MCMC samplers for models fit in `spOccupancy`

Jeffrey W. Doser

September 15, 2021

Contents

1 Introduction

This vignette provides statistical details on the MCMC algorithms used to fit each occupancy model in `spOccupancy`. We provide detailed descriptions of the joint posterior distributions for each model, how each parameter is updated in the model fitting process, and provide relevant citations to more specific documentation of the approaches where necessary. We also provide information on the composition sampling algorithms used for each model.

2 Single species occupancy model

2.1 Model description

Let z_j be the true presence (1) or absence (0) of a species at site j, with $j = 1, \ldots, J$. We assume this latent occupancy process arises from a Bernoulli process following

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j), \\ \text{logit}(\psi_j) &= \boldsymbol{x}_j' \cdot \boldsymbol{\beta}, \end{aligned} \tag{1}$$

where ψ_j is the probability of occurrence at site j, which is a function of site-specific covariates $\boldsymbol{X}$ and a vector of regression coefficients ($\boldsymbol{\beta}$).

We do not directly observe z_j and rather we observe an imperfect representation of the latent occurrence process. Let $y_{j,k}$ be the observed detection (1) or nondetection (0) of a species of interest at site j during replicate k for each of $k = 1, \ldots, K_j$ replicates at each site j. We envision the detection-nondetection data as arising from a Bernoulli process conditional on the true latent occurrence process:

$$\begin{aligned} y_{j,k} &\sim \text{Bernoulli}(p_{j,k} \cdot z_j), \\ \text{logit}(p_{j,k}) &= \boldsymbol{v}'_{j,k} \cdot \boldsymbol{\alpha}, \end{aligned} \tag{2}$$

where $p_{j,k}$ is the probability of detecting a species at site j during replicate k (given it is present at site j), which is a function of site and replicate specific covariates $\boldsymbol{V}$ and a vector of regression coefficients ($\boldsymbol{\alpha}$).

This model (after specifying appropriate priors), completes the standard single species occupancy model. Traditionally, when estimation occurs in a Bayesian framework, the regression coefficients for occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) must be updated using Metropolis updates, which can lead to slow convergence times and bad mixing of MCMC chains (Clark and Altwegg 2019). Instead, we introduce a Polya-Gamma latent variables (Polson, Scott, and Windle 2013) for both the occurrence and detection portions of the model, which induces Gibbs updates for all parameters in the single species occupancy model.

More specifically, let $\omega_{j,\beta} \sim \text{PG}(1, 0)$, which indicates $\omega_{j,\beta}$ is a random variable with a Polya-Gamma distribution with parameters 1 and 0. Given this latent variable, we can express the Bernoulli process of z_j as

$$\begin{aligned} \psi_j^{z_j}(1-\psi_j)^{1-z_j} &= \frac{\exp(\boldsymbol{x}'_j\boldsymbol{\beta})^{z_i}}{1+\exp(\boldsymbol{x}'_j\boldsymbol{\beta})} \\ &= \exp(\kappa_j \boldsymbol{x}'_j\boldsymbol{\beta}) \int \exp(-\frac{\omega_{j,\beta}}{2}(\boldsymbol{x}'_j\boldsymbol{\beta})^2) p(\omega_{j,\beta} \mid 1, 0) d\omega_{j,\beta}, \end{aligned} \tag{3}$$

where $\kappa_j = z_j - 0.5$ and $p(\omega_{j,\beta})$ is the probability density function of a Polya-Gamma distribution with parameters 1 and 0 (Polson, Scott, and Windle 2013). We define $\omega_{j,k,\alpha} \sim \text{PG}(1, 0)$ as a Polya-Gamma latent variable for the detection portion of the occupancy model, which results in a simliar re-expression of the Bernoulli likelihood for $y_{j,k}$ as we showed above for z_j. These re-expressions of the Bernoulli processes results in Gibbs updates for both the occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) regression coefficients.

