Ecological Modelling 283 (2014) 62–69

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Review

Bayesian model selection: The steepest mountain to climb

Simone Tenan[a,b,c,*], Robert B. O'Hara[d], Iris Hendriks[e], Giacomo Tavecchia[a]

[a] Population Ecology Group, IMEDEA (CSIC-UIB), Miquel Marqués 21, 07190 Esporles, Mallorca, Spain
[b] Sezione Zoologia dei Vertebrati, MUSE – Museo delle Scienze, Corso del Lavoro e della Scienza 3, I-38123 Trento, Italy
[c] DSTA – Dipartimento di Scienze della Terra e dell'Ambiente, Università di Pavia, Via Adolfo Ferrata 9, I-27100 Pavia, Italy
[d] Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany
[e] Global Change department, IMEDEA (CSIC-UIB), Miquel Marqués 21, 07190 Esporles, Mallorca, Spain

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form 24 March 2014
Accepted 25 March 2014

Keywords:
Bayesian analysis
BUGS language
Hierarchical modelling
Hypothesis testing
Model selection
Variable selection

ABSTRACT

Following the advent of MCMC engines Bayesian hierarchical models are becoming increasingly common for modelling ecological data. However, the great enthusiasm for model fitting has not yet encompassed the selection of competing models, despite its fundamental role in the inferential process. This contribution is intended as a starting guide for practical implementation of Bayesian model and variable selection into a general purpose software in BUGS language. We explain two well-known procedures, the product space method and the Gibbs variable selection, clarifying theoretical aspects and practical guidelines through applied examples on the comparison of non-nested models and on the selection of variables in a generalized linear model problem. Despite the relatively wide range of available techniques and the difficulties related to the maximization of sampling efficiency, for their conceptual simplicity and ease of implementation the proposed methods represent useful tools for ecologists and conservation biologists that want to close the loop of a Bayesian analysis.

Contents

* Corresponding author. Current address: Sezione Zoologia dei Vertebrati, MUSE – Museo delle Scienze, Corso del Lavoro e della Scienza 3, I-38123 Trento, Italy. Mobile: +39 349 777 5003.
E-mail address: simone.tenan@muse.it (S. Tenan).

http://dx.doi.org/10.1016/j.ecolmodel.2014.03.017
0304-3800/

1. Introduction

With the recent advance in computational statistics and the refinement of ecological questions there is an increasing interest in a hierarchical approach to ecological modelling (e.g., Clark, 2005; Royle and Dorazio, 2008; Halstead et al., 2012; Kéry and Schaub, 2011). This approach puts emphasis on the distinct components, i.e., processes of ecological systems, leading to the hierarchical models that explicitly incorporate variances from the multiple levels of the information (Royle and Dorazio, 2008; Gelman et al., 2003). The development of Markov chain Monte Carlo (MCMC) framework (Robert and Casella, 2004) and the advent of MCMC engines and the BUGS language (Lunn et al., 2000; Plummer, 2003) have simplified the computational problems caused by the complex structure of the model, and contributed to the fast-growing use of Bayesian hierarchical models in ecology and conservation biology (e.g., Halstead et al., 2012; Kéry and Schaub, 2011; Schaub and Kéry, 2012). However, model validation and selection procedures for Bayesian hierarchical models, which represent a fundamental part of the inferential process (Burnham and Anderson, 2002; Link and Barker, 2006), have not generated the same enthusiasm as model building or fitting. Rather the opposite. Widely used and easy to implement information criteria like DIC (Spiegelhalter et al., 2002) cannot be safely used to compare hierarchical models (Millar, 2009; Kéry and Schaub, 2011). At present, and possibly unknown to many enthusiastic ecologists, model or variable selection in hierarchical models is complex, computationally challenging and no consensus has emerged in the literature on a single approach (Link and Barker, 2006). The problem is not new to statisticians and there are examples of possible alternatives for Bayesian model choice (e.g., Sisson, 2005; Congdon, 2006; Lunn et al., 2009; Ando, 2010; Ntzoufras, 2002), and an exhaustive comparison of the performance of different methods implementable in BUGS is provided by O'Hara and Sillanpää (2009).

Ecological applications are generally available on publications that require a good statistical background (e.g., King, 2009; Link and Barker, 2006; Royle and Dorazio, 2008; Spiegelhalter et al., 1996), and willing ecologists have to fill the gap between exhaustive guidelines to model formulation and fitting (e.g., Kéry, 2010; Kéry and Schaub, 2011; Parent and Rivot, 2012) and highly technical indications on procedures for Bayesian multimodel inference (e.g., Ando, 2010).

Here, conscious of the practical needs of ecologists approaching the Bayesian methods as a mode of analysis and inference, we provide an example to address the problems of both Bayesian model and variable selection using two Gibbs sampler based strategies, taken from the wider survey of methods already implemented using BUGS software (O'Hara and Sillanpää, 2009; Table 1).

This contribution is neither intended to overview existing methods nor to present new ones, but to provide an example application that can aid ecologists to move the first steps into Bayesian multimodel inference.

Examples shall concern two common problems in ecological inference, the comparison of non-nested hierarchical models and the selection of predictors in generalized linear models. We proposed the use of two methods based on the so-called trans-dimensional Markov chains sampling frameworks (Sisson, 2005 for a general review) that permit the construction of Markov chains which simultaneously traverse both parameter and model space.

Details for model implementation are also given in the supplemented R and BUGS codes. All models were implemented in JAGS (Plummer, 2003) through the R (R Core Team, 2012) package R2jags (Su and Yajima, 2012).

