from years t to $t+1$ (provided it was present at site j in year t, $z_{i,t} = 1$) and $\gamma_{i,t}$ denote the probability that site i is colonized in year $t+1$ (provided it was not present at site i in year t, $z_{i,t} = 0$), we then computed the probability of occupancy at site i in subsequent years as

$$\psi_{i,t+1} = \phi_{i,t} * z_{i,t} + \gamma_{i,t} * (1 - z_{i,t})\,. \tag{2}$$

First, to improve efficiency, filter over latent states to calculate model likelihoods in order to limit MCMC sampling to top-level parameters. We then use two computational approaches to improve the efficiency of this model 1) dynamic blocking of the parameters (Turek *et al.*, 2016), and 2) a custom MCMC specification where a slice sampler is used for all parameters.

Multi species, single season occupancy model

The last model is from (Zipkin *et al.*, 2010). It is a multi-species, single season occupancy model examining the effect of wildlife management and habitat characteristics on bird communities (Zipkin *et al.*, 2010). The species-specific coefficients for the effect of basal tree area, understory foliage and deer management where bound together by a common distribution with an estimated variance. The model is similar to Eq. 1, except each for species k, we let $z_{i,k}$ denote its true occupancy state at site i.

To improve the efficiency of this model, we first filtered over latent states to calculate model likelihoods in order to limit MCMC sampling to top-level parameters. We also vectorized all calculations that would have require for loops in BUGS or JAGS. We then tried two approaches to speed sampling of the top-level parameters 1) dynamic blocking of the parameters (Turek *et al.*, 2016), and 2) a custom blocking scheme were the parameters of each species are blocked together.

Model assessment

We implemented a procedure to calculate calibrated posterior predictive p-values (Hjort *et al.*, 2006). After the parameters have been fit to the model, a sample of the posterior is used to simulate data from the model. A discrepancy measure, which we chose to be the model likelihood, is then calculated, and the posterior p-value is the number of simulated p-values that fall below the observed. To "calibrate" the distribution of posterior p-values, the MCMC is rerun on the simulated data to refit the model. THEN?

Model selection

Cross-validation is one of the most fudamental procedures in model selection, but, because it requires iternativly re-fitting the model, is computationally intensive (Hooten & Hobbs, 2014). In cross validation, we exclude a subset of the data (y_k) from model fitting, then use the fitted model to predict y_k. The preidtion error is summarized by coparing the simulated y_k to the true y_k.

In the multi species occupancy models (Section), practitioners are often interested in determining whether a model including a species random effect for explanatory variables is a better fit than a model without the random effect. We implemented a cross-validation procudure for this model where the detection data for species is left out, the model refitted, and the fitted model used to predict the occurence of that species. The predictive error of the model included a random effect of species is then compared to a model where no species random effects were included.

Results

Discussion

Acknowledgments

References

(????) OpenBUGS. http://www.openbugs.net/w/FrontPage.

(????) WinBUGS. http://www.mrc-bsu.cam.ac.uk/software/bugs/.

Bayarri, M. & Berger, J. (2000) P values for composite null models. *Journal of the American Statistical Association*, **95**, 1127–1142.

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M. *et al.* (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–673.

Hines, J., Nichols, J., Royle, J., MacKenzie, D., Gopalaswamy, A., Kumar, N. & Karanth, K.

(2010) Tigers on trails: occupancy modeling for cluster sampling. *Ecological Applications*, **20**, 1456–1466.

Hjort, N.L., Dahl, F.A. & Hognadottir, G. (2006) Post-processing posterior predictive p values. *Journal of the American Statistical Association*, **101**, 1157–1174.

Hooten, M.B. & Hobbs, N.T. (2014) A guide to Bayesian model selection for ecologists. *Ecological Monographs*, pp. in press, online early.

Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, **29**, 97–106.

MacKenzie, D., Nichols, J., Royle, J., Pollock, K., Bailey, L. & Hines, J. (2006) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, Burlington, Massachusetts, USA.

M'Gonigle, L., Ponisio, L., Cutler, K. & Kremen, C. (2015) Habitat restoration promotes pollinator persistence and colonization in intensively-managed agriculture. *Ecol Appl*, **25**, 1557–1565.

Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & DeWan, A. (2014) Guidelines for a priori grouping of species in hierarchical community models. *Ecology and evolution*, **4**, 877–888.

Plummer, M. *et al.* (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pp. 20–22.

Robins, J., van der Vaart, A. & Ventura, V. (2000) Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, **95**, 1143–1156.

Royle, A. & Kéry, M. (2007a) A bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–1823.

Royle, A. & Kéry, M. (2007b) A bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–1823.

Turek, D., de Valpine, P. & Paciorek, C.J. (2016) Efficient markov chain monte carlo sampling for hierarchical hidden markov models. *arXiv preprint arXiv:160102698*.

Zipkin, E.F., Royle, J.A., Dawson, D.K. & Bates, S. (2010) Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation*, **143**, 479–484.