We assume multivariate normal priors for the occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) regression coefficients to complete the Bayesian specification of the model. Our full joint posterior for a single species occupancy model, thus takes the following form, where N stands for the multivariate normal distribution:

$$\begin{aligned} [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha \mid \boldsymbol{Y}] \propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} &\text{Bernoulli}(y_{j,k} \mid p_{j,k} \cdot z_j) \times \\ &\text{Bernoulli}(z_j \mid \psi_j) \times \\ &\text{PG}(\omega_{j,\beta} \mid 1, 0) \times \\ &\text{PG}(\omega_{j,k,\alpha} \mid 1, 0) \times \\ &N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \\ &N(\boldsymbol{\alpha} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \end{aligned}$$

2.2 MCMC sampler

The Polya-Gamma data augmentation induces a Gibbs update for all parameters in the single species occupancy model. We first sample the occurrence and detection auxiliary variables from

$$\begin{aligned}\omega_{j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j'\boldsymbol{\beta}),\\ \omega_{j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}'\boldsymbol{\alpha}),\end{aligned} \tag{4}$$

respectively. We next sample the occurrence regression coefficients $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \mid \cdot \sim N\Big([\boldsymbol{\Sigma}_\beta + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}[\boldsymbol{X}'(\boldsymbol{z} - 0.5\mathbf{1}_J) + \boldsymbol{\Sigma}_\beta\boldsymbol{\mu}_\beta], [\boldsymbol{\Sigma}_\beta + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}\Big), \tag{5}$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,\beta}, \ldots, \omega_{J,\beta})$. Similarly, we sample the detection regression coefficients $\boldsymbol{\alpha}$ from

$$\boldsymbol{\alpha} \mid \cdot \sim N\Big([\boldsymbol{\Sigma}_\alpha + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}[\tilde{\boldsymbol{V}}'(\tilde{\boldsymbol{y}} - 0.5\mathbf{1}_{J^*}) + \boldsymbol{\Sigma}_\alpha\boldsymbol{\mu}_\alpha], [\boldsymbol{\Sigma}_\alpha + \tilde{\boldsymbol{V}}'\boldsymbol{S}_\alpha\tilde{\boldsymbol{V}}]^{-1}\Big). \tag{6}$$

The detection regression coefficients $\boldsymbol{\alpha}$ are only informed by the locations where $z_j = 1$, since we assume no false positive detections in the standard occupancy model. We define J^* as the total number of sites at the current iteration of the MCMC with $z_j = 1$. $\boldsymbol{S}_\alpha$ is a diagonal matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,1,\alpha}, \ldots, \omega_{J^*,K_{J^*},\alpha})$. The matrix $\tilde{\boldsymbol{V}}$ is the matrix of detection covariates associated with the sites where $z_j = 1$. Similarly, $\tilde{\boldsymbol{y}}$ is a vector of stacked detection-nondetection data values at the entries associated with $z_j = 1$.

Finally, z_j is set to 1 for all sites where there is at least one detection, and thus we only need to sample z_j at sites where there are no detections. Thus, for all locations with no detections, we sample z_j according to

$$z_j \mid \cdot \sim \text{Bernoulli}\left(\frac{\psi_j \prod_{k=1}^{K_j}(1 - p_{j,k})}{1 - \psi_j + \psi_j \prod_{k=1}^{K_j}(1 - p_{j,k})}\right). \tag{7}$$

2.3 Prediction

Prediction for a nonspatial single species occupancy model is a simple composition sampling algorithm. Given a set of occurrence covariates at a set of non-sampled locations ($\boldsymbol{X}_0$), we can derive the latent occurrence probability and the latent occurrence state at each non-sampled site $j = 1, \ldots J_0$ for each posterior sample s of the MCMC sampler following

$$\begin{aligned}\text{logit}(\psi_j^{(s)}) &= \boldsymbol{x}_{0,j} \cdot \boldsymbol{\beta}^{(s)},\\ z_j^{(s)} &\sim \text{Bernoulli}(\psi_J^{(s)}).\end{aligned} \tag{8}$$