Table 1
Approaches to variable and model selection implemented in BUGS language (following O'Hara and Sillanpää, 2009) with the related main reference.

Indicator model selection	
(a) Gibbs variable selection (GVS)	Dellaportas et al. (2000)
(b) Unconditional Priors for variable Selection	Kuo and Mallick (1998)
Stochastic search variable selection (SSVS)	George and McCulloch (1993)
Adaptative shrikage	
(a) Jeffreys' prior	Hobert and Casella (1996)
(b) Laplacian shrinkage	Park and Casella (2008)
Model space approach	
(a) Reversible jump MCMC	Green (1995)
(b) Product space	Carlin and Chib (1995)
(c) Composite model space (CMS)	Godsill (2001)

2. Illustrative real data

We present our applications by using the data from Hendriks et al. (2012), where capture-mark-recapture and individual body size data on a large bivalve (the noble pen shell, *Pinna nobilis*) were used. We considered a subset of the original data, gathered on five sites along the coast of the islands Majorca and Cabrera (Balearic Islands, Spain). In each site, transects of 30 m length were randomly positioned underwater and each transect was randomly assigned to a team of two divers. Capture–recapture data were collected along this line. Each diver marked all noble pen shells found along a side of the transect line using a metal peg with a unique alphanumeric code. Once at the end of the transect line, divers switched side and searched for already marked noble pen shells marked by the previous diver ('re-capture'). The shell width of each marked individual was measured. On subsequent surveys, diver teams changed randomly to minimize a possible 'diver' effect on recapture probability. We considered data for 234 marked individuals, with an average shell width of 15.28 ± 4.84 cm (mean $\pm$ SD), ranging from 2.3 to 26.6 cm. The number of marked individuals changed across the five sites as follows: 48, 63, 14, 12, 97. For further details see Hendriks et al. (2012, 2013).

3. Comparison of non-nested models

Ecologists are frequently faced with the problem of testing non-nested hypotheses, such as the comparison of the fit of Poisson and negative binomial models to count data (Lindén and Mäntyniemi, 2011). Bayesian model selection can be viewed as an extension of the Bayesian inference we already know: in this case models are the unknown quantities and we just want to make inference about them on the basis of their posterior distribution, given the data $\boldsymbol{y}$ (Link and Barker, 2009). In the problem of choosing between K models for the observed data, where each model corresponds a distinct parameter vector $\boldsymbol{\theta}_j$, with $j = 1, \ldots, K$, our main interest lies in the posterior probabilities of each of the K models, that is $p(M_j|\boldsymbol{y})$. In principle the process of calculating posterior model probabilities is quite simple: for $M = j$ we firstly choose prior model probability $p(M_j)$ independently of the data, and priors for model parameters $p(\boldsymbol{\theta}_j|M_j)$. Then, the observed data contribute to the posterior model probabilities through the so-called marginal likelihood $p(\boldsymbol{y}|M_j)$. In the simple case of two competing models (M_1, M_2) their relative probability is

$$\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \frac{p(M_1)}{p(M_2)} \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)} \quad (1)$$

where the second ratio on the right hand side is the Bayes factor (B_{12}), that is the ratio of the two marginal likelihoods (Kass and Raftery, 1995). We can therefore read Eq. (1) as "posterior model odds = B_{12} $\times$ prior model odds". Thus, the Bayes factor can be interpreted as the change in the model odds resulting from observing the data (Lodewyckx et al., 2011), providing a mechanism for

converting prior model probabilities to posterior model probabilities (Link and Barker, 2006). We can naturally base the choice of the best supported model or produce a model-averaged prediction just on posterior probabilities, bearing in mind that the probability of a model is always conditional on the model set, and can be interpreted as a relative degree of support within that set.

3.1. Example: shell width differences among sites

Using shell width data collected by Hendriks et al. (2012) at five different sites, we could be interested in whether populations differ. The problem can be treated as a random-effects ANOVA, assuming that the expected shell width in the five populations is not independent. In the case of positive continuous response variables, like shell width, we can assume that the logarithm of the original values is normally distributed, which is equivalent to using the log-normal distribution for the original response variable. However, other distributions such as the Gamma can be adopted (Ntzoufras, 2009). Thus, we focus our attention on the comparison of two non-nested models, a log-Normal and a Gamma random-effects ANOVA, through the product space method. For each pen shell i of a given width, measured at site s (with $s = 1, \ldots, 5$) the models under comparison are:

$$M_1: \quad width_i \sim \log N(\mu_i, \sigma^2), \quad \mu_i = \alpha_{s(i)}, \quad \alpha_{s(i)} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$M_2: \quad width_i \sim \Gamma(\mu_i \tau, \tau), \quad \log(\mu_i) = \gamma_{s(i)}, \quad \gamma_{s(i)} \sim N(\mu_\gamma, \sigma_\gamma^2).$$

3.2. The product space method

We compared the two non-nested models by means of the product space method, which is based on using MCMC to estimate the posterior model probabilities. At each iteration of the MCMC sampler, the parameters of all models are simulated, but only those for one model are allowed to affect the likelihood. The BUGS model script (without prior specification to improve readability) is as follows:

```
   data{
   C < - 100
   for (i in 1:n.ind){
   zeros[i] < - 0
   }
   }
   model{
   ######### MODEL INDEX
   M ~ dcat(p[])
   p[1] < - prior1
   p[2] < - 1-prior1
   postr1 < - 2-M
   postr2 < - M-1
   ######### PRIORS and PSEUDOPRIORS
   # (omitted; see supplementary material)
   ######### LIKELIHOOD
   for (i in 1:n.ind){
     zeros[i] ~ dpois(zeros.mean[i])
     zeros.mean[i] < - -loglik[M,i] + C
     # LogNormal model
     loglik[1,i] < - loglik1[i]
     loglik1[i] < - -0.5 * log(2 * 3.14159)-
log(width[i])
     - 0.5 * log(sigmaLN * sigmaLN)
     - 0.5 * pow((log(width[i])-mu[i]), 2) / (sigmaLN
* sigmaLN)
     # Gamma model
     loglik[2,i] < - loglik2[i]
     loglik2[i] < - (mu[i]*tauG) * log(tauG) +
     ((mu[i]*tauG)-1) * log(width[i]) - tauG *
width[i]
     - loggam((mu[i]*tauG))
     mu[i] < - (2-M) * alphaLN[site[i]] + (M-1) *
exp(alphaG[site[i]])
     }
   }
```

The model indicator parameter M can be updated to define which model is affecting the likelihood (Fig. 1). Under the simplest product space method, when a parameter is not affecting the likelihood, it is updated from its prior. But if the prior is diffuse, the probability the sampled value will be in the parameter space (that gives a reasonably large likelihood) is small. Thus, only the model currently affecting the likelihood (and thus having its parameters sampled from the posterior distribution) will get a large posterior probability, and thus it can be very difficult to leave that model. Carlin and Chib (1995) solved this problem by pointing out that when a parameter is not affecting the likelihood, it does not affect the posterior distribution, so it can be arbitrary. Thus, by making the prior distribution for a parameter conditional on the model, when it is not affecting the likelihood it can be generated from any distribution that simulates reasonable values close to its conditional posterior distribution. Mathematically, the only effect of this "pseudoprior" is on the probability of selecting the model, i.e., on the proposal distribution of the indicator variable M. In practice, the effect of the pseudoprior is to affect the mixing of the MCMC sampler: a good choice of pseudopriors (i.e., ones close to the conditional posterior) will give good mixing, and hence more efficient estimation of the posterior model probabilities and thus the Bayes factor B. In regard to B, the prior of the model index M corresponds to the prior model odds, while the posterior of M corresponds to the posterior model odds (Eq. (1)), in this case estimated through a Gibbs sampler. In practice, the posterior model probability for model j is estimated as

$$\hat{p}(M_j|\boldsymbol{y}) = \frac{\text{number of occurrence of } M = j}{\text{total number of iterations}} \tag{2}$$

We estimated posterior model probabilities and the corresponding log Bayes factor to quantify the relative evidence between the competing models. Prior model probabilities (`prior1` for model 1 and `(1-prior1)` for model 2 in the model script above) were calibrated manually, by running a Markov chain sampler several times and adjusting the prior model weights (Link and Barker, 2009). This operation can be time-consuming when one or more models are strongly supported compared to the others. This means that less favoured models can be almost never selected by the Gibbs sampler. Therefore, we need to find prior model probabilities that boost the achievement of approximately equal number of posterior model activations. As in our example, these priors can be strongly asymmetric, and hence difficult to spot. In consequence, it can be difficult (or even impossible) to get equal posterior model probabilities, which would let the MCMC sampler estimate their probabilities efficiently. Fortunately, the Bayes Factor does not depend on the prior probabilities, so these can be tuned to improve sampling. Chosen priors ($p(M_j)$) and observed posterior model probabilities ($\hat{p}(M_j|\boldsymbol{y})^{obs}$) are easily transformed into corrected posterior model probabilities as follows

$$\hat{p}(M_j|\boldsymbol{y})^{corr} = \left(\frac{\hat{p}(M_j|\boldsymbol{y})^{obs}}{p(M_j)}\right) \Big/ \sum_{j=1}^{K} \left(\frac{\hat{p}(M_j|\boldsymbol{y})^{obs}}{p(M_j)}\right).$$

This can be done in R by using the posterior samples for model 1 (`postr1` in the BUGS code that follows). We can summarize in a matrix the prior probabilities for the two models (`pr`), the observed posterior model probabilities for each chain (`ps1`, `ps2`, `ps3`), and the