3 Single species spatial occupancy models

3.1 Gaussian Process formulation

3.1.1 Model description

We extend the previous single species occupancy model to incorporate a spatial Gaussian Process that accounts unexplained spatial variation in species occurrence across a region of interest. The species-specific occurrence probability at site j, ψ_j, now takes the form

$$\text{logit}(\psi_j) = \boldsymbol{x}_j' \cdot \boldsymbol{\beta} + \text{w}_j, \tag{9}$$

where w_j is a realization from a zero-mean spatial Gaussian Process, i.e.,

$$\mathbf{w} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})). \tag{10}$$

We define $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})$ as a $J \times J$ covaraince matrix that is a function of the distances between any pair of site coordinates $\boldsymbol{s}$ and $\boldsymbol{s}'$ and a set of parameters ($\boldsymbol{\theta}$) that govern the spatial process. The vector $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta} = \{\sigma^2, \phi, \nu\}$, where σ^2 is a spatial variance parameter, ϕ is a spatial decay parameter, and ν is a spatial smoothness parameter. ν is only specified when using a Matern correlation function.

The detection portion of the occupancy model remains unchanged from the non-spatial occupancy model and follows Equation (2). Formulation of Polya-Gamma latent variables is also exactly analogous to the nonspatial model (Equation (3)), with all references to ψ_j now including the latent spatial random effects in addition to the site-level covariates.

Following standard recommendations for point-referenced spatial data (Banerjee, Carlin, and Gelfand 2003), we assign an inverse-Gamma prior to the spatial variance parameter and uniform priors to the spatial decay and spatial smoothness parameters. Our full joint posterior distribution takes the following form, where IG stands for the inverse-Gamma distribution:

$$\begin{aligned}
[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\omega}_\beta, \boldsymbol{\omega}_\alpha, \mathbf{w}, \boldsymbol{\theta} \mid \boldsymbol{Y}] \propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} & \text{Bernoulli}(y_{j,k} \mid p_{j,k} \cdot z_j) \times \\
& \text{Bernoulli}(z_j \mid \psi_j) \times \\
& N(\mathbf{w} \mid \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})) \times \\
& \text{PG}(\omega_{j,\beta} \mid 1, 0) \times \\
& \text{PG}(\omega_{j,k,\alpha} \mid 1, 0) \times \\
& N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \\
& N(\boldsymbol{\alpha} \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \times \\
& \text{IG}(\sigma^2 \mid a_{\sigma^2}, b_{\sigma^2}) \times \\
& \text{Uniform}(\phi \mid a_\phi, b_\phi) \times \\
& \text{Uniform}(\nu \mid a_\nu, b_\nu)
\end{aligned}$$

3.1.2 MCMC sampler

We first sample the occurrence and detection auxiliary variables from

$$\begin{aligned}
\omega_{j,\beta} \mid \cdot &\sim \text{PG}(1, \boldsymbol{x}_j'\boldsymbol{\beta} + \text{w}_j), \\
\omega_{j,k,\alpha} \mid \cdot &\sim \text{PG}(1, \boldsymbol{v}_{j,k}'\boldsymbol{\alpha}),
\end{aligned} \tag{11}$$

The Polya-Gamma scheme induces a Gibbs update for the occurrence regression coefficients, which are updated at each iteration according to

$$\boldsymbol{\beta} \mid \cdot \sim N\Big([\boldsymbol{\Sigma}_\beta + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}[\boldsymbol{X}'(\boldsymbol{z} - 0.5\mathbf{1}_J - \boldsymbol{S}_\beta\mathbf{w}) + \boldsymbol{\Sigma}_\beta\boldsymbol{\mu}_\beta], [\boldsymbol{\Sigma}_\beta + \boldsymbol{X}'\boldsymbol{S}_\beta\boldsymbol{X}]^{-1}\Big), \tag{12}$$

where $\boldsymbol{S}_\beta$ is a diagonal $J \times J$ matrix with diagonal entries equal to the latent PG variable values $(\omega_{1,\beta}, \ldots, \omega_{J,\beta})$.

The full conditional for the detection regression coefficients is the same as in the non-spatial model shown in Equation (6).

The spatial variance parameter, σ^2, is sampled via a Gibbs update of the form

$$\sigma^2 \mid \cdot \sim \text{IG}\Big(\frac{J}{2} + a_{\sigma^2}, \frac{\mathbf{w}'\boldsymbol{R}^{-1}\mathbf{w}}{2} + b_{\sigma^2}\Big), \tag{13}$$

where $\boldsymbol{R}$ is a $J \times J$ spatial correlation matrix.