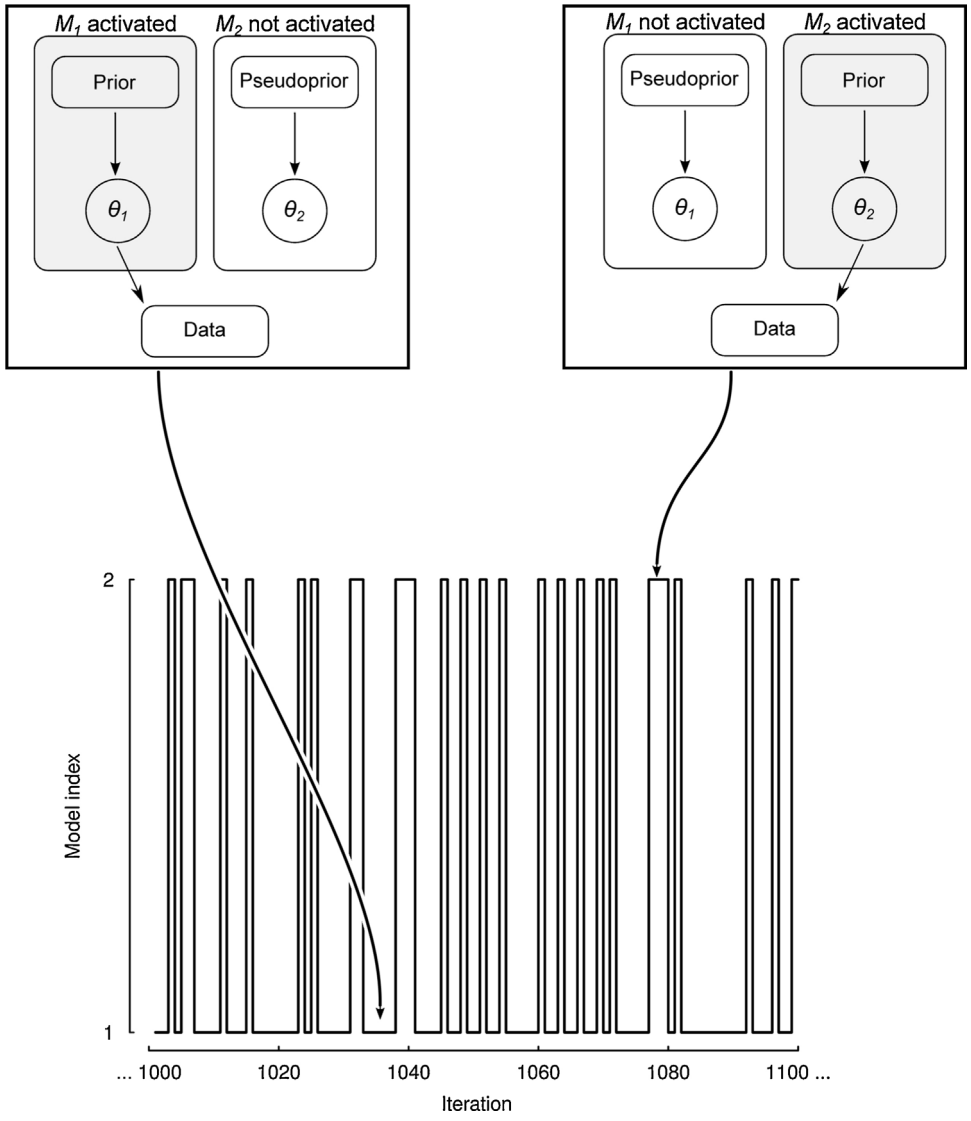

Fig. 1. Schematic representation of the functioning principle of the product space method, for the comparison of two models (M_1 and M_2). At each MCMC iteration only one of the two models is activated by the model index (`M` in the `BUGS` script). The two possible events are reported in this example, with model 1 activated on the left-hand side of the graph, and model 2 activated on the other side. Only the parameters of the activated model are assigned a prior distribution and are connected to the data, while the parameters of the non-activated model are assigned a pseudoprior distribution.

corrected posterior model probabilities (`psc`). In addition, we can derive the log Bayes factor (`logb`).

```
Mpost < - out$BUGSoutput$sims.array[,''postr1'']
M < - 2-Mpost
prob < - matrix(,2,5,dimnames=list(c(''M1'',''M2''),
       c(''pr'',''ps1'',''ps2'',''ps3'',''psc'')));
prior < - c(bugs.data$prior1, 1-bugs.data$prior1)
prob[,1] < - prior
nchains < - 3
for(m in 1:2){
  for(ch in 1:nchains){
    prob[m,ch+1]=mean(M[,ch]==m)
  }
}
postr < - rowMeans(prob[,2:4])
######### CORRECTED POSTERIOR MODEL PROBABILITIES
prob[,5] < - (postr/prior)/sum(postr/prior)
######### LOG BAYES FACTOR(12)
postr < - prob[,5]
logb < - log(postr[1])-log(postr[2])
```

Corrected posterior model probabilities represent the values we would obtain if we set a uniform prior for model index (Lodewyckx et al., 2011). In our specific case the best prior model probabilities had very extreme values, $p(M_2) = 10^{-9}$, 10^{-9}, 10^{-9}, 10^{-13}, and 10^{-14} for five sets of priors, which means prior probabilities near one for model 1 ($p(M_1)$). The five sets of prior distributions were chosen to evaluate sensitivity of the Bayes factors to prior assumptions (Table 2; see the specific section below for a discussion about prior sensitivity). In particular, for μ_α and μ_γ we considered a uniform prior between −5 and 5, a weakly informative normal prior with mean equal to 2.5 and variance 100, and two vague priors with mean of zero and variance equal to 1000 and 10^6. The first two priors were sufficiently uninformative to yield posterior distributions for parameters similar to those arising from the use of the vague priors. In addition, the log Bayes factors indicated very strong support in favour of the Gamma model M_2 (Table 2; see e.g., Kass and Raftery, 1995 for an interpretation scheme for values of the Bayes factor).

Table 2
Prior assumptions for model parameters used in the product space method. In addition, the log Bayes factor $\log(B_{12})$ related to each prior set is reported.

		Set 1	Set 2	Set 3	Set 4	Set 5
M_1	σ	$U(0, 10)$	$U(0, 10)$		$U(0, 10)$	$U(0, 10)$
	σ^2			$\Gamma^{-1}(0.001, 0.001)$		
	μ_α	$N(2.5, 100)$	$U(-5, 5)$	$U(-5, 5)$	$N(0, 1000)$	$N(0, 10^6)$
	σ_α	$U(0, 10)$	$U(0, 10)$		$U(0, 10)$	$U(0, 10)$
	σ^2_α			$\Gamma^{-1}(0.001, 0.001)$		
M_2	$\sigma = \tau^{-0.5}$	$U(0, 10)$	$U(0, 10)$		$U(0, 10)$	$U(0, 10)$
	σ^2			$\Gamma^{-1}(0.001, 0.001)$		
	μ_γ	$N(2.5, 100)$	$U(-5, 5)$	$U(-5, 5)$	$N(0, 1000)$	$N(0, 10^6)$
	σ_γ	$U(0, 10)$	$U(0, 10)$		$U(0, 10)$	$U(0, 10)$
	σ^2_γ			$\Gamma^{-1}(0.001, 0.001)$		
$\log(B_{12})$		-26.5	-25.2	-26.3	-28.0	-29.2

4. Selecting predictors in GLMs

Another use of the product space approach is to select variables in a regression, or similar model, by placing priors on the individual covariates. The variable selection procedure can be seen as one of deciding which of the regression parameters are equal to zero (O'Hara and Sillanpää, 2009). Assume that we want to explain an outcome y_i for individual i ($i = 1, \ldots, N$) using p covariates. Clearly, these variables can be continuous or discrete, and are candidate for the inclusion in the linear predictor. For a generalized linear model, for which we assume distribution, link function and variance function known, the linear predictor can be written as

$$\boldsymbol{\eta} = \sum_{j=0}^{p} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j \tag{3}$$

where $\boldsymbol{X}_j$ and $\boldsymbol{\beta}_j$ are the design matrix and the parameter vector for the jth term.

If we assume that model uncertainty is restricted to variable selection, each model M can be represented by a vector of binary indicators $\boldsymbol{\gamma} \in \{0, 1\}^p$. This vector indicates which of the possible sets of covariates are present in the model. In other words, we can use an auxiliary indicator variable γ_j, where $\gamma_j = 1$ indicates presence and $\gamma_j = 0$ absence of covariate j in the model, with $\boldsymbol{\theta}_j = \gamma_j \boldsymbol{\beta}_j$. The posterior probability that a variable is "in" the model can simply be calculated as the mean value of the indicator γ_j.

4.1. Example: what affects detectability of noble pen shell

Hendriks et al. (2012) used logistic regression models where the recaptured outcome (1 = recaptured, 0 = missed) was the response variable. As potential predictors of the variability in the probability of recapture, they considered the difference across sites and the shell width as a continuous individual covariate (centred for the analysis). We thus considered the following full model

$$y_{is} \sim Bernoulli(p_{is}), \quad \log\left(\frac{p_{is}}{1 - p_{is}}\right) = \alpha + \beta\, width_i + \epsilon_s \tag{4}$$

for a pen shell i, of a given width, at site s (with $s = 1, \ldots, 5$), where p_{is} is the site-specific recapture probability and ϵ_s is the random effect for the sth site, assuming $\epsilon_s \sim N(0, \sigma^2_\epsilon)$.

4.2. Gibbs variable selection

The model formulation we used is a product space method with particular assumptions about the prior probabilities of the models, i.e., the probability for any one $\boldsymbol{\theta}_j = \gamma_j \boldsymbol{\beta}_j$ is conditionally independent (and hence, we do not need to explicitly include all of the 2^p models in the model specification). The prior distributions of indicator γ_j and effect $\boldsymbol{\beta}_j$ are instead assumed to depend on each other. Gibbs Variable Selection (GVS, Dellaportas et al., 2000) procedure samples the effect size $\boldsymbol{\beta}_j | (\gamma_j = 0)$ from a pseudoprior. In fact, for prior and pseudoprior model parameters we can use a mixture of Normal distribution as follows

$$p(\boldsymbol{\beta}_j | \gamma_j) = (1 - \gamma_j)\, N(\bar{\mu}_j, S_j) + \gamma_j\, N(0, \Sigma_j) \tag{5}$$

where user-defined hyperparameters $\bar{\mu}_j$ (`mu.beta.ps`, in the following `BUGS` code) and variance S_j (`1/tau.beta.ps`) may be obtained from a pilot run of the full model, whereas Σ_j is the fixed prior variance of $\boldsymbol{\beta}_j$. In the script referred to model in Eq. (4), where site random effects are modelled in addition to shell width, $\Sigma_j^{-1} = 0.001$ and only one β is present:

```
model {
######### PRIORS
alpha ~ dnorm(0, 0.001)
beta ~ dnorm(beta.mean.prior, beta.tau.prior)
# GVS priors for beta
beta.mean.prior < - (1-g[1]) * mu.beta.ps
beta.tau.prior < - (1-g[1]) * tau.beta.ps + g[1] *
0.001
for(g in 1:n.sites){
  eps[g] ~ dnorm(0, tau.site)
}
tau.site < - 1/(sigma.site * sigma.site)
sigma.site ~ dunif(0, 5)
# Priors for variable indicators
g[1] ~ dbern(0.5)
g[2] ~ dbern(0.5)
######### LIKELIHOOD
for (i in 1:n.ind){
  C[i] ~ dbern(p[i])
  p[i] < - 1 / (1 + exp(-lp[i]))
  lp[i] < - alpha + g[1] * beta * width[i] + g[2]
* eps[site[i]]
}
######### MODEL CODE
mdl < - 1 + g[1] * 1 + g[2] * 2
######### VECTOR WITH MODEL INDICATORS
for (j in 1:nmodels){
  pmdl[j] < - equals(mdl, j)
}
}
```

The priors for the inclusion indicators (`g[1]` and `g[2]` in the script) were defined as $\gamma_j \sim Bernoulli(0.5)$, with $j = 1, 2$, with γ_1 related to the β parameter and γ_2 to the random effects. We used 10 sets of priors for model parameters, in the light of the related sensitivity of posterior model probabilities. More specifically we chose five different priors for regression coefficients α and β, represented by a weakly informative prior on the covariate effect ($\beta \sim N(0.2, 100)$), in addition to four prior sets where α and β were assumed drawn from a $N(0, \sigma^2)$ for $\sigma^2 \in \{10, 100, 1000, 10^6\}$. In

Table 3
Posterior variable inclusion probabilities $p(\gamma_j = 1|\,\boldsymbol{y})$, obtained under five different prior sets for regression coefficients α and β, and for two prior distributions on random effect hyperparameter σ_ϵ. γ_1 and γ_2 are the inclusion indicators.