The full conditional distributions for the spatial range parameter, ϕ, and spatial smoothness parameter, ν, are not available in closed form, and thus we use random walk Metropolis updates (e.g., Robert and Casella (2013)) to update the parameters. We use a random-walk Metroplis step with a multivariate normal proposal distribution (either of dimension 1 or of dimension 2 if Matern covariance function is used). To use the normal distribution as a proposal distribution, we transform the parameters to have a support spanning the entire real line, including a Jacobian adjustment for the Metropolis step. Tuning parameters are adaptively updated using Adaptive MCMC following Roberts and Rosenthal (2009).

The Polya-Gamma data augmentation scheme also enables a Gibbs update for the latent spatial Gaussian process ($\mathbf{w}$), as opposed to a traditional spatial occupancy model that requires a Metropolis update for the latent spatial process. The spatial process is updated according to

$$\mathbf{w} \mid \cdot \sim N\Big([\boldsymbol{S}_\beta + \boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{z} - 0.5\mathbf{1}_J - \boldsymbol{S}_\beta \boldsymbol{X}\boldsymbol{\beta}], [\boldsymbol{S}_\beta + \boldsymbol{\Sigma}^{-1}]^{-1}\Big). \tag{14}$$

Finally, for all sites with no detections, the latent occurrence values z_j are updated following Equation (7).

3.1.3 Prediction

Prediction for spatial occupancy models requires use of standard results for conditional multivariate normal distributions (Banerjee, Carlin, and Gelfand 2003). To predict latent occurrence and occurrence probability at non-sampled sites, we first need to predict the spatial process at the unobserved locations. Let $\mathbf{w}_0$ denote the spatial process at J_0 non-sampled locations. We assume that $\mathbf{w}_0$ and $\mathbf{w}$ (the spatial process at observed locations) arise from a multivariate normal distribution following

$$\begin{bmatrix}\mathbf{w} \\ \mathbf{w}_0\end{bmatrix} \mid \cdot \sim N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right), \tag{15}$$

where $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{12}$ is the $J \times J_0$ cross-covariance matrix between $\mathbf{w}$ and $\mathbf{w}_0$, and $\boldsymbol{\Sigma}_{22}$ is the variance-covariance matrix for $\mathbf{w}_0$. Using conditional multivariate normal theory, this results in the following posterior predictive distribution for the spatial process at nonsampled locations

$$\mathbf{w}_0 \sim N(\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{w}, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \tag{16}$$

We can use composition sampling to sample from this posterior predictive distribution by using the values for $\mathbf{w}$ at each sample q of the posterior distribution. This will generate a full predictive posterior sample which we can summarize with full uncertainty quantification.

Predicting all J_0 locations jointly can be expensive when J_0 is large. Thus, we perform independent individual predictions of the spatial process at each non-sampled location $j = 1, \ldots, J_0$. Thus, to predict the latent occurrence and latent occurrence probability at each non-sampled site j, we perform the following steps for each posterior sample q.

1. Sample $\text{w}_{0,j}^{(q)}$ from Equation (16), substituting in the current values at sample q of the spatial parameters and latent spatial process at the observed locations.

2. Compute the latent occurrence probability $\psi_j^{(q)}$ as $\text{logit}^{-1}(\boldsymbol{x}_{0,j}\boldsymbol{\beta}^{(q)} + \mathbf{w}_{0,j}^{(q)})$.
3. Sample the latent occurrence from $z_j^{(q)} \sim \text{Bernoulli}(\psi_j^{(q)})$.

References

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2003. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman; Hall/CRC.

Clark, Allan E, and Res Altwegg. 2019. "Efficient Bayesian Analysis of Occupancy Models with Logit Link Functions." *Ecology and Evolution* 9 (2): 756–68.

Polson, Nicholas G, James G Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–49.

Robert, Christian, and George Casella. 2013. *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Roberts, Gareth O, and Jeffrey S Rosenthal. 2009. "Examples of Adaptive Mcmc." *Journal of Computational and Graphical Statistics* 18 (2): 349–67.