	Priors				
α	$N(0, 1)$	$N(0, 10)$	$N(0, 100)$	$N(0, 1000)$	$N(0, 10^6)$
β	$N(0.2, 100)$	$N(0, 10)$	$N(0, 100)$	$N(0, 1000)$	$N(0, 10^6)$
			$\sigma_\epsilon \sim U(0, 5)$		
γ_1	0.994	0.998	0.995	0.981	0.621
γ_2	0.048	0.055	0.054	0.057	0.066
			$\sigma_\epsilon \sim half-Cauchy(1)$		
γ_1	0.994	0.998	0.994	0.980	0.614
γ_2	0.141	0.150	0.146	0.139	0.173

combination with the five sets of priors for regression coefficients, we adopted two different prior distributions on the random effect hyperparameter σ_ϵ, a uniform $(0, 5)$ and a half-Cauchy with scale 1 (see Gelman, 2006 for a justification of this prior, also in relation to the small number of levels for the site effect). Posterior variable inclusion probabilities (Table 3) and the derived posterior model probabilities (Table 4) strongly indicated presence of only shell width effect in the linear predictor.

5. Relevant points for practical implementation

In order to facilitate the implementation of the two methods, we can pinpoint some practical aspects.

5.1. Comparing non-nested models

The likelihoods were defined in the `BUGS` script using the so-called "zeros trick" (Lunn et al., 2012; Spiegelhalter et al., 2007, in section: Tricks: Advanced Use of the BUGS Language). This allows the use of any form likelihood and does not restrict in the number of distributions available. For further details and examples on the use of the trick of zeros and ones to select competing models see, e.g., Katsis and Ntzoufras (2005).

As we previously said, pseudopriors do not have influence on posterior model probabilities but their choice is important for sampling efficiency. We estimated pseudopriors' parameters by running the models separately and then using the posterior samples (see lines 91–101, 169–179, and 219–258 in the supplemented code). Alternatively, fitting a model using a frequentist approach might help to faster derive pseudopriors.

To achieve good quality of posterior model probability estimates we have to combine approximately equal posterior model activation (of model index `M`) with frequent model switching. For the categorical parameter `M`, the lack of model switches in its Markov chains is comparable to a high level of autocorrelation for the Markov chain of a continuous parameter. Some possible solutions to improve switching of model index are (i) adjusting prior model probabilities if one or more candidate models are much more supported by the data than the others, (ii) reparameterizing models to increase the number of shared parameters, (iii) improving pseudoprior estimation, (iv) running the sampler for longer. If we set parameters in common to the competing models we have to be sure that their interpretation is the same in the different models (see example in Carlin and Chib, 1995), and we need to check whether their posterior distributions have enough overlap.

5.2. Selecting predictors

The likelihood for the full model is specified as usual in `BUGS`, with the only difference being the incorporation of the binary inclusion indicators γ_1 and γ_2 in the linear predictor (`g[1]` and `g[2]` in the code). When we are dealing with models using explanatory variables that do not involve interactions, the latent variable γ_j can be treated as *a priori* independent, as we did. We then adopted independent priors for model parameters using a mixture of independent normal distributions as in Eq. (5). When categorical predictors are considered in the model the definition of the prior distribution for model parameters can be complicated. To this end, Ntzoufras (2002) and Dellaportas et al. (2000) provide detailed guidelines and examples.

We directly calculated posterior model probabilities in `BUGS` by adding a few lines of code. This is feasible when the number of models under consideration is small (and equals to 2^p if we consider all possible models). In the case of large model spaces this approach is not recommended since it involves a large amount of values stored, slowing down the `BUGS` software (Ntzoufras, 2009). Therefore we can save only the model indicator and export it into `R` to obtain its frequency tabulation. The model indicator (noted as `mdl` in the code) was calculated by transforming $\boldsymbol{\gamma}$ with the following formula (Ntzoufras, 2002) that converts numbers defined in the

Table 4
Posterior model probabilities derived by the Gibbs variable selection approach (see also Table 3), under five different prior sets for regression coefficients α and β, and for two prior distributions on random effect hyperparameter σ_ϵ. γ_1 and γ_2 are the inclusion indicators.

				Priors				
			α	$N(0, 1)$	$N(0, 10)$	$N(0, 100)$	$N(0, 1000)$	$N(0, 10^6)$
			β	$N(0.2, 100)$	$N(0, 10)$	$N(0, 100)$	$N(0, 1000)$	$N(0, 10^6)$
	γ_1	γ_2	Model			$\sigma_\epsilon \sim U(0, 5)$		
M_1	0	0	α	0.006	0.001	0.004	0.017	0.345
M_2	1	0	$\alpha + \beta\, width_i$	0.946	0.944	0.941	0.926	0.589
M_3	0	1	$\alpha + \epsilon_s$	0.001	0.000	0.001	0.002	0.034
M_4	1	1	$\alpha + \beta\, width_i + \epsilon_s$	0.048	0.055	0.054	0.055	0.032
						$\sigma_\epsilon \sim half-Cauchy(1)$		
M_1	0	0	α	0.005	0.002	0.005	0.016	0.301
M_2	1	0	$\alpha + \beta\, width_i$	0.854	0.849	0.849	0.845	0.526
M_3	0	1	$\alpha + \epsilon_s$	0.001	0.000	0.001	0.004	0.085
M_4	1	1	$\alpha + \beta\, width_i + \epsilon_s$	0.140	0.149	0.145	0.135	0.088

binary numerical system to the corresponding numbers in decimal numerical system

$$m(\boldsymbol{\gamma}) = 1 + \sum_{j=1}^{p} \gamma_j 2^{j-1} \quad (6)$$

Posterior model probabilities were then calculated with the BUGS command equals, by tracing whether a specific model is visited at each MCMC iteration.

6. Prior sensitivity

Posterior model probabilities are often more sensitive to the prior specification than the posterior distribution of parameters themselves. In the presence of model uncertainty the priors on the parameters $p(\boldsymbol{\theta}|M)$ need to be specified with care, and a sensitivity analysis should always be performed and discussed for a number of sensible priors (King, 2009). This sensitivity of posterior model probabilities and Bayes factors on priors, related to the so called "Lindley-Bartlett paradox", is probably the main disadvantage of Bayesian variable and model selection and we must pay particular attention to that. In particular, the specification of prior variance σ^2_θ is hard since, in non-informative cases, must be large to avoid prior bias within each model but not large enough to activate the above mentioned paradox and fully support the simplest model. The same issue appears in any model selection problem and it is more evident in nested model comparisons (Ntzoufras, 2009). Bayes factors are hence unstable in the presence of improper, non-informative priors for model parameters, but the problem extends also to the use of vague proper priors (Berger and Pericchi, 1996). Furthermore models having more parameters allow greater prior uncertainty in the range of the data to be produced, which means more sensitivity of the Bayes factor. As Link and Barker (2006) pointed out, assessing the reasonableness of selecting priors is inevitably a subjective process. Therefore, there does not seem to be a general solution to this issue, but it should be admitted in any analysis and priors should be clearly stated (Royle and Dorazio, 2008). We illustrated this sensitivity to prior by repeating both the product space and the GVS analysis using different prior distribution sets that were sufficiently uninformative to yield very similar posterior distributions for model parameters. Note that in the GVS example, the use of an almost fully uninformative prior (with variance equals 10^6) tends to increase the evidence in favour of the simplest model (Table 4), as in more complex models the vague prior have a greater prior uncertainty. This effect would have been even more evident if we used a larger prior variance. To get a sense of this effect see the example of Table 1 in Link and Barker (2006).

7. Conclusions

We presented two possible procedures for hierarchical variable and model selection in a Bayesian framework, with emphasis on the practical aspect of the approaches. Though other effective procedure are available, e.g., the reversible jump MCMC (Green, 1995) is implemented in the WinBUGS add-on package jump (Lunn et al., 2009), GVS algorithm can be easily implemented by all the freely available MCMC pieces of software in the BUGS language (WinBUGS, OpenBUGS, JAGS). Furthermore, this approach can be used not only in relation to variable selection problems (where the models concerned differ only in the form of the linear predictor) but also to compare models of different distributional form (e.g., Poisson and negative binomial or Generalized Poisson; Katsis and Ntzoufras, 2005), as we did with the Carlin and Chib's method. GVS can also be easily modified into the simpler (indicator model selection) method proposed by Kuo and Mallick (1998), frequently used by Royle and Dorazio (2008) in their book which represents a benchmark for ecologists. The key differences between the GVS and the product space method, but also more broadly between the other Gibbs sampler based variable selection strategies, are in their requirements in terms of priors and/or pseudopriors. Product space method and GVS both require pseudopriors just to improve the efficiency of the sampler. GVS is less expensive in requirements of pseudopriors, but correspondingly less flexible (Dellaportas et al., 2000). The main drawback of the product space method is the unavoidable specification of, and generation from, many pseudoprior distributions. Even if the pseudopriors do not enter the marginal likelihood distributions, generation from $K-1$ pseudopriors at each iteration is required and is computationally demanding (Dellaportas et al., 2002), especially when the number of models K is large.

The conceptual simplicity of the two approaches can represent an appealing feature for non-statisticians, however we have to be aware of the difficulties related to the maximization of sampling efficiency, an issue that is in any case common to all trans-dimensional MCMC methods. Furthermore, bearing in mind the sensitivity of the Bayes factors to the choice of priors on parameters, we need to articulate the reasons for prior assumptions and evaluate the related sensitivity of inferences (Link and Barker, 2006).

Though the procedures we adopted are not novelties in the world of quantitative ecology, for their ease of implementation they represent useful tools for ecologists and conservation biologists that fit their models with the widely used software packages in BUGS language.

Acknowledgements

We thank the three anonymous referees for constructive comments on previous versions of this manuscript. We also thank Aaron Lemma for IT assistance. Funds were partially provided by the Regional Government of Balearic Islands and FEDER funding. ST was funded by a PhD grant from the Muse – Museo delle Scienze (Trento, Italy) in collaboration with the University of Pavia.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2014.03.017.

References

Ando, T., 2010. Bayesian model selection and statistical modeling. Taylor & Francis US, Boca Raton, USA.
Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. J. Am. Statist. Assoc. 91, 109–122.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Inference. A Practical Information-Theoretic Approach, 2nd edition. Springer, New York, USA.
Carlin, B.P., Chib, S., 1995. Bayesian model choice via markov chain monte carlo methods. J. Ro. Statist. Soc. Series B (Methodol) 57, 473–484.
Clark, J.S., 2005. Why environmental scientists are becoming bayesians. Ecol. Lett. 8, 2–14.
Congdon, P., 2006. Bayesian Statistical Modelling. John Wiley & Sons, Chichester, UK.
Dellaportas, P., Forster, J., Ntzoufras, I., 2000. Bayesian variable selection using the Gibbs sampler. In: Dey, D., Ghosh, S., Mallick, B. (Eds.), In: Generalized Linear Models: A Bayesian Perspective, vol. 5. CRC Press, New York, USA, pp. 273–286.
Dellaportas, P., Forster, J.J., Ntzoufras, I., 2002. On bayesian model and variable selection using MCMC. Statist. Comput. 12, 27–36.
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1, 1–19.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis. CRC Press, New York, USA.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Statist. Assoc. 88, 881–889.
Godsill, S.J., 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Computat. Graph. Statist. 10, 230–248.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Halstead, B.J., Wylie, G.D., Coates, P.S., Valcarcel, P., Casazza, M.L., 2012. Exciting statistics: the rapid development and promising future of hierarchical models for population ecology. Anim. Conserv. 15, 133–135.
Hendriks, I., Deudero, S., Tavecchia, G., 2012. Recapture probability underwater: predicting the detection of the threatened noble pen shell in seagrass meadows. Limnol. Oceanogr. Methods 10, 824–831.
Hendriks, I.E., Tenan, S., Tavecchia, G., Marbà, N., Jordà, G., Deudero, S., Álvarez, E., Duarte, C.M., 2013. Boat anchoring impacts coastal populations of the pen shell, the largest bivalve in the Mediterranean. Biol. Conserv. 160, 105–113.
Hobert, J.P., Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J. Am. Statist. Assoc. 91, 1461–1473.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Statist. Assoc. 90, 773–795.
Katsis, A., Ntzoufras, I., 2005. Bayesian hypothesis testing for the distribution of insurance claim counts using the Gibbs sampler. J. Comput. Methods Sci. Eng. 5, 201–214.
Kéry, M., 2010. Introduction to WinBUGS for Ecologists. A Bayesian Approach to Regression, ANOVA, Mixed Models and Related Analyses. Academic Press, Amsterdam.
Kéry, M., Schaub, M., 2011. Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective. Academic Press, Amsterdam.
King, R., 2009. Bayesian Analysis for Population Ecology. CRC Press, Boca Raton, USA.
Kuo, L., Mallick, B., 1998. Variable selection for regression models. Sankhya Series B 60, 65–81.
Lindén, A., Mäntyniemi, S., 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92, 1414–1421.
Link, W.A., Barker, R.J., 2006. Model weights and the fundations of multimoldel inference. Ecology 87, 2626–2635.
Link, W.A., Barker, R.J., 2009. Bayesian Inference: With Ecological Applications. Academic Press, San Diego, USA.
Lodewyckx, T., Kim, W., Lee, M.D., Tuerlinckx, F., Kuppens, P., Wagenmakers, E.J., 2011. A tutorial on Bayes factor estimation with the product space method. J. Math. Psychol. 55, 331–347.
Lunn, D., Best, N., Whittaker, J., 2009. Generic reversible jump MCMC using graphical models. Stat. Comput. 19, 395–408.
Lunn, D., Jackson, C., Spiegelhalter, D.J., Best, N., Thomas, A., 2012. The BUGS Book: A Practical Introduction to Bayesian Analysis, vol. 98. CRC Press, Boca Raton, USA.
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. Statist. Comput. 10, 325–337.
Millar, R.B., 2009. Comparison of hierarchical bayesian models for overdispersed count data using DIC and Bayes' factors. Biometrics 65, 962–969.
Ntzoufras, I., 2002. Gibbs variable selection using bugs. J. Statist. Softw. 7, 1–19.
Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. John Wiley and Sons, Hoboken, USA.
O'Hara, R.B., Sillanpää, M.J., 2009. A review of Bayesian variable selection methods: What, how and which. Bayesian Anal. 4, 85–118.
Parent, E., Rivot, E., 2012. Introduction to Hierarchical Bayesian Modeling for Ecological Data. CRC, Chapman and Hall, Boca Raton, USA.
Park, T., Casella, G., 2008. The Bayesian Lasso. J. Am. Statist. Assoc. 103, 681–686.
Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods. Springer, New York, USA.
Royle, J.A., Dorazio, R., 2008. Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities. Academic Press, San Diego, USA.
Schaub, M., Kéry, M., 2012. Combining information in hierarchical models improves inferences in population ecology and demographic population analyses. Anim. Conserv. 15, 125–126.
Sisson, S.A., 2005. Transdimensional Markov chains: A decade of progress and future perspectives. J. Am. Statist. Assoc. 100, 1077–1089.
Spiegelhalter, D., Thomas, A., Best, N., Lunn, D.J., 2007. WinBUGS User Manual, Version 1.4.3.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Statist. Soc. Series B (Statist. Methodol.) 64, 583–639.
Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R., 1996. BUGS Examples Volume 2, Version 0.5, (version ii). Technical Report. MRC Biostatistics Unit, Cambridge.
Su, Y.S., Yajima, M., 2012. R2jags: A Package for Running Jags from R. R Package Version 0. 03-